Sampling elusive populations: Applications to studies of child labour

Vijay VERMA

Department of Statistics and FPRW/IPEC
International Labour Organization
4, route des Morillons
CH-1211 Geneva 22
Switzerland
Tel.: +41 (0) 22 799 8181
Fax: +41 (0) 22 799 8771
ipec@ilo.org
www.ilo.org/ipec
Sampling elusive populations: Applications to studies of child labour

Vijay VERMA

FPRW/International Programme on the Elimination of Child Labour


International Labour Office Dept. of Statistics; ILO International Programme on the Elimination of Child Labour

child labour / data collecting / survey - 13.01.2

ILO Cataloguing in Publication Data

NOTE

This publication was elaborated by Vijay Verma for the Department of Statistics and the International Programme on the Elimination of Child Labour of the Fundamental Principles and Rights at Work Branch (FPRW/IPEC), and coordinated by Mustafa Hakki Özel from Department of Statistics and previously with FPRW/IPEC.

Funding for this publication was provided by the United States Department of Labor (Project GLO/09/56/USA). This publication does not necessarily reflect the views or policies of the United States Department of Labor, nor does mention of trade names, commercial products, or organizations imply endorsement by the United States Government.

The designations employed in ILO publications, which are in conformity with United Nations practice, and the presentation of material therein do not imply the expression of any opinion whatsoever on the part of the International Labour Office concerning the legal status of any country, area or territory or of its authorities, or concerning the delimitation of its frontiers.

The responsibility for opinions expressed in signed articles, studies and other contributions rests solely with their authors, and publication does not constitute an endorsement by the International Labour Office of the opinions expressed in them.

Reference to names of firms and commercial products and processes does not imply their endorsement by the International Labour Office, and any failure to mention a particular firm, commercial product or process is not a sign of disapproval.

ILO publications and electronic products can be obtained through major booksellers or ILO local offices in many countries, or direct from ILO Publications, International Labour Office, CH-1211 Geneva 22, Switzerland. Catalogues or lists of new publications are available free of charge from the above address, or by email: pubvente@ilo.org or visit our website: www.ilo.org/publns.

Visit our website: www.ilo.org/ipec

Design and printing by the International Training Centre of the ILO, Turin – Italy
to the memory of Leslie Kish

Acknowledgement

Farhad Mehran has made important contributions to this manual by developing numerical illustrations elucidating the various sampling procedures, and by undertaking a thorough review of the material. Thanks are also due to a number of colleagues for reviewing and assisting in the preparation of the manuscript, in particular to Bijoy Raychaudhuri, Debi Mondal, Gianni Betti, Gillian Verma, Francesca Gagliardi, Giulio Ghellini, Lorenzo Fattorini, and the ILO technical editor. Mustafa Hakkı Özel has helped greatly by consistently pursuing and supporting this work with the assistance of Ayşegül Tuğçe Beycan. I would like to dedicate this book to the memory of Leslie Kish, my teacher and friend.

Vijay Verma
Siena, 31 October, 2013.
Sampling has now become an absolutely necessary instrument for collecting the vast amount of information required for understanding the functioning of our society. It provides a solid basis for estimating unknown values and ratios, and for testing the validity of presumed relationships in different areas of science. The remarkable aspect of probability sampling is its ability to produce not only valid estimates of the parameters of interest but also under broad conditions of their margin of errors due to sample variability. This feature has no doubt greatly contributed to the wide acceptance of sampling as an objective tool of measurement among analysts and the public at large.

The present publication has a dual function. It complements the earlier volume on Sampling for household-based surveys of child labour also prepared by Vijay Verma within the framework of the ILO Statistical Information and Monitoring Programme on Child Labour (SIMPOC). While the earlier volume dealt with sampling issues in conventional, broad-based household surveys, the present volume deals with non-standard issues involved in the sample design of child labour in targeted sectors and activities, such as child street vendors or child domestic workers.

The second function of the present document is its contribution to survey methodology that goes beyond the particular subject of child labour. Thus anyone interested in issues and practical solutions to problems such as sampling from imperfect frames, or sampling difficult populations because they are relatively few or mobile or secretive, would benefit from the contents of this publication. It puts together in one document a vast amount of materials on various sampling schemes such as multiplicity sampling, adaptive cluster sampling, controlled selection and balanced sampling, snowball sampling, and capture-recapture sampling. In most cases, novel ideas are brought on the underlying theory and the methodologies are illustrated with real-life numerical examples.

The ILO Department of Statistics and the ILO Fundamental Principles and Rights at Work (FPRW) hope that national statistical offices as well as researchers and analysts would use the materials presented here to improve national data collection programmes on child labour and as a result help to eliminate child labour across the world. The ILO could not be more pleased if this manual would also serve the statistical community at large and the survey statisticians in particular.

Constance Thomas
ILO Fundamental Principles and Rights at Work

Rafael Diez de Medina
ILO Department of Statistics
# Contents

Acknowledgement........................................................................................................ iii
Preface ........................................................................................................................ iv
Abbreviations ............................................................................................................... xviii

## Chapter 1

SURVEYING CHILD LABOUR ...................................................................................... 1

1.1 Introduction ......................................................................................................... 1

1.2 Household-based vs. other types of child labour survey ....................... 2
1.2.1 Surveying child labour through households ............................................. 2
1.2.2 Surveying working children not living at home ....................................... 3

1.3 Elusive populations of labouring children.................................................... 4
1.3.1 The concept of elusive populations .............................................................. 4
1.3.2 Characteristics of elusive populations of labouring children ................. 4

1.4 Executive summary: Chapter-by-chapter outline of the contents ......... 7
1.4.1 Substantive and technical background ...................................................... 7
1.4.2 Sampling from imperfect frames ................................................................. 9
1.4.3 Rare populations .......................................................................................... 11
1.4.4 Mobile populations .................................................................................... 14
1.4.5 Reclusive populations ................................................................................. 16
1.4.6 Annexes ....................................................................................................... 19

## I. Substantive and technical background

## Chapter 2

CHILD LABOUR SITUATIONS, DATA NEEDS AND SOURCES ................................. 23

2.1 Introduction ....................................................................................................... 23
2.1.1 Child labour situations and data needs ...................................................... 23
2.1.2 Sources of data on child labour (Sections 2.12-2.15) ............................. 25

2.2 Scope and nature of the information sought on child labour ............... 25
2.2.1 ILO Convention No. 138 .......................................................................... 25
2.2.2 UN Convention on the Rights of the Child (CRC) .................................. 26
2.2.3 ILO Convention No. 182 .......................................................................... 26
2.2.4 International comparisons ......................................................................... 28
2.2.5 Children in hazardous work ...................................................................... 30
2.2.6 Household chores (unpaid household services) ....................................... 31
2.2.7 Framework distinguishing forms of child labour .................................... 32

2.3 Child labour in domestic work ................................................................. 33
2.3.1 Nature of the activity ............................................................................... 33
### 2.4 Agriculture, especially commercial agriculture ........................................ 40
- 2.4.1 Nature of the activity ........................................................................... 41
- 2.4.2 Some examples .................................................................................. 43
### 2.5 Fishing and aquaculture ....................................................................... 44
- 2.5.1 Sampling considerations ..................................................................... 44
- 2.5.2 Common tasks and hazards in fishing ................................................. 44
- 2.5.3 Example: Child labour in fishing, El Salvador .................................... 45
### 2.6 Mining and quarrying .......................................................................... 45
- 2.6.1 Sampling considerations ..................................................................... 45
- 2.6.2 Common tasks and hazards in mining ................................................. 46
- 2.6.3 Some examples .................................................................................. 47
### 2.7 Manufacturing ...................................................................................... 47
- 2.7.1 Sampling considerations ..................................................................... 47
- 2.7.2 Common tasks and hazards in manufacturing .................................... 48
- 2.7.3 Some examples .................................................................................. 49
### 2.8 Construction .......................................................................................... 49
- 2.8.1 Sampling considerations ..................................................................... 49
- 2.8.2 Common tasks and hazards in construction ....................................... 50
- 2.8.3 Example: Child labour baseline survey in Uganda ................................ 51
### 2.9 Street children ....................................................................................... 51
- 2.9.1 Diverse situations of street children .................................................... 51
- 2.9.2 The condition of street children ......................................................... 53
- 2.9.3 Examples ........................................................................................... 56
### 2.10 ‘Idle’ children ....................................................................................... 57
### 2.11 Worst Forms of Child Labour (WFCL) other than Hazardous Work... 57
- 2.11.1 Surveying worst forms of child labour other than hazardous work ... 57
- 2.11.2 Trafficking in children for labour and sexual exploitation ................. 59
- 2.11.3 Commercial sexual exploitation of children ..................................... 60
- 2.11.4 Child bonded labour and forced labour ............................................ 61
- 2.11.5 Children and War ............................................................................. 62
- 2.11.6 Children engaged in drug trafficking ................................................. 63
### 2.12 Sources of information on child labour .............................................. 63
- 2.12.1 Household-based surveys .................................................................. 63
- 2.12.2 Sources supplementing household-based child labour surveys ....... 64
- 2.12.3 Establishment surveys ..................................................................... 66
- 2.12.4 Rapid assessments .......................................................................... 67
- 2.12.5 Baseline surveys and studies (BLSs) .................................................. 68
- 2.12.6 Complementary nature of the sources ............................................. 71
### 2.13 Surveying child labour through households ....................................... 72
- 2.13.1 Child Labour Survey (CLS) and Labouring Children Survey (LCS)...... 72
Chapter 3

BASIC SAMPLING AND ESTIMATION PROCEDURES

3.1 Probability sampling ................................................................. 84

3.2 Departures from simple random sampling ................................ 85

3.2.1 Stratification ........................................................................ 85

3.2.2 Clustering or multi-stage sampling ...................................... 86

3.2.3 Unequal selection probabilities .......................................... 86

3.2.4 A typical design ................................................................. 87

3.3 Probability proportional to size (PPS) sampling ...................... 88

3.3.1 Self-weighting sample ......................................................... 88

3.3.2 Some common variations ................................................... 89

3.3.3 Different choice of size measures ...................................... 91

3.3.4 Systematic sampling .......................................................... 93

3.4 Weighting of sample data and estimation from the sample ....... 96

3.4.1 Sources of information for weighting .................................. 96

3.4.2 Step-by-step procedure for weighting ................................ 97

3.4.3 Estimation from sample data ............................................. 107

3.4.4 Estimating from a sample: numerical illustration ............ 110

3.5 Practical variance estimation procedures ............................... 114

3.5.1 Estimating variance from comparison among primary selections ... 114

3.5.2 Estimating variance from comparison among replications of the full sample based on repeated re-sampling ............. 116

3.5.3 Jackknife Repeated Replication (JRR) .............................. 116

3.6 Analysis of design effects ......................................................... 118

3.6.1 The concept ....................................................................... 118

3.6.2 Components of design effect ............................................ 118

3.6.3 Estimating the effect of weighting \(d_w\) .......................... 119

3.6.4 Estimating the effect of clustering within ultimate sampling units \(d_u\) ............................................................. 120

3.6.5 Estimating the effect of clustering, stratification and other design complexities \(d_x\) ...................................................... 121

3.7 Brief description of errors in survey data ............................... 122

3.7.1 A typology of errors ........................................................ 122

3.7.2 Errors in measurement ..................................................... 123

3.7.3 Errors in estimation ......................................................... 126

3.7.4 Item non-response ............................................................ 127
II. Imperfect frames

Chapter 4
SAMPLING FROM IMPERFECT FRAMES ......................................................... 131

4.1 Basic concepts .................................................................................. 132
  4.1.1 The target population ................................................................. 132
  4.1.2 The sampling frame ................................................................. 134

4.2 Area frames ...................................................................................... 138
  4.2.1 Basic requirements ................................................................. 138
  4.2.2 Common problems with area frames ...................................... 139
  4.2.3 Desirable frame properties ..................................................... 139

4.3 Imperfect list frames ........................................................................ 142
  4.3.1 Basic consideration: the relationship between sampling and analysis units ........................................... 142
  4.3.2 Imperfections which increase variance, but without introducing bias .................................................. 143
  4.3.3 Imperfections resulting in potential bias .................................... 148

4.4 Illustrations from national practice .................................................... 151
  4.4.1 Set of five baseline surveys in Bangladesh, 2002-2003 ............ 151
  4.4.2 Observations on the baseline surveys ...................................... 159
  4.4.3 Establishment-based survey in metropolitan areas, Bangladesh (2003a) ............................................. 161

4.5 Types of linkage between sampling and analysis units ................. 163
  4.5.1 Direct sampling ........................................................................ 164
  4.5.2 Sampling with multiplicity ...................................................... 165
  4.5.3 Link-trace sampling ................................................................ 168

4.6 The multiplicity estimator ............................................................... 170
  4.6.1 Commonly used estimation procedures .................................. 170
  4.6.2 Technical details of the multiplicity estimator ....................... 173
  4.6.3 Note on the weight-share method of estimation ..................... 176
  4.6.4 Note concerning ‘indirect sampling’ ........................................ 178
  4.6.5 A formal framework ............................................................... 179

Chapter 5
SAMPLING ESTABLISHMENTS EMPLOYING CHILDREN .......................... 183

5.1 Introduction ..................................................................................... 183
  5.1.1 Scope of the chapter ................................................................ 183
  5.1.2 Design strategy ....................................................................... 185
  5.1.3 Sampling stages in a sample of children working in establishments .................................................. 186

5.2 Stratification ..................................................................................... 187
  5.2.1 Important general considerations ........................................... 187
  5.2.2 Stratification criteria ............................................................... 188
  5.2.3 Take-all strata ........................................................................ 189
  5.2.4 Limiting the number of strata and optimal stratification ......... 190
5.3 Sample allocation ........................................................................ 191
   5.3.1 Proportionate allocation ...................................................... 191
   5.3.2 Optimal allocation .............................................................. 195
   5.3.3 Neyman allocation .............................................................. 196
   5.3.4 Optimal stratification .......................................................... 197
   5.3.5 Scaling unit size measures to facilitate sample allocation
       and selection ........................................................................... 198
   5.3.6 Application in practice ....................................................... 199

5.4 Single-stage sampling of medium and large establishments
      employing children ................................................................. 201
   5.4.1 Selection of units within strata: πPS versus stratified-SRS ....... 202
   5.4.2 Some specific sampling schemes .......................................... 204
   5.4.3 Design criteria ................................................................. 207

5.5 Selecting child workers within establishments ......................... 209
   5.5.1 Selecting establishments and child workers: some two-stage
       design options ........................................................................ 209
   5.5.2 Numerical illustration of two-stage designs: selection
       of establishments and of working children within establishments... 212

5.6 Small and informal sector establishments employing children...... 221
   5.6.1 Child labour activity in the informal sector ............................ 221
   5.6.2 Characteristic features of small and informal sector establishments... 223
   5.6.3 Consequences for survey design .......................................... 224
   5.6.4 Most commonly used area-based design ............................... 226
   5.6.5 Using existing sources of information .................................... 227

5.7 Basic design issues .................................................................. 228
   5.7.1 Separate single-sector versus integrated multi-sectoral surveys.... 228
   5.7.2 Stand-alone survey, versus a module attached to another survey... 229

5.8 Sample allocation through ‘strata of concentration’ .................... 231
   5.8.1 Description of the basic design ........................................... 231
   5.8.2 Constructing and using ‘strata of concentration’ (StrCon) ...... 233

5.9 Efficiency of the allocation procedure; possible improvements ...... 236
   5.9.1 The effect of ‘random’ weights ............................................ 236
   5.9.2 Meeting sample size requirements ....................................... 237
   5.9.3 Numerical illustrations ....................................................... 238
   5.9.4 More flexible models: an empirical approach ....................... 244
   5.9.5 Note on sample selection procedure .................................... 246

5.10 Selection of establishments within sample areas ....................... 247
   5.10.1 Segmentation and listing ................................................... 247
   5.10.2 Listing units ................................................................. 248
   5.10.3 Cost and quality of the listing operation ......................... 250
   5.10.4 Selecting establishments .................................................. 251

5.11 Some examples of country surveys of small
      and informal sector establishments ......................................... 253
Sampling elusive populations: Applications to studies of child labour

5.11.1 Baseline survey on child labour in glass-bangles industry, Pakistan ..... 253
5.11.2 Child Labour and the urban informal sector, Uganda .................. 255

5.12 Remarks concerning sample implementation .................. 260
5.12.1 Units requiring special treatment .................................... 260
5.12.2 Departures from probability sampling .......................... 261

III. Rare populations

Chapter 6

SAMPLING RARE POPULATIONS .......................................................... 265

6.1 Introduction: making use of clustering and uneven distribution . 265
6.2 Locating concentrations of the rare population using existing large-scale sources ........................................... 268
   6.2.1 Use of economic censuses and surveys .......................... 268
   6.2.2 Use of population-based censuses and surveys ............ 269
   6.2.3 Other sources .......................................................... 270
6.3 Separating out near-empty clusters ........................................ 271
   6.3.1 Introduction ......................................................... 271
   6.3.2 The sequential test .............................................. 271
   6.3.3 Numerical illustrations of the sequential test ............. 273
   6.3.4 Using incomplete lists for identifying areas containing the rare population ............................................. 283
6.4 Gain from separating out empty clusters: numerical illustration .......................................................... 284
6.5 Identifying strata with differing concentrations .................... 290
   6.5.1 The procedure ...................................................... 290
   6.5.2 Illustration of the procedure .................................. 292
6.6 Oversampling strata with higher concentrations .................... 296
   6.6.1 Introduction: when is oversampling useful? ......... 296
   6.6.2 Estimation of size of the rare population ......... 299
   6.6.3 Estimating means and similar statistics for the rare population ...... 305
   6.6.4 Obtaining large enough sample sizes for in-depth analysis ............................. 313
   6.6.5 Compromise allocation ........................................ 313
6.7 Counting, listing, screening, two-phase sampling .................... 318
   6.7.1 Distinct but closely-related operations ................. 318
   6.7.2 Screening ................................................. 319
   6.7.3 Two-phase sampling .................................... 323
6.8 Cumulation over time ............................................................. 324

Chapter 7

MULTIPLICITY SAMPLING .............................................................. 327

7.1 Basics of multiplicity sampling ........................................... 327
Chapter 8

MULTI-FRAME SAMPLING ................................................................. 357

8.1 Multiple frames in the context of child labour surveys .......... 357
  8.1.1 Potential for use of multiple frames .................................. 357
  8.1.2 Characteristic features of the situation ............................. 358

8.2 Some examples ................................................................. 359
  8.2.1 Supplementary area frame for a rare population ................ 359
  8.2.2 Combining several list frames ........................................ 361

8.3 Dealing with multiple sampling frames ............................... 362

8.4 Constructing non-overlapping frames .................................. 363

8.5 Estimation using overlapping frames .................................. 367
  8.5.1 Commonly used estimation procedures ............................ 367
  8.5.2 Further comments on the multiplicity estimator for multi-frame sampling ............................................................... 369

8.6 Illustration 1. A three-frame example ................................. 371
  8.6.1 An illustrative scenario .................................................. 371
  8.6.2 Numerical illustration for the three-frame example ............ 373

8.7 Illustration 2. A refined frame with a part of the overlap removed 376
  8.7.1 ‘Refining’ the frames by removing overlap .......................... 376
  8.7.2 A numerical illustration of the ‘refined’ frame .................... 378
  8.7.3 Estimating multiplicity (or overlap between frames) ............ 380

8.8 Choosing among options ..................................................... 383
# Chapter 9

## ADAPTIVE CLUSTER SAMPLING ................................................................. 385

### 9.1 Introduction ................................................................. 385
- 9.1.1 The approach ........................................................................ 385
- 9.1.2 Potential advantages .......................................................... 386
- 9.1.3 Limitations ............................................................................ 387

### 9.2 Basics of the technique ............................................................... 388
- 9.2.1 Constructing adaptive clusters or networks ......................... 388
- 9.2.2 Clarifying terminology ......................................................... 389
- 9.2.3 Basic estimation procedure .................................................. 391

### 9.3 Relative efficiency of adaptive versus conventional sampling ...... 392
- 9.3.1 Variance ................................................................................. 393
- 9.3.2 Cost ....................................................................................... 394
- 9.3.3 Factors influencing relative efficiency .................................. 395

### 9.4 Estimation ................................................................. 396
- 9.4.1 Horvitz-Thompson (H-T) estimator ......................................... 396
- 9.4.2 Hansen-Hurwitz (H-H) estimator ........................................... 397
- 9.4.3 Multiplicity estimator ............................................................ 397
- 9.4.4 Weight-share estimator ......................................................... 398

### 9.5 Numerical illustration ............................................................... 401
- 9.5.1 Illustrative population .......................................................... 401
- 9.5.2 Simple random sample of area units .................................... 406
- 9.5.3 Adaptive cluster sampling with stratification ....................... 411

### 9.6 Further technical aspects .......................................................... 414
- 9.6.1 Unequal unit selection probabilities ..................................... 414
- 9.6.2 Stratification ......................................................................... 415
- 9.6.3 Multistage sampling .............................................................. 416
- 9.6.4 Multivariate and auxiliary criteria ....................................... 419
- 9.6.5 Order statistics ...................................................................... 420
- 9.6.6 Imperfect detectability .......................................................... 420

### 9.7 Implementation issues .............................................................. 421
- 9.7.1 When should adaptive sampling be introduced? ..................... 421
- 9.7.2 Criteria, definitions, rules for adaptive sampling ................... 422
- 9.7.3 Size of the initial sample ........................................................ 423
- 9.7.4 Controlling sample size .......................................................... 423

### 9.8 Variations on the standard adaptive cluster sampling design ...... 425
- 9.8.1 Constructing adaptive clusters in waves .................................. 425
- 9.8.2 Estimation with overlapping adaptive clusters ...................... 426
- 9.8.3 Illustration 1. Constructing adaptive clusters in waves ........... 430
- 9.8.4 Illustration 2. Adaptive clusters formed with immediate neighbours .. 432
- 9.8.5 Illustration 3. Estimation when adaptive clusters in the sample overlap .......................................................... 439
9.9 Some guidelines and recommendations................................. 440
  9.9.1 Concerning the approach................................................... 440
  9.9.2 Concerning design............................................................ 441
  9.9.3 Concerning implementation .............................................. 441
  9.9.4 Concerning analysis ........................................................ 441
  9.9.5 Concerning evaluation ..................................................... 441

IV. Mobile populations

Chapter 10
SAMPLING MOBILE POPULATIONS ........................................... 445

  10.1 Issues and framework ........................................................... 445

  10.2 Location sampling ............................................................... 448
    10.2.1 Introduction ..................................................................... 448
    10.2.2 Strong assumptions ....................................................... 449
    10.2.3 Single-stage sampling of persons at locations.................. 450
    10.2.4 Numerical illustrations ............................................... 456
    10.2.5 Sampling of locations ................................................... 461

  10.3 Time-location sampling ....................................................... 462
    10.3.1 Basic characteristics ..................................................... 462
    10.3.2 Applications in diverse fields ........................................ 466

  10.4 Sampling mobile populations in more difficult situations: examples ... 468
    10.4.1 Introduction ..................................................................... 468
    10.4.2 A survey concerning sexual behaviour of young persons (YSMS) .... 469
    10.4.3 A survey of migrants in Brazil ....................................... 472

  10.5 Probability of selection and sample weights of a mobile individual .... 477
    10.5.1 The issue .......................................................................... 477
    10.5.2 Some solutions ............................................................... 478

Chapter 11
CAPTURE-RECAPTURE SAMPLING ............................................ 485

  11.1 Introduction .......................................................................... 485
    11.1.1 Applications of the procedure in different fields .............. 485
    11.1.2 Brief outline of the approach ......................................... 487
    11.1.3 Sort of situations in which capture-recapture sampling
           might be useful ................................................................. 489
    11.1.4 An illustration: a profile of Cairo street children ............ 490

  11.2 Assumptions in the capture-recapture method ................................ 493
    11.2.1 Basic Assumptions ........................................................ 493
    11.2.2 Closed population ........................................................ 494
    11.2.3 Representative samples ............................................... 498
    11.2.4 Linking of sources at the micro-level ............................ 499
    11.2.5 Assumption concerning sample design and selection probabilities . 499
11.3 Numerical illustrations of capture-recapture sampling
(schemes “M₀, M₁, M₂, M₃”) ................................................................. 502
11.3.1 Table 11.1 (scheme “M₀”) ......................................................... 502
11.3.2 Table 11.2 (scheme “M₁”) ......................................................... 503
11.3.3 Table 11.3 (scheme “M₂”) ......................................................... 503
11.3.4 Table 11.4 (scheme “M₃”) ......................................................... 503
11.3.5 Table 11.5 Control through stratification (scheme “M₃” cont.) ...... 504
11.3.6 Note of the selection method for the simple illustrations .......... 504

11.4 Separate estimates for major domains or population subgroups ....... 514
11.4.1 Need for separate domain estimates ......................................... 514
11.4.2 Analysis by domain .................................................................. 515
11.4.3 Using sample weights at the level of individual units ................. 517

11.5 Weighting for individuals visiting study locations with different
frequencies .......................................................................................... 518
11.5.1 A simple procedure ................................................................. 518
11.5.2 Numerical illustration .............................................................. 519

11.6 Capture-recapture with multistage sampling ............................... 520
11.6.1 Observed overlap with samples drawn from the same locations... 520
11.6.2 Illustrations ............................................................................. 521

11.7 Estimation of sample weights in the more general situation ......... 524
11.7.1 Reason for weighting sample data ........................................... 524
11.7.2 Outline of a methodology ....................................................... 526
11.7.3 A unit’s selection into m₂ versus selection into n₂ ................. 529
11.7.4 Variations in selection probabilities arising from design
and implementation ........................................................................... 530
11.7.5 Unknown and inaccessible population ................................... 532

11.8 Weight components derived from the sample overlap m₂ .......... 533
11.8.1 Overall weight as a function of unit characteristics ............... 533
11.8.2 Separating scaling factor (W₁(2)) and haphazard component (W₀(2)) ... 536
11.8.3 Carry over from m₂ to n₂; possibly to n₁ .................................. 536

11.9 Use of more than two data sources ............................................. 538

11.10 Some guidelines and recommendations ..................................... 540
11.10.1 Identification of the target population .................................. 540
11.10.2 Using administrative lists ....................................................... 541
11.10.3 Survey-based approach .......................................................... 542
11.10.4 Choice of sample size ............................................................. 542
11.10.5 Sample design ....................................................................... 546

11.11 Simple illustration of the various procedures discussed ............ 547
11.11.1 Children selling flowers at traffic lights ................................. 547
11.11.2 Departure from sampling assumptions .................................... 552
Chapter 14

RESPONDENT-DRIVEN SAMPLING .............................................. 667

14.1 Methodology of respondent-driven sampling ..................... 667
  14.1.1 Introduction ................................................................. 667
  14.1.2 Underlying concepts: homophily, equilibrium, degree, network structure........................................... 670

14.2 Aspects of implementation ............................................... 673
  14.2.1 Development of design and procedures ......................... 673
  14.2.2 Seeds ............................................................................. 675
  14.2.3 Recruitment by previous respondents .......................... 677
  14.2.4 Incentives and coupons .................................................. 679
  14.2.5 Sample size .................................................................... 680

14.3 Examples of use of respondent-driven sampling .................. 680
  14.3.1 A survey of street children in Bamako and Accra ........... 681
  14.3.2 A survey of female sex workers in Brazil ...................... 685
  14.3.3 A survey of gay and bisexual men in Kampala, Uganda .... 688
  14.3.4 A survey to research into the community impact of hate-crimes in Canada .................................................... 691

14.4 Statistical modeling of the RDS procedure .......................... 692
  14.4.1 Basic estimation procedure ............................................. 692
  14.4.2 Assumptions of the RDS model ...................................... 694
  14.4.3 Estimation of unit selection probabilities and sample weights ...... 695
  14.4.4 Analysis tools .................................................................. 698

14.5 Note on alternative approaches ......................................... 698
  14.5.1 Targeted Sampling ......................................................... 698
  14.5.2 Indigenous fieldworker sampling .................................... 700

14.6 Assessment: comparisons with alternative sampling approaches... 701
  14.6.1. Comparison of different approaches in a survey in Brazil .......... 702
  14.6.2 Comparison of RDS with indigenous fieldworkers sampling (IFS) ... 705

14.7 Assessing validity of the model assumptions ....................... 708
  14.7.1 Review of RDS field experience in an international setting .... 709
  14.7.2 A common problem: sluggishness in recruitment .............. 711
  14.7.3 Simulations to assess the impact of network structure .......... 714
  14.7.4 Simulations to assess sensitivity to conditions of implementation .. 717
Annexes

Annex A Illustrations from studies of the worst forms of child labour .......... 721

A.1 Child domestic work ................................................................. 721
A.2 Agriculture; commercial agriculture ........................................ 726
A.3 Fishing and aquaculture ............................................................ 730
A.4 Mining and quarrying ............................................................... 732
A.5 Manufacturing and handicrafts .................................................. 734
A.6 Construction ..................................................................... 738
A.7 Street work and the informal sector ......................................... 740
A.8 Unconditional Worst Forms of Child Labour (UWFCL) ............. 751

Annex B Efficiency of oversampling strata of concentration: a case study .... 757
Two-stage, two-phase sampling for Child Labour Survey in Cambesia .... 757
B.1 Cambesia ........................................................................ 757
B.3 Targeted two-phase sampling .................................................. 767

Annex C Applying sampling procedures in combination: A case study ........ 774
C.1 Lessons from a baseline survey of child labour in Jamaica .......... 774
C.2 Possible scenarios for a full-scale survey of child labour .......... 782
C.3 Survey of the fishing sector ....................................................... 783
C.4 Household-based component .................................................. 786
C.5 Establishment-based survey component ................................... 787
C.6 Child vendors .................................................................. 787
C.7 Children on the street ............................................................. 788

References

Bibliography ............................................................................. 793
Author Index ............................................................................. 809
Subject Index ........................................................................... 813
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIDs</td>
<td>Acquired Immune Deficiency Syndrome</td>
</tr>
<tr>
<td>BLSs</td>
<td>baseline surveys and studies</td>
</tr>
<tr>
<td>BRR</td>
<td>balanced repeated replication</td>
</tr>
<tr>
<td>CAPMAS</td>
<td>Central Agency for Public Mobilization and Statistics (Egypt)</td>
</tr>
<tr>
<td>CDW</td>
<td>child domestic workers</td>
</tr>
<tr>
<td>CLS</td>
<td>child labour survey</td>
</tr>
<tr>
<td>CPS</td>
<td>Child Protective Services</td>
</tr>
<tr>
<td>CR</td>
<td>capture-recapture (sampling)</td>
</tr>
<tr>
<td>CRC</td>
<td>Convention on the Rights of the Child</td>
</tr>
<tr>
<td>CWIN</td>
<td>Child Workers in Nepal (NGO)</td>
</tr>
<tr>
<td>DA</td>
<td>drug abuse</td>
</tr>
<tr>
<td>EAs</td>
<td>census enumeration areas</td>
</tr>
<tr>
<td>EB</td>
<td>establishment blocks</td>
</tr>
<tr>
<td>ECHP</td>
<td>European Community Household Panel</td>
</tr>
<tr>
<td>ECPAT</td>
<td>“End Child Prostitution, Child Pornography, and Trafficking of Children for Sexual Purposes” International</td>
</tr>
<tr>
<td>EU-SILC</td>
<td>EU Statistics on Income and Living Conditions</td>
</tr>
<tr>
<td>FSW</td>
<td>female sex workers</td>
</tr>
<tr>
<td>FWs</td>
<td>field workers</td>
</tr>
<tr>
<td>GREG</td>
<td>general regression estimator</td>
</tr>
<tr>
<td>GWSM</td>
<td>generalised weight-share method</td>
</tr>
<tr>
<td>H-H</td>
<td>Hansen-Hurwitz (estimator)</td>
</tr>
<tr>
<td>H-T</td>
<td>Horvitz-Thompson (estimator)</td>
</tr>
<tr>
<td>HIV</td>
<td>Human Immunodeficiency Virus</td>
</tr>
<tr>
<td>HRH</td>
<td>high-risk heterosexual</td>
</tr>
<tr>
<td>ICLS</td>
<td>International Conference of Labour Statisticians</td>
</tr>
<tr>
<td>IDUs</td>
<td>injection drug users</td>
</tr>
<tr>
<td>IFS</td>
<td>indigenous fieldworkers sampling</td>
</tr>
<tr>
<td>ILC</td>
<td>International Labour Conference</td>
</tr>
<tr>
<td>ILO</td>
<td>International Labour Organisation</td>
</tr>
<tr>
<td>IPEC</td>
<td>International Programme for the Elimination of Child Labour</td>
</tr>
<tr>
<td>IWGDMF</td>
<td>International Working Group for Disease Monitoring and Forecasting</td>
</tr>
<tr>
<td>JRR</td>
<td>jackknife repeated replication (for variance estimation)</td>
</tr>
<tr>
<td>LCS</td>
<td>labouring children survey</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>LFS</td>
<td>labour force survey</td>
</tr>
<tr>
<td>MDMA</td>
<td>an “empathogenic drug”, commonly referred to as “ecstacy”, “mandy” or “molly”, depending on its form, degree of purity, etc.</td>
</tr>
<tr>
<td>MMP</td>
<td>Mexican Migration Project</td>
</tr>
<tr>
<td>MoS</td>
<td>measure of size</td>
</tr>
<tr>
<td>MSEs</td>
<td>micro and small enterprises</td>
</tr>
<tr>
<td>MSM</td>
<td>men who have sex with men</td>
</tr>
<tr>
<td>NGO</td>
<td>non-governmental organization</td>
</tr>
<tr>
<td>NSAC</td>
<td>Nepal South Asia Centre</td>
</tr>
<tr>
<td>PBDEs</td>
<td>polybrominated diphenyl ethers</td>
</tr>
<tr>
<td>PD</td>
<td>psychiatric diagnosis</td>
</tr>
<tr>
<td>PPS</td>
<td>probability proportional to size (sampling)</td>
</tr>
<tr>
<td>PSF</td>
<td>primary sampling frame</td>
</tr>
<tr>
<td>PSU</td>
<td>primary sampling unit</td>
</tr>
<tr>
<td>QCS</td>
<td>quick count survey</td>
</tr>
<tr>
<td>RA</td>
<td>rapid assessment</td>
</tr>
<tr>
<td>RDS</td>
<td>respondent-driven sampling</td>
</tr>
<tr>
<td>RDSAT</td>
<td>RDS Analytical Tool (software for analysing RDS-based data)</td>
</tr>
<tr>
<td>RMSE</td>
<td>root mean square error</td>
</tr>
<tr>
<td>SES</td>
<td>socio-economic status</td>
</tr>
<tr>
<td>SIPP</td>
<td>Survey of Income and Program Participation</td>
</tr>
<tr>
<td>SLID</td>
<td>Survey of Labour and Income Dynamics</td>
</tr>
<tr>
<td>SRS</td>
<td>simple random sampling</td>
</tr>
<tr>
<td>SRSWOR</td>
<td>simple random sampling without replacement</td>
</tr>
<tr>
<td>SSFs</td>
<td>secondary sampling frames</td>
</tr>
<tr>
<td>STI</td>
<td>sexually transmitted infection</td>
</tr>
<tr>
<td>StrCon</td>
<td>stratum of concentration</td>
</tr>
<tr>
<td>SWs</td>
<td>sex workers</td>
</tr>
<tr>
<td>TLS</td>
<td>time-location sampling</td>
</tr>
<tr>
<td>UN</td>
<td>United Nations</td>
</tr>
<tr>
<td>UNICEF</td>
<td>United Nations Children’s Fund</td>
</tr>
<tr>
<td>USU</td>
<td>ultimate sampling unit</td>
</tr>
<tr>
<td>UWFCL</td>
<td>Unconditional Worst Forms of Child Labour</td>
</tr>
<tr>
<td>VDT</td>
<td>venue-day-time (sampling)</td>
</tr>
<tr>
<td>WFCL</td>
<td>Worst Forms of Child Labour</td>
</tr>
</tbody>
</table>
Chapter 1
Surveying child labour

1.1 Introduction

Child labour is a global issue of great concern. Detailed and up-to-date statistics on working children are needed in order to determine the magnitude and nature of the problem, identify the factors leading to child labour, reveal its consequences, generate public awareness of the related constellation of issues, and formulate policies and projects to combat it (International Labour Organisation, 2004a).

Data on child labour may be obtained from diverse sources, often used in combination. These include national population censuses and secondary and administrative sources for baseline information. Data on child labour also come from general household-based sample surveys such as surveys on the labour force, living conditions, household income and expenditure, demography and health. Such surveys normally do not produce detailed data on child labour, but they can yield information that is useful for analysis of the situation concerning child labour. Attaching child labour modules to such household-based surveys is also a potential source of information. More comprehensive and pertinent information on child labour requires special studies and surveys focussed on the subject. Apart from household-based child labour surveys, different instruments include rapid assessments, establishment surveys, school-based surveys, community-level enquiries, street children surveys, and baseline studies undertaken in the context of specific intervention projects. The various sources are generally complementary.

Collecting comprehensive data on child labour is a challenging task, and no single survey method can satisfy all data needs (ILO, ibid). This applies with particular force when surveying what are called the worst forms of child labour, especially the unconditionally worst forms. It can be extremely difficult to make contact with children engaged in such forms of labour for collecting the necessary information. Worst forms of child labour usually remain hidden and the necessary sampling frames do not exist for their enumeration. Nor can the required samples be designed and selected without prior information on the location, characteristics and circumstances of the children engaged in worst forms of labour. Consequently, regular household-based surveys are largely ineffective for this purpose, and special sampling and enumeration procedures must be employed.

Sampling strategies for household-based child labour surveys have been elaborated in the ILO book Sampling for Household-based Surveys of Child Labour (Verma, 2008). For many purposes, such surveys must be supplemented by data collection from other sources.

The present book addresses the special considerations involved in the design of non-household based surveys aimed at estimating the prevalence and nature of child labour, especially surveys focussed on particular sectors or activities of worst forms of child labour.

In fact, the sampling techniques elaborated in the following pages have a much wider application than the field of child labour alone. They are relevant and applicable to
surveys in many other areas and circumstances. Nevertheless, the work has been
developed in the context of and for specific applications to child labour studies promoted
and supported by the ILO.

Section 1.2 briefly reviews sampling aspects of household-based versus child labour
surveys of other types, noting circumstances in which alternative approaches are
required. Surveys very different from normal household-based surveys may be required
to enumerate working children not living at home: children living away without close
contact with their family; children working at mobile, difficult to locate and identify, or
even unknown places; and working children who are not directly approachable.

Section 1.3 considers the concept and characteristics of elusive populations of labouring
children, and their consequences for the sampling process: for the frame, for sample
selection and for sample enumeration. In the context of sampling methodology, elusive
refers to populations for which, for diverse reasons, it is not possible to define, draw
or implement adequate samples using the normal procedures of general population
sampling. Such elusive populations tend to be ill-defined, heterogeneous, isolated,
rare, mobile, and often hidden from and sensitive to contact. These characteristics
make all aspects of the sampling process difficult. We identify four primary aspects
of elusiveness: lacking a sampling frame; the population and traits being rare; the
population being mobile; and populations being reclusive, i.e. being invisible, or being
unwilling or unable to co-operate in statistical surveys. These aspects of elusiveness
provide the framework for organising the various sampling methodologies developed in
the following chapters.

Section 1.4 provides an executive summary, a chapter-by-chapter overview of the
contents of this book.

1.2 Household-based vs. other types of child labour
survey

1.2.1 Surveying child labour through households

Household-based sampling provides an efficient approach to estimating the prevalence
and characteristics of the predominant forms of child labour for children living in private
households, irrespective of whether the work is performed at home or outside.

The main objective of a household-based child labour survey is to provide information
on the phenomenon of child labour – its prevalence, distribution, forms, economic
sectors etc. as well as its conditions, characteristics and consequences. The target
of these surveys is children, and also their parents or guardians living in the same
household.

There can be large differences of scope, content and methodology between household-
based surveys of different types.

Such surveys may be conducted as stand-alone surveys, or as separate but linked
operations, or simply as modules attached to other national household-based surveys
such as a labour force survey. The household survey content may be detailed and
specialised, providing information on the dynamics of child labour or gross flows between different child labour categories; or it may be confined to a few basic characteristics of working children. The primary objective of the survey may be to measure the prevalence of child labour, or it may be to investigate in-depth the circumstances, characteristics and consequences of child labour. The survey may focus mainly on child labour, or on child activity more broadly, or may be even wider in scope to include more general information about children beyond their economic and non-economic activity. All these variations have important sampling implications.

We comment further on these aspects in Section 2.13. For a detailed elaboration of the sampling techniques, see Verma (2008) referred to above.

1.2.2 Surveying working children not living at home

The appropriate survey approach depends on the circumstances of child labour – specifically, on the relationship of the child and of child labour with the household. Several scenarios may be identified.

(1) Children living and working at home.

(2) Children living at home but working away, such as selling goods, portering, doing day labour.

(3) Children living at home and working away, but with insufficient information available to their parents/guardians on the location or type of work they are doing.

(4) Working children not living at home, but with sufficient connection to their household to be selected, identified and traced from there.

(5) Children not living at and insufficiently connected to the household, but working at some fixed, more or less identifiable location.

(6) Children not living at and insufficiently connected to the household, who are working at mobile, difficult to locate and identify, or even unknown places.

(7) Working children approachable at best only through intermediaries – through adults or other children.

Situations (1)-(3), and possibly a part of (4) may be covered through household-based surveys. These cover a large majority of labouring children, but not necessarily of those engaged in the most hazardous or the worst forms of child labour, or living or working in the most adverse circumstances. Surveying situations (5)-(7), and possibly also (4), generally require data collection approaches other than the conventional household survey. These situations and methods are characterised by their diversity, arising from the need to deal with the diverse characteristics, circumstances and problems of child labour and to meet different survey objectives. Alternative approaches are also required when the objective is to focus on specific sectors or types of child labour activity, when their concentrations cannot be captured efficiently or the required information cannot be obtained with sufficient detail and accuracy through general household-based surveys. For example, it may be necessary to enumerate working children at the place of their work based on a sample of such places.
1.3 Elusive populations of labouring children

1.3.1 The concept of elusive populations

Many populations of labouring children have certain characteristics which makes them elusive to the sampling process. By elusive populations we mean populations for which - by virtue of their characteristics, or of the lack of suitable sampling frames, or difficulties in obtaining the required information - adequate samples cannot be defined, drawn or implemented using the normal procedures of general population sampling. Such problems arise when no frame, or at best very partial frames or lists, are available for sampling, or when many units are not available or willing to participate in the survey. A large proportion of the units stay away from the sampling frame, or from enumeration if selected for the survey. Such problems appear in a wide variety of surveys, including surveys on child labour.

Both stages – that of selecting a representative sample, and of successfully enumerating the units selected – can present severe problems and require special procedures. Problems can arise from various sources. For instance, certain categories of persons may not be covered in the sampling frames used for sample selection (such as recent arrivals, homeless persons, other ‘hidden’ populations); there may be difficulties in locating the selected units due, for instance, to their mobility or a lack of information about their whereabouts in the frame; the identification of persons of interest, even after they have been located, may require information beyond what is available in the frame; language differences and possibly also religious or other cultural differences can present obstacles. And of course, the individuals may not wish to be located, contacted or interviewed for various reasons. Such difficulties are often highly selective – the most elusive population groups can be precisely the ones of the greatest policy concern. We may also include in this group surveys of what may be considered ‘normal’ populations, but surveys which are subject to serious problems of under-coverage or of non-response in particular cases. These may result from the nature of the units, from the type of information sought in the survey, or the particular conditions under which the survey is conducted.

Addressing sampling issues requires the identification of the special characteristics and circumstances of the populations to be sampled. Kish (1991) in a review paper enumerates over 30 categories of elusive population and organises them into ten types or classes on the basis of shared characteristics in relation to the sampling process. Motivated by that effort, below we have identified and reorganised the ones most relevant in the context of child labour surveys. We include in the concept of elusiveness systemic problems in sample selection as well as in implementation of the selected sample (cf. Kish, ibid).

1.3.2 Characteristics of elusive populations of labouring children

(1) Ill-defined population

This refers to the existence of a fundamental vagueness in the definition of the population, beyond the usual problems relating to its precise demarcation in content, space and
time, and the usual problems of coverage errors to which almost all sample surveys are subject. Of particular importance in labouring children surveys is the major effect of the definition adopted as to what constitutes ‘child labour’. The precise definitions of the populations of interest can be vague. For example, we can distinguish between labouring children vs. child activity vs. children’s surveys. Among these the most restrictive are surveys of labouring children, which focus on the study of conditions and consequences of child labour, but even these vary depending on the precise definition of what is included as ‘child labour’.

(2) Great heterogeneity

This refers to populations involving diverse types of unit to be covered in the same survey – for example units differing greatly in size, sector and nature of economic activity. These different units may require different kinds of data, different degrees of detail, different survey periodicities and sample sizes, etc. Even different data sources and modes of collection may apply. The units may be covered by different sampling frames, with possible overlaps. As a consequence of the above, the required sampling procedures may also differ for units of different types. A particularly challenging sampling problem concerns control over sample size and distribution for very different unit types in the same survey.

There are several important consequences of this heterogeneity. Most importantly, it is unlikely that a single approach or methodology will suffice to cover the diverse types of child labour. Normally, a mixture of approaches will be necessary. Secondly, it is very likely that in practice it is impossible to capture all the diverse subpopulations of interest: often it will be necessary (and wise) to concentrate resources and effort over fewer, most important groups. Of course, ‘most important’ does not automatically mean the largest. Often smaller groups may present problems of greater policy concern.

(3) Small/rare population

Here the characteristic feature is that sampling the whole population with ‘normal’ procedures (such as equal probability sampling of elements) does not yield a sample of adequate size for the subpopulations or domains of interest because of their small size. Procedures are required for more intensive and targeted sampling for the purpose. This covers several different situations, including the following common ones.

(i) Rare traits. The objective is to estimate small proportions possessing certain specified traits (e.g. children who are engaged in a certain type of labour). It is not the base population which is necessarily small, but the population possessing the characteristic of interest which forms the numerator of the (small) proportion.

(ii) Rare populations. Here the objective is to estimate measures (proportions, mean values etc.) for a small population (e.g. some characteristic of children engaged in a certain type of labour). In this case it is the base population (the denominator of the statistic) which is small.

(iii) Small areas and other small domains. Here estimates are required not for some selected subpopulations but more generally for small partitions making up the whole. Samples of any reasonable size (or even of some maximum feasible size) cannot meet these reporting requirements. The characteristic feature is the need to use less complex
or up-to-date but much larger data sets from alternative sources (registers, censuses) in conjunction with data from the more complex but smaller sample survey. A variety of small area estimation techniques have been developed for the purpose (Rao, 2003).

(4) Uneven distribution, patchy concentrations

This refers to the situation when the subpopulation of interest exists in large and uneven concentrations, with limited or no prior information on the location or structure of those concentrations. Typically, large parts of the total population or space have a very low density of the subpopulation of interest, or are even completely empty of it. A general sample design, not able to take into account the patterns of concentration, is unlikely to provide an adequate sample. Also, it cannot guarantee prior to the survey that a sample of sufficient size would be obtained for the subpopulation of interest. The unevenly distributed population of interest may not necessarily be very small in relation to the total, but being small would further increase the problems of sampling.

A common characteristic of many groups of labouring children is their clustering together and highly uneven distribution in the host population. Depending on the patterns of distribution – and on whether information is available on these patterns before sample selection – it may become difficult, or even impossible, to obtain good samples for the population of interest using normal procedures. Special techniques such as adaptive cluster sampling are likely to prove useful in such circumstances (Thompson and Seber, 1996). Alternatively, general spatial sampling may have to be abandoned in favour of sampling according to (possibly transient) points of concentration, using techniques such as time-location sampling.

(5) High mobility, temporal instability

Many labouring children do not have a fixed place of work. Patterns of movement are very diverse depending on the situation and nature of work. Special but also very specific procedures are required to catch reasonably representative samples of them. Under-coverage and double-counting can both be problems, one not precluding the other. Another dimension worth mentioning is that many children move in and out of child labour depending on the season, vacation period, etc.

(6) Hidden segments of the population, difficult to identify and locate

This is reflected in the absence, or serious incompleteness, of the sampling frame. Typically, available frames do not list labouring children directly. At best, only higher-stage units containing children, such as areas, establishments or households, can be directly identified. Additional information involving a listing or interviewing operation is usually required for identifying children. Completeness of coverage of the target population depends on the quality of such information. The situation is worse when the higher stage units themselves remain hidden.

(7) Reclusive and sensitive character of units

Reclusiveness implies shying away from contact with outsiders. In the context of a statistical survey, we use the term reclusive to refer to a tendency among units in the target population to stay away from participation. This may result from different aspects
of the unit or of the survey situation. A unit may be reclusive because it is unwilling to participate, is afraid to participate, believes that it will lose too much by participating, lacks freedom to participate, or is elusive simply as a result of its isolated or hidden existence. Sensitivity of the information can also contribute to reclusiveness of the respondent. Sensitive refers to nature of the information sought, or more precisely, to the reaction of the respondent to the enquiry. One common reason for this is the threat or shame the respondent may feel in providing the information. Special procedures need to be employed to recruit and access a reclusive respondent. The respondent may need to be both encouraged and helped in providing the information sought.

Dimensions of elusiveness

The above population characteristics affect all aspects of sampling: the frame, sample selection, and sample implementation. Elusive populations, seen from the point of view of the consequences for sampling methodology, can be said to have four aspects: (1) populations lacking an adequate sampling frame; (2) rare populations and traits; (3) mobile populations; and (4) reclusive populations.

These aspects are distinct though often overlapping. Furthermore, they tend to form a hierarchy: while problem (1), lack of an adequate sampling frame, may exist in surveying any type of population, surveying rare populations tends to be subject to both problems (1) and (2), surveying mobile populations to problems (1)-(3), and surveying reclusive populations to all the four problems (1)-(4). Such a hierarchy is a very common scenario, though not necessarily the pattern in all situations.

We need to explore the feasibility of special procedures which have been developed and applied in order to capture elusive populations of labouring children. Among the potentially useful methods described later are multiplicity sampling, adaptive cluster sampling, time-location sampling, capture-recapture sampling, snowball sampling, and respondent-driven sampling. However, the practicality of some of these procedures in the typical context of official statistical agencies - as distinct from more informal (flexible) research settings – needs to be demonstrated.

1.4 Executive summary: Chapter-by-chapter outline of the contents

The following chapter-by-chapter outline of the contents is aimed at providing the reader with useful pointers to the issues covered.

1.4.1 Substantive and technical background

Chapter 2: Child labour situations, data needs and sources

Child labour situations and problems are diverse and specific to the particular situation. A good understanding of the situation to be studied is essential for the choice of an appropriate survey methodology and sampling strategy. The chapter illustrates the variety of situations and types of child labour in order to provide the necessary background for the diverse sampling techniques discussed in subsequent chapters.
Various worst forms of child labour are described, covering diverse sectors: child domestic work; agriculture including commercial crops; fishing and aquaculture; mining and quarrying; manufacturing including handicrafts; construction; street work and the informal sector; and also various ‘unconditionally worst forms of child labour’ including child trafficking, commercial sexual exploitation, forced or bonded labour, engaging or living in armed conflict, and children’s involvement in illicit activities, in particular in drug trafficking.

In each case, the type of child labour involved, the condition of children working in the sector, and the main hazards involved in the work are described, noting their implications for sampling and related aspects of data collection methodology. Details of the national studies illustrated are given in Annex A. Two types of illustration are included: description of specific situations, and a more general review of problems in the type of child labour being considered.

Then the chapter summarises data sources for different types of child labour. Sources of information on child labour include: household-based surveys; supplementary sources or surveys (school-based surveys, community-level inquiries, general national household surveys, censuses, other secondary sources); employers’ surveys; establishment surveys; baseline surveys and studies; and rapid assessments. A description is provided of household-based surveys and rapid assessment studies. Household-based surveys of child labour generally use large probability samples of the general population; the information collected tends to be extensive rather than in-depth. By contrast, rapid assessment studies are small-scale but intensive; they may collect information from many different sources. National household-based surveys of child labour on the one hand, and typical rapid assessment studies on the other, form two ends of the range of application of the various sampling techniques addressed in the following chapters.

Chapter 3: Basic sampling and estimation procedures

The primary objective of the chapter is to provide a reminder of some basic principles concerning sample design and selection which underlie the more specialised techniques discussed in this book. These include principles of probability sampling, common departures from simple random sampling (stratification, clustering, unequal selection probabilities), probability proportional to size (PPS) sampling, and systematic sampling. A typical, commonly used type of stratified multi-stage design is described to bring out salient points concerning practical sample design.

The chapter reviews basic principles concerning weighting of sample data and estimation from a sample. Weighting of sample data is an essential step in survey analysis. We review sources of information for weighting and present a step-by-step procedure for weighting – computation of design weights, adjustment for non-response, calibration against external standards, and trimming and scaling of the weights. Numerical illustrations of the procedure are provided. The chapter continues with a discussion of practical variance estimation procedures and the concept and analysis of design effects, and concludes with presentation of a typology of errors in survey data.
1.4.2 Sampling from imperfect frames

Chapter 4: The sampling frame

This chapter has two themes. The first is to describe characteristics and common shortcomings of sampling frames. Starting from basic concepts (the survey population, the sampling frame for single-stage and for multi-stage sampling), the chapter identifies the essential requirements for any sampling frame, as well as desirable properties which a frame should have in order to yield a reasonable probability sample.

An example, from a set of baseline surveys in Bangladesh, is provided to illustrate the concepts defined. The surveys cover selected sectors; most of the establishments in the sectors are small and a high proportion employ child labour. This example illustrates several common aspects concerning sampling frames, such as the problem of surveying a population in the absence of an existing sampling frame; the cost and quality implications of the quantity of information to be collected for each unit during the operation; the possibility of economising by sharing the costs between different surveys; use of the listing operation for making substantive estimates; and special problems related to the type of units in the frame (e.g. establishments versus other locations where working children are found). We also make some critical remarks on the approach followed in this particular illustration in order to bring out some practical points.

We discuss basic requirements and desirable quality, efficiency and cost-related properties of area frames. Common problems with area frames include: failure to cover the population of interest exhaustively; errors and changes in area boundaries; inappropriate type and size of units; lack of information on size and other characteristics of the units; and high cost.

Next we consider uses and problems of list frames. A crucial aspect of sample selection from lists is the correspondence between listing, sampling and analysis units. We have a ‘perfect list’ when there is perfect one-to-one correspondence between these units. In practice, lists are subject to imperfections, and in the worst case, we may have no frame at all. The chapter discusses these problems from a practical perspective.

The second theme of the chapter is to develop and explain the following concepts. Notwithstanding frame imperfections noted above, in most applications of conventional sampling we have one-to-many or one-to-one (or sometimes also one-to-none) correspondence between sampling and analysis units: any analysis unit is associated with at most one sampling unit. We term this direct sampling. By implication, its complement is ‘indirect sampling’ when such correspondence between sampling and analysis units is lacking. There are two forms of indirect sampling. First, a widespread situation encountered in surveying elusive populations is the presence of many-to-one and many-to-many links between sampling and analysis units - an analysis unit being associated with more than one sampling unit. This gives rise to sampling with multiplicity. The multiplicity estimator is the common link between many of the sampling techniques discussed in this book. We discuss technical details of this estimator at some length. The concept and formal framework of the related ‘weight-share estimator’ is also introduced.
Then there are situations when the sample has to be obtained by exploiting links between analysis units themselves, rather than primarily between analysis units and sampling units. This is *link trace sampling*, and we develop its applications in later chapters.

**Chapter 5: Sampling establishments employing children**

Most of the establishments using child labour tend to be small and informal sector establishments. However, children may also be employed by larger establishments. The sampling considerations and procedures differ between the two types of situation.

In this chapter we first discuss sampling aspects which apply equally to both categories of establishments. These concern issues of stratification and sample allocation, and also procedures for selecting children within establishments included in the survey.

The main difference in sampling procedures between establishments of the two types concerns the selection of establishments. For large and medium-sized establishments, samples are often selected directly from lists. The chapter describes sampling procedures for selecting establishments from list frames, explaining specific sampling schemes such as collocated and sequential Poisson sampling.

The emphasis of the chapter, however, is on the second type of design: special considerations which arise in the design of samples for surveys of small and informal sector establishments. Such units, like households, are small-scale, numerous and widely dispersed in the population. While they may be very unevenly distributed in the population, usually their patterns of distributions are not hidden, and they are usually sufficiently numerous not to present the special problems of sampling rare populations. Also, a majority of informal sector workers in developing countries have little to conceal and therefore can be surveyed using the direct method of enquiry used in statistical surveys of many other types. Furthermore, in this chapter we confine the discussion to issues in sampling small and informal sector units employing children which are based in households or some other fixed premises, even if the persons working are themselves mobile. Rare, mobile and hidden units present additional problems addressed in subsequent chapters.

The commonly used samples for small and informal sector establishments are area-based and involve two (or sometimes more) stages of sampling, the last stage involving listing and enumeration of establishments in the areas selected into the sample.

The technical issues discussed include:

- characteristic features of small and informal sector establishments and their consequences for survey design;
- the choice between integrated multi-sectoral and separate single-sector surveys;
- stand-alone versus surveys attached as modules to other surveys;
- the construction and use of ‘strata of concentration’ of different types and sectors of establishments to control distribution of the sample;

---

1 There are also examples in national surveys of child labour where list samples have been used even for small and informal sector establishments; usually such applications have to be preceded by major listing operations.
1.4 Executive summary: Chapter-by-chapter outline of the contents

- procedures for selecting establishments within sample areas; and
- issues in survey implementation.

A number of illustrations from surveys of establishments employing children in developing countries are provided. These include a multi-sectoral survey on hazardous child labour in Bangladesh, and informal sector surveys in Pakistan and Uganda. Critical comments are offered based on evaluation of experience with this type of sampling, noting that in many situations, the experience has, unfortunately, been less than positive.

1.4.3 Rare populations

Chapter 6: Sampling rare populations

The characteristic feature of a rare population is that sampling the whole population with normal procedures does not yield a representative sample of adequate size for the subpopulations of interest because of their small size. The chapter discusses procedures for more intensive and targeted sampling in order to capture rare populations more effectively and efficiently.

In surveying different types of child labour, the rare populations of interest - working children - are generally unevenly distributed among the general population of children. A successful sampling strategy involves identifying and making use of the pockets of concentration and patterns of distribution of the rare population of interest. There are five aspects of the strategy, which are distinct but are often used in combination for greater effectiveness. The chapter discusses various useful techniques concerning each of these aspects.

(1) Locating concentrations of the rare population using existing large-scale sources.

(2) Partitioning the frame according to the degree of concentration of the rare population, using techniques such as:
   - employing sequential tests to remove areas nearly empty of the rare population;
   - using lists to separate out areas containing the rare population from other areas;
   - screening areas to sort them into strata with different concentrations of the rare population.

(3) Oversampling strata of concentration, making use of the patterns of concentration identified.

(4) Listing, screening and two-phase sampling, aimed at the identification and sampling of the final elements (households, children).

(5) Special procedures to increase selection probabilities of units in the rare population and thereby increase the achieved sample size. In this chapter, we summarise one particular technique, namely cumulating information over a number of surveys. There are a number of other procedures, discussed in subsequent chapters, which can increase selection probabilities, such as multiplicity sampling, multi-frame sampling and adaptive cluster sampling. The common link between these procedures is that they involve sampling with multiplicity.
Chapter 7: Multiplicity sampling

In this chapter we describe and illustrate fundamental features of the multiplicity sampling approach, and indicate situations where the approach may be useful in surveying rare populations of labouring children. The basis of multiplicity sampling is the relationship between sampling units and analysis units. Sampling with multiplicity arises when an analysis unit is linked to more than one sampling unit. A number of examples are provided to illustrate the use of such sampling. Discussed also are important principles in choosing multiplicity 'counting rules'.

An attempt is made in the chapter to identify some conditions which may make multiplicity rules preferable to conventional rules, emphasising that consideration must be given to the joint effect of sampling, coverage and measurement errors in selecting the counting rules for multiplicity sampling. Potential advantages and uses of multiplicity sampling are discussed, identifying situations where multiplicity sampling may be useful.

The chapter also notes the limitations and problems of the method. Successful implementation of the multiplicity sampling procedure depends on salience of the family and social relationships on which its counting rules are based, and on salience of events and of the objects of the enquiry for the respondent. Reporting biases are often larger for multiplicity counting rules than for ordinary unitary counting rules. Another concern is the increased complexity. There can also be serious ethical, confidentiality and privacy concerns in using the method.

Next, the chapter explains procedures for estimation with multiplicity sampling. The standard multiplicity estimator takes the weights as inversely proportional to the unit's multiplicity. As a detailed numerical illustration of the multiplicity estimation procedure, we consider a household survey designed to estimate the total number of child domestic workers from a probability sample of households. Child domestic workers who were living and/or were working in a sample household were taken into the sample, rather than only those who were currently residents of the selected households.

Chapter 8: Multi-frame sampling

This chapter discusses the use of multiple sampling frames in the context of child labour surveys. When no single sampling frame can provide a complete representation of the target population, the use of multiple frames can reduce coverage errors.

Generally, the multiple sampling frames overlap and procedures are needed to deal with this, for example by constructing a new single frame without duplicates or by accounting for the duplicates in the estimation procedure. Both these possibilities are discussed, describing the main methods of removing the duplicates and constructing non-overlapping frames, as well as the main procedures for accounting for duplications and estimation from overlapping frames. The procedure is illustrated numerically, involving multiple frames targeting working children.

Experience with multi-frame sampling, particularly in the field of child labour, is very limited in developing countries. Nevertheless, it is a technique which has the potential to be usefully employed in this field, especially when the situation has the following characteristics: the main frame for the survey is a frame of area units; the area frame
provides reasonably complete coverage of the target population, even if only implicitly; and lists of analysis units can be found, each with a high concentration of particular subgroups in the rare population but not necessarily accounting for a large proportion of those subgroups. The lists may overlap, or may contain only incomplete information for the identification and linking of individuals.

The use of multiple frames can involve multiplicity in the selection of units at two levels: a unit may appear in more than one frame; and within any of those frames, the unit may appear more than once. This is not an unusual situation. We develop and discuss the multiplicity estimator in a general form to cover such situations. Finally, the chapter considers practical aspects of implementing this procedure in the context of a child labour survey, illustrating the procedure by two numerical examples worked out in detail.

**Chapter 9: Adaptive cluster sampling**

Adaptive sampling is a technique designed to obtain more adequate and efficient samples for a population which is rare and very unevenly distributed. Starting from a conventional initial sample, the technique involves selecting an additional sample in the neighbourhood of points where a concentration of the population of interest is found during implementation of the initial sample. The technique is most effective when the population of interest tends to be concentrated in relatively few and large clusters but little information is available on the extent, location and patterns of its concentration. Examples of such populations include street children, children engaged in street trades and child beggars.

This chapter identifies the sort of situations in which adaptive sampling might be useful, provides examples, and notes potential advantages and limitations of the adaptive sampling procedure. The technique is described, again with detailed numerical illustrations.

A number of further technical aspects are touched upon, such as: unequal unit selection probabilities; stratification with adaptive sampling; multistage sampling; multivariate criteria for adaptive sampling; adaptive sampling using ‘order statistics’; arbitrary rules for stopping the adaptive process; problem of imperfect detectability; and aspects of the estimation procedures with adaptive sampling.

The chapter also discusses a number of implementation issues: whether to introduce adaptive sampling; criteria, definitions, rules for adaptive sampling; choosing size of the initial sample; controlling sample size, etc. Guidelines and recommendations are made concerning the approach, the design, implementation in practice, and data analysis and evaluation.

Finally, the procedure is illustrated in detail on the basis of an artificially constructed small population (60 small area segments). Areas containing more than a few (5) labouring children form potential seeds for adaptive additions. The procedures are clarified by developing a graphical illustration of the geographical order of the area units. The illustration demonstrates how adaptive sampling can help in locating large concentrations of the population of interest by increasing the chance of their appearing in the sample, and hence also in obtaining a larger number of elements of interest (such as children working in a particular sector), thus providing a basis for more detailed
1. Surveying child labour

surveys of these. We also show how adaptive additions may be divided into waves, permitting termination of the process at different stages if desired.

1.4.4 Mobile populations

Chapter 10: Sampling mobile populations

This chapter addresses special problems and issues which arise in sampling mobile populations. In the context of sampling, the concept of ‘mobile population’ is more general than simply not having a fixed place of residence or work. It refers to situations when it is necessary or preferable to sample and enumerate units through their mobility (movement).

The population of interest can be mobile and difficult to locate and identify to varying degrees. The difficulties of enumerating such populations concern the identification of (i) who are the eligible respondents for the survey, and (ii) where and (iii) when to find them; also (iv) what information to ask them concerning their mobility, and (v) how to obtain the information. Another set of issues concerns (vi) how to use sample data to produce valid estimates for the population, and (vii) how to assess variances and biases to which those estimates are subject.

Procedures are required to establish links between mobile units and fixed reference points through which a probability sample of the units could be obtained. The chapter develops a framework to organise the variety of circumstances, problems and solutions encountered in sampling mobile populations. Four important concepts in the framework are: sampling locations, observation points, time segments, and ‘time-location primary sampling units’.

Location sampling is widely used to sample mobile populations. This type of sampling has been used, for instance, for sampling a variety of populations that tend to congregate in certain places by sampling those places and then persons present at those places. Data collection and estimation procedures for location sampling (also termed ‘centre sampling’, ‘intercept point sampling’, etc.) are described with numerical illustrations.

Next, we discuss procedures for time-location sampling. Such sampling is designed to monitor the flow of individuals through fixed locations during specified time segments. The flow may comprise individuals passing through chosen vantage points, arriving or leaving some place or institution, visiting a facility, etc. Examples of application in diverse fields are available mostly from developed countries. Also presented in some detail are a few examples of surveys of mobile populations in more difficult situations.

Finally, the chapter develops procedures for estimating the probability of selection and sample weight of a mobile individual. When the sample is taken at an ‘observation point’ at a given ‘time segment’, the probability of an individual being included in the sample is affected by the person’s frequency and pattern of visits to the locations and times covered in the survey population. The situation can be complex. Complex situations require a lot of information to be collected in order to take fully into account the variations in selection probabilities, but collecting all the required information is often not possible. In view of this, it is necessary to have an alternative, simplified and more practical strategy. Quantitative expressions are developed in the chapter for
variations in individual selection probabilities in a number of commonly encountered situations.

**Chapter 11: Capture-recapture sampling**

The chapter reviews the main stages of a capture-recapture survey, from identification of the population to estimation from survey data, with the objective of providing guidelines and recommendations concerning practical aspects.

The capture-recapture sampling technique involves taking two (or more) independent samples from the same population and using the overlap found between the samples to estimate the selection probabilities applied to obtain those samples and the total population size. Capture-recapture applications in the social field are usually based – in developing countries in particular – on a combination of sample surveys and administrative sources.

The chapter outlines the basic approach. Situations in which capture-recapture sampling might be useful are noted. An illustration is provided from a study of street children in Cairo, based on two independent closely conducted sample surveys covering the same population. The objective of the illustration is to bring out practical concerns in applying the methodology.

The capture-recapture procedure is based on certain assumptions about the population and the manner in which the samples have been drawn from it. The chapter discusses these assumptions and their possible impact. Even though these assumptions are not always valid, the usefulness of the capture-recapture technique arises from two factors: often the procedure is rather robust against departures from the assumed statistical model; and furthermore, statistical procedures have been developed to control the effect of certain departures from the original simple model. We touch briefly the procedures for the use of more than two data sources.

Simulated data are used to provide detailed numerical illustrations of capture-recapture sampling. A number of situations concerning capture probabilities are demonstrated: when the probabilities are constant; when they vary with time; when they vary by behavioural response to capture; and when the probabilities vary by individual unit.

The chapter outlines procedures for analysis by domain through developing and using sample weights at the level of individual units. A simple weighting procedure for individuals visiting study locations with different frequencies is developed and numerically illustrated.

The samples may be selected in multiple stages, with the two (capture and recapture) samples sharing some or all higher stage units. Basic methodology of capture-recapture with multistage sampling is developed and numerically illustrated.

A major technical contribution of the chapter concerns the development of procedures for the estimation of sample weights in a more general situation. The reason for weighting sample data for producing estimates is to compensate for variations in unit selection probabilities. These include: variations in selection probabilities arising from survey conditions (conditioning, unknown population size, incomplete frame, haphazard differences in coverage and ‘catchability’); and variations in selection probabilities arising from survey design and implementation (samples dependent by
design, differences in sampling rates, inherent differences in unit selection probabilities, non-response). Procedures for putting together all these effects are explained.

Chapter 12: Controlled selection and balanced sampling

Controlled selection is a procedure to control the structure of the sample beyond what is possible with ordinary independent selection within strata. Controlled selection permits extra control while conforming to the requirements of probability sampling. Surveys, in particular of mobile and other difficult-to-access populations, often have to be restricted to a limited area and to a small number of primary units. Controlled selection is a very useful technique when one has to select a small sample of primary units, but at the same time ensure that it is ‘balanced’ and ‘representative’ of the population in terms of many characteristics (or control variables).

In this chapter three detailed numerical illustrations are presented and discussed with the objective of explaining the procedure in concrete terms. Following a simple illustration of controlled selection, we provide examples to illustrate in greater detail implementation of the controlled selection procedure when the sample is stratified in terms of a number of variables.

Controlled selection can also be viewed in the context of the modern theory of balanced sampling, thus providing the possibility of dealing with a wider range of issues and more efficient sampling algorithms. Balanced sampling is a more general technique than controlled selection. Its objective is to control the distribution of the achieved sample according to some specified set of auxiliary (control) variables. These control variables may include one or more stratification variables, which correspond to controlled selection. In the present chapter, our interest is in the use of balanced sampling procedures as a tool for achieving controlled selection. The same above-mentioned three examples are reformulated in terms of the balanced sampling procedure. Examples of balanced samples generated by the procedure are shown.

We offer some concluding remarks on the relationship between the controlled selection and balanced sampling. In the context of stratification, the two procedures are the same, even though the steps involved in their application may differ in form. Both require an understanding and clear specification of the stratification controls required, as is clear from the detailed illustrations presented.

1.4.5 Reclusive populations

Chapter 13: Snowball sampling

This chapter discusses snowball sampling in the practical context of surveying reclusive populations of labouring children. In this context, the term snowball sampling is taken to refer to a convenience sampling mechanism in settings characterised by the lack of a serviceable sampling frame. We begin with a discussion of the concept and methodology. With snowball sampling a unit of the target population can enter the sample through direct selection into the initial sample, or by being identified (‘named’) for inclusion by someone already in the sample. There are a number of parameters which define the design of a snowball sample: the number of waves, number of contacts to request, and criteria for including a participant in the sample. If the aim of a study is primarily
1.4 Executive summary: Chapter-by chapter outline of the contents

exploratory, snowball sampling offers many practical advantages. However, over the years snowball sampling has advanced to become a more sophisticated method of sampling and data analysis.

Next, we discuss pros and cons of the approach. There is a range of potential advantages in favour of the approach. The primary advantage of the method is its success in identifying individuals from unknown populations and from small, hidden groups dispersed within a large population; it provides a means of accessing social groupings which are vulnerable or difficult to penetrate, and a means for obtaining respondents in settings where trust is required to initiate contact.

The most important deficiency of the snowball method concerns the quality of the data, and in particular selection bias which limits the validity of the sample. The snowball method is generally not an effective tool for producing reliable estimates of total population size or population aggregates of other variables, and is of limited use in estimating proportions and means.

The chapter outlines a simplified procedure for estimating size of a hidden population. In this procedure assumptions are made about probability nature of the procedures of selecting the initial sample and following up on the basis of referrals. These are rather unrealistic assumptions, giving unrealistic results.

In order to convey an impression of implementation aspects, the chapter provides a number of illustrations from surveys in diverse settings, including surveys concerning homeless people in Hungary, Japanese migrants in Brazil, Mexican migrants to the US, and persons with drug abuse and psychiatric disorder in Brazil. The examples contain both positive and negative experiences in applying the method.

A major effort has been made in the chapter to construct numerical simulations for demonstrating some important aspects of the method: how the snowball sample develops and how it might be influenced and controlled. Two aspects in particular are examined in detail: how a proportionate sample may be developed, and how the snowball procedure may spread to include more and more hidden segments of the target population.

The chapter concludes with a review of the practical aspects in implementing a snowball sample.

Chapter 14: Respondent-driven sampling

This chapter begins with a discussion of the concept and methodology of the respondent-driven sampling (RDS) method. The procedure is an improved variant of the usual snowball sampling. Both the procedures are types of chain-referral sampling. As with snowball sampling, a unit of the target population can enter the sample through direct selection into the initial sample, or by being identified for inclusion by someone already in the sample. In both cases, the process starts with a small number of peers, usually chosen non-randomly. However, as an improvement over ordinary snowball sampling, RDS is designed to produce a closer approximation to probability sampling. It incorporates features such as the direct recruitment of peers by their peers, a dual system of incentives (for participation and for recruiting), and recruitment quotas (e.g. a maximum of three recruits per respondent).
In order to illustrate implementation of the respondent-driven sampling procedure, the chapter provides a number of illustrations from surveys in diverse settings, for example surveys concerning: street children in Ghana; female sex workers in Brazil; gay persons in Uganda; and a study into the community impact of hate-crimes in Canada. The objective of the presentation is to illustrate how (and whether) the studies actually implemented the special features of the RDS procedure which distinguish it from ordinary snowball sampling.

Next, we clarify the basic assumptions of the RDS model, which are used to justify the application of Markov chain theory to the procedure. In the case of recruitment in the RDS method, the theory implies that after a sufficient number of waves, the characteristics of the individuals in the final sample are independent of the seeds' characteristics. Under the hypothesis that the probability of selecting each participant is proportional to his/her social network size, supposedly unbiased estimates for the population of interest can be produced and confidence intervals can be constructed around these estimates.

Concerning assessment of the RDS procedure, the chapter analyses the experience with two sets of studies. The first set involves performance comparisons of the RDS and alternative sampling approaches. The examples include:

- a study comparing snowball sampling, time-location sampling and RDS applications to surveys of men who have sex with men in Fortaleza, Brazil;
- a theoretical review of relative merits of time-location sampling, targeted sampling and RDS; and
- a study of the spread of injecting drug use and associated HIV infection in the Russian Federation and Estonia, comparing RDS to a sampling procedure using indigenous fieldworkers.

In the second set of examples, studies concerning assessment of RDS examine how well its procedures are implemented in terms of the theoretical assumptions of the model. Examples of studies undertaken for assessing validity of the RDS model assumptions include the following:

- a wide-ranging review of experience of RDS applications in developing countries (128 HIV surveillance studies from more than 28 developing and transition countries), providing insights into the requirements, problems and achievements of the RDS sampling methodology;
- a study (of young adult MDMA/ecstasy users in Ohio) illustrating a common problem in implementing link-trace sampling such as RDS, namely the difficulty in getting fast enough sufficient numbers of recruits;
- a study investigating the performance of RDS by simulating sampling from 85 known network populations; and
- a simulation study to assess the performance of existing RDS estimators in three areas concerning sensitivity of the method, namely sensitivity to the convenience sampling procedure used to select seeds for the initial sample, respondent behaviour in bringing in new respondents for subsequent waves, and the assumption that the sampling involved is with-replacement sampling.
1.4.6 Annexes

Annex A: Illustrations from studies of the worst forms of child labour

Child labour problems are very diverse and situation-specific. This annex complements Chapter 2, “Child labour situations, data needs and sources”. It provides a wide range of illustrations from national practices, covering different types of child labour activity in different settings. The illustrations have been compiled from original study reports.

Annex B: Efficiency of oversampling strata of concentration: a case study

Efficiency in sampling a rare population can be improved by stratifying the sampling frame according to the degree of concentration of the rare population, and then oversampling strata with higher concentrations. These procedures are developed in Chapter 6. The primary objective of this case study is to numerically illustrate the procedure, and provide a quantitative indication of the gain in efficiency which might be obtained.

It can also be useful and interesting for the reader to have hands-on experience of repeatedly selecting different samples to apply and test the procedures, and examine the estimates of sampling bias and variance so produced. For this purpose, this case study is accompanied by a data file providing a list frame of census enumeration areas, with a set of variables for each area which are useful for sample selection and estimation and for comparing sample estimates with the true population values as present in the frame.

Annex C: Applying sampling procedures in combination: a case study

Starting from a real situation as identified from a baseline survey in a community, this case study discusses scenarios for the development of a set of surveys and studies aimed at providing a comprehensive assessment of the child labour situation in the community. This scenario calls upon several of the sampling procedures discussed in this book, to be chosen from and articulated. Chapter 10 on sampling mobile population provides the main technical background for this discussion. For this case study, we draw on the report of one of three baseline surveys on child labour in Jamaica, namely a 2001 study of child labour covering the fishing communities of Rocky Point, Claredon and Old Harbour Bay, St Catherine.
I. Substantive and technical background
Chapter 2
Child labour situations, data needs and sources

2.1 Introduction

This chapter provides a description of diverse child labour situations and data needs (Sections 2.2-2.11), and of the main sources of data to meet those needs (Section 2.12-2.15). The concluding Section 2.15 has the specific objective of clarifying the role of the sampling procedures addressed in the following chapters towards filling gaps in the existing data sources.

2.1.1 Child labour situations and data needs

A. Objectives

As distinct from surveys of the general population of children or of working children, the purpose of this book is to describe sampling procedures appropriate for special groups, situations and types of child labour, for example:

- particular sectors of child labour activity;
- rare, mobile or reclusive populations of working children;
- child labour involving different degrees of hazard; and
- the worst forms of child labour going beyond simply hazardous work.

A good understanding of the situation to be studied is absolutely essential for the choice of an appropriate sampling and survey strategy. To this effect, the chapter identifies and illustrates the variety of special groups, situations and types of child labour in order to provide the necessary background for, and make more concrete and comprehensible, the diverse sampling techniques discussed subsequently.

B. Illustrations

Child labour situations and problems are very diverse and very specific to the particular situation. Hence it is necessary and useful to provide a wide range of illustrations from national studies, covering different types of activity in different settings.

For the convenience of the reader, we have moved examples from country studies of child labour to Annex A. In compiling these illustrations we draw liberally from the original sources such as research papers, national survey reports, official documents and in particular, publications of ILO. The originally published descriptions have been shortened, edited and paraphrased as necessary for the present purpose, but often they have also been quoted verbatim. We have considered it neither necessary nor helpful to rewrite published descriptions where they are already sufficiently clear and concise for our purpose. In each survey or item described in the illustrations, the original source of the material has been identified. Given the above, we have considered it unnecessary
to distinguish exact quotation from paraphrased text by the use of quotation marks etc. In any case, it would be cumbersome to maintain such distinction consistently and strictly.

The objective of these illustrations is not to describe or comment on the technical design of the particular surveys or studies cited. Rather, the objective is to convey the type of set-up and the child labour situation represented in each illustration. These factors dictate appropriate sampling strategies to be used for different purposes in different situations. The illustrations thus provide the context for subsequent discussion of the sampling techniques.

It should be remembered that “child labour is, in every instance, a highly localised phenomenon – located in specific places and at specific times, and conditioned and shaped by local circumstances” (Hindman, 2009, pp. xxvii). Nevertheless, there are commonalities between different situations, and the large variety of illustrations below hopefully captures some of those commonalities. Also, as noted in Section 2.2.5 below, “although the number of children in hazardous work is large, some of the most dangerous types of child work are concentrated in specific localities, specific occupations, specific tasks and specific age groups. Focusing energies on these pockets could go a long way towards generating the momentum needed to make progress” in understanding and combating child labour.

C. Outline of contents (Sections 2.2-2.11)

After introducing the concepts relating to child labour as formulated in the basic international conventions, especially those of hazardous work and other worst forms of child labour as two components of the worst forms of child labour (WFCL), each section of the chapter deals with one of the main sectors of hazardous child labour, concluding with comments on the various types of WFCL. The sectors are:

- domestic work
- agriculture, including commercial crops
- fishing and aquaculture
- mining and quarrying
- manufacturing, including handicrafts
- construction
- street work and the informal sector
- worst forms of child labour going beyond simply hazardous work, including: (1) child trafficking; (2) children in forced or bonded labour; (3) children engaged or living in armed conflict; (4) commercial sexual exploitation of children; and (5) children involved in illicit activities, in particular in drug trafficking.

For each sector, the following information is summarised:

- the type of child labour involved, the condition of children working in the sector, and the main hazards involved in the work; and
- examples from national studies (presented in Annex A) illustrating the above.
Each section considers the implications of the above for the sampling and related aspects of data collection methodology.

The above issues are discussed in greater detail for the sector concerning child domestic work (Section 2.3). This is because child labour in domestic work is one of the most widespread forms of child labour. In some aspects, a survey on child labour in domestic work lies between a general household-based child labour survey and a sector-specific survey.

2.1.2 Sources of data on child labour (Sections 2.12-2.15)

Child labour data may come from different sources. We summarise in Section 2.12 the main sources. More detailed descriptions may be found in other publications, in particular ILO (2004a). General household-based surveys of child labour and the intensive but restricted rapid assessment studies identify, as the two end points, a wide range of situations and issues to which the sampling techniques discussed in the present work apply. This is emphasised and clarified in Section 2.15, with the preceding two sections providing more details on the methodology of household-based surveys (Section 2.13) and rapid assessment studies (Section 2.14). Full information may be found in, respectively, Verma (2008) and ILO-UNICEF (2005).

2.2 Scope and nature of the information sought on child labour

This section introduces the basic international conventions which define child labour, and in particular worst forms of child labour (WFCL). These cover hazardous child labour and other worst forms of child labour. The various sampling methods discussed in the following chapters are aimed primarily at studying worst forms of child labour.

The principal policy instruments defining the basic concepts of child labour are the following:

- Minimum Age Convention 138 (ILO, 1973)
- Convention 182 (ILO, 1999).

2.2.1 ILO Convention No. 138

ILO Convention No. 138 takes an age-based approach, seeking to prohibit or regulate the intensity (hours) and other conditions of work at various age thresholds. It provides that the minimum age for admission to employment “shall not be less than the age of completion of compulsory schooling and, in any case, shall not be less than 15 years”.

---

2 With regard to domestic work, C.189 provides definitions of “domestic work” and of “domestic worker”. Therefore when dealing with the participation of children, i.e. persons below the age of 18 in domestic work, both in permissible or non-permissible situations, it is necessary to refer to those definitions to determine if the children do perform domestic work and if they can be considered as domestic workers.
2. Child labour situations, data needs and sources

A higher minimum age, eighteen years, is established for work “which by its nature or the circumstances in which it is carried out is likely to jeopardise the health, safety or morals of young persons”. Further, the Convention provides that national laws or regulations “may permit the employment or work of persons 13 to 15 years of age on light work which is not likely to be harmful to their health or development; and not such as to prejudice their attendance school”. Thus, economic activity carried out by a child below the minimum age (generally or for light work, if permitted nationally) is categorized as “child labour”.

While Convention No. 138 allows that developing countries may initially exclude some economic sectors from the application of its standards, only very few countries have done so and even then, the Convention must cover at a minimum: “mining and quarrying; manufacturing; construction; electricity, gas and water; sanitary services; transport, storage and communication; and plantations and other agricultural undertakings mainly producing for commercial purposes”.

2.2.2 UN Convention on the Rights of the Child (CRC)

The CRC takes a rights-based approach, articulating a wide array of fundamental rights of children aged under eighteen, including freedom of expression, thought, conscience, religion and association; the rights to the highest attainable standard of health and to a standard of living adequate for the child’s physical, mental, spiritual, moral and social development; right to rest and leisure, and to engage in play and recreational activities appropriate to the child’s age. The convention addresses child labour by recognizing “the right of the child to be protected from economic exploitation and from performing any work that is likely to be hazardous or to interfere with the child’s education, or to be harmful to the child’s health or physical, mental, spiritual, moral or social development”, while leaving the details to the relevant provisions of other international instruments (such as ILO Conventions). Further, the CRC calls for children to be protected from a variety of other related exploitations, including use of children in trafficking of illicit drugs, child sexual abuse including commercial sexual exploitation, the abduction, sale, or trafficking of children for any purpose, and the use of children in armed conflicts. The convention calls for “measures to promote physical and psychological recovery and social reintegration” of children who have fallen victim to such exploitation.

2.2.3 ILO Convention No. 182

A. Worst forms of child labour

The primary motivation for ILO Convention No. 182 is to prioritize and target the worst forms of child labour for more immediate action for eradication. It defines the worst forms of child labour as follows.

(1) All forms of slavery or practices similar to slavery: the sale and trafficking of children; debt bondage and serfdom; and forced or compulsory labour, including forced or compulsory recruitment of children for use in armed conflict.

(2) The use, procuring or offering of a child for prostitution, for the production of pornography or for pornographic performances.
(3) The use, procuring or offering of a child for illicit activities, in particular for the production and trafficking of drugs.

(4) Work which, by its nature or the circumstances in which it is carried out, is likely to harm the health, safety or morals of children.

Categories (1)-(3) of the worst forms of child labour on the one hand, and category (4), hazardous work, on the other, have slightly different characteristics, as the latter requires national specification and may depend on the conditions of work.

B. The worst forms of child labour other than hazardous work

This category of WFCL comprises situations which are defined internationally and does not depend on national determination, includes:

(1) Sale and trafficking of children. Boys are most commonly trafficked for labour exploitation, especially in commercial farming, but also for forced labour in petty crimes and the drug trade. Girls are most commonly trafficked for commercial sexual exploitation or domestic work. Most trafficked children end up in another worst form of child labour.

(2) Children in forced and bonded labour.

(3) Children in armed conflict. While boys clearly dominate, substantial numbers of girls are also involved.

(4) Commercial sexual exploitation of children in prostitution and pornography. Children affected by commercial sexual exploitation are predominately girls, but many boys are affected as well. All regions of the world are implicated. Unlike most other worst forms of child labour, which are mostly confined to developing nations, commercial sexual exploitation of children is widespread also in the more developed regions.

(5) Children engaged in illicit activities, especially in drug production and trafficking. Like commercial sexual exploitation of children, this is a problem found in all regions of the world.

Supplementing Convention No. 182, the ILO Recommendation No.190 recommends that nations make these activities criminal offences so that those responsible for involving children in them are subject to penal sanctions. Since these activities are illegal or criminal, it is especially difficult to arrive at reliable estimates of the numbers involved in these worst forms of child labour.

C. Hazardous work

Work, which by its nature or the circumstances in which it is carried out, is likely to jeopardise or harm the health, safety or morals of children is referred to as hazardous work. There is no uniform international list that can be consulted to determine which activities are hazardous and which are not: both of the ILO Conventions (Nos. 138 and 182) obliges each country to specify work it will treat as hazardous. Certain types of work, however, are nearly universally considered hazardous, such as work in mining and construction, work with heavy machinery, and exposure to pesticides. In mining and quarrying in particular, many children work in small-scale, informal-sector mines.
and quarries.\textsuperscript{3} In general terms, boys are more likely to be engaged in such hazardous work than girls. Conventions Nos. 138 and 182 call on ratifying states to specify in their national laws or regulations those types and conditions of work considered hazardous and prohibit them for children under 18 years of age. In determining hazardous types and conditions of work, it is recommended that consideration be given to:

- work which exposes children to physical, psychological or sexual abuse;
- work underground, under water, at dangerous heights or in confined spaces;
- work with dangerous machinery, equipment and tools, or which involves the manual handling or transport of heavy loads;
- work in an unhealthy environment which may, for example, expose children to hazardous substances, agents or processes, or to temperatures, noise levels, or vibrations damaging to their health; and
- work under particularly difficult conditions such as work for long hours or during the night or work where the child is unreasonably confined to the premises of the employer.

\textbf{2.2.4 International comparisons}

Relevant concepts and definitions concerning hazardous work, and indeed concerning all forms of child labour, may differ from country to country. At the international level, criteria defining child labour have to be standardized if national data are to be used for the purposes of international comparison and the establishment of global estimates.

Consider for example the ILO global child labour estimates of 2002 (ILO-IPEC, 2002). The production of these estimates was guided by benchmark ages reflected in the provisions of ILO resolutions and the relevant international instruments. Assuming a minimum age of 12 years for light work and a minimum age of 15 years for admission into employment, ILO estimated the global incidence of child labour using a measure that included the following children:

- those between the ages of 5 and 11 engaged in any economic activity;
- all working children aged between 12 and 14 years except those in light work; and
- all children aged 15 to 17 in hazardous work and other WFCL.\textsuperscript{4}

\textsuperscript{3} Hazardous child labour may also be a result of child trafficking or in bonded labour. They are not mutually exclusive, though extremely difficult to disaggregate.

\textsuperscript{4} Of course, children in the 5-11 and 12-14 can be in child labour due to their age but in addition, they can be also subject to hazardous work or the other WFCL.
2.2 Scope and nature of the information sought on child labour

Figure 2.1. International standards on child labour statistics

<table>
<thead>
<tr>
<th>Children (5-17 years old) in productive activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children in employment</td>
</tr>
<tr>
<td>-------------------------</td>
</tr>
<tr>
<td>CHILD LABOUR</td>
</tr>
<tr>
<td>Worst forms of child labour</td>
</tr>
<tr>
<td>Hazardous work by children</td>
</tr>
<tr>
<td>Exposure to physical, psychological or sexual abuse</td>
</tr>
<tr>
<td>Underground, under water, dangerous heights, confined spaces</td>
</tr>
<tr>
<td>Dangerous machinery, equipment or tools, heavy loads</td>
</tr>
</tbody>
</table>
| Unhealthy environment, hazardous substances, temperatures, noise levels or vibrations damaging to health | Current conceptual framework of the ILO global estimation of child labour is shown in Figure 2.2.


Light work was defined as non-hazardous work by girls and boys 12 years and older performed for fewer than 14 hours per week (on average two hours per day). Hazardous work included work performed for 43 hours or more per week as well as work in mining, construction, and selected occupations considered hazardous in many countries.

It is interesting to note that the definition of child labour as adopted for the global estimates published in 2002 (ILO-IPEC, 2002), has been updated by Resolution adopted by the 18th International Conference of Labour Statisticians (ICLS), Geneva, 2008, as shown in Figure 2.1.

Current conceptual framework of the ILO global estimation of child labour is shown in Figure 2.2.
2.2.5 Children in hazardous work

Children in hazardous work are in many respects the silent majority within child labour. Hitherto too few policies or programmes have been geared to the special needs of children who do hazardous work. “There are solid reasons for giving this issue urgent attention: (i) the scale of the problem – estimates place the current total of children in hazardous work at 115 million; (ii) the recent rise in hazardous work among older children – an increase of 20 per cent within 4 years; and (iii) the growing evidence that adolescents suffer high rates of injury at work, in comparison with adult workers” (ILO-IPEC, 2011).
The above-mentioned study looks at the scientific data with respect to seven sectors: crop agriculture, fishing, domestic work, manufacturing, mining and quarrying, construction, and street and service industries, noting that these sectors “were selected not because they are necessarily the ‘worst’, but in order to demonstrate the importance of knowing and understanding the risks inherent in an industry ... Although the number of children in hazardous work is large, some of the most dangerous types of child work are concentrated in specific localities, specific occupations, specific tasks and specific age groups. Focusing energies on these pockets could go a long way towards generating the momentum needed to make progress” (ibid).

2.2.6 Household chores (unpaid household services)

Children may be engaged in three types of household-based work:

- performing chores within their own households;
- undertaking economic activity for the benefit of their household, such as working in a household enterprise; and
- doing domestic work in the household of someone else.

It is important to distinguish between ‘household chores’ or ‘housework’, performed in one’s own household, which is considered nonmarket and noneconomic work, and ‘child domestic work’ which refers to situations where children under 18 years of age are engaged to perform work for a household or households within an employment relationship, i.e. in the home of a third party or employer, and is therefore considered economic activity (ILO-IPEC 2007). Many children do housework, both for other households for payment in cash and/or in kind, and within their own households, unpaid, by participating in chores. The former also covers the case of the child living in a household where he/she is employed to work.5

Hazardous household chores6

When household chores are performed by a person under 18 for long hours, in an unhealthy environment with unsafe equipment or heavy loads, in dangerous locations, and so on, they could be considered as ‘unsuitable unpaid household services’ and be studied in statistical surveys on child labour. “When determining ‘long hours’, it is important to consider the compounding effect of chores plus possible other work

---

5 C.189 in Article 1 contains the following definitions:
(a) the term domestic work means work performed in or for a household or households;
(b) the term domestic worker means any person engaged in domestic work within an employment relationship;
(c) a person who performs domestic work only occasionally or sporadically and not on an occupational basis is not a domestic worker.

6 An appropriate term for this would be “unacceptable household chores”, for the following reason. “... it is important to note that the term ‘hazardous’ in the context of unpaid household services (as found in the 18th ICESL resolution paragraph 15 (c)) be understood as wider than ‘hazardous’ economic activity (or ‘hazardous work’ corresponding to the category covered by the same resolution paragraph 17 (d) in the context of ILO Convention Nos. C138 and C182). Thus, while the former might comprise a situation (say, household chores in excess of 20 hours weekly) that ‘has negative effect and is unsuitable’ for the child’s development (say, education) and is determined as a case fit to be identified as ‘unacceptable household chores’ to be targeted for elimination, the same situation cannot be termed as ‘hazardous’ in the sense of requiring the minimum working age of 18 years under C138 and included as a worst form of child labour prohibited by C182.”
activities on a child’s education” (ICLS, 2008). However, for comparability, ICLS Resolution calls for the separation of these figures from economic activity figures.

2.2.7 Framework distinguishing forms of child labour

A more complete and up-to-date framework distinguishing forms of child labour, covering also unpaid household services, is shown in Figure 2.3.

**Figure 2.3. Framework for statistical identification of child labour**

<table>
<thead>
<tr>
<th>Age group</th>
<th>SNA production</th>
<th>Non-SNA production</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children below the minimum age specified for light work (for example, 5–11 years)</td>
<td>Employment below the minimum age for light work</td>
<td>Employment below the general minimum working age</td>
</tr>
<tr>
<td>Children within the age range specified for light work (for example, 12–14 years)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Children at or above the general minimum working age (for example, 15–17 years)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1 (3a) is applicable where the general production boundary is used as the measurement framework for child labour.
2 Age-group limits may differ across countries depending upon the national circumstances.
3 Where applicable at the national level.
4 Children in employment other than those covered under columns (1a), (2a) and (2b).

Denotes child labour as defined by the resolution.

Denotes activities not considered child labour.

Note: “Worst forms of child labour other than hazardous work” were sometimes also termed “Unconditional worst forms of child labour”. The term “Hazardous” unpaid household services may be replaced by “Unsuitable” since it does not imply the same notion as “hazardous work” within SNA production boundary.

2.3 Child labour in domestic work

Child labour in domestic work, i.e. in a household other than the child’s own is one of the most widespread forms of child labour.

“Domestic work” is now defined as work performed in or for a household; and a domestic worker is anyone engaged in domestic work within an employment relationship under ILO Convention No.189 (ILC, 2011). Child labour in domestic work can therefore be defined as domestic work undertaken in the household of a third person by a child below the relevant legal minimum age for work or employment, in hazardous work or in a slavery like situation (e.g. trafficking, sale of children, debt bondage, serfdom or forced labour).

2.3.1 Nature of the activity

For an earlier study on child domestic labour, see ILO (2004b). The following description is based on ILO-IPEC (2011) and Blagbrough (2009).

A child domestic worker may do a variety of domestic chores, including caring for children, running errands, and helping employers run small businesses. Included in this group are children who ‘live in’ and those who live separately from their employers, and children who are paid for their work as well as those who are not paid in cash but may receive in-kind benefits, such as food and shelter. The distinction may not be easy, especially where the child in question is considered as being entrusted under foster care, or informally adopted by that household (because, if the child is a family member, the “household chores” may become acceptable, unless it is excessive amount or hazardous tasks).

A. Child labour in domestic work

Domestic work is not always recognized as a form of economic activity, and so it becomes, in effect, an invisible form of work. This is even more so in the case of domestic work by children. Consequently, there is widespread reluctance on the part of institutions to address the issue with specific policies and laws. It also accounts, in some cases, for ignorance of or disregard for the hazards to which a child may be exposed. Children as young as seven years old are routinely pressed into domestic work. Some people have argued in the past that what child domestic workers do is not actually ‘work’ per se, but they are providing ‘helping hands’ to their employers’ households. The same sources consider that it was subject to debate whether child domestic workers should really be called ‘workers’, as well as whether they deserve full rights as workers (Matsuno and Blagbrough, n.d.). The recent adoption of C.189 and R.201 has ended with all these discussions. C.189 (Art. 4.2) clearly states that “Each Member shall take measures to ensure that work performed by domestic workers who are under the age of 18 and above the minimum age of employment does not deprive them of compulsory education, or interfere with opportunities to participate in further education or vocational training; furthermore, R.201 (§5.2) states that “When

---

7 According to a recent ILO report “…child domestic work” refers to children’s work in the domestic work sector in the home of a third party or employer. Where that work is performed by children below the relevant minimum age (for light work, full-time non-hazardous work and hazardous work respectively) or in a slavery-like situation that work is referred to as “child labour in domestic work”.

---
regulating the working and living conditions of domestic workers, Members should give special attention to the needs of domestic workers who are under the age of 18 and above the minimum age of employment as defined by national laws and regulations, and take measures to protect them ( )". What children performing domestic work do is work and therefore, they are “domestic workers”.

Worldwide, it is estimate that 15.5 million children are in paid or unpaid domestic work in the home of a third party or employer. Of these, 10.5 million are in child labour, of which 8.1 million are in hazardous work (52 per cent of all child domestic workers); and 5 million, children under the age of 18 and above the minimum age of employment are in permissible domestic work. In addition, undetermined numbers of children are in domestic work as result of forced labour and trafficking. Of all child domestic workers, 72 per cent are girls; and 47 per cent are below 14 years: 3.5 million aged 5 to 11 and 3.8 million aged 12 to 14 (ILO, 2013, pp. 19-25).

Statistics on child labour in domestic work are limited due to the hidden nature of the work, especially because it takes place within individual households with the right to privacy. The high incidence of informal and undeclared working relationships leads to under-reporting.

B. Possible causes

Child labour in domestic work has multiple causes. Push factors include poverty, discrimination due to low social status, displacement and migration, lack of education, and lack of alternative employment opportunities. The need to sustain themselves and their families arising from poverty is commonly the reason why children begin in domestic work. For example, according to reporting by children in an interview survey in India, a number of children worked to repay loans. In Peru and the Philippines older children spoke of their decision to seek work in the city in order to pursue their education. A quarter of participants in Tanzania recounted that they were forced into domestic work because family members had died due to HIV/AIDS, and they had no reliable relatives to take care of them. Often family problems were the catalyst for children to begin work. Family breakups and physical and sexual abuse in the children’s own families were common causes, as were issues such as alcoholism. In India, several children cited alcoholic fathers as the reason they had left home to undertake domestic work. Children were also pulled into domestic work by siblings and friends already working as domestic workers, and because of employers’ demands for younger domestic workers.

Pull factors include the common perception that domestic work offers an opportunity for better living conditions and education (especially for young children from rural areas), and that it will lead to other opportunities. Children were also pulled into domestic work by siblings and friends already working as domestics, and by employers’ demands for younger workers. With the increasing participation of women in the labour force, there is more demand for affordable child domestic workers.

Child labour in domestic work has been socially acceptable in many cultures and viewed positively as a protected kind of work. This is particularly so in the case of girls: working with a family, at someone’s home, is perceived as a positive and safe place for girls. Hence while children become domestic workers primarily due to poverty, it is also because the practice is seen as normal and indeed beneficial, especially for girls who
will one day become wives and mothers. Powerful and enduring myths surround the
practice, which encourage it to continue. Employers of child domestic workers, far from
seeing themselves as exploiters, often consider that they are helping the children and
their families by taking them in. In many cases employers believe that they are treating
these children as ‘part of the family’.

C. Worst form of child labour in domestic work

Domestic work as such is not defined as a worst form of child labour under Convention
No. 182. It urges, however, states to take effective and time-bound measures “to take
account of the special situation of girls”. (Art 7.2e and Recommendation No. 190
includes a call for programmes giving special attention to “the problem of hidden work
situations, in which girls are at special risk”.) Child domestic workers are large in
numbers, yet they remain invisible and marginalized both economically and socially.
A wide range of abuses – including physical, verbal, and sexual violence – routinely
accompanies this type of work. Their isolated situation, coupled with their ambiguous
role in the employers’ household, makes child domestic workers particularly vulnerable.
When violence occurs, their dependency on the employers for basic needs and their
acceptance of the violence as an occupational hazard make them far less likely to
report it. Examples of physical violence include beating, kicking, whipping, pinching,
scalding, overwork, and denial of food. Sexual violence against child domestic workers
is also widespread due to the children’s vulnerability and isolation in the home of
their employer. If girls become pregnant, they are often thrown out of the house and
are forced to fend for themselves on the streets. In these instances, domestic work
may lead to prostitution, as the girls and young women have few options available.
In Bangladesh, for example, an NGO interviewing children working in commercial
sexual exploitation in Dhaka found that all of them had previously worked as child
domestic workers and had been sexually abused by members of their employing family.
Traffickers of children into the sex trade routinely deceive children and their families by
promising them attractive jobs as domestic workers.

Aside from violence, child domestic workers are exposed to a variety of other dangers,
including hazardous household chemicals such as cleaning fluids, kitchen knives, irons,
boiling water, and unfamiliar household appliances which can cause serious injuries
and even death — especially among younger children and those already exhausted from
a full day’s work. There are also likely to be long-term health consequences of chronic
sleep deprivation, being ‘on call’ twenty-four hours a day, or having to perform heavy
tasks such as water collecting. Furthermore, understanding the psychosocial effects
of child domestic work is vital to forming a comprehensive picture of their condition.
In Kenya, one of the few field studies that specifically looked at the psychological
impact found that child domestic workers experienced significantly more psychological
problems than other working or nonworking children. Similarly in Indonesia, interviews
with child domestic workers who had been working for a long time, or who began work
at a young age, indicated that their self-esteem was eroded.

A number of factors conspire to make a child domestic worker particularly vulnerable to
abuse and exploitation, and these relate mainly to the inequality of the relationship that
the child has with members of the employing household. Child domestic workers have
reported that the daily experience of discrimination and their isolation in the employer’s
household are the most difficult part of their burden.
2. Child labour situations, data needs and sources

2. Child labour situations, data needs and sources

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Hazards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooking, cleaning, ironing and other</td>
<td>Sharp blades; hot pans; stoves and other tools in poor repair;</td>
</tr>
<tr>
<td>household chores</td>
<td>toxic chemicals</td>
</tr>
<tr>
<td>Gardening</td>
<td>Sharp objects; heavy loads; hot weather; stinging insects;</td>
</tr>
<tr>
<td></td>
<td>toxic pesticides and fertilizers</td>
</tr>
<tr>
<td>Gathering fuel, water, groceries</td>
<td>Heavy loads; traffic and other urban hazards; long distances on foot</td>
</tr>
<tr>
<td>All tasks out of public view</td>
<td>Inadequate food and shelter; long hours; no privacy; physical, verbal</td>
</tr>
<tr>
<td></td>
<td>and sexual abuse; humiliating or degrading treatment</td>
</tr>
<tr>
<td>All tasks when working alone</td>
<td>Isolation; separation from family and peers</td>
</tr>
</tbody>
</table>


In summary, the most common hazards for children in domestic work are: long working hours, which create fatigue; lack of public scrutiny, which can provide opportunities for sexual exploitation; and isolation, inhibiting normal social and intellectual development. Additional risks are physical and sexual violence, and being subject to carrying out heavy and dangerous tasks.

Certainly not all child domestic workers suffer abuse, neglect or exploitation. The work itself is not automatically considered dangerous or inhumane (it is subject to national determination). However, because of their ‘invisible’ nature, lack of negotiating power against employers and lack of awareness of their rights as children and as workers, children in domestic work are at risk. Because child domestic workers toil within the confines of a house, it is extremely difficult to know how they are treated or to reach them with help if they have any problem with the employers. This obviously makes obtaining and enumerating good samples of them difficult.

2.3.2 Some examples

Please refer to Annex A for some examples from national studies of child domestic work. Extracts are provided from the following studies.

A. Child domestic work in South-east and East Asia (Matsuno and Blagbrough, n.d.).

B. Child domestic work in Côte d’Ivoire (Diallo, 2009).

C. “Small maids” in Morocco (Sommerfelt, 2009).

---

8 For conceptual clarification, we may note the following:
(i) The list of hazardous activities for children is to be determined at the national level;
(ii) C.182 3(d) refers to work that by its nature or the circumstances in which it is carried out;
(iii) R.201 says: “...Members should identify types of domestic work that, by their nature or the circumstances in which they are carried out, are likely to harm the health, safety or morals of children, and should also prohibit and eliminate such child labour.”
This implies that it is up to the national authorities to determine if domestic work, by its nature or the circumstances in which it is carried out is or not dangerous. It also implies that domestic work is not automatically considered dangerous or inhumane, but certainly there are places where domestic work by its nature is considered just as dangerous (to mention a few: Brazil, Panama, Peru, Togo), i.e. no one below the age of 18 can perform domestic work. In some other places only certain types of domestic work are considered as dangerous, such as domestic work performed at night, in live-in conditions, when the children take care of the sick, of other children.
2.3 Child labour in domestic work

2.3.3 Some implications for sampling and related aspects

A. Use of household-based surveys of the population

Domestic work is one of the larger sectors of child labour. The work is performed in private households, which tend to be fairly widely scattered in the population. Hence, to a considerable extent, child labour in domestic work may be investigated on the basis of a survey of private households, with many features and procedures similar to ordinary household surveys of the general population.

However, the survey design would need to take into account certain special features of the target population of households containing child domestic workers. Firstly, in many situations only a very small proportion of households employ child domestic workers. Secondly, this small proportion is often quite unevenly distributed in the general population. Consequently, a general household survey, not specifically focused on the target population of households employing child domestic workers, may fail to capture a representative sample of that population efficiently, or even fail to capture a representative sample at all.

B. An illustration: The 2005 Baseline Survey on Child Domestic Labour in Bangladesh.

The following information has been extracted from the survey report (Bangladesh, 2006a).

The baseline survey covered a population of around 126 million in the urban and rural areas of the country excluding Dhaka City Corporation area. The survey population was stratified into four domains: (1) rural; (2) small municipalities with population under 100 thousands; (3) large municipalities with population of 100 thousands or more; and (4) city corporations. These strata accounted for, respectively 88, 6, 3 and 3% of the survey population - as much as 88% of the population is rural and only 12% urban. A sample of 734 primary sampling units (PSUs) was allocated to the domains as shown in Table 2.1. The PSUs consisted of administratively defined areas, called ‘mauza’ in rural areas and ‘mahalla’ in urban areas.

<table>
<thead>
<tr>
<th>Domain</th>
<th>No. of sample PSUs</th>
<th>% distribution</th>
<th>Total</th>
<th>Sample</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Rural area</td>
<td>64 districts, 2 per district</td>
<td>128</td>
<td>17.4</td>
<td>87.9</td>
<td></td>
</tr>
<tr>
<td>(2) Small municipalities</td>
<td>64 districts, 4 per district</td>
<td>256</td>
<td>34.6</td>
<td>6.0</td>
<td></td>
</tr>
<tr>
<td>(3) Large municipalities</td>
<td>20 municipality, 5 per municipality</td>
<td>100</td>
<td>13.6</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>(4) City corporations</td>
<td>5 corporations, 50 per corporation</td>
<td>250</td>
<td>34.1</td>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td>(2)-(4) All urban areas</td>
<td>606</td>
<td>82.6</td>
<td>12.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>734</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

No. of sample PSUs actually enumerated 725
Within each domain, PSUs were selected with probability proportional to size (PPS) of the population. Each selected PSU was divided into approximately equal clusters of size 125 households. Then one of these segments, in which the concentration of child domestic workers was thought to be high, was automatically taken into the sample; a second segment was selected at random from the remaining segments in the PSU. Thus two segments covering approximately 250 households were selected from each sample PSU, giving a total sample of over 725x250=180,000 households.

### Table 2.2. A few selected characteristics of child domestic workers

<table>
<thead>
<tr>
<th>Percentage of households employing domestic workers (any age)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult workers only</td>
<td>7.2</td>
</tr>
<tr>
<td>Child workers only</td>
<td>1.0</td>
</tr>
<tr>
<td>Both adult and child workers</td>
<td>0.1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>8.3</strong></td>
</tr>
</tbody>
</table>

Children account for 12.7% of all domestic workers, and adults for 87.3%.

<table>
<thead>
<tr>
<th>Distribution according to type of place</th>
<th>Percentage of households employing children</th>
<th>Distribution of child domestic workers in the population</th>
<th>Distribution of child domestic workers in the sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Rural area</td>
<td>0.8</td>
<td>63.6</td>
<td>5.0</td>
</tr>
<tr>
<td>(2) Small municipalities</td>
<td>2.7</td>
<td>14.6</td>
<td>30.0</td>
</tr>
<tr>
<td>(3) Large municipalities</td>
<td>3.5</td>
<td>9.5</td>
<td>15.0</td>
</tr>
<tr>
<td>(4) City corporations</td>
<td>4.4</td>
<td>12.3</td>
<td>50.0</td>
</tr>
<tr>
<td>(2)-(4) All urbal areas*</td>
<td>3.3</td>
<td>36.4</td>
<td>95.0</td>
</tr>
<tr>
<td><strong>Total</strong>*</td>
<td>1.1</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

*Excluded Dhaka city corporation, not covered in the survey.

#### Variation among the 5 city corporations in the percentage of households employing child domestic workers

<table>
<thead>
<tr>
<th>Maximum</th>
<th>Minimum</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.6%</td>
<td>2.2%</td>
<td>4.4%</td>
</tr>
</tbody>
</table>

#### Characteristics of child domestic workers

| % full-time | % part-time | 6.2 |
| % girls | % boys | 22.4 |
From the results presented in Table 2.2, three important characteristics of the population of child domestic workers in Bangladesh may be noted:

(1) It is a ‘rare’ subpopulation: only a very small proportion (merely 1.1%) of households in the country employ a child domestic worker. Furthermore, such households are scattered in the general population and no special list frames exist to separate them out.

(2) This subpopulation is unevenly distributed: proportionately four times as many urban households employ child domestic workers compared to rural households (though the actual proportion employed is very small in both cases, respectively 0.8% and 3.3%, as noted).

(3) There is variation within individual localities, e.g. the percentage of households employing child domestic workers varies from 2.2% to 9.6% among the five city corporations. Actually, it is quite likely that much larger variations in this percentage exist not only among localities, but also within localities such as different neighbourhoods in a city.

Efficient design in such a situation requires oversampling of areas which contain a higher concentration of the target population of interest – in this case the proportion of households employing child domestic workers. More precise procedures for this purpose will be developed and illustrated in Chapter 6 (with a further illustration in Annex B). In the present illustration the design does in fact try to achieve this, even if in a far from optimal way. For example, while rural areas contain 88 per cent of the total population of households, they account for only 17-18 per cent of the households in the sample (Table 2.1). This is in line with the fact that engaging children for domestic work is rare in the rural sector – only 0.8 per cent of households containing a child domestic worker (Table 2.2). By contrast, city corporations contain only 3 per cent of this population but 34 per cent of the sample households. Child domestic work in this domain is more prevalent – with an average of 4.4 per cent of households containing a child domestic worker. The contrast is sharper still when we consider the sample of child domestic workers (rather than of households or the general population). The rural domain contains approximately only 5 per cent of the sample of child domestic workers, while roughly 50 per cent of the sample of such children comes from the city corporation domain (Table 2.2).

The fact that, overall, child domestic workers are a rare population requires special procedures to enhance efficiency of the sample. This is obvious in the case of rural areas: a PSU with 250 households enumerated is expected to contain no more than 2-3 households with a child domestic worker.
C. Sampling units

Child domestic workers may be enumerated according to where they live: (i) at the employers’ household where they work; (ii) at their own household; or (iii) at some other location where they live, if they do not live at the employers’ or their own household. Alternatively, sometimes it is also possible and convenient to identify them through a location other than where they live: for example at the household of their family even if they are for the time being living with the employer or elsewhere. The locations through which they are selected constitute the effective sampling units. It is important to note that a child domestic worker may be associated with more than one sampling unit. The choice of the type of units to be used as sampling units has to be made on the basis of technical and operational considerations in the particular conditions of the survey.

Special care is also needed to properly deal with part-time child domestic workers, who can be very common in certain situations (see for instance the example from the Philippines in Annex A). Frequently, such child domestic workers do not live with their employer, and may in fact work for more than one employer at the same time.

D. Problems in locating and enumerating child domestic workers

From the above description of the conditions under which child domestic workers live and work, it is clear that such children may be often hidden from and be inaccessible to the survey. For example, it is very common in some countries for parents to send away their children to live, within the framework of the extended family, under the ‘care’ of a relative or a member from the same community in a better situation. This can blur the distinction between being an ‘ordinary’ member of the latter household versus being a child domestic worker there. Contrast the situation in the examples from Côte d’Ivoire and Morocco given in Annex A. Employers may hide the presence of child domestic workers in their household or refuse permission to contact them, for example because of legal reasons, or because of the poor living and working conditions of the children concerned.

This tends to make the target population of child domestic workers a ‘reclusive’ one in certain circumstances. Problems and possible technical solutions to such problems will be addressed in Chapters 13 and 14.

2.4 Agriculture, especially commercial agriculture

There are two major types of activity in the agricultural sector: traditional, usually family-based, crop and live-stock farming; and specialised commercial crops where individuals or families work largely as employees.

In developing countries, traditional cultivation generally constitutes the largest and most widely dispersed sector, often in the form of ‘family farms’ where children work as unpaid family helpers. In surveying this sector, the appropriate sample design is similar to that of a normal household-based survey of child labour. The main difference is that agriculture is essentially confined to rural areas. The real problem is the potential under-reporting of child labour on the family farm.
Specialised commercial cropping is often confined to more-or-less well-defined areas of concentration, though there are many exceptions to this. In any case, such units usually constitute a relatively small subpopulation of all economic units. Special procedures are therefore required to sample them. The sampling procedures would differ according to the type of commercial crop and the type of production unit involved. For example, in large estates producing commercial crops such as tea, households of the workers attached to (usually living at) the estate may form the appropriate sampling units. In the case of production based on smaller enterprises employing hired labour, the enterprise or establishment may itself be the main sampling unit.

In surveying child labour in agriculture – whether conventional farming or specialised commercial cropping – a problem needing attention is the under-coverage due to hidden nature of the children’s work. On the family farm, children are commonly engaged as unpaid helpers. In large estates as well, often the child provides labour as a part of the employment contract of the adults in the household.

### 2.4.1 Nature of the activity

**A. Commercial agriculture as a worst form of child labour**

As summarised by Fyfe (2009):

> “Agriculture is a complex and diverse economic sector and it is where most child labour is found. … The vast majority of working children are toiling in fields and fisheries, not in factories. In fact, the number of children working in agriculture is nearly ten times the number involved in factory work …. This situation is by no means confined to the developing world. Entire families of migrant labourers, as in the case of Mexican migrant workers in the United States, help plant and harvest the rich world’s fruit and vegetables. In the United States, there are an estimated 300,000 children working in agriculture. They account for 8% of working children but suffered 40% of work-related fatalities. … This basic fact about child labour is often ignored in favour of an urban and industrial image of the problem … This neglect of agricultural child labour, linked to an unquestioned assumption that children working on farms and in fisheries are at less risk than their urban counterparts, still prevails today and distorts our view of the problem …..

> “Agriculture is a diverse economic sector comprised of a number of subsectors … highly industrialized and mechanized commercial operations at one end of the spectrum and traditional small-scale, family-based subsistence farming at the other. However, this distinction is slowly eroding under commercial and globalization effects promoting export-oriented agriculture. Children working in agriculture are engaged in undertakings of all kinds ranging from family farms to corporate-run farms, plantations, and agro-industrial complexes. At one end of the spectrum children may work with basic equipment, low levels of mechanization, and few agricultural inputs such as pesticides and fertilizers; at the other extreme, they may engage in intensive, highly organized, highly capitalized, commercial production systems.
“Risks increase where children are contracted to a third party outside the family, as in the case, for example, of labour contracted to commercial farms and plantations. In extreme cases, children may be victims of bonded labour or forced labour practices. Minorities and migrant workers are often the most vulnerable to exploitation. For example, the children of migrant workers are often classed as helpers, hired as part of a family unit by contractors or subcontractors, thus enabling farm and plantation owners to deny responsibility for under-age workers ..... 

“Millions of the world’s child labourers in agriculture engage in hazardous work ... The family farm — that universal element in agriculture that is so bound up with culture and tradition — renders the work of children largely invisible, thereby making it difficult to acknowledge that children can be at risk in such a setting. The world over, children working on family farms, and in agricultural activities generally, are viewed as helping out and engaged in family solidarity. ... Child labour regulation is much weaker in agriculture than in other sectors. However, it is precisely because of these factors - large numbers, hazardous work, lack of regulation, invisibility, and loss of education - that agriculture should be a priority sector for the elimination of child labour. Children become agricultural workers at an early age ... This is particularly the case for girls, who tend to start work at five, six, or seven years of age ...”.

However, in balance it must also be pointed out that not all agricultural work is hazardous to the child (ILO-IPEC, 2011):

“Agriculture is the sector with the most child labourers. It is also the sector with the most potential for decent work for rural children and young adolescents who have reached the legal minimum age of employment. Although we must keep in mind that work in field and tree crops has a number of serious risks for children, one must not assume that these risks make the whole industry off-limits to children of all ages. Indeed, many types of farm work can be positive for children, providing them with experience and technical skills. Because of the numbers involved and the value of the industry to the family as well as to the national economy, it is absolutely crucial to discern which tasks, which working conditions, which products and which tools are hazardous and to help parents and policy-makers alike to take the necessary steps to protect children from those hazards”.

B. Common tasks and hazards in crop agriculture

The following comments and the list of hazards in children’s work in agriculture are quoted from ILO-IPEC (2011), abbreviated for brevity.

With globalization, the profile of child work in crop agriculture is rapidly changing. Subsistence farmers in developing countries are adopting the chemicals used by big plantations, but they may have little training in their use and product warning labels may not be in a language they understand. Increasingly growers are producing for the international market, as in the case of sugar, bananas, flowers, cocoa, tobacco, tea and coffee. In such production, child workers may be exposed to toxic substances particular
to that industry (such as methyl bromide, a particularly toxic pesticide used in flower production), or to the use of sharp tools (which, although traditional, are being used at an industrial pace, as in cocoa production); in some cases, contact with plants themselves may cause injury or illness.

In crop agriculture, much of the recent research has centred on the health impacts of exposure to pesticides. Pesticide poisonings are under-reported because farmers (and health workers) often do not recognize the symptoms. “The evidence we do have comes mainly from national surveillance systems in industrialized nations; however, what is shocking is that while developing countries use only 25 per cent of the world’s pesticides, it is estimated that they have 99 per cent of pesticide-related fatalities, implying that lack of knowledge of proper use is a critical factor”. It is not only pesticides that may pose such risks. A field study of child labour on tobacco farms reported that children as young as 5 years old suffer from Green Tobacco Sickness, a type of poisoning that occurs when nicotine from tobacco leaves is absorbed through the skin.

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Hazards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loading and carrying produce or water</td>
<td>Heavy loads</td>
</tr>
<tr>
<td>Climbing trees to harvest fruit</td>
<td>Dangerous heights; slippery surfaces; unstable ladders</td>
</tr>
<tr>
<td>Collecting fodder</td>
<td>Heavy loads; thorns and sharp objects</td>
</tr>
<tr>
<td>Preparing, handling manure</td>
<td>Bacteria, parasites and other micro-organisms</td>
</tr>
<tr>
<td>Weeding and harvesting</td>
<td>Thorns; bending for long hours; insects and animals; biological toxins</td>
</tr>
<tr>
<td>Caring for farm animals</td>
<td>Kicks; bites; brucellosis, anthrax and other bacterial exposure such as E. coli and salmonella; rabid animals</td>
</tr>
<tr>
<td>Handling agrochemicals</td>
<td>Toxic chemicals</td>
</tr>
<tr>
<td>Using motorised or sharp farm tools</td>
<td>Tools in poor repair; lack of safety features; sharp blades; heavy weights; loud noise; vibrations; faulty power supply; poor ventilation</td>
</tr>
<tr>
<td>Driving tractors or other farm machinery</td>
<td>Vehicle or machine in poor repair; lack of safety features; fast moving blades; moving belts; uneven ground or mud</td>
</tr>
<tr>
<td>Outdoor work in general</td>
<td>Exposure to extreme weather; sun; insects; wild animals; parasites; lack of drinking water</td>
</tr>
</tbody>
</table>


### 2.4.2 Some examples

Please refer to Annex A for some examples from national studies of child labour in agriculture, especially commercial agriculture. Extracts are provided from the following studies.

A. The size and trend of rural child labour in some countries of South Asia (de Groot, 2009).

B. Pakistan Baseline Survey 1996 (5-14 years old).

C. Children in India’s tea industry (Ashraf, 2009a).

D. Child labour on sugarcane plantations in the Philippines (de Boer, 2009).
2.5 Fishing and aquaculture

2.5.1 Sampling considerations

The fishing sector involves a large variety of conditions and tasks of child labour. Consequently, the sampling approach has to be devised in view of specific situations. Normally the conditions of work are hazardous and children work for adult fishermen. They may also be involved in other tasks for which fishing provides an opportunity. In some cases this could include drug trafficking, as shown in an illustration from Jamaica presented later in the case study in Annex C. The implications are that in many cases, child labour in the fishing sector has to be sampled from units other than households, such as from fishing enterprises or boats; and also that care and special methods are needed in order to contact and interview the children involved.

2.5.2 Common tasks and hazards in fishing

Fishing encompasses a range of tasks, from hazardous work to light work that could be suitable for children under certain conditions. On land there is dock work, such as lifting nets and fish cases, repairing nets, maintaining vessels and cleaning and processing fish. On the water there is basic crew work, hauling nets, line fishing and diving. Children are engaged in all of these. These activities may entail long hours, extreme temperatures and harsh weather conditions. Girls tend to be more involved in post-harvest work, while boys undertake most of the work related to catching fish. This gender division of labour is very marked in fishing. Hence hazards and risk tend to differ for girls and boys, and attention has to be paid to gender differences in risk assessment.

As to hazards of the work, “some of the more dangerous practices that children engage in are: jumping off moving boats at night to position nets; working 10–12 hours a day without protection from the sun; and living and working for weeks at a time on small fishing platforms positioned several kilometres out at sea. Children in the fishing industry get stabbed by bones, hooks and sharp fins; risk poisoning by venomous fish or sea snakes; may become caught in net-lifting winches, nets, snares or lines; or get hit by rudder blades” (ILO-IPEC, 2011).
### 2.5.3 Example: Child labour in fishing, El Salvador

The illustration provided in Annex A concerns an investigation using the rapid assessment methodology into the worst forms of child labour in fishing in El Salvador. It is based on Godoy (2002), *El Salvador Child Labour in Fishing: A Rapid Assessment*.

### 2.6 Mining and quarrying

In terms of a proportion of the total child labour, children working in mines constitute a very small category, but the working environment and working conditions of the children involved are exceptionally hazardous. These factors have important consequences for the sampling and data collection strategy.

#### 2.6.1 Sampling considerations

The target population tends to be localised, i.e. confined to areas where mining activity is concentrated. It is necessary to know all of these locations in order to achieve good coverage. Often mining sites are in remote locations, and present harsh survey conditions, though by no means as harsh as those to which the children working there are subject. In some situations, mining is a small-scale family business, and working children may be sampled through their households. Working children may be living in mining camps, rather than in their family households. In such cases it may be possible to take all or a sample of camps, and list and select working children directly within each camp. However, often children have to be sampled at their work location.

Mobility of the working children – either because their families move looking for mining opportunities, or because the children have to work at different locations – requires special sampling procedures, of the type discussed in Chapter 10. In modern
open-cast mines, children from neighbouring communities are often displaced by the development of mines in their area, and are reduced to working as scavengers. This also requires sampling procedures designed to capture mobile populations.

The difficulty and intensity of the survey may require the sample to be restricted to a small number of locations. Yet there may be many dimensions (types and sizes of mines, diverse conditions, etc.) to be represented in the survey. The controlled selection (balanced sampling) procedures of the type discussed in Chapter 12 may be useful.

### 2.6.2 Common tasks and hazards in mining

Mining and quarrying are forms of work dangerous to children in every way.

> “Mining and quarrying are physically dangerous because of the heavy and awkward loads, the strenuous work, the unstable underground structures, the heavy tools and equipment, the toxic dusts and chemicals and the exposure to extremes of heat and cold. The work is morally and psychologically risky, too, given that mining often takes place in remote areas where law, schools and social services are unknown, where family and community support may not exist and where ‘boom or bust’ conditions foster alcohol abuse, drugs and prostitution” (ILO-IPEC, 2011).

Children’s work in and around informal mines is varied. They might accompany their parents to the site, especially when there is no alternative means of looking after them during the working day. Almost all child miners work in small-scale mines. They meet the same risks as adults but, because their bodies and judgement are still developing, injuries are more likely to happen, and they are more likely to fall victim to the free-wheeling lifestyle common in mining camps. Evidence from various surveys and research studies demonstrates that mining is by far the most hazardous sector for children with respect to fatal injuries. Like fishing, mining presents some particularly horrific examples of child labour. Compressor mining is one of these. The chemical of major concern in mining is mercury, because it is used commonly. Lead is also a concern in mining, as is cobalt, which has the potential to damage the heart, thyroid and lungs, and can exacerbate occupational asthma. Children’s health risks are compounded by the contaminated environment in which they live. Clean drinking water and health services are often lacking, especially in the remote mining areas. Some children in mining areas are coaxed or forced into prostitution.

> “In summary, it is true that hazardous work of children is on the increase in certain mining areas, as the price of gold rises and where child soldiers are trying to escape into another life; true, these are hard-to-reach places – the mountains of the Andes, the deserts of western Africa. On the other hand, child labour in mining is not widespread; it is concentrated in particular places and the numbers of children involved – fewer than 1 million – suggest that determined efforts could eliminate this form of work” (ibid).
### 2.6.3 Some examples

In Annex A, extracts are provided from the following national studies.

A. Child labour in small-scale gold mining in the Philippines (Espino, 2009).


### 2.7 Manufacturing

#### 2.7.1 Sampling considerations

Manufacturing covers a wide range of activities, covering both traditional and modern sectors, involving formal as well as informal sector establishments, and also work performed at home by family units or by individual members of the household. The prevalence and conditions of child labour differ by all these factors, and these influence the choice of the methods of sampling and data collection.

A large number of sector-specific surveys, often in the form of *child labour baseline surveys*, have been conducted. Just to give a couple of examples, Bangladesh has conducted child labour surveys covering automobile establishments, battery recharging and recycling, among others. In Pakistan, surveys have covered the glass bangles industry, surgical instrument manufacturing industry, and tanneries, among others.

Many of these surveys are establishment based, i.e. the establishment is taken as the basic sampling unit, within which child workers are identified, sampled and interviewed. Some industries may involve medium-to-large sized establishments, but child labour is predominantly concentrated in small, often very small, establishments. Some of these establishments may be more or less ‘formal’ economic units, but many of them are clearly in the ‘informal’ sector. Often the sampling frames available for such units

---

### Selected list of common tasks and hazards in mining

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Hazards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tunnelling; diving into muddy wells</td>
<td>Drilling equipment; explosives; confined spaces; faulty supports; stagnant air; poisonous gases; dust; darkness; dampness; radiation</td>
</tr>
<tr>
<td>Digging or hand-picking ore, slabs, rock or sand</td>
<td>Heavy tools; heavy loads; repetitive movements; dangerous heights; open holes; falling objects; moving vehicles; noise; dust</td>
</tr>
<tr>
<td>Crushing and amalgamating; sieving, washing and sorting</td>
<td>Lead, mercury and other heavy metals; dust; repetitive movements; bending, squatting or kneeling</td>
</tr>
<tr>
<td>Removing waste or water from mines</td>
<td>Heavy loads; repetitive movements; chemical and biological hazards; dust</td>
</tr>
<tr>
<td>Transporting materials</td>
<td>Heavy loads; large and unwieldy vehicles</td>
</tr>
<tr>
<td>Cooking and cleaning for adults</td>
<td>Physical and verbal abuse; unsafe stoves; explosive fuels</td>
</tr>
<tr>
<td>Selling goods and services to miners</td>
<td>Physical and verbal abuse</td>
</tr>
<tr>
<td>Mining and quarrying in general</td>
<td>Remote locations; lawless atmosphere; poor sanitation; contaminated drinking water; stagnant water (and mosquitoes); inadequate nutrition; recruitment into sex trade; gambling, drugs and alcohol</td>
</tr>
</tbody>
</table>

*Source: ILO-IPEC, 2011.*
are grossly imperfect, sometimes non-existent. Issues in sampling establishments, and especially establishments in the informal sector, in the presence of imperfect sampling frames, are discussed in detail in Chapters 4 and 5.

As shown in the example of child workers in India’s glass industry summarised in Annex A, even in a single sector, different stages in the production process may involve very different modes of organisation and different compositions of the work force involved. Consequently, somewhat different sampling and data collection procedures may be required at different production stages in the same sector. The prevalence of subcontracting practices – subcontracted by larger establishments to smaller, often informal sector, establishments, and by both these categories to households and individuals - complicates the picture. The survey has to cover both the children employed directly in the establishment, and children working through various subcontracting arrangements.

2.7.2 Common tasks and hazards in manufacturing

Children are employed in a great diversity of manufacturing tasks. These include manufacturing in both the formal and informal sectors, in large- and small-scale enterprises, from vast garment factories to home-based workshops doing piecework or traditional crafts. Many national laws concerning occupational safety and health in the formal economy do not apply to micro-enterprises or home-based work. The most hazardous conditions are often found in the thousands of smaller-scale operations. In many cases, children work at home alongside a parent or guardian who is a home-based worker.

“Manufacturing is a sector for which we do have high-quality studies. Evidence shows that children involved in manufacturing suffer from substantially higher health-related risks than comparable non-working children. Among many others, this includes risks such as of experiencing pain of various kinds, especially back pain, ... skin, eye and ear problems, injuries, and disturbed sleep. Manufacturing enterprises often have toxic substances present, such as organic solvents used in furniture work, shoe-making and automobile repair” (ILO-IPEC, 2011).

Studies show that children exposed to solvents face a very high risk of developing long-term neuro-behavioural impairments – children complaining of headaches, performing poorly on motor dexterity and memory tests, showing loss of concentration and memory deficits and overall tending to be irritable, angry and confused. Leather-tanning is another form of work that poses so many risks that it is likely to be hazardous in all respects.

“In summary, the wide scope of manufacturing and repair industries makes it impossible to make any judgements as to the relative risk of one versus another. But the studies point out that certain ones – and they are easy to spot – are so dangerous that immediate action is called for” (ibid).
CHAPTER 2

2.8 Construction

Selected list of common tasks and hazards in manufacturing

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Hazards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Textile: dyeing; weaving; sewing; embroidery; cleaning</td>
<td>Harmful dyes; awkward postures; repetitive movements; sharp tools; close work</td>
</tr>
<tr>
<td>Leather: dehairing; tanning; sewing; cleaning</td>
<td>Harmful dyes, solvents and other chemicals; fumes; sharp tools</td>
</tr>
<tr>
<td>Footwear: shoe manufacturing and repair</td>
<td>Dangerous solvents, adhesives and other chemicals; sharp tools</td>
</tr>
<tr>
<td>Crafts: jewellery-making; stonework; woodwork</td>
<td>Awkward postures; repetitive movements; close work; dust; sharp tools</td>
</tr>
</tbody>
</table>


2.7.3 Some examples

Please refer to Annex A for some examples from national studies of child labour in various branches of industry. Extracts are provided from the following studies.

A. Children in India’s carpet industry (Ashraf, 2009b).

B. Child labour in ‘subcontracting sectors’ of Indonesia’s garment and footwear industries (Tjandraningsih, 2009).


D. Children in India’s glass industry (Ashraf, 2009c).

2.8 Construction

Generally, construction is a small sector in regard to the numbers of children involved, but it seems to be prevalent in a number of countries, such as in South-Asia. Working in construction is extremely hazardous for children everywhere.

2.8.1 Sampling considerations

Like some other sectors such as plantations and informal sector manufacturing, children are often engaged in construction work through the employment of the parents. In principle the sampling procedures in these situations can be similar, such as obtaining a sample of children through a sample of households. The added complexity is the temporary nature of the construction work sites.

Experience in conducting sector-specific surveys focused on construction is limited. In some cases, numbers of child workers and conditions of work have been estimated from general household-based surveys of child labour, as shown by the example of a baseline survey in Uganda in Annex A. The results cannot be very reliable when construction is a small sector and is not well distributed in the general population. More targeted surveys, possibly using special methods such as adaptive cluster sampling (Chapter 9), are required for more accurate information.
2. Child labour situations, data needs and sources

2.8.2 Common tasks and hazards in construction

The construction industries rank among the most dangerous for children.

“The risks that adult construction workers face are well known: dangerous machinery, dangerous heights, dangerous materials and ubiquitous dust. And yet children are found on construction sites around the world, performing support work in this hazardous environment. In many countries, particularly in South Asia, children (including girls) are recruited to carry, stack and align heavy bricks for long hours. In industrialized countries, adolescents of legal working age work with tools designed for full-grown adults. Similar risks abound in the industry of brick manufacturing. Added to the hazards, however, are the extreme temperatures and airborne ash created by the kilns. Children often younger than 10 years old haul bricks – each weighing about 2 kilograms – from one place to another all day long, breathing air thick with dust” (ILO-IPEC, 2011).

Many countries have banned the practice of employing children at construction sites, yet child labour in construction remains prevalent. Because of the undocumented nature of construction work, there is a lack of high-quality studies. It has been estimated by ILO that over 100,000 workers are killed on site every year in construction. Workers in the construction sector also experience a high rate of work-related non-fatal injuries. Brick kilns are common throughout Asia, but are also found in the Middle East, Africa and Latin America.

“A study of child labourers in Cambodia showed that child workers in brick factories suffered more from work-related health effects than child scavengers, children working in fish processing centres and child car-washers. … A cross-sectional survey of brick-workers in Pakistan found that chronic bronchitis, asthma and tuberculosis were much higher in the brick-kiln workers than in the control group. Girls working in the brick sector often had poor nutrition and the numbers of girls with low weight was double that of boys. Alarmingly, 68% of boys and 76% of girls between the ages of 10 and 14 were not attending school and were classified as illiterate” (ibid).

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Hazards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Construction: hauling and stacking materials; carpentry; masonry</td>
<td>Heavy loads; dangerous heights; falling objects; sharp objects; power tools; live wires; moving vehicles; loud machines; exposure to extreme weather; dust</td>
</tr>
<tr>
<td>Brick-making: toting, stacking</td>
<td>Heat from kilns and ovens; flying ash; heavy loads; dropped bricks; dust; exposure to extreme weather; remote locations; poor sanitation; moving vehicles</td>
</tr>
</tbody>
</table>

2.8.3 Example: Child labour baseline survey in Uganda

For a summary of the methodology and findings of the study, see Annex A. The description is based on report of The Child Labour Baseline Survey conducted by Uganda Bureau of Statistics (Uganda, 2010).

2.9 Street children

2.9.1 Diverse situations of street children

The term ‘street’ refers to an unspecified place other than a family accommodation or an institution of reception and stay. Street children are defined as minors who earn their living working on the streets, including children who live on the street full- or part-time. UNICEF estimates there are more than 100 million children who live and work on streets in the developing world. Latin America is home to 40 million street children (Ferguson 2004).

A. Typology

Often two situations of street children are distinguished: ‘children of the street’ who live and work on the street; and ‘children on the street’ who work (or otherwise spend most of their waking hours) on the street, but live (at least sleep) in a household or other place of habitation. UNICEF (2001) for example made the following distinction. (1) Children of the street are homeless children who live or sleep on the street in urban areas. For these children, family ties may exist, but they are tenuous and maintained occasionally. (2) Children on the street, on the other hand, earn their living or beg for money on the street and return to their home at night. They are likely to hand over all or part of their earnings to the family, thus contributing to the economic survival of the family unit. The parents often encourage their being in the street. The distinction between the two groups is important because children on the street have families and homes to go to, whereas children of the street are alone and lack the emotional and psychological support normally provided by the parent.

Although it may be helpful to maintain these two main categories, the situation is often more complex:

“While the concepts of children of the street and children on the street usefully reflect the different circumstances [that] children are living under, the complexity of the phenomenon means that overlaps and grey areas remain. … the extent to which children have contact with their family varies considerably. Some children of the street are abandoned and rejected by their families; other children of the street left their family due to prevailing circumstances, but maintain regular contact and may visit the family for a while before returning to the street. … Meanwhile, the category of children on the street includes a grey area of children who sometimes sleep on the street and sometimes sleep at home. There are also children within this category who are staying with distant relatives or employers. Children on
2. Child labour situations, data needs and sources

the street often live in poor households, and many of these children are candidates for becoming children of the street.” (Hatløy and Huser, 2005). A third category which needs to be added to the ‘of the street’/’on the street’ dichotomy is that of children of street families. These are children who are living on the street with one or both of their parents. They have either moved to live on the street with their family, or in fact were born on the street – the ‘second generation street children’.

A more elaborate and realistic classification, such as the one proposed in Table 2.3, is desirable, as it can have important implications for the sampling and data collection methodologies. (National reports do not necessarily follow a uniform or consistent classification, as can be seen from some examples below and further examples in Annex A.) In these illustrations, different categories of the above schema have been identified as important in different situations. This schema is based largely on two criteria: living arrangements and contact with the family. Sometimes additional or different categories are identified based on the children's condition or circumstances, such as ‘children in difficult situations’, ‘children in need of special protection’, ‘vulnerable children’, ‘children with broken families’, ‘children in conflict with the law’ (because of involvement in crime or other illicit activities or indeed for other reasons), ‘drifting children’, and so on.

<table>
<thead>
<tr>
<th>Children of the street</th>
<th>Have no family, are orphaned, or have been abandoned by the family</th>
</tr>
</thead>
<tbody>
<tr>
<td>No contact with the family; not living with or supervised by adult family members</td>
<td>Run-away children, usually having at least the potential to return</td>
</tr>
<tr>
<td>Have contact with their family - whether regular or tenuous</td>
<td>Living with their ‘street family’</td>
</tr>
<tr>
<td>Mostly not living with family, but have contact</td>
<td>Have left the family (usually in a rural area), to temporarily live on the street (usually in a city), for a limited period, normally for a preconceived objective</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Children on the street</th>
<th>Living not at home, but also not on the street – such as at an institution, with distant relatives, with employer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequently or mostly living at home, but often staying on the street</td>
<td>At-risk children living at home in absolute poverty, such as slum children and others spending most of their day on the street; may work for or with the family</td>
</tr>
</tbody>
</table>

9 ‘Some children of the street are abandoned and rejected by their families’. To this we must add ‘... or by their employers’. See for instance *Ending child labour in domestic work* (ILO, 2013, p. 36): “Similar findings have been reported in El Salvador. In cases where girls become pregnant, they are often thrown out of the house and forced to fend for themselves on the streets, as the shame of their situation makes it difficult for them to return home. Many families reject these “spoiled girls” because their behaviour has brought dishonour to the family. In these instances, domestic work typically becomes a precursor to sexual commercial exploitation, as those concerned have few other available options.”
B. Sampling implications

In any case, the categorisation of street children can have implications for sampling and other aspects of survey methodology, the following being among the important ones.

(1) Different categories of children may, and often do, cluster in different types of activity, sometimes requiring different types of sector-specific survey.

(2) Living arrangements determine what sort of sampling design or designs are suitable. Different categories of children may be clustered in different ways. The appropriate units for sampling and accessing them may be different.

(3) When specialised methods such as snowball or respondent driven sampling (Chapters 13-14), capture-recapture sampling (Chapter 11), adaptive cluster sampling (Chapter 9), or informal sector establishment sampling (Chapter 5) are used, it is critical to take into account the heterogeneity of the target population. This applies both to the design (for example how the initial sample is ‘expanded’ with snow-ball, respondent-driven or adaptive sampling through links among children), and also in estimation (such as with capture-recapture sampling).

Specialised sampling strategies are needed because, despite being visible, children working in the street are difficult to study. They are not found within households or schools, and therefore cannot be covered through normal surveys. Furthermore, most of these children remain on the move from one place to another during daytime, and sleep outside buildings at night. Specially designed street children surveys are needed to collect the relevant information. Purposive or convenience sampling approaches are often used, both in selecting the areas to be covered, and in conducting random interviews with children regarding their living and working conditions and interviews with employers in the informal sector employing children. The challenge is to make samples more representative of the target population, and obtain more precise quantitative estimates from the surveys. Techniques such as capture-recapture sampling and respondent-driven sampling have been successfully used in some settings (see, for instance, Hatløy and Huser, 2005). Other approaches are also possible. We will elaborate a number of these sampling techniques in subsequent chapters.

The following subsections describe conditions of street children, with quite a number of illustrations from country studies given in Annex A.

2.9.2 The condition of street children

A. Street children and street trades

The following account is based on Maidment (2009), Worst Forms of Child Labour: Street Children and Street Trades.

There are no ‘typical’ street children. Numbers are notoriously difficult to assess — anything up to 120 million has been estimated by UNICEF. Children are found on the streets of developed and developing countries, many driven by events that cause them to see street life as preferable to other options. About 85% of street children are boys, but increasing numbers of girls are coming to the streets, often escaping abuse
suffered while working as domestic servants. In a few countries a high percentage are girls. Most children first come onto the streets around the age of ten to twelve, but in certain countries, especially India, and in African countries ravaged by HIV/AIDS, it is not uncommon to find children as young as four or five; typically street children are only 7-8 years old.

Street-Working Children (Children on the Street)

The majority of street children spend their day on the street but return to their families at night. The largest group are children from the city shantytowns and slums who remain in their neighbourhoods and spend the day assisting their families to augment their income, often dropping out of school as a result. Then there are children who commute to city centres from the suburban slums. Such commuting children undertake a wide variety of jobs - shoe shining, vending activities, collecting and reselling bottles and recyclable materials, windscreen washing, cleaning, and casual labour in hotels and restaurants. Sometimes older children, adult criminals, or gang leaders organize territories for trading. Slum children (as well as the children of street-living families) are heavily involved in activities such as scavenging for refuse that can be sold for recycling. Ragpicking is a variation of this activity that occupies many young children. The environment in which this activity takes place is particularly dangerous for children - the filth and open sewers, plus the dangers from infected items, expose these children to both chronic and acute diseases.

Street-Living Children (Children of the Street)

It has been estimated that 5 to 10% of street children are abandoned, orphaned, or runaway children with little or no contact with their families. For them, ‘family’ consists of other street children loosely formed in gangs or small groups that often demonstrate considerable loyalty to their members. These children sleep where they can - in shop doorways, in covered bus and railway stations, under market stalls, in parks, and on beaches. They often seek out adults who can provide them with food or protection in return for work - work such as running errands, assisting at street vending, collecting and sorting for recycling. Many of these activities are dangerous, especially for younger children. For instance, Indian street children cluster around railway stations, jumping on and off moving trains. They acquire seats for passengers, carry luggage and clean train compartments, and scavenge on the railway tracks. As a result, many children are injured, some fatally. A further risk is that children become dependent on adults for whom they operate. Children unable to earn money may resort to scavenging for leftover food in trains and trash cans, around restaurants, and in marketplaces. Children unable to work at all may resort to begging or become totally dependent on an exploitative adult or group of older children. Some get involved in illegal activities in order to survive.

Children living on the street often cite a beating by someone for whom they worked, or from a father or stepfather, as the catalyst for their leaving home. Children who are thought to have cash are susceptible to protection rackets from adult criminals, street gang leaders, and corrupt police, with threats of violence if they resist. Because the public’s perception of street children is negative - they are seen as dirty, delinquent, at best a nuisance, at worst a danger, especially when in gangs - the children are at
risk of abuse by those in authority who should be protecting them. Children are often rounded up by the police under laws against vagrancy or loitering. They are arrested on suspicion of committing petty crimes.

As most street children have suffered rejection or abuse from their families and suffer further rejection from society, many form poor opinions of their self-worth. If neglect, rejection, and abuse persist, the experiences of children of the street and on the street can lead to extremes of antisocial behaviour, such as the youth gang culture in parts of Central and South America, South Africa, and Eastern Europe. Yet many street children survive these abuses. It is often the strongest children, physically and mentally, who decide to opt for the streets to avoid worse fates. There they learn to fend for themselves.

B. Common tasks and hazards in street work

Carrying goods, street vending and waste picking are some of the common ‘on the street’ activities of labouring children. In fact many service activities take place on the street: serving as a bus attendant, running errands, transporting goods, shining shoes, washing and guarding cars, vending everything from food to flowers, working in markets and collecting recyclables, entertaining, etc. Some activities are illegal, such as prostitution, selling drugs and begging, yet these are all common methods of income generation for many of the world’s children. The following account is based on ILO-IPEC (2011).

Porters

Similar to the construction industry, a primary hazard for children involved in portering and transport is the heavy load they often have to carry, resulting in bone, joint and muscle problems. Studies have reported child porters carrying loads heavier than their own weight. In addition to musculoskeletal problems, children experience abuse of various kinds, injury, nutritional deficiencies and low body weight resulting from long periods away from home.

Street vendors

In studies of youth street vendors, large majorities report having experienced headaches, stress, fatigue, and discomfort caused by heat, noise and repetitive work. Significant minorities report occupation-related accidents.

Waste-pickers

Waste-pickers experience significant rates of gastrointestinal, respiratory and skin diseases, as well as life-threatening tetanus. Other hazards on garbage sites are violence, drug use especially of alcohol, and coercion from armed groups. A cross-sectional study in Nicaragua reported that children who were working as waste-pickers had the highest-ever reported levels of polybrominated diphenyl ethers (PBDEs) from breathing the dust at the garbage sites. PBDEs are toxic to the liver and thyroid, and inhibit neurodevelopment. These studies also showed elevated levels of persistent organic pollutants, mercury, lead and cadmium. A series of small studies in the Philippines observed that child scavengers’ learning ability progressively deteriorated with the amount of time working on the garbage dump.
### 2. Child labour situations, data needs and sources

#### 2.9.3 Examples

Street children participate in a great variety of activities, and under very diverse conditions in different settings. In Annex A we provide a number of examples from national studies of labour of street children. Extracts are provided in the annex from the following studies.


B. *Street children in Mexico* (Peralta, 2009).


E. *Child street vendors in Brazil* (Kassouf and Ferro, 2009).


H. *Children of Kolkata slums* (Bagchi, 2009).

I. *Children of Delhi slums* (Mitra, 2009).

---

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Hazards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portering and transport</td>
<td>Heavy loads; long hours; unsanitary conditions; poor access to food; long distance from home; violence; exposure to alcohol, cigarettes, drugs and adult language and situations; sexual exploitation</td>
</tr>
<tr>
<td>Outdoor shop work, such as vehicle repair, scrap yards</td>
<td>Toxic fumes and liquids; acids from batteries; clutter; slippery floors; sharp objects; heavy and dangerous machinery; loud noises</td>
</tr>
<tr>
<td>Scavenging; trash and recycling collection</td>
<td>Sharp objects; contaminated objects; moving traffic; vehicle exhaust; bending; heavy loads; long hours; extreme weather; street crime</td>
</tr>
<tr>
<td>Street-corner services such as shoe-shine or food sales</td>
<td>Street crime; harmful chemicals; hot surfaces; long hours; extreme weather</td>
</tr>
<tr>
<td>Street environment in general</td>
<td>Exposure to violence, crime, illicit drugs, tobacco, alcohol and sex</td>
</tr>
</tbody>
</table>

2.10 ‘Idle’ children

Data sets from developing countries that provide information on children’s activities consistently show a significant group of ‘idle’ children, who neither attend school, nor participate in economic activity. Idle children constitute an important policy concern – they are not only denied schooling but are also the category of children most at risk of becoming child labourers if the economic situation of their household deteriorates. While considerable research attention has been accorded to child labourers, the often large group of children absent from both school and economic activity has been the subject of very little research.

In a study by Biggeri, Guarcello, Lyon and Rosati (2009), data from six countries - Brazil, Cameroon, Guatemala, Nepal, Turkey, and Yemen – were analysed in an attempt to address this issue. For a brief review of the study results, see Annex A.

2.11 Worst Forms of Child Labour (WFCL) other than Hazardous Work

2.11.1 Surveying worst forms of child labour other than hazardous work

The worst forms of child labour refer to forms of child labour that are so fundamentally contrary to children’s basic human rights that they must be absolutely prohibited for all persons under the age of 18.

Five major types of activities have been identified under the rubric of WFCL going beyond simply hazardous work. These have been referred to in previous publications as ‘unconditional’ WFCL (‘UWFCL’).

(1) Commercial sexual exploitation of children.
(2) Children involved in drug trafficking or other illicit activities.
(3) Children working under slavery-like conditions, or in other modes of forced labour.
(4) Children subject to labour as a result of child trafficking or debt bondage.
(5) Children directly engaged in armed conflict, or living and working under war conditions.

Defining characteristics

We may also identify four defining characteristics which make a child’s labour take the form of ‘UWFCL’.

(1) Nature of the activity. Certain kinds of activities of children are inherently ‘UWFCL’, irrespective of why and how the children came to engage in them. Examples are: children subject to commercial sexual exploitation; those involved in drug trafficking or other illicit activity; children engaged in extremely hazardous work, or in work resulting in disability or other long term damage.
(2) The process of entry. A child’s labour can amount to a UWFCL determined by the nature of the process by which child came to be in that labour situation, irrespective of the nature of the child labour activity itself. Examples are child’s entry as a result of trafficking or debt bondage, or being forcibly recruited to perform a task, for example in an armed conflict or other violent activity.

(3) Lack of freedom concerning working. It is categorised as UWFCL when the child is forced to work against his/her will, forced to do certain kinds of work, forced to work under slavery-like or otherwise very restrictive conditions, or is prevented from leaving the work.

(4) Extreme hostility of the work environment. This refers to the child being subject to, for instance, violence, severe abuse, harsh punishment, being isolated, or prevented from contacting family and friends.

These defining characteristics are distinct dimensions, but can often overlap, such as child trafficking leading to commercial sexual exploitation of the child (often in the case of girls), or trafficking leading to forced labour (often in the case of boys).

In practice it is extremely difficult to identify and quantify UWFCL because of reasons such as illegal or hidden nature of the activity, inability or unwillingness of the victim child to cooperate with statistical enquiries, the danger and personal pain involved in the process, lack of knowledge and understanding on the part of the statistical investigator, and the resulting difficulty of selecting and enumerating representative samples. The challenge is to reduce and partly overcome these difficulties in obtaining the required information.

**Methods of assessment**

A diversity of methods may be used to estimate the magnitude of different types of UWFCL. The following seven are among them.

(1) Use of existing data bases. These usually involve counts from records of actual cases. In general, the results are subject to gross under-estimation because of incomplete registration.

(2) Anecdotal estimates (sometimes called ‘guesstimates’) based on isolated and selected items of information from different sources, plus subjective assessment.

(3) Estimation by analogy, for example extrapolation from known information from other countries, regions, etc. in similar circumstances.

(4) Estimation through covariates. The idea is to identify and use ‘covariates’, i.e. variables which can be measured and are related to the incidence of UWFCL. One possible schema can be as follows. Subgroups of children who are at risk of being subject to the UWFCL of interest may be identified on the basis of their circumstances and characteristics. First, representative sampling may be used to identify the sizes of these subgroups. Then, small intensive studies may be used to estimate the proportion subject to UWFCL in each subgroup. Putting these two items together estimates the proportion or number subject to UWFCL in the total population of children.
2.11.1 Worst Forms of Child Labour (WFCL) other than Hazardous Work

(5) Indirect estimation. Here the two components put together are: (a) an estimate of the prevalence of the activity for a different group, where such estimates are expected to be more reliable; and (b) an estimate of the ratio of expected prevalence in the group of children of interest, to expected prevalence in that different group. The first component is based on a more representative sample, but the second is derived on the basis of small scale studies, by analogy, etc. An example is the estimation of commercial exploitation of girls from an estimate of its prevalence among adult females, and an estimation of the (child/adult) ratio of this prevalence.

(6) Direct estimates, where they can be based on valid statistical methodology. Examples are using methods such as capture-recapture design and respondent-driven sampling (developed, respectively, in Chapter 11 and Chapter 14).

(7) ‘Triangulation’. By which is meant putting the results from different sources or methods together to arrive at the most plausible or apparently reasonable estimate. This is desirable because each method has its own strength and weakness.

Our concern in this book is of course with procedures leading to (6), namely direct estimates.

2.11.2 Trafficking in children for labour and sexual exploitation

The UN Protocol to Prevent, Suppress and Punish Trafficking in Persons, adopted in 2000, defines trafficking in children as the “recruitment, transportation, transfer, harbouring or receipt of a child for the purpose of exploitation,” whether or not coercion, fraud, deception, or the giving or receiving of payments is involved; the consent of the child is irrelevant.

Children are vulnerable to trafficking for many of the same reasons as adults.

“Children are recruited through various means, such as false promises of work in hotels or homes, or in the modelling and entertainment sectors. Once children are involved, the mechanisms by which they are kept in the control of traffickers for sexual and other forms of exploitation can include movement to another location within the same country or across country borders, placement in unfamiliar surroundings, social isolation, the use or threat of violence to the child or family members, forced indebtedness, removal of identity documents, and falsification of the child’s identity. Children of all ages can become victims of trafficking. Younger children are often trafficked for adoption, begging, use in petty crime, or child labour, while adolescents are more often trafficked for sexual exploitation (ECPAT 2007).

For an illustration, see Annex A. The example is based on Albania (2004), Rapid Assessment of Trafficking in Children for Labour and Sexual Exploitation in Albania.
2.11.3 Commercial sexual exploitation of children

Although it is difficult to quantify the number of sexually exploited children due to the clandestine nature of these crimes, ILO has estimated that there are some 1.2 million children trafficked worldwide every year. While not all children are trafficked for commercial sex, many are.

The following account is based on Madrinan (2009), *Worst Forms of Child Labour: Commercial Sexual Exploitation of Children*.

Commercial sexual exploitation is a crime manifest in the form of child prostitution, child sex tourism, child pornography, and trafficking in children for sexual purposes. It may also include marriages where remuneration for access to children as sexual partners is involved. The sexual exploitation of a child for profit embodies a planned and organized effort. Networks may operate locally or across national and international borders. Often entrapment is facilitated by agents from the child’s own local environment: neighbours, relatives, caretakers, other authorities, taxi drivers, police, etc.

About 80% of the victims of human trafficking, the majority of whom are women and young girls, are trafficked into prostitution. The remaining 20%, who are mostly men and boys, face forced labour.

The prostitution of children is often organized in environments such as brothels, bars, clubs, and hotels, but it also takes place in private residences and on the street. It can involve individuals or groups working on a small scale or in large criminal networks. Children can be entrapped in prostitution as they seek to meet needs such as food, shelter, clothing, or drug dependency, or in exchange for physical or social safety on or off the street. They may also be prostituted in exchange for higher school grades, extra pocket money, or goods and favours otherwise out of their reach. Prior violence or sexual abuse suffered in their immediate environment is a common example of circumstances that can lead to the prostitution of children.

Misunderstandings surrounding the term ‘child prostitution’ have led to use of the term ‘commercial sexual exploitation of children’, as the latter is a more encompassing description of the full range of commercial sexual violations perpetrated against children. Nevertheless, ‘child prostitution’ remains in common usage even though it carries problematic connotations that can result in blaming the victims of these crimes. In addition, public understanding of ‘prostitution’ and ‘prostitute’ worldwide has been shifting as a result of the introduction of the term ‘sex worker’, intended to raise the perceived status and protection of the rights of adult women in prostitution. However, in relation to children, reference to ‘sex work’ is wholly misleading and harmful, as it downplays the criminal exploitation of the child and can suggest that the child has somehow made a choice to follow a ‘profession’ in ‘sex work’.

Annex A provides a description of the following two illustrations based on ILO (2004c), *Girl Child Labour in Agriculture, Domestic Work and Sexual Exploitation: Rapid assessments on the cases of the Philippines, Ghana and Ecuador*: (1) Children in commercial sexual exploitation in the Philippines; (2) Commercial sexual exploitation of girls in Ghana.
2.11 Worst Forms of Child Labour (WFCL) other than Hazardous Work

2.11.4 Child bonded labour and forced labour

Child bonded labour refers to situations where a child’s labour services are offered in exchange for a loan. In some cases, the labour of the child alone, or of the entire family, is directly offered. In other cases, bondage is intergenerational: once a parent is no longer able to work, debts are passed down from parent to child.

The following description is based on Genicot (2009), *Worst Forms of Child Labour: Child Bonded Labour*.

Bonded labour has been widespread since ancient times, but the extent to which it persists today is not always remembered. Bonded labour is outlawed by the 1956 UN Supplementary Convention on the Abolition of Slavery, the Slave Trade, and Institutions and Practices Similar to Slavery. However, millions of people are still held in bonded labour around the world, including 15 million children in India alone (Human Rights Watch/Asia 1996).

At the origin of bonded labour lies a loan that a family takes from an employer (often to repay another loan or to pay for food, health care, a wedding, or a funeral). Having no other asset, the family may pledge the labour of some of its members - adults or children - to the employer-cum-lender in exchange for the loan. These family members are then forced to work until the debt is paid off with their wages. Employers generally charge very high interest rates, pay low wages, deduct payments for equipment used, charge fines for faulty work, and sometimes take advantage of the debtor’s illiteracy and lack of basic math skills. As a result, in many cases the debt actually increases instead of decreasing as the family member works.

Working conditions are typically difficult. Persons employed as bonded labourers work long hours over many years, in some cases in dangerous occupations. The existence and persistence of child bonded labour raises many questions. Indeed, this institution combines many elements that have been banned by governments around the world: child labour, long-term pledge of labour services including bonded labour and contract labour, and inheritability of debt.

Some additional issues arise concerning bonded labour in the particular case of children. Sometimes a child’s labour is directly pledged as a security for a loan. Allowing other family members to be responsible for repayment of a debt can clearly improve credit access, and this method is often used informally or formally. Incentives to default may be lower for community members who can expect retaliation on their offspring (La Ferrara, 2003). During the borrower’s life, depositing a child, or a wife, with a creditor as security for repayment of debt seems a common practice in many regions of the world. South Asia is often cited as the region where child bonded labour is most prevalent. But use of children as ‘debtor-pawns’ – children given as household servants by debtors - is also well documented in other parts of the world.
As to forced labour, in 2012 the ILO produced its Global Estimate of Forced Labour (ILO, 2012; see also ILO, 2013). Of the total number of 20.9 million forced labourers, women and girls represent the greater share of total forced labour – 11.4 million victims (55 per cent), compared to 9.5 million (45 per cent) men and boys. Children aged 17 years and below represent 26 per cent of all forced labour victims (or 5.5 million children). While the specific number of children in forced labour and trafficking for domestic work remains unknown, evidence points to the existence of significant numbers of children in debt bondage, victims of trafficking and in servitude situations.

### 2.11.5 Children and War

Like adults, children experience wars in many ways: as targets, indirect victims, bystanders, witnesses, and perpetrators of violence. War brings children new vulnerabilities as a result of displacement, forced recruitment, death of family members, and the destruction of physical, economic, or political infrastructure. Depending on their age, gender, cultural background, and personal characteristics, children respond differently to wars, often assuming new identities and functions.

#### A. Child soldiers

The following description is based on Marriage (2009), *Worst Forms of Child Labour: Children and War.*

Some years ago, the picture was as follows. The ILO estimated that there were around 300,000 child soldiers fighting in thirty-six countries around the world, with children serving as combatants, spies, camp followers, wives, porters, and minesweepers in regular troops, paramilitary, and opposition forces. The majority of contemporary wars were taking place in Africa, around half the child soldiers in the world serving in this region. “The context for these wars is the post-Cold War dominance of neo-liberalism in global politics and the fallout from the AIDS epidemic.... Alongside [the neo-liberal] economic and political developments, AIDS has reshaped demographics over the past two decades. Sub-Saharan Africa has the highest regional incidence of AIDS, and particularly high sero-prevalence rates [proportions in a population who test positive based on blood test] are recorded in armed groups that are mobile, use commercial sex workers, and often do not practice safe sex. AIDS decreases life expectancy, placing greater responsibility on younger members of society in all forms of employment. In the light of the decentralized structures, this employment - including work as soldiers - is frequently informal and unprotected, may not afford a subsistence level of return, and regularly involves violence or other forms of abuse. ... The influence of neo-liberalism, experienced through structural adjustment and unequal labour conditions, and the AIDS epidemic, means that many people in sub-Saharan countries contend with dangerous levels of poverty irrespective of the onset of fighting...”.

#### B. Children caught-up in the war situation

War affect the conditions of child labour much more widely than simply through the recruitment of child soldiers.
“War exacerbates the conditions of poverty - through the further loss of assets, houses, and farmland, or through loss of public entitlements or market opportunities. Children are made homeless through war, can be internally displaced, or can become refugees in another country, all of which affect their labour opportunities. The death of parents and the destruction of social structures during war propel children into roles as bread winners, often providing for other children. Child-headed households face particular obstacles in terms of land rights, access to education and health services, and legal protection. … This leads to increased violence and coercion in trading arrangements, the most extreme form of which is abduction for recruitment, forced labour, or slavery” (ibid).

War gives cover for these practices, which often target young people. Boys are taken into agricultural work or mining, while girls are more generally used for domestic chores or for sexual exploitation.

But it may also be noted that while coercion into armed groups can increase the vulnerability of children, it can, in other circumstances, improve their security. An ILO study found that two-thirds of children fighting in wars in central Africa were volunteers (Dumas and de Cock 2003).

2.11.6 Children engaged in drug trafficking

Generally drug trafficking is a very difficult sector of child labour to assess. In Annex A we have summarised the methodology and main findings reported in de Souza e Silva and Urani (2002), Brazil Children in Drug Trafficking: A Rapid Assessment. Using the rapid assessment methodology, the study compiled and organized data concerning living standards of children working in drug trafficking schemes in several low-income communities in Rio de Janeiro.

2.12 Sources of information on child labour

2.12.1 Household-based surveys

The household is generally the most appropriate unit for identifying children and their families, and measuring their socio-economic and demographic characteristics and conditions. Specifically, household surveys provide a versatile instrument for obtaining information regarding the circumstances that force or motivate children to work, and the conditions of their work in household-based activities. Regular child labour surveys are, therefore, household-based national sample surveys whose target are children and their parents/guardians living in the same household.

Household surveys of child labour may use a variety of designs and organizational structures. The main factors determining the design are: (i) the substantive objectives concerning the content, complexity and periodicity of the information sought; (ii) practical aspects, such as the survey conditions and the resources available. These substantive and practical requirements determine features of the survey structure such as its timing, frequency, reference period and sampling arrangements. In practice, the
household survey content may be detailed and specialised, providing information on the dynamics of child labour or gross flows between different child labour categories; or it may be confined to a few basic items concerning child labour. The choice depends on the data needs, available resources, and the arrangements and circumstances under which the survey is conducted. The key respondents in a child labour survey are the working and potentially working children and their parents or guardians. Also, the survey may be a continuing survey (normally designed to obtain regular time-series data on current levels and trends), or it may be an occasional survey (normally designed to obtain benchmark and structural information). Most of the recent national child labour surveys have been occasional or one-time surveys.

While household surveys based on probability sampling of the general population provide an effective instrument for estimating the prevalence and characteristics of child labour, there are certain important limitations to their effectiveness. A survey with a broad coverage may not be able to capture particular sectors of child labour, especially sectors which are small or are not well-distributed in the general population. General household surveys are also unlikely to be able to capture certain special child labour categories such as children living on the street or those engaged in hidden forms of child labour. The main limitation of household surveys is that generally they are not a suitable instrument for collecting information from children engaged in the worst forms of child labour.

We will elaborate further on the methodology of household-based surveys of child labour in Section 2.13. This is useful in view of the considerations noted later in Section 2.15.

2.12.2 Sources supplementing household-based child labour surveys

A. School-based surveys

There are at least two types of school-based survey, each type serving a specific purpose.

(1) The primary aim of the first type is to determine the impact of work on school attendance and performance of children, and to assess attitudes of working children towards studying. The survey may also attempt to assess school-related factors such as the quality of education available to the children. Interviews are conducted with children, teachers, school management, and parents. Normally, schools attended by a subsample of working children identified in the main child labour survey form the sample for the school survey. Such a sample is not necessarily a representative or efficient sample of all the schools in the area of study. Normally a sample of non-working children is also covered as a control group, preferably from the same schools as attended by the surveyed working children.

(2) The other type of school survey is based on a more representative sample of schools. The primary objective here is to estimate the numbers of children enrolled at and actually attending school, classified by the children’s age, sex, activity status, and other characteristics. Comparison with the total population of children in the area covered gives an estimate of the numbers of children who are not enrolled or not attending school, and hence are exposed to the risk of child labour. A sample of schools may be selected from the list of schools serving the population covered in the
main child labour household survey. The two samples may be drawn independently, or they may be linked in terms of higher-stage units, for instance schools and households in the survey coming from the same sample areas. However, normally the samples are not linked at the micro-level (i.e. in terms of individual children in the two samples), unlike in the above-noted first type of school survey.

B. Community-level inquiries

Community-level inquiries may be conducted as independent investigations collecting data on particular child labour situations. Alternatively, they may supplement other survey types such as regular household or establishment-based surveys, or special operations like rapid assessment and baseline surveys. Community-level inquiries can be useful for identifying the main variables related to the prevalence of child labour in the community. They usually collect information from administrators and other community leaders on the cultural and socio-economic profile of the community and other characteristics which may be related to child labour in the area. This may include information in the area regarding the average household income level, poverty, major economic activities, seasonal unemployment, literacy, and availability of public services and utilities such as schools, medical facilities, transport system, water, electricity and recreational facilities.

C. Alternative sources

Data on child labour may be obtained from alternative sources that are not specifically focussed on child labour. These sources, often used in combination, may include existing national household surveys, population censuses, and other secondary sources.

General national household surveys

Many countries collect socio-economic and demographic data through periodic household-based sample surveys, including surveys on the labour force, living conditions, household income and expenditure, demography and health. Such surveys normally do not produce detailed data on child labour, but they can yield information that is useful for analysis of the child labour situation. Moreover, attaching child labour modules to such surveys is a potential source of information.

Censuses

Although few national population censuses provide data on the prevalence of child labour, information from censuses serves as an essential basis for the interpretation and analysis of data on child labour from other sources. The population census is also the basic source of sampling frames for child labour and similar surveys.

Secondary sources

A wide range of institutions which, while not primarily concerned with child labour, produce useful information pertaining to it. Examples are periodic school reports compiled by ministries of education, school surveys and inspection reports, statistical reports by national statistical offices, and surveys and research conducted by international development organizations (ILO, 2004a).
D. Employers’ surveys

Finally, another supplementary source can be surveys of employers (or establishments employing children) identified through the household-based survey of child labour. This is the first type of establishment survey described below.

2.12.3 Establishment surveys

As in the case of school-based surveys, establishment-based surveys can also be of two types, each type serving a specific purpose.

A. Establishments associated with household-based sample of working children

Surveys of the first type involve interviewing establishments selected through their link with working children already identified through a household-based child labour survey. The survey is conducted to collect statistics on the nature and conditions of children’s work in more detail than is normally possible in a more general household-based survey. Interviews may be conducted with children as well as employers. Information may be collected on children’s views and attitudes towards the work they perform. The survey may also attempt to assess establishment-related factors, such as the size, location and physical conditions of work. The survey provides additional variables on children linking them to the establishments. It does not necessarily yield in itself a representative or efficient sample of the total population of establishments employing children.

B. Representative sample of establishments and children working in them

The other type of survey is based on a more representative sample of establishments. The conventional definition of an establishment is as follows (UN, 1975): “the store, shop, office, or other single location at which a combination of resources and activities is directed by one owner in operating one kind of business; [the] establishment includes associated ancillary units, e.g. subsidiary warehouses, garages, and offices in its vicinity”. Most sources of information refer to the ‘supply side’ (surveys of children supplying the labour). To collect information on the ‘demand side’ (the recipients of child labour), establishment or workplace based surveys are required.

The sample of establishments is selected from a frame of establishments likely to be employing children for the production of estimates: (i) for the population of establishments employing children; and (ii) for the associated population of working children. Establishment survey questionnaires are normally administered at the workplace – whether at a factory or a home-based production unit that engages hired workers – and seek to obtain information concerning the production unit and the nature of its workforce, with a focus on child workers. Children’s wages and other benefits, nature of activity, hours of work, other working conditions, and injuries and illnesses at work are among the items of information sought. These data can then be compared with parallel data on adult workers.

For estimates pertaining to establishments, information can also be sought regarding employers’ perception of reasons for using child labour, the advantages and drawbacks they see in using child workers, recruitment methods used, etc. The main source
providing such information is likely to be the employer or a representative of the employer.

Concerning estimates pertaining to working children as the units of analysis, normally a sample of child labourers in the establishments would be interviewed to obtain additional variables. Sampling procedures for establishment-based surveys of child labour are elaborated in Chapter 5. Technical procedures for selecting establishments can be quite different depending on establishment size; details are discussed for large and medium sized establishments in Section 5.4, and in Sections 5.6-5.10 for small and informal sector establishments. Some aspects are common to any type of establishment, such as stratification (Section 5.2) and sample allocation (Section 5.3). Section 5.5 describes procedures for selecting working children within establishments.

### 2.12.4 Rapid assessments

Producing qualitative estimates of child labour relies on a variety of instruments, one of the most important among these being the Rapid Assessment (RA) studies. The RA methodology has been elaborated jointly by ILO and UNICEF (ILO-UNICEF, 2005). This methodology is primarily intended to provide information relatively quickly and inexpensively, of the type needed for creating a general awareness of problems concerning child labour, formulating specific projects, or as a base leading to in-depth research. The methodology has been specifically designed to use largely qualitative techniques to gather in-depth information on hidden or invisible worst forms of child labour. Its application is usually concentrated in particular geographical areas. RAs use a participatory approach to obtain information on working and living conditions of children involved in particular activities or occupations which are otherwise difficult to identify and characterise. RAs also obtain some quantitative data, but the approach is not directed at providing quantitative estimates of the prevalence of child labour. Often, the usefulness of qualitative information from a RA study can be enhanced by complementing it with more representative, probability-based surveys.
Table 2.4. Examples of rapid assessment studies in countries

<table>
<thead>
<tr>
<th>Rapid assessment topic</th>
<th>Example of surveys conducted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commercial agriculture</td>
<td>Tanzania (coffee) Tanzania (tea)</td>
</tr>
<tr>
<td>Ecuador (flower plantations)</td>
<td>Tanzania (horticulture)</td>
</tr>
<tr>
<td>Bolivia (sugarcane)</td>
<td>El Salvador (sugarcane)</td>
</tr>
<tr>
<td>Lebanon (tobacco)</td>
<td>Tanzania (tobacco)</td>
</tr>
<tr>
<td>Other agriculture</td>
<td>Ecuador Ghana Philippines</td>
</tr>
<tr>
<td>Commercial sexual exploitation</td>
<td>Costa Rica Ecuador El Salvador Ghana Jamaica</td>
</tr>
<tr>
<td>Madagascar Philippines Sri Lanka Tanzania Vietnam</td>
<td></td>
</tr>
<tr>
<td>Domestic work</td>
<td>Brazil Ecuador El Salvador Ethiopia Ghana</td>
</tr>
<tr>
<td>Nepal S. Africa Sri Lanka Thailand</td>
<td></td>
</tr>
<tr>
<td>Trafficking for labour and sexual exploitation</td>
<td>Albania Bangladesh Estonia Moldova</td>
</tr>
<tr>
<td>Nepal Romania Thailand Ukraine</td>
<td></td>
</tr>
<tr>
<td>Drug trafficking/use</td>
<td>Brazil Estonia Philippines Thailand</td>
</tr>
<tr>
<td>Garbage sorting and rag picking</td>
<td>El Salvador Guatemala Nepal</td>
</tr>
<tr>
<td>Informal sector</td>
<td>El Salvador Tanzania Nepal</td>
</tr>
<tr>
<td>Bonded labour</td>
<td>Nepal</td>
</tr>
<tr>
<td>Fishing</td>
<td>El Salvador</td>
</tr>
<tr>
<td>Mining</td>
<td>Tanzania</td>
</tr>
<tr>
<td>Child porters</td>
<td>Nepal</td>
</tr>
<tr>
<td>Street children</td>
<td>Turkey</td>
</tr>
<tr>
<td>War, child soldiers</td>
<td>Philippines</td>
</tr>
</tbody>
</table>

In view of the considerations noted in Section 2.15, it is useful to elaborate on the methods, merits and limitations of the RA approach, which are presented in Section 2.14. In Table 2.4 we list examples of RA studies which have covered a wide range of worst forms of child labour in many countries (the list is not meant to be exhaustive). Many of these studies have involved some form of street surveys of children.

2.12.5 Baseline surveys and studies (BLSs)

Baseline surveys and studies are useful for identifying target populations and their characteristics, and for analysing the determinants and consequences of child labour in particular sectors or settings. In principle, such surveys or studies may involve one or more of the available methods of data collection, for example a combination of quantitative and qualitative techniques such as a sample survey supplemented by participatory research (ILO, 2004a). In fact, both in scope (objectives) and in methodology, the term ‘baseline survey’ or ‘baseline study’ has been used for a variety of instruments, covering the following five among them.

A. Impact assessment

In the context of project or programme impact assessment, the objective of a baseline survey is to capture the initial picture of the situation, against which the picture at the conclusion of the intervention can be compared in order to assess the impact – the final and sustainable effect of the intervention. Impact assessment requires information...
not only on what has changed over the course of the intervention, but also on what part of that change may be attributed to the intervention as such and what part is likely to be the result of external changes during that period. It also seeks information on what factors or mechanisms – specifically those related to the intervention – are likely to have contributed the achieved impact. These broad requirements have several implications for the design of the baseline survey. (1) In order to measure the impact, the baseline must be conducted prior to the start of the intervention, measure quantitatively a set of relevant indicators defining the sought for impact, and must use a repeatable methodology for that measurement so that comparable figures can be produced on conclusion of the intervention. (2) In order to partition the impact between the intervention and external factors, information must be obtained on relevant external factors. (3) In order to understand what factors or mechanisms may explain the impact, quantitative measures have to be supplemented by an assessment of more in-depth qualitative aspects.

B. Project design and monitoring

Alternatively, the baseline survey may be designed for the broader context of project design, refinement, targeting, monitoring and evaluation – as a tool for gathering data on conditions existing before the intervention, and as a planning and implementation tool for projects concerned with improving the targeting of interventions. In fact, a series of surveys may collect data for use at each stage of the programme cycle: on initial (baseline) conditions, as well as on subsequent project design, implementation and monitoring. Practical constraints may imply that these broad objectives replace, or at least push to a subsidiary role, the objective of an impact assessment as such. However, incorporation of that objective as well is not precluded in principle.

C. Supplementing rapid assessment studies

In the context of the RA methodology, a baseline survey has been described as a quantitative supplement to the essentially qualitative RA (ILO-UNICEF, 2005). A primary objective of the baseline survey in this context, like that of the RA, is to generate empirical data on child labour with special emphasis on the worst forms.

To conduct a baseline survey for this purpose, an appropriate choice of the unit where the working child is to be found and observed is crucial. The conventional definition of an establishment (see Section 2.12.3B) is too limited for the purpose of a baseline survey aimed at worst forms of child labour. This is because “the types of workplace can be as diverse as the types of WFCL. In the WFCL, workplaces can include factories and other formal workplaces, as well as workplaces lacking fixed premises” (ILO-UNICEF, 2005). Consequently, the survey sites can range from household-based to workplace-based to other location-based sites. The child’s workplace, rather than the establishment, is the appropriate survey unit. Location-based site may be used as the general term to refer to “non-standard units where children engaged in the WFCL are sampled. These locations are usually their workplaces, and they can range from a street intersection to a fishing boat to plantations to dumpsites to a variety of other sites” (ibid).
D. Surveys of particular sectors

However, it appears that hitherto, most baseline surveys have followed the conventional cross-section design of quantitative surveys, though often supplemented by qualitative information from special categories of respondents. The main application of this type of baseline survey has been to investigate the size and characteristics of particular sectors of child labour activity, focussed on hazardous forms of child labour. Often, though not always, such surveys have been establishment-based surveys. Usually, the surveys have been confined to limited geographic areas, but there are also many applications at the broad regional or national level.

In Table 2.5 are listed a few examples of sectoral and multi-sectoral surveys, which have been published as ‘baseline surveys’. A number of surveys among these are described in more detail as illustrations of national studies, in Sections 4.4 and 5.11 in the context, respectively, of sampling frames and sampling for establishment surveys.

E. Multiple-sector surveys

Finally, some baseline surveys have covered multiple sectors. For example, the list Table 2.5 includes an integrated baseline survey covering 45 hazardous sectors (Bangladesh, 2006b), a street children survey (Bangladesh 2003d), a household-based baseline survey (Uganda, 2010), and informal sector surveys in these two countries (Bangladesh 2003a, and Uganda 2004).
2.12 Sources of information on child labour

Table 2.5. Examples of baseline surveys covering various sectors

<table>
<thead>
<tr>
<th>Country</th>
<th>Sector(s)</th>
<th>Units</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bangladesh</td>
<td>Welding establishments</td>
<td>establishments</td>
<td>national</td>
</tr>
<tr>
<td></td>
<td>Automobile workshops</td>
<td>establishments</td>
<td>national</td>
</tr>
<tr>
<td></td>
<td>Battery recharging and recycling</td>
<td>establishments</td>
<td>national</td>
</tr>
<tr>
<td></td>
<td>Road transport sector</td>
<td>establishments</td>
<td>national</td>
</tr>
<tr>
<td></td>
<td>Integrated survey covering 45 hazardous sectors</td>
<td>establishments</td>
<td>national</td>
</tr>
<tr>
<td></td>
<td>Street children survey</td>
<td>area units, children</td>
<td>national, excluding areas with no street children</td>
</tr>
<tr>
<td></td>
<td>Children working in small and informal sector establishments</td>
<td>locations/sites</td>
<td>6 metropolitan cities</td>
</tr>
<tr>
<td>Pakistan</td>
<td>Glass bangles</td>
<td>establishments</td>
<td>Hyderabad district</td>
</tr>
<tr>
<td></td>
<td>Coal mines</td>
<td>establishments</td>
<td>3 districts</td>
</tr>
<tr>
<td></td>
<td>Tanneries</td>
<td>establishments</td>
<td>Kasur district</td>
</tr>
<tr>
<td></td>
<td>Surgical instrument manufacturing</td>
<td>establishments</td>
<td>Sialkot district</td>
</tr>
<tr>
<td>Uganda</td>
<td>Informal sector</td>
<td>households, establishments, (mobile) sites</td>
<td>4 districts</td>
</tr>
<tr>
<td></td>
<td>All sectors</td>
<td>enumeration areas (EAs), households</td>
<td>3 districts</td>
</tr>
<tr>
<td>Turkey</td>
<td>Cotton harvesting</td>
<td>villages, households</td>
<td>Karatás district</td>
</tr>
<tr>
<td>Jamaica</td>
<td>Mainly fishing, with limited coverage of some other activities</td>
<td>Various</td>
<td>2 fishing communities</td>
</tr>
</tbody>
</table>

2.12.6 Complementary nature of the sources

It is important to note that the various sources of information on child labour are generally complementary, in the sense of providing information on different aspects of the child labour situation. Of course, different sources can also overlap in the type of information provided, and may not always be consistent with each other.

The practical implication of the sources being complementary is that all of the information requirements cannot, and need not, be met with a single data source or survey approach. As noted in ILO (2004a), “experience has demonstrated that collecting comprehensive data on child labour is an exceedingly challenging task, and no single survey method may in itself satisfy data needs”.

This is firstly because of complex data needs: “Children are found working in a vast array of circumstances, and no single technique can be devised to survey all of these situations. Furthermore, policy analysis and targeted project intervention require information from a variety of potential respondents who may influence the life and development path of the child. These include the children themselves, parents or guardians, employers, school teachers, community leaders, child peers, and siblings. Circumstances in the home, school, workplace, and larger community to which the child belongs all bear on child labour outcomes and characteristics. To collect all relevant data from all relevant parties by means of a single survey or on a single occasion is impossible”. 
Secondly, we need to face the special problems in enumerating worst forms of child labour. It can be very difficult to make contact with children engaged in WFCL in order to collect the necessary information. As WFCL usually remain hidden, adequate sampling frames do not exist for their enumeration. Nor can the required samples be designed and selected without prior information on the location, characteristics and circumstances of the children engaged in the worst forms of child labour. Consequently, conventional surveys are inadequate for this purpose, and special sampling and enumeration procedures must be employed.

2.13 Surveying child labour through households

Sampling strategies for household-based child labour surveys have been elaborated in the ILO book *Sampling for Household-based Surveys of Child Labour* (Verma, 2008). In the following we summarise some important aspects.

Household-based sampling provides an efficient approach for estimating the prevalence and characteristics of the predominant forms of child labour for children living in private households, irrespective of whether the work is performed at home or outside. As noted in Sections 2.12.1–2, such surveys may be conducted as stand-alone surveys, or as separate but linked operations, or simply as modules attached to other national household-based surveys such as a labour force survey. The statistics generated by these surveys may include economic activities and non-economic activities (such as household chores) of children, working hours, nature of the tasks performed, health and safety issues including injuries at work, and also background variables such as demographic and social characteristics of household members and other basic characteristics of the household. The household survey content may vary from being detailed and specialised, to being confined to a few basic items. Data from household-based surveys may be augmented by supplementary (often linked) sources such as surveys of schools, employers, communities, informed individuals, etc.

As a generic term, we use ‘a household-based child labour survey’ to indicate a household sample survey, the main objective of which is to provide information on the phenomenon of child labour – its prevalence, distribution, forms, economic sectors etc. as well as its conditions, characteristics and consequences. While retaining this general and descriptive use of the term ‘child labour surveys’, it is useful to keep in mind two quite different survey types, which differ in their objectives, or at least in the emphasis given to different objectives.

2.13.1 Child Labour Survey (CLS) and Labouring Children Survey (LCS)

The defining factor in this distinction is the relevant base population for which the survey estimates are generated. Essentially, all children within specified age limits are included for the CLS, and only those considered to be in child labour are included for the LCS.
A. Child Labour Survey

The CLS refers to surveys where the primary objective is to measure the prevalence of child labour. The surveys may also study variations in this prevalence by geographical location, type of place (urban-rural), household type and characteristics, the household’s employment and income situation, children's age and gender, and similar factors. The target population of a survey with these objectives is the total population of children exposed to the risk of child labour. This base population is defined essentially in terms of age limits, and therefore tends to be well-distributed in the general population. The size and structure of the sample is determined largely by the size and distribution of the population of all children, or more commonly by its approximation – the size and distribution of the general population.

B. Labouring Children Survey

We have a different type of survey when the primary objective is to investigate circumstances, characteristics and consequences of child labour: what type of children are engaged in work-related activities, what types of work children do, the circumstances and conditions under which children work, the effect of work on their education, health, physical and moral development, and so on. The objectives may also include investigating the immediate causes and consequences of children falling into labour. The relevant base population in the LCS is the population of working children. What is meant by the LCS concept is that, when the objective is to determine the conditions and consequences of child labour, as distinct from its prevalence among all children, then it is appropriate that the size and structure of the sample is determined primarily by the size and distribution of the population of working children.

The concept of a ‘labouring children survey’ does not imply that the ultimate units enumerated in the survey be only labouring children. On the contrary, it is normally necessary in such a survey to enumerate comparable groups of children not engaged in labour, so as to provide a control group for comparison with the characteristics and circumstances of child labourers. Nevertheless, this is a secondary consideration in determining the sample size and design of a LCS.

C. Labouring children vs. child activity vs. children’s surveys

In practice, the target populations of interest can be more diverse than the all children versus labouring children distinction of the CLS and LCS. The focus of all LCSs is on the details of child labour, or more generally, on children’s work-related activities. However, some surveys collect a broader range of information on children. The following three situations may be distinguished.

(1) Child labour survey: a majority of LCSs have as their main focus the study of conditions and consequences of child labour.

(2) Child activity survey: often the scope is broader, and covers all types of activity of children, including economic and non-economic activities, education, leisure, and even non-activity.

(3) Children's survey: occasionally, the scope is even broader to include more general information about children beyond their economic and non-economic activity, such as information on children’s health or housing conditions.
These variations have important sampling implications. For surveys focused on working children, option (1), the LCS samples should reflect closely the patterns of concentration of child labour. With their broader and more defused scope, children’s surveys, option (3), would require a design similar to that of the CLS, or may even incorporate the latter. However, even for this type of survey, and especially for a child activity survey, option (2), the measurement of characteristics and conditions of economic activity of children as such is likely to remain an important objective. A compromise design is therefore desirable which covers both working and non-working children but with greater weight given to the former. Such a compromise design would of course require a CLS-type operation preceding it, so as to identify – even if with limited precision – the level of child labour in the survey areas.

D. CLS vs. LCS: sampling implications

Child Labour Surveys and Labouring Children Surveys can differ in major ways in terms of sampling requirements.

(1) As noted, the target population (and hence the required structure and distribution of the sample) of the CLS tends to be very similar to that of a survey of the general population, in particular to the Labour Force Survey (LFS), to which the CLS is very similar in concepts, definitions and survey content. The target population of the LCS – whether the population of children engaged in any work-related activity, or defined more narrowly as these engaged in specific forms of child labour – is, by comparison, smaller and more unevenly distributed, often in areas of concentration. Consequently, the sample design required is generally also different from that of LFS or CLS.

(2) The two survey types differ in their size and complexity. The CLS is normally less intensive and requires larger sample sizes. The primary statistical consideration dictating its sample size is the precision with which the proportion of children engaged in child labour is to be estimated and the reporting domains requiring separate estimation. By contrast, for investigating the detailed conditions and consequences of child labour, the LCS is more intensive in data collection, often involving interviewing the guardians as well as the children concerned separately, collecting attitudinal and other qualitative data, and carrying out associated enquiries such as at the school or the place of work of children in the sample. Consequently, the appropriate sample size for a LCS is likely to be much smaller than that for a CLS in similar circumstances. For an intensive survey such as the LCS, large sample sizes are often unnecessary from the statistical point of view, and in any case are precluded by practical and cost considerations. Having too large a sample in an intensive survey can in fact damage the quality and value of the information collected, in so far as it hinders close control over the survey operation.

E. Linkage of CLS to a base operation

Normally the CLS is linked with the operation preceding it. There are three dimensions: (i) the preceding operation may be a household listing operation, or it may be a large-scale survey such as the LFS; (ii) the CLS may be combined with that operation, or be conducted subsequently as a separate operation; or (iii) the CLS may be conducted on
the same sample, or on a sub-sample of the preceding operation. These options may be combined in various ways. National surveys provide a variety of arrangements.

F. CLS-LCS linkage

A labouring children survey (LCS) requires a prior operation which identifies, explicitly or implicitly, a sample of working children. Normally this is the function of the CLS component. Two forms of the relationship between CLS and LCS may be identified: (i) an integrated CLS-LCS operation; and (ii) linked CLS and LCS operations.

The distinction discussed above between the two survey types, the CLS and the LCS, by no means implies that they must be organised as two separate operations. In fact, in a majority of the national child labour surveys conducted so far, they have been completely integrated into a single survey operation – the same survey covering the two different objectives. The LCS questions form additions to the CLS questionnaire, which become applicable if the child concerned is found to be engaged in work-related activity. This is an integrated design, in which information for the CLS and LCS components is collected as a single interview operation.

An alternative to the integrated design is to conduct the CLS and LCS components as separate operations. However, these two cannot be stand-alone (i.e. entirely separate) surveys, but must be linked to each other in some way. The LCS sample can be identified on the basis of the CLS results in different ways, and this provides different forms of linkages between the two surveys. The best solution depends on the particular situation and objectives, but primarily on two factors: (i) the time gap between the two surveys; and (ii) the quality of screening provided by the CLS in identifying the presence of working children. A diversity of options in CLS-LCS linkage is possible, and is found in national designs. Many examples and details of sample selection procedures for different arrangements may be found in Verma (ibid).

2.13.2 Household-based vs. child labour surveys of other types

The appropriate survey approach depends on the circumstances of child labour, specifically, on the relationship of the child and of child labour with the household. Several scenarios may be identified. Please refer to Section 1.2.2 for a typology.

Generally, child labourers living at (their own) home - and possibly also some of those not living at home but still in touch with their family - can be enumerated through household-based surveys. Working children in other circumstances generally require data collection approaches other than the conventional household survey. Sampling approaches for dealing with such situations are elaborated in subsequent chapters.
2.14 On methodology of rapid assessment (RA) studies

Rapid assessments collect information from many sources, from respondents of different types, and also on many complex variables from the same respondents so that the variables can be related to each other. This gives rapid assessment studies the potential to throw light on ‘how things work’ concerning (at least some aspects of) the complex reality of child labour, and also the potential to access children subject to the worst forms of child labour. The information obtained is essentially qualitative, and not statistically representative of the target population in the sense that it cannot be generalised to this population. A major part of the emphasis in rapid assessment studies is on case studies, which investigate and describe the condition of individual child labourers or of other individuals connected to those children. This by no means implies that the value of the information is confined only to the particular individual cases discussed in a RA report. Usually, limited generalisations to a broader context are possible, if made with due caution.

In selecting individuals to be included in the study, an attempt is made to include cases which are ‘typical’ in some sense of the various categories of child workers of interest. In so far as this is successfully achieved, the study observations can be taken to be indicative of those categories. Extension beyond that to other parts of the target population depends on some assumed similarity of conditions and mechanisms.

As noted in Section 2.12.4, the rapid assessment methodology is intended to provide relevant information in a relatively quick and cost-efficient way. It is a research method that applies various strategies to achieve a relatively rapid understanding of a specific problem or issue while employing limited resources of time and money compared to other forms of research such as large scale surveys. Detailed methodology and implementation guidelines have been published in ILO-UNICEF (2005), preceded by a clear but briefer exposition in ILO (2004a). There is also a very useful ILO-IPEC study on the lessons learnt, based on experiences from selected rapid assessments carried out during 2000 to 2002, by Fee (2005). In the following subsections we briefly summarise rapid assessment methods and their merits and limitations.

2.14.1 Methods

RA is usually used in a local or regional context or in urban settings, where populations with problems are known to exist. The methodology uses several methods in combination, incorporating them in a child-centred, participatory approach to data collection. These methods are ideally suited for obtaining detailed knowledge of the working and life circumstances of children by means of discussions and interviews. They have proved to be effective ways to gather information on hard-to-access worst forms of child labour. The methods used include:

- compiling background information from diverse sources;
- observation;
- focus group discussions;
in-depth interviews and conversations with key informants, employers, parents and children; and

- interviews using semi-structured or structured questionnaires.

In principle, the information collected can be part descriptive and part statistical, and the mix depends on the particular methodologies used. The output is primarily qualitative and descriptive.

### A. Background information

Sources such as the following may provide relevant information: historical accounts (population, industry, labour force); national censuses and surveys; reports from international organizations, NGOs, government agencies; public officials; trade unions, employer and business associations; industry and company reports; research studies of specific occupations or population categories; community activists, women’s organizations, social-work and charitable associations; schools; newspapers and magazines; and information on national and local legal framework as relevant.

Special areas of study demand specific background information, as for instance in relation to armed conflict.

### B. Information from the field

Some of the sources of information used by RA researchers are described briefly below, summarized from ILO-UNICEF (2005) and other above-mentioned sources. Ethical issues must be considered in collecting field information. Any information coming from multiple sources needs to be cross-checked. The specific research components drawn upon in any given RA will vary depending on the objectives, and the resources and possibilities that are available in a given situation. These may include:

- systematic visual observation of child workers and of work places in various parts of the study area, seeking information about activities and working conditions;
- making sketch maps of the area, showing its physical layout and locations where children congregate for work;
- canvassing (i.e. contacting and conducting brief interview with) households door-to-door in selected areas;
- conducting in-depth discussions with key informants;
- conducting focus group discussions (formally structured or spontaneous) with small groups of adults and children;
- conducting interviews and conversations with children’s employers, parents, teachers, and with others relevant to the children’s work and life; and
- conducting individual interviews with children in the sample selected, and possibly also with some non-working children.

Focus groups refer to discussion groups assembled to talk about a particular subject, issue or problem, usually led by a facilitator. They are widely used in social research and marketing and have also proven useful in child labour rapid assessment studies. The group members can be drawn from various categories of individuals (e.g. older
2. Child labour situations, data needs and sources

working and non-working children, parents or guardians, neighbours, local officials, labour inspectors, employers, teachers).

C. Primary informants

These may include the following: government officials, including at the district or local level; government labour inspectors; police officers; trade union officials; teachers; employers; community leaders and members; representatives of NGOs and international organizations who have worked with children in the area; former child labourers and their families; parents of current child labourers; and parents of boys and girls who do not work.

Some of the more helpful among them may become ‘key informants’.

D. Type of work place

The choice of places which might be considered for observing and finding children is very specific to the study focus and local conditions. Child workers may be found concentrated at some of these locations, while other working children may be dispersed over the study area. Observations may be made at locations such as the following (the list is by no means exhaustive):

- places of production activity - factories, workshops, informal establishments, home-based establishments, rural enterprises including carpet-weaving loom sheds, small-scale mining, etc.;
- marketplaces for child vendors, porters, assistants, and domestic workers;
- transportation depots - ferry landings and ports, railway stations, bus terminals - for child beggars, vendors, porters, or victims of sexual exploitation;
- garbage disposal and dump areas, for child scavengers and ragpickers working individually or alongside their families;
- urban commercial areas for child street workers;
- locations where commercial sexual transactions occur - streets, parks, bars, hotels, restaurants, tourist areas, dance halls, brothels;
- tourist areas - restaurants and hotels, beaches, parks, cafés - especially in the tourist season;
- city parks, squares and religious places where, for example, immigrant and child domestic workers may gather;
- drop-in centres and rehabilitation and child-protection facilities, often frequented by child street workers or sexually exploited children;
- informal education centres, which are often attended by working children;
- roadside restaurants, especially the back areas where child helpers work;
- seaside and port areas and villages, for children who work in fishing and related activities;
- agricultural or crop cultivation areas, especially where migrant or other hired families may be working with their children; and
- village water sources, especially for child domestic workers.
E. Sample for interviewing children

The purpose of using samples is to ensure that those actually interviewed are in some definable way representative of the target group being studied.

The overall sample size of RA studies is constrained by the available resources, both budgetary and personnel. Usually a quota type of sample is used. This involves identifying relevant categories of children in the target population and, if possible, obtaining some idea of the size of these categories. This gives a basis for determining the target sample size for each category. Children may be included in the survey simply as they are found and agree to participate, so long as the quota sample size allocated to each category is reasonably met.

If access to individuals to interview is very difficult - for example in investigating children engaged in armed conflict, sexually exploited children, and children working in other hard-to-reach areas – it may not be possible to control the distribution of the sample by category as above. It is desirable if it can be done at least roughly. Otherwise, one has to include children as they become available for the survey without involving any formal sample selection.

Often, the problem is finding enough children to participate in the survey. Special procedures may have to be used to augment the cases directly found. For instance, the snowball sampling technique, or its development called respondent-driven sampling, can provide a way of accessing new informants through other informants as they become known during the field work (see Chapters 13 and 14). Such techniques can be used with or without controlling the sample quotas according to the identified categories of working children.

2.14.2 Potential uses of an RA

RA methodology offers great potential for uncovering rich information about the population and delving deeply into the issues involved. Its findings serve many purposes, such as the following (ILO-UNICEF, ibid):

- RAs have been especially effective in publicising the conditions of children in WFCL, and to induce governments to consider adhering to the international conventions and agreements aimed at improving the situations of child workers. They can also raise awareness and publicise conditions of working children in general, and bring about shifts in perceptions of child labour, and have in many cases directly influenced the creation of public policies.

- RAs can provide inputs for the design and implementation of policies and programmes. They can also provide guidance for mobilizing resources, targeting funding, and planning preventive measures and alternatives.

- Good quality RA reports can stimulate further research, sometimes in other communities and in other sectors of child labour. The research process can also serve to strengthen the technical capacity of researchers, whether from universities, government agencies, NGOs, trade unions or employers’ organisations.
If their coverage of the target population is not too restrictive, RAs – and also similar studies such as small-scale baseline surveys – can provide a picture of the child labour situation and indicative estimates of its important characteristics. Such information can provide the basis for developing scenarios for more representative surveys and studies with the objective of producing more comprehensive and precise estimates of the size and characteristics of child labour in the population. A possible scenario of this type is illustrated in the case study presented in Annex C.

### 2.14.3 Limitations of an RA

One commonly cited drawback of RA is the fact that it applies to small populations in limited areas, and its results cannot be generalised to other populations of working children, even those working in the same occupation in nearby provinces or regions. Unlike a national survey, which uses samples meant to be statistically representative of the population being investigated, RA produces research on a much narrower basis. Although researchers may strive to carefully select their locations, occupations and working child populations to be as representative as possible, it must still be demonstrated that they actually succeed in doing so. The assumption cannot be automatically made that living and working conditions, family situation, schooling and literacy rates, or the push-pull factors that encouraged children to work are similar in different geographical areas or in different sectors of economic activity. In a nutshell, RA results essentially apply only to the population surveyed, and can be extrapolated or generalised only on the basis of some additional assumptions.

Another limitation of RA research is the subjectivity of the researchers. Because rapid assessments rely upon observation and conversations as well as semi-structured interviews for much of their information, the researchers’ and assistants’ own assumptions, biases and cultural blind spots can cloud the findings. Methods must be developed to assess and reduce this problem. One solution is to implement careful ‘sensitivity training’ to research and field workers, including training in gender sensitivity. Another is to present and discuss research findings frequently among members of the research team while the research is in progress, so that the team acts as a check on individual researchers’ interpretations. A procedure that is more or less standard is to verify the accuracy of data by crosschecking information from different sources and procedures. Still another is the use of control groups to permit comparison between working and non-working children.

### 2.15 Household surveys to rapid assessment studies: bridging the gap

National household-based surveys of child labour on the one hand, and typical rapid assessment studies on the other, form two ends of the range of application of the various sampling techniques addressed in the following chapters. Household-based surveys of child labour generally use representative (i.e. relatively large and probability) samples of the general population. The information collected tends to be extensive rather than in-depth. By contrast, rapid assessment studies are small-scale, but intensive.
Figure 2.4 displays the range between large-scale probability sampling (typified by a household-based survey of child labour), and limited case studies (typified by a rapid assessment). The graphical representation is of course a simplification, but hopefully is useful in communicating the intended place of the various sampling techniques described in the following chapters in bridging the gap between extensive quantitative and intensive qualitative data sources on child labour.

**Figure 2.4. The range from household-based surveys to rapid assessment studies**

Starting from large-scale, probability-based sampling, the rectangular boxes in the diagram represent increasingly difficult sampling situations A to D (see text). The placement of a box along the y-axis indicates the extent to which probability sampling is achieved, 0 = purely purposive, haphazard sampling, 1 = fully representative probability sampling. The position of a box along the x-axis indicates how far in-depth qualitative information is collected. Generally these two aspects are negatively related. The objective of improved sampling and data collection procedures is to move upwards (toward more representative sampling) and to the right (obtaining richer qualitative information), as indicated by boxes in dotted lines. Various improved sampling procedures are described in the following chapters as indicated under the boxes in the diagram.
From left to right, the text boxes represent increasing difficulty in sampling the population: starting with (A), the sampling frame being imperfect; to (B), imperfect frame plus the added problem of the population being rare; to (C), where in addition to the above problems we have the problem of the population being mobile; and (D), where added to all those is the problem of the population also being reclusive, in the sense defined in Chapter 1.

In Figure 2.4, the height of the boxes along the y-axis is a representation of the degree to which probability sampling is achieved, starting with 0 per cent at the bottom to 100 per cent at the top. The dotted boxes, above the original box in each case, represent the aspired improvement in the probability basis of the sample by using better sampling procedures, for example of the type described in the chapter numbers indicated below each box in the diagram.

Consider, for example, the studies of street children, which has been the topic of many rapid assessment studies. In Cairo, Egypt, the capture-recapture procedure (Chapter 11) has been used to produce more reliable quantitative estimates of the number of street children (CAPMAS 2009); and in Bamako and also in Accra, respondent-driven sampling method (Chapter 14) has been used for the same purpose (Hatløy and Huser, 2005).

The position of the boxes from left to right is meant to indicate the degree to which in-depth qualitative information is obtained. We do not discuss in this book data collection methodologies which may be used for enhancing such information over the wide range of situations shown, except for noting here that the rich experience of the RA studies would certainly form the basis for developing such methodologies.

Moving a box upwards and to the right indicates that both the probability basis and the qualitative information collected have been improved.

At the bottom of the figure, some examples of surveys of different child labour situations have been placed from left to right to suggest (even if vaguely) where they may lie in the increasing level of difficulty in achieving probability sampling, and at the same time, in the increasing role of qualitative information. See further explanatory notes at the bottom of Figure 2.4.
Chapter 3

Basic sampling and estimation procedures

This chapter is aimed at clarifying some basic sampling procedures involved in the most common designs used for child labour studies, with a view to providing useful background (and reminders) for the technical discussion in subsequent chapters.

With sampling, we draw inferences about the whole population based on information only from some of the units in the population. The validity of these inferences depends on the manner in which the sample is drawn and the size of the sample. The design of samples involves determining:

(1) the sample size, that is the number of units of analysis to be selected for the survey;

(2) sample structure, i.e. how those units are to be selected; and

(3) estimation procedures, i.e. how the results from the sample are to be used to draw inferences about the population from which the sample was selected.

Sample design cannot be isolated from other aspects of survey design and implementation. In practice the sample design must take into account numerous considerations other than sampling theory. Sample size and structure affects the magnitude not only of sampling error, but also of most components of non-sampling errors. And of course it determines the cost of the survey.

Section 3.1 introduces the basic concept of probability sampling, noting how probability sampling may be achieved and its dependence on the availability of an adequate sampling frame.

In Section 3.2, we describe common departures from simple random sampling in real surveys – such as clustering, stratification, and unequal unit selection probabilities. Features of some most commonly used designs for child labour and similar surveys are noted.

Section 3.3 describes a very commonly used selection procedure – sampling with probability proportional to some measure of unit size (PPS), and the use of appropriately scaled size measures to facilitate its application. The procedure of systematic sampling is also described in this context.

Section 3.4 considers in detail issues concerning weighting of sample data. A step-by-step procedure is described for computation of unit weights, taking into account variations in design selection probabilities, non-response and calibration against external standards.

Practical methods of estimating sampling errors in the presence of a complex sample design are outlined in Section 3.5. Linearization and Jackknife Repeated Replication procedures are described briefly.
Analysis of design effects is addressed in Section 3.6 – introducing the concept of design effect and its components, and describing procedures for their estimation.

Sampling errors are not the only quality concern in surveys. Non-sampling errors are also important – often more critical in surveys dealing with complex topics. A typology of survey errors is presented in Section 3.7, drawing the important distinction between errors of measurement and errors of estimation in survey data.

### 3.1 Probability sampling

Inferences from the sample to the whole population can be drawn on a scientific basis only if the sample is composed of units selected using a randomised procedure which gives a known non-zero chance of selection to every unit in the target population. A sample drawn in this way is called a **probability sample**. The term *random sample* is commonly used to mean the same thing. The major strength of probability sampling is that the probability selection mechanism permits the application of statistical theory to obtain estimates of population values from the sample observations in an essentially objective manner, free from arbitrary assumptions or subjective judgement.

A similar but more demanding and inclusive concept is that of *measurability*. A sample is said to be measurable if it provides estimates not only of the required population parameters (as does a probability sample), but also of their sampling variability.

The design of a random sample specifies the type of randomised procedure applied in sample selection, and how the population parameters are to be estimated from the sample results. As to the *selection procedure*, many designs are possible and used in practice. The procedure may for example give the same (equal) chance of appearing in the sample to all elements in the population, or some units may be given a higher chance than others. We may select the elements individually, or first group them into larger clusters and apply the selection procedure to those clusters. We may partition the population into strata and then apply any of the above procedures separately within each stratum. The *estimation procedure* involves the statistical or mathematical forms in terms of sample values and possibly also of information from other sources external to the sample; it provides estimates of population parameters of interest. The procedure also includes the estimation of measures of uncertainty (sampling error) to which the sample results are subject.

#### A. Obtaining a probability sample

In order to obtain a probability sample, certain proper procedures must be followed at the selection, implementation and estimation stages. These include the following.

1. Each element in the population (each analysis unit) must be represented explicitly or implicitly in the frame from which the sample is selected. For this we ideally need a **complete sampling frame**.

2. The type of units in the frame from which a sample is selected may be the same or may be different from the type of analysis units of which a sample is required. In any case, it is necessary to have unambiguous rules of association which link sampling and analysis units.
The sample must be selected from the frame by a process involving one or more steps of automatic randomisation, which gives each unit a specified probability of selection.

At the implementation stage, all selected units - and only those units - must be included in the survey and successfully enumerated. This means avoiding non-response, and not permitting any substitution for non-respondents.

In estimating population values from the sample, the data from each unit in the sample should be weighted in accordance with the unit’s probability of selection, and to take into account other factors resulting from imperfections in sample implementation.

In practice, some approximations in the implementation of the ideal requirements of probability sampling are often unavoidable due to reasons such as under-coverage, sample selection errors, non-response, and biases in the estimation procedure. It is a matter of practical judgement as to the level of errors up to which a sample may still be considered effectively a probability sample.

### B. Sampling frame

The population to be surveyed has to be represented in a physical form from which samples of the required type can be selected. A sampling frame is such a representation. In the simplest case, the frame is simply an explicit list of all units in the population, from which a sample of the units concerned can be selected directly. With more complex designs, the representation in a part of the frame may be only implicit, but still accounting for all the units. A frame may be constructed from a single source, or may have to be compiled by combining information from a number of sources. Different types and/or sources of frames may be used for different parts of the population. It is also possible to use multiple frames in combination to represent the same population more adequately. The use of multiple frames raises special issues in determining the units’ selection probabilities. Issues and problems relating to sampling frames are discussed in Chapter 4 and are often referred to in subsequent chapters.

### 3.2 Departures from simple random sampling

The simplest design is one in which every possible set of \( s = 1 \) to \( n \) units from a population of \( N \) units receives the same chance of selection. This is a 'simple random sample' (SRS) of size \( n \). There is no preference as to which units or which combination of units appear in the sample. The sample design may depart from simple random sampling in a number of ways, the three common and important ones being the following.

#### 3.2.1 Stratification

Stratification refers to partitioning the population before sample selection. Within each part, a sample is selected separately (independently). In each part or stratum, the design may involve other complexities such as multi-stage sampling, and may differ from one stratum to another.
3. Basic sampling and estimation procedures

In so far as the strata represent relatively homogeneous groupings of units, the resulting sample is made more efficient by ensuring that units from each grouping are appropriately represented in a controlled way. Control through stratification reduces the danger of getting a poorly distributed sample by chance. Apart from statistical efficiency, it may be necessary or convenient to introduce stratification for other purposes. For instance, when data are required separately for sub-divisions of the population, it is desirable to treat each subdivision as a separate stratum, to be sampled independently. Stratification permits flexibility in the choice of the design separately for each part of the population with different sampling requirements and problems. Stratification may also be introduced for administrative convenience. Also, a sample well-controlled in its distribution across different parts of the population is likely to be more easily accepted by the general public.

### 3.2.2 Clustering or multi-stage sampling

Concentrating the units to be enumerated into clusters reduces costs in the preparation and maintenance of the sampling frame. It also reduces travel and other costs of data collection. For the same reason, it can improve the coverage, supervision, control, follow-up and other aspects determining quality of the data collected. Administrative convenience in implementation of the survey can be another important advantage.

The major disadvantage of clustered or multi-stage sampling is the increase in sampling error compared with that in a simple random sample of the same size. There can also be some loss in flexibility in the sample design and in the targeting of the sample to populations with particular characteristics since the selection of ultimate units of any given type cannot be controlled separately, in so far as units of different types are mixed within the same clusters. Complexity of the design also increases the complexity of analysis of the survey data.

### 3.2.3 Unequal selection probabilities

There can be various reasons for introducing unequal selection probabilities for units in the final sample. We may divide them into three broad types.

1. Different reporting domains or population subgroups may be sampled at different rates in order to meet specific reporting requirements.

2. Then there are ‘unnecessary’ departures from equal selection probabilities, unnecessary in the sense that they are not dictated by the survey objectives but arise from the particular survey procedures or conditions, such as shortcomings in the sampling frame. For example, effective sampling probabilities of units in a sampling domain are reduced if the sampling frame for the domain is incomplete (misses out a proportion of units), and no compensation is or can be made for this shortcoming. Often frame under-coverage – or at least the degree of under-coverage present – is not known.

3. Variation in sampling probabilities may be introduced to make the sample design more efficient (reduce variances and/or costs). An example is ‘optimal allocation’ involving oversampling of strata with units which are more diverse and/or less costly to enumerate. Similarly, ‘calibration’ weights, which make the sample
conform to some more reliable external control totals and distributions, may be introduced with the aim to reduce sampling variance and bias.

Variations in unit selection probabilities of the type (1) and (2) are external and in that sense essentially arbitrary, not connected to levels and variances in the population. Such variations generally increase sampling error in the total sample. By contrast, variations of type (3) can result in reduced variance or bias, but not necessarily always or significantly.

### 3.2.4 A typical design

#### A. Primary and other higher-stage sampling units in a multi-stage design

Most samples for child labour surveys are selected in multiple sampling stages. For instance, the whole study area may be divided into smaller area units (localities, census enumeration areas, etc.), and a sample of these area units selected at the first stage. PSUs of other types may be involved instead, such as institutions, locations, workplaces, even ‘time-location’ units of the type discussed in Chapter 10. Units selected at the first stage are called primary sampling units (PSUs). The sample is selected from a list of such units, lists covering the entire population exhaustively and without overlaps and also providing information for the selection of units efficiently. Such a frame is called the primary sampling frame (PSF). These aspects will be discussed more fully in Chapter 4.

The next stage may consist of dividing each of the PSUs selected at the first stage into smaller areas such as blocks, and then selecting one or more of these units. As above, units of other types may be involved, for instance institutions or establishments within area units selected at the first stage. A frame is required only for the units lying within the PSUs selected at the first stage. In many designs, no such intermediate stage is involved at all: it is common to use two-stage samples.

#### B. Ultimate units

The final stage involves the selection of ultimate sampling units (USUs). The sample, and hence the frame required for their selection, is confined to the higher stage units selected at the preceding stage. This frame is normally in the form of lists of the units to be selected. Children or working children are the main ultimate units of analysis for surveys of child labour. The USUs may refer to these ultimate units themselves, or may be small clusters of these units, such as households or establishments containing children. In the latter case, we have essentially four options in selecting children from the USUs:

1. Take all children (or all working children) in each larger unit selected.
2. Take a sample of children in each larger unit selected.
3. Make a combined list of all the children found in the selected sample of larger units, and select a sample of children from it directly, without making further reference to the larger units individually.
4. Dispense altogether with intermediate level units such as households or establishments, and go directly to listing and screening of children in the study area.
Each scheme has its advantages and drawbacks in terms of statistical efficiency and costs. This will be discussed more fully in Section 5.5. Schemes (1) and (2) are more commonly used than the other schemes.

C. Single-stage designs

In statistically more developed countries, even large-scale surveys such as the LFS are increasingly using sampling of households or persons in a single-stage design. This is due to the increasing use of telephone interviewing (though normally some face-to-face interviewing remains, e.g. for the initial interview, for non-response follow-up, for persons without telephone, etc.).

In surveys concerning child labour, especially in developing countries, single-stage designs may be used when the target population is found in large concentrations, such as in large establishments, institutions, worksites, or other places where children congregate. Normally all such locations are included in the sample automatically. Actually, large economic establishments are not common units in child labour surveys, since such labour occurs predominantly in small establishments. Another situation likely to involve direct single-stage sampling is when the scope of the study is restricted to a few pre-determined locations or worksites, as is the case, for instance, in many rapid assessment studies. Elements are selected directly within those few locations.

When the target population consist mainly of small establishments, a sample of those may be selected in a single stage. A common design is to take all working children in the smallest establishments, but introduce subsampling of working children within selected establishments among establishments employing more than a certain number of such children.

3.3 Probability proportional to size (PPS) sampling

3.3.1 Self-weighting sample

A most common technique for selecting multi-stage samples is to select the PSUs (possible also other higher stage units if involved) with probability proportional to some measure of size of the unit \((x_i)\), and within each selected unit, to select ultimate units with probability inversely proportional to the size measure. In Equation (3.1), the summation \(\sum()\) is over all PSUs in the population. Parameter \(a\) refers to the number of PSUs selected. The actual number of ultimate units of interest present in the PSU is \((x_i^I)\), as distinct from the PSU measure of size \(x_i\) used for its selection at the first stage.

A. Selection equations

Selection of PSUs

The first stage selection probability is

\[
p_{i1} = \left( \frac{a}{\sum x_i} \right) x_i = \frac{x_i}{I} \quad \text{with} \quad I = \left( \frac{\sum x_i}{a} \right).
\]
3.3 Probability proportional to size (PPS) sampling

Selection of ultimate units within selected PSUs

In a two-stage design, the second stage selection probability is

\[ p_{2i} = \left( \frac{b}{x_i} \right) \]

where \( b \) refers to workload or ‘sample-take’ per PSU. \( \text{ (3.2) } \)

Overall selection probability of an ultimate unit

\[ p_i = p_{1i}p_{2i} = \left( \frac{b}{x_i} \right) = p, \text{ a constant. } \] \( \text{ (3.3) } \)

Sample size for a PSU

\[ b_i = x'_i p_{2i} = b \left( \frac{x'_i}{x_i} \right). \] \( \text{ (3.4) } \)

If \( x'_i = x_i \) strictly, then constant \( b \) is the number of ultimate units selected from any PSU, with the resulting sample size \( n = ab \). Hence, in the ideal case, such a design yields the dual advantage of: (i) control over sample size and fixed workload \( b \) per PSU; and (ii) a uniform overall sampling probability \( p = (b/l) \) for all ultimate units (for households, children, etc.).

B. Imperfect size measures

In practice, it is unlikely that both – or even one – of these features can be achieved exactly. Most often, the size measures available in the frame are outdated and may not correspond well with the actual (current) sizes of the area units. The size measures \( x_i \) determining selection probabilities of units are often based on incomplete and approximate information from the past, and differ from the actual current size \( x'_i \) of the unit.

3.3.2 Some common variations

A. Constant probability design

With size measures \( x_i \) taken as a constant, say \( x \), we get a design with uniform probabilities at both stages: \( p_{1i}=(x/l), p_{2i}=(b/x) \). Hence the overall probability is uniform, \( p = p_{1i} p_{2i} = (b/l) \) as before, but the sample take from a unit varies in proportion to the actual size of the unit:

\[ b_i = (b/x) x'_i \]

B. ‘Fixed take’ design

A commonly used variant of the PPS self-weighting design is to take a prefixed size of the sample from each selected PSU:

\[ b_i = b, \] \( \text{ (3.4a) } \)

so that the selection probabilities become
As before, and
\[ p_{1i} = \frac{x_i}{l}, \quad p_{2i} = \left( \frac{b}{x_i} \right)^t, \quad (3.1a), (3.2a) \]

Hence the overall selection probability of an ultimate unit becomes
\[ p_i = p_{1i}p_{2i} = \left( \frac{x_i}{l} \right) \left( \frac{b}{x_i} \right)^t = p \left( \frac{x_i}{x_i^t} \right), \quad (3.3a) \]

Taking sample of a fixed, predetermined size for each sample PSU is unavoidable or preferred in situations such as the following.

1. The size measures available before sample selection \((x_i)\) are outdated and do not reflect well the actual unit sizes \((x_i')\) found subsequently during data collection. Hence the ratios \((x_i'/x_i)\), and therefore the PSU sample takes \(b_i = b (x_i'/x_i)\) under the self-weighting design vary too much. A fixed-take design suppresses this variation, at the expense, of course, of making the overall unit selection probabilities variable. Using fixed-take rather than self-weighting designs seems to be very common in practice. This may be for the reason that excessive variation in work-load between sample areas is a serious inconvenience.

2. Fixed-take designs may also be preferred in very ‘heavy’ surveys (e.g. involving very lengthy interviews, frequent and repeated visits to each household, or elaborate physical measurements), when even minor variations in the number of sample cases is felt to result in unacceptable variations in interviewer workloads.

3. There are situations when we are able to apply a more or less random selection procedure, but there is no list available of units in the population, and even the total number of units in the population may not be known. For instance, at entrances to a sample of locations, a random sample may be taken of persons passing through those entrances during selected time intervals; or mobile populations visiting specified locations may be sampled at random. In such situations, it may be simpler and more convenient to take an \textit{a priori} fixed number of units at each location.

C. ‘Take-all’ sampling, possibly with a cut-off

There are situations when it is appropriate to take all the ultimate units (households, children) in each unit selected at the preceding stage. In the context of child labour surveys, take-all sampling often arises from the need to ensure that sufficient numbers of working children are obtained for the survey from a limited number of sample PSUs.

For surveys of labouring children, it is usually necessary to concentrate the sample in areas with a higher concentration of such children. This means sampling the areas with size measures defined in terms of the number of working children expected in the area, and then taking all or most of the households (or households with working children) in each area selected. The probabilities of selection of the ultimate units vary according to their area: they are higher for children living in areas with a higher concentration of child labour. This and related techniques, useful for sampling rare populations, are discussed in detail in Chapter 6.

A commonly used option in such designs is to impose a limit on the maximum number of elementary units which will be selected from any PSU. If the unit has more elements...
3.3 Probability proportional to size (PPS) sampling

than that limit, subsampling will be introduced in order to keep the sample size at that limit.

The same considerations apply in selecting children from establishments, as discussed in Section 5.5.

D. Dealing with very large units

A ‘very large’ unit in the context of PPS sampling means that the unit is so large that its implied probability of selection \( p_{1i} \) in equation (3.1) exceeds 1.0, i.e. a unit for which the size measure \( x_i > I \). There are several methods of dealing with this problem, such as: segmenting the unit into smaller units of size not exceeding \( I \); assuming the unit to be automatically selected and taking it out of the selection procedure (such units are usually referred to as ‘self-representing units’); or simply redefining its size measure so that it does not exceed \( I \).

E. Dealing with very small units

A ‘very small’ unit in the context of PPS sampling means that the unit is so small that its implied probability of selection \( p_{2i} \) in equation (3.2) exceeds 1.0, i.e. a unit for which \( x_i < b \). Again, there are several methods of dealing with this problem, such as: grouping the units into bigger units; or redefining the size measures of small units so that the minimum value is \( b \). Very small units, if together they form only a small part of the population, are sometimes excluded from the survey altogether.

For further discussion of procedures for dealing with units of very large or very small size in the above sense, see Verma (2008, Sections 3.9-3.11, 4.4-4.7).

3.3.3 Different choice of size measures

The size measures \( x_i \) used for selecting units need not be the same as the actual unit sizes \( x_i' \), even as an approximation. Firstly, the two often refer to units of different types. For instance, in order to obtain a sample of households or of households containing children from sample areas, \( x_i' \) may refer to the number of such households in the area, whereas the frame available (say from the population census) gives \( x_i \) values in terms of the total population of individuals. The application of the selection procedure described above is then based on the expectation of a high correlation between the two measures (e.g. between the number of households and of persons in each area).

Secondly, for some purposes, we may wish to have a design different from the strict ‘probability proportional to size’ design considered above. For instance, we may wish to select areas with probability proportional to square root of unit size. For this case, we may define size measures as \( s_i = \sqrt{x_i} \), to approximate \( \sqrt{x_i'} \). The selection equations would then be as follows.

**Selection of PSUs**

\[
p_{ii} = \left( \frac{a}{\sqrt{\sum x_i}} \right) \sqrt{x_i} = \left( \frac{a}{\sum s_i} \right) s_i = \frac{s_i}{I}, \text{ say.} \quad \text{(3.1b)}
\]
Selection of ultimate units within selected PSUs

\[ p_{2i} = \left( \frac{b}{\sqrt{x_i}} \right) = \frac{b}{s_i}, \quad \text{where } b \text{ is some constant.} \]  

(3.2b)

Overall selection probability of an ultimate unit

\[ p_i = p_{ii} = p_{2i} = \left( \frac{b}{l} \right) = p, \quad \text{a constant.} \]  

(3.3b)

Sample size for a PSU

\[ b_i = x_i' p_{2i} = b \left( \frac{x_i'}{s_i} \right). \]  

(3.4b)

In the above, ignoring the difference between \( x_i \) and \( x_i' \), the PSU sample takes vary as \( \sqrt{x_i} \). Note that the size measure \( x_i \) may be scaled arbitrarily.

Finally, we may also note that parameter \( b \) may be given a clearer meaning by writing the constant in Equation (3.2b) in the following form, simply by introducing an additional constant \( \sum s_i / \Sigma x_i \):

\[ p_{2i} \]  

(3.2c)

This gives the sample take of PSU \( i \) as

\[ b_i \]  

Since the cluster has been selected with probability \( p_{1i} = (as_i / \sum s_i) \) (giving \( \Sigma p_{1i} = a \)), the expected value of sample take per PSU in the sample simplifies to:

\[ \bar{b} = (\Sigma b_i p_{1i} / \Sigma p_{1i}) = b (\sum x_i' / \sum x_i) = b / S, \]  

where \( S = (\sum x_i / \sum x_i') \) is the overall scaling of size measures \( x_i \) relative to the actual sizes \( x_i' \). Since the actual sizes are normally known only for the units selected and enumerated in the sample, we may use its sample value instead, \( s = (\sum x_i' / \sum x_i') \), where \( \sum (\cdot) \) indicates the sum over the sample, while \( \Sigma (\cdot) \) represented the sum over the whole population. Hence with parameter \( b \) as defined in Equation (3.2c), we have \( b = \bar{b} s \), so that parameter \( b \) can be viewed as a measure of the average sample take per PSU, except for the (essentially arbitrary) scaling of the size measures used for sample selection in relation to the actual unit sizes.

It can be seen that exactly the same interpretation applies to parameter \( b \) in Equations (3.1)-(3.4) in the standard PPS design. Indeed, the above applies with \( s_i \) defined as any arbitrary function of \( x_i \), \( s_i = f(x_i) \); the scaling factor \( S \) (or \( s \)) is not affected by such transformation in a PPS sample of the type given in Equations (3.1b-3.4b).
3.3.4 Systematic sampling

A. Basic procedure

This is a common and convenient method of selecting a sample. In a multi-stage design, it can be applied to any stage of selection.

Let us describe the procedure briefly as applied for the selection of a probability proportional to size (x_i) sample. The same procedure would apply with some other measure of size, such as s_i in the preceding subsection. Selection at a uniform rate is just a special case of this with x_i equal to 1.0, or some other constant.

To select units with probability proportional to some measure of unit size, let

x_i be the size measure of unit i,

\[ X_i = \sum_{j=1}^{i} x_j \] the cumulation of x_i values for units 1 to i in the population, ordered in some meaningful way,

X is the above sum over all units in the population, and

a is the number of units to be selected with PPS.

The sample from the list is selected by applying a sampling interval to it. The interval to be applied to the cumulative size measures X_i is

\[ I = X/a. \] (3.5)

A random number r in the range 0 < r ≤ I identifies the first unit selected: it is the first unit whose cumulative size equals or exceeds r. This is the unit the sequence number \( i \) of which satisfies the relationship

\[ X_{i-1} < r \leq X_i. \]

Then, starting with r, the selection point is increased each time by I, giving a sequence like

\[ r' = r + I, \quad r + 2I, \ldots, \quad r + (n-1)I, \]

and for each value of r' in this sequence, the unit \( i \) with cumulative size measure satisfying the relationship

\[ X_{i-1} < r' \leq X_i, \]

is taken into the sample. The chance that a selection point falls on a particular unit is proportional to

\[ X_i - X_{i-1} = x_i \]

i.e. to the unit's measure of size, as required.
B. Small adjustments to size measures for convenience

Interval $I = X/a$ may of course not be an integer. One can work with fractional intervals, but it may be more convenient to adjust the scale of size measures such that $I$ turns out to be an integer. This is achieved by defining the interval as $I = \left( \sum x_i / a \right)$, rounded to the nearest integer, and the modified size measures as

$$s_i = \left( \frac{a I}{\sum x_i} \right) x_i, \text{ so that by definition } \Sigma s_i = a I.$$

Applying integral interval $I$ to cumulative list of size measures $s_i$ gives exactly $a$ selections, as required. The procedure can involve a small adjustment to the unit selection probabilities (but not affecting their relative values) in order to ensure that the sample size is fixed, irrespective of the value of the starting random number $r$.

A convenient particular case of the above is to rescale the size measures such that the required PPS sampling interval equals 1. This is achieved by defining

$$s_i = \left( \frac{a}{\sum x_i} \right) x_i, \text{ giving } \Sigma s_i = a.$$

As another example, consider a population divided into two strata, (1) and (2), with their total size measures, with obvious notation, being $\sum(1) x_i + \sum(2) x_i = \sum x_i$. If the same sampling interval for PPS selection is applied to both the strata, the expected number of selections each will receive (out of a total of $a$ selections) is proportional to its total size measure: $(a \sum(1) x_i / \sum x_i)$ and $(a \sum(2) x_i / \sum x_i)$, respectively. Now suppose that we wish to sample the strata at different rates, the desired number of selections being $(a_{(1)}, a_{(2)})$ respectively, with $a_{(1)} + a_{(2)} = a$. One can select the sample separately for each stratum, but (especially with many strata to deal with) it may be convenient to adjust the scale of size measures such that the same sampling interval can be applied to both the strata after the adjustment. The rescaled size measures are:

For stratum (1),

$$s_{(1)i} = I \left( \frac{a_{(1)}}{\sum(1) x_i} \right) x_i, \text{ with } I = \left( \frac{\sum x_i}{a} \right).$$

These sum to $\Sigma s_{(1)i} = a_{(1)} I$ over stratum (1).

For stratum (2),

$$s_{(2)i} = I \left( \frac{a_{(2)}}{\sum(2) x_i} \right) x_i, \text{ which sum to } \Sigma s_{(2)i} = a_{(2)} I \text{ over stratum (2), as required.}$$

Again, a convenient particular case of the above is to scale the size measures such that the required PPS sampling interval equals 1. This is achieved by defining\(^\text{10}\)

$$s_{(1)i} = \left( \frac{a_{(1)}}{\sum(1) x_i} \right) x_i, \text{ giving } \Sigma s_{(1)i} = a_{(1)}$$

---

\(^{10}\) The idea of adjusting the scaling factor so as to be able to use the same sampling interval in systematic sampling in multi-strata design has also been evoked in, for instance, Tillé (2006; p. 124).
3.3  Probability proportional to size (PPS) sampling

\[ s_{(2)i} = \frac{a_{(2)}}{\sum_{(2)} x_{i}} \] giving \( \sum s_{(2)i} = a_{(2)} \).

C. An illustration

Table 3.1 provides a simple illustration of the size adjustment and the PPS selection procedure. The column headings and explanation are provided at the bottom of the table. Column (10) gives cumulation of the adjusted size measures, and an example of the sample selected. We begin with a random number from 1 to \( I = 100 \); let this number be \( r = 54 \). The first unit selected is the one for which the cumulated size measure equals or exceeds \( r = 54 \) for the first time; the next unit selected is the one for which the cumulated size measure equals or exceeds \( r + I = 154 \); the third unit is the one in which the cumulated size measure equals or exceeds \( r + 2I = 254 \), and so on.

<table>
<thead>
<tr>
<th>Stratum (1)</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
<th>(9)</th>
<th>(10)</th>
<th>Selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>unit 1</td>
<td>21</td>
<td>21</td>
<td>22.5</td>
<td>27.1</td>
<td>27.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>27.1</td>
<td></td>
</tr>
<tr>
<td>unit 2</td>
<td>32</td>
<td>53</td>
<td>34.3</td>
<td>41.3</td>
<td></td>
<td>68.4</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unit 3</td>
<td>40</td>
<td>93</td>
<td>42.9</td>
<td>51.6</td>
<td>120.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unit 4</td>
<td>26</td>
<td>119</td>
<td>27.9</td>
<td>33.5</td>
<td>153.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unit 5</td>
<td>36</td>
<td>155</td>
<td>38.6</td>
<td>166.2</td>
<td>1.66</td>
<td>2.00</td>
<td>200.0</td>
<td>2.00</td>
<td>200.0</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Stratum (2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unit 6</td>
<td>43</td>
<td>43</td>
<td>46.1</td>
<td>39.4</td>
<td>239.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unit 7</td>
<td>49</td>
<td>92</td>
<td>52.5</td>
<td>45.0</td>
<td>284.4</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unit 8</td>
<td>34</td>
<td>126</td>
<td>36.5</td>
<td>31.2</td>
<td>315.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unit 9</td>
<td>42</td>
<td>168</td>
<td>45.0</td>
<td>38.5</td>
<td>354.1</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unit 10</td>
<td>50</td>
<td>218</td>
<td>53.6</td>
<td>233.8</td>
<td>2.34</td>
<td>2.00</td>
<td>200.0</td>
<td>2.00</td>
<td>400.0</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>373</td>
<td>400.0</td>
<td>4.00</td>
<td>4.00</td>
<td>400.0</td>
<td>4.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I (with ( a = 4 ))</td>
<td>93.3</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes to columns
(1) given size measures
(2) cumulation of size measures
(3) rescaled size measures, so as to sum to 400, giving interval \( I = 400/4 = 100 \), to obtain \( a = 4 \) selections
(4) rescaled size measures, total by stratum
(5) expected number of selections, by stratum, using interval \( I = 100 \) throughout
(6) desired number of selections
(7) rescaled size measures for this purpose
(8) rescaled size measures, total by stratum
(9) number of selections achieved, using common selection interval \( I = 100 \).
(10) cumulation of (7), and indicator of whether a unit is selected (interval \( I = 100 \); assumed random number between 1-100, \( r = 54 \))
Given the stratum totals of the original size measures shown in column (4), in order to obtain the required number (=2) of selections in each stratum as shown in column (6), the sampling intervals required would be as follows.

Stratum 1 \( I_{(1)} = \frac{166.2}{2} = 83.1 \),

Stratum 2 \( I_{(2)} = \frac{233.8}{2} = 116.9 \).

With re-scaled size measures shown in column (7), the required number of selections (=2) for each stratum is obtained by applying the same interval \( I = 100 \) to both strata.

### 3.4 Weighting of sample data and estimation from the sample

In most situations sample data have to be weighted to produce estimates for the population of interest. When the sample data are to be weighted, it is highly desirable - as a matter of practical convenience - to attach its weight to each individual record as a variable in the micro data file. Most of the required estimates, such as proportions, means, ratios and rates, can then be produced in a very straightforward way without any further reference to the structure of the sample. Variance estimation also becomes simplified: most practical methods of computing sampling errors simply require weighted aggregates at the level of primary sampling units, along with the identification of PSUs and the strata in which they lie.

There are major advantages in following certain basic standards and a systematic approach in computing sample weights. Some important aspects are brought out below.

#### 3.4.1 Sources of information for weighting

In determining the sample weights, full use should be made of the information available, both internal to the sample and from external sources. The primary role is given to information internal to the survey; external information is introduced to the extent judged useful for further improving the representative nature of the sample. There are various information sources which can be used in a systematic manner to develop a step-by-step weighting procedure:

- sample design, that is, the design probabilities of selecting the ultimate unit in the sample;
- sample implementation, that is, response rates and information on non-respondents;
- the sampling frame, which may provide additional information on responding and non-responding units in the sample selected, and sometimes also information on under-coverage;
- other, significantly larger surveys (such as the LFS) with better coverage, higher response rates and more reliable information on certain characteristics of households and/or persons, compared to the child labour survey;
3.4  Weighting of sample data and estimation from the sample

- external sources - current registers, administrative records, population projections, etc. -, providing information on population size and characteristics.

When the same or similar information is available from more than one source, priority should be given to the source internal to the survey; for instance weighting to compensate for differences in selection probabilities and known incidence of non-response should always be applied before corrections on the basis of external data are introduced. In using external information, it is necessary to ensure that: (i) the information is significantly more reliable than that which is available internally to the survey; that (ii) the items of information used are defined and measured in a comparable way in the two sources; and that (iii) the coverage and scope are the same.

In using external information as the standard for adjusting survey data, the above-mentioned considerations imply that usually external information based on similar sample surveys should be given priority over similar information from other sources such as administrative records and registers. This is because data coming from sources of the same type tend to be more compatible, compared with data coming from sources of different types.

### 3.4.2 Step-by-step procedure for weighting

To achieve common standards, as well as for clarity and convenience, it is desirable that a step-by-step procedure be adopted which separates out the different aspects of weighting. As a rule, each step should be applied separately so that its contribution to the final weights can be identified. The following describes the main steps.

#### A. Design weights

The design weights are introduced to compensate for differences in the probabilities of selection into the sample. Each ultimate unit in the sample is weighted in inverse proportion to the probability with which it was selected. If \( p_j \) is the overall sampling probability of unit \( j \), and \( n \) the number of units successfully enumerated in the sample, the design weights \( w_j^{(A)} \) are:

\[
 w_j^{(A)} = \left( \frac{n}{\sum(1/p_j)} \right) \cdot \frac{1}{p_j},
\]

where the sum \( \sum \) is over the \( n \) units in the sample. The factor in parentheses is simply a constant, chosen to scale the average value of the weight per unit enumerated in the survey to be 1.0, since \( \sum w_j^{(A)} = n \). Such scaling can be convenient in practice. Superscript (A) has been used to indicate that the reference is to design weights.

Note also that the weights in (3.6) are not affected by the scaling of \( p_j \) values: the \( p_j \) values need to be only proportional to (within some arbitrary constant), rather than equal to, the unit selection probabilities.

#### A simple illustration of calculating design weights

Let us assume that the survey units are households, with the size measures in column (7) of Table 3.1 being some function of household size and composition.
A simple numerical illustration is provided in Table 3.2, using the data from Table 3.1. Suppose that the units selected in the previous table are the analysis units for which population parameters are to be estimated. For the sample of \( n = 4 \) units selected in Table 3.1, Table 3.2 illustrates the procedure for computing design weights. Columns (1)-(5) of the table are described at the foot of the table.

### Table 3.2. Illustration of calculating design weights

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>household 2</td>
<td>41.3</td>
<td>0.413</td>
<td>2.42</td>
<td>1.031</td>
</tr>
<tr>
<td>household 5</td>
<td>46.5</td>
<td>0.465</td>
<td>2.15</td>
<td>0.917</td>
</tr>
<tr>
<td>household 7</td>
<td>45.0</td>
<td>0.450</td>
<td>2.22</td>
<td>0.947</td>
</tr>
<tr>
<td>household 9</td>
<td>38.5</td>
<td>0.385</td>
<td>2.60</td>
<td>1.105</td>
</tr>
<tr>
<td>Total</td>
<td>171.2</td>
<td>1.7</td>
<td>9.4</td>
<td>4.0</td>
</tr>
</tbody>
</table>

### Parameters

- Number of households in population (\( A \)) = 10
- Total size measure in the population (\( S \)) = 400
- Number of households selected (\( a \)) = 4
- Sampling interval for systematic PPS (\( I = S/a \)) = 100

### Notes to columns

1. unit identification number;
2. the (rescaled) size measure \( s_i \) from column (7) of Table 3.1;
3. the unit selection probability with PPS sampling, \( p_i = s_i/I \), with \( I = 100 \) being the interval for systematic sampling;
4. the sample weight, \( w_i = 1/p_i \), being the inverse of the unit selection probability; and
5. the same weight, rescaled to average 1.0 per unit in the sample, using Equation (3.6).

### Analysis units and sampling units

It is important to note that the weights in Equation (3.6) are required for analysis units in the survey: subscript \( j \) refers to an analysis unit in a sample of \( n \) such units; \( p_j \) is the probability of analysis unit \( j \) appearing in the sample. Analysis units refer to the units for which survey estimates are produced from a sample of those units. The sample selection process is actually applied to what we call sampling units. Therefore, the probabilities of selection are defined as such for these sampling units. Analysis units come into the sample through their association with sampling units.

In many survey designs encountered in practice, the link between the two classes of unit is relatively straightforward: the same set of units appears in both classes, or every analysis unit is associated with only one sampling unit. In this situation, probability \( p_j \) of an analysis unit appearing in the sample is the same as the selection probability of the sampling unit it is linked to. For instance, consider a sample of households in which information is obtained on all working children in the sample households. For analysis of working children living in sample households, the probability of a working child (analysis unit) appearing in the sample is the same as the selection probability of the associated household (sampling unit).

However, the sampling situation, especially in the context of sampling elusive populations of labouring children, is often more complex than direct sampling in the sense described.
3.4 Weighting of sample data and estimation from the sample

above. An analysis unit can be associated with more than one sampling unit, and can therefore appear in the sample through the selection of any of the associated sampling units. For instance, we may obtain a sample of labouring children by using multiple overlapping frames, so that a child can be selected from any of the frames in which he/she appears. (The child is the analysis unit; representations of the child in the frames are the associated sampling units.) The same applies in the case of children working at more than one location during the same period when they are selected through a sample of locations, or in the case of children making repeated visits to a facility when they are selected through a sample of visits as the sampling units. Special estimation procedures are required to determine the probability of selection \( p_j \), and hence the design weight \( w_j \) of an analysis unit, given the selection probabilities or weights of the (possibly multiple) sampling units with which the analysis unit is associated. These procedures involve “sampling with multiplicity”, and are described in Sections 4.5-4.6, and applied to diverse forms of sampling in subsequent chapters.

Weighting for coverage error

In certain circumstances it is useful (and necessary) to incorporate into the design weights a correction for known exclusion or gross under-coverage of some parts of the study population, which may have occurred as a result of defects in the sampling frame or other reasons to do with the procedures of sample selection and implementation. Under-coverage can be a major problem in surveys concerning child labour where the frame available for sample selection may be incomplete, or it may not be possible to enumerate certain parts of the population. It is better to estimate (even if crudely) the degree of incompleteness and to incorporate it as an adjustment in the estimation, than to ignore it altogether in reporting the results.

One way to apply such a correction would be to deflate values of the design probabilities of selection (that is, to increase the computed weights correspondingly) in proportion to the coverage rates in the affected domains. Sometimes it is possible to incorporate a compensation for the excluded domains into other domains covered in the survey which are similar to the ones excluded.

Here are a couple of simple illustrations of adjusting design weights for under-coverage.

Suppose that a population of children working in a particular sector has been divided into two domains: (i) children living at home; and (ii) children with other living arrangements, including street children and children with no identifiable living place at all. List frames are available from which a proportionate sample is drawn from the two domains. The sample is in principle self-weighting by design; on this assumption the design weights for units selected from the two domains would be in the ratio 1:1. However, suppose that there is external evidence to believe that the listing for the domain of children not living at home is only 40 per cent complete, while that for the domain of households is nearly 100 per cent complete. As a result, the sample from the former domain would turn out to be only 40 per cent in size compared to what it should have been according to the chosen design. The weight of every selected unit in this domain should be inflated by the factor \( 1/0.40 = 2.5 \) to compensate for this shortfall in the realised sample. Such a compensation of course does not reduce the impact of selectivity in missing a part of the frame. The chance of a child missing from the listing depends on the situation and characteristics of the child. But the
compensation does correct for the effect of under-coverage in distorting the relative sampling rates of the two strata.

As another example, suppose that a population of establishments employing children is divided into four strata according to some measure of establishment size: large, medium, small, and very small establishments; and that these parts account for, respectively 20%, 50%, 23% and 7% of working children in the target population. Now suppose that establishments in the ‘very small’ stratum are numerous, scattered and are difficult to find and contact, and consequently are too difficult and costly to enumerate. However, rather than dropping them from the survey altogether for the above reasons, a better option is to compensate for them by appropriately increasing the sample size of the ‘small’ establishment stratum, in specific terms by the factor \( e = (23 + 7)/23 \). This adjustment assumes that ‘very small’ establishments tend to be similar to ‘small’ establishments, and differ from ‘medium’ and ‘large’ establishments. It also assumes that the relative size (here 7 per cent) of the excluded stratum is not too large. With this adjustment in the example, we increase the selection probabilities of all units in the ‘small’ stratum by the factor (30/23), but we do not adjust (i.e. reduce by the factor 23/30) the weights of units in this stratum: the weights are kept unchanged at their original (higher) value in order to provide a compensation for the excluded ‘very small’ stratum.

An alternative would be not to increase the sample size of the ‘small’ domain, but to inflate the design weight of all selected units in that domain by the factor (30/23). Generally, this second option is less efficient, but is also less costly.

B. Non-response weights

These are introduced to reduce the effect of differences in response rates achieved in different parts of the sample. These weights can only be estimated in relation to characteristics that are known both for the responding and the non-responding units. Weighting for non-response is particularly important when rates of non-response are high and vary from one part of the sample to another.

Weighting classes or cells

Weighting for non-response involves the division of the sample into appropriate ‘weighting classes’, and within each weighting class, the weighting-up of the responding units in inverse proportion to the response rate in the class. This is done in an attempt to ‘make up’ for the non-responding cases in that class. The weighting classes should be constructed so as to make them as homogeneous as possible, and to maximise the differences in unit characteristics and in response rates across those classes.

It is obvious that weighting classes can be defined on the basis of only those characteristics that are available for both the responding and non-responding units. In a multi-stage design involving units at different levels, the characteristics of units at any of the levels may be used to define the weighting classes.

In a multi-stage sample, we may distinguish between characteristics of two types, differing in the ease and sources of their availability: characteristics pertaining to ‘macro’ units, such as sample areas, locations, institutions, or large establishments; and
3.4 Weighting of sample data and estimation from the sample

Characteristics pertaining to ‘micro’ units, such as households, small establishments, children or other individuals.

Area-level characteristics for instance refer to characteristics relating to areas or other aggregates, such as geographic location (administrative divisions), type of place of residence (urban-rural classification), and various socio-economic characteristics of the areas. Some such information is practically always available from the sample design itself (specifically, the geographic location and other information used for stratification of sampling areas). Additional information may come from the sampling frame or external sources, such as local area statistics from the census or administrative sources.

At a minimum, geographic location and other information used for stratification of the sample areas should be used as the basis for defining appropriate weighting classes for non-response adjustment. For instance, sample areas within each stratum may be arranged according to geographic location or some other criteria judged to be related to the survey variables, then formed into groups defining the weighting classes. If the sample areas have been selected using systematic sampling from ordered lists, then grouping according to that order would generally be the appropriate choice.

As an example of ‘micro’ units, household-level characteristics refer to characteristics relating to individual households, such as its household size and type, characteristics of the household head or reference person, income level, socio-economic status, tenure of accommodation, the number of working persons, and other characteristics which may be related to the level of household living conditions and economic activity. Age and sex are other important characteristics concerning children and other persons. For establishments, normally the most important factors include size and sector of activity of the establishment.

There are several potential sources of such information on micro-level units. In situations where the sample is drawn from lists which include relevant information for the classification of the selected units, the information in the frame can be used to provide common classifications for responding and non-responding units. In some situations, where the necessary access to micro-level data is possible, an added source of information can be the linkage of sample units with external sources such as administrative records or previous surveys involving the same individual units. When such external information is used, the same source should be used for the classification of both non-responding and responding units, even if for the latter the same type of information is also available from the survey itself. This is necessary in order to retain consistency in the classification.

Finally, it may also be mentioned that, especially in complex surveys prone to high rates of non-response, it may be worthwhile to make a special effort to collect at least a few basic items of information on each unit selected into the sample, irrespective of whether or not the unit is successfully enumerated in the main survey.

Concerning the appropriate number and size of classes to be used for the purpose, the use of many weighting classes has the possible advantage of reducing non-response bias by creating relatively small and homogeneous weighting categories within which characteristics of respondents and non-respondents can be assumed to be similar. On the other hand, the use of many small weighting classes can result in the application of large and variable weights that can greatly increase the variance of the sample.
estimates. A compromise is therefore required. The choice will depend on how variable the response rates are across different parts of the sample, and how these variations are related to characteristics of the units.

An average of around 100 sample cases per weighting class may be a reasonable choice, as a general rule of thumb.

**Computing response rates**

In principle, the computation of a response rate is straightforward: it is the ratio of the number of units successfully enumerated, to the number of units originally selected. For instance, for a weighting class $k$ the response rate $R_k$ is computed as the ratio of the number of units successfully enumerated (say $m_k$) to the number selected (say $n_k$) in the weighting class:

$$R_k = \frac{m_k}{n_k},$$

with the required non-response weight being

$$w_{ij(k)}^{(w)} = \frac{1}{R_k}.$$  \hspace{1cm} (3.7)

The notation $j(k)$ refers to all units $j$ in weighting class $k$ – all these units receive the same non-response weight, determined by the response rate in their class. Superscript (B) has been used to indicate that the reference is to weights arising from non-response.

It is often convenient to scale the weights to average 1.0 in the sample, which is obtained by defining them as

$$w_{ij(k)}^{(B)} = \frac{\bar{R}}{R_k},$$  \hspace{1cm} (3.8)

where $\bar{R}$ is the average of the $R_k$ values over the sample.

When units in a weighting class have different design weights, $w^{(A)}$, then the real impact of non-response is better captured by using weighted numbers of responding and selected cases in defining the response rate:

$$m_k^{(w)} = \sum_{j=1}^{m_k} w_j^{(A)} \quad \text{over the } m_k \text{ responding units in weighting class } k,$$

$$n_k^{(w)} = \sum_{j=1}^{n_k} w_j^{(A)} \quad \text{over all } n_k \text{ selected units in the class},$$

$$R_k = \frac{m_k^{(w)}}{n_k^{(w)}},$$

and as before, $w_{ij(k)}^{(B)} = \frac{1}{R_k}$ or $\frac{\bar{R}}{R_k}$.

In actual computations, care has to be taken to correctly define the denominator in the above expression. For example, it should include only valid units, e.g. occupied addresses in the sample, excluding addresses found to be empty, non-existent, inaccessible, lost, or otherwise containing no eligible unit.
Complications in computing response rates can also arise for other reasons. One is that, often, sampling units and analysis units are not units of the same type. For instance, we may select a sample of establishments and then select a sample of children working in each of those establishments. Non-response may occur at both stages: in enumerating the selected establishments, and in enumerating children within the successfully enumerated establishments. Let us consider the following quantities in a weighting class $k$:

$n_k$  number of establishments selected
$m_k$  number of establishments successfully enumerated
$n'_k$  number of working children found in the $m_k$ enumerated establishments
$m'_k$  among those working children, the number successfully enumerated
$n''_k$  number of working children expected in the $n_k$ selected establishments

The last-mentioned quantity is not known for establishments which could not be enumerated. But it forms the denominator for the required overall response rate in enumerating working children: $R_k = m'_k / n''_k$. We can estimate the denominator if it is reasonable to assume that the average number of working children per establishment is the same for the responding and non-responding establishments, giving

$$\frac{n''_k}{n_k} = \frac{n'_k}{m_k} \text{ giving } R_k = \frac{m_k}{n_k} \cdot \frac{m'_k}{n'_k}.$$ 

The first factor in $R_k$ is the response rate at the stage of enumerating establishments; the second factor is the response rate at the stage of enumerating children within the enumerated establishments. The overall response rate for children is the product of the two factors.

**C. Weights based on more reliable external information**

After the sample data have been adjusted for differential sampling probabilities and response rates, the distribution of the sample according to the number and characteristics of the units will usually still differ from the same distributions available from more reliable external sources such as the population census, projections, registers or other large-scale surveys. Normally, the precision of the estimates is improved by further weighting the sample data so as to make the sample distributions agree with the external information. The application of this step does not require matching of the sample and the external source at the level of individual households or persons. The weighting adjustments are made on the basis of comparison of sample and external distributions at the aggregate level.

**Example**

Suppose that for a certain characteristic classifying the sample, proportion $c_i$ lies in class $i$ of the classification. If the data have already been assigned design and non-response weights for instance, then this proportion is computed using weighted data (see Section 3.4.3 below).
3. Basic sampling and estimation procedures

Let the same proportion from a more reliable external source be $C_i$. Then the additional weights to be applied to all sample units $f(i)$ in class $(i)$ are simply:

$$w_{f(i)}^{(C)} = \frac{C_i}{c_i}. \quad (3.9)$$

With these weights (applied in addition to any design and non-response weights), the weighted sample proportions will obviously become identical to the external standard proportions $C_i$.

**Calibration to external data sources**

Weighting in order to ensure that weighted sample sums of specified control variables become equal to the known population totals for those variables is called calibration. Calibration can be particularly useful if the sample is subject to gross distortion because of high levels of non-coverage and non-response.

In many situations it is not sufficient to consider distribution by just a single characteristic; it is desirable to control all important characteristics simultaneously. This, however, can result in too many controls, and consequently in small adjustment cells and large variations in the resulting weights. Often the practical approach is to control simultaneously only by marginal distribution of each of the variables, rather than to insist on controlling by detailed cross-classification of the different variables.

As a simple example, suppose that we have two control variables: classification by type of place (3 categories), and by sex and age of the child (6 categories), and that external information on distribution of the population by each of these variables is available which is believed to be more reliable than what can be obtained from our small sample. Using adjustments of type (3.9) we may calibrate the sample in terms of the external distribution by the 3 categories of the first control variables, and also in terms of the distribution by the 6 categories of the second control variable – i.e. for the marginal distribution of each of the two variables. More stringent would be to control the distribution by $3 \times 6 = 18$ cells of the cross-classification of the two variables (i.e. controlling sex-age distribution within each type of place separately). This second option is often less practical or feasible. For instance: (i) external information may be available only for the marginal distributions but not for individual cells of the full cross-tabulation; and (ii) full cross-tabulation may yield too many small control categories.

**Raking procedure**

A robust and convenient method of adjusting the weighted sample distribution to a number of external controls simultaneously is the classical iterative proportional fitting or raking (Deming, 1943). The basic idea is to re-weight the sample to make the sample distribution agree with the external distribution for each control variable in turn, and then to repeat the whole process until sufficiently close agreement is obtained for all the variables concerned simultaneously. In outline the procedure is as follows. For a sample classified by a number of variables, consider the marginal distributions of the sample taking one classifying variable at a time. Suppose that in each case, the marginal distribution is known (or is known more reliably) for the population from some external source. Using the procedure outlined above (Equation (3.9) for example), we can alter the weights of sample cases to make the sample distribution according to one of the
classifying variables agree with the known marginal distribution. After incorporating these weights, we can repeat the same procedure for each of the remaining variables in turn, incorporating the change in weights at each step. The whole procedure can then be repeated until convergence is achieved, that is, further repetition does not materially affect the weights of the sample cases.

Using special algorithms, upper and lower bounds can be imposed on the weight adjustment during calibration. (The limits, however, cannot be made too narrow as the iterative procedure involved becomes slow in converging, and may fail to converge altogether beyond a certain limit.)

Both macro and micro-level control variables are useful in calibration. Useful variables tend to be similar to those used for non-response adjustment, assuming availability: geographical location, tenure status of the household’s accommodation, household type and size, age-sex composition of the population, etc. In some situations, additional variables may be available because the information is required only at the aggregate level.

In our experience the classical raking procedure outlined above, with simple trimming of any weight values outside desirable limits, provides a practical and acceptable approach in many situations. (Concerning trimming, see below).

As an example of further development of the calibration procedure, we may mention that Deville and Särndal (1992) develop a family of calibration estimators of which the standard general regression estimator (GREG) is a first approximation.

**Basic requirement before considering calibration**

While calibration can be useful in principle, a number of practical requirements must be examined before the decision is taken to use external information for weighting.

1. First, it is necessary to establish that such weights are needed, as may be the case if the sample is small, or there are obvious shortcomings in its design and implementation – especially serious departures from probability sampling. These considerations may often apply to intensive labouring children surveys.

2. It is also necessary that such weighting be relevant and effective in improving the representativeness of the survey results.

3. It is necessary that the external information is clearly more reliable than what is available from the survey itself.

4. The crucial requirement in calibration is to ensure that the external control variables are strictly comparable to the corresponding survey variables, the distribution of which is being adjusted. That is, the variables used in the adjustment should be defined and measured in the same way in the sample and in the external source(s).

5. To the extent possible, the external information should cover diverse variables. Consistency is also required between the external sources when more than one source is involved.

Often in statistically less developed countries, good and up-to-date external data are not available. It is necessary therefore to be cautious in applying external weights to
‘correct’ the sample results, and to avoid application of the correction at a level of classification that is too detailed. Indiscriminate and elaborate adjustment of sample results on the basis of external data of insufficient quality has, unfortunately, been the practice in many surveys.

In such situations, the primary concern should be to identify and try to rectify major distortions in the achieved sample – such as in its composition by sex, age, geographic stratum and sector of activity of working children.

D. Trimming of weights

It is desirable to avoid assigning extremely large weights to any units in the sample. The use of large and variable weights, even if it affects only a small part of the sample, can result in a substantial increase in variance (Kish, 1992b). It is a common practice therefore to trim extreme weights to be within some maximum and minimum values to limit the associated increase in variance. The justification for this procedure is that the effect of any bias introduced due to arbitrary trimming of extreme weights is likely to be smaller than the benefit arising from reduced variance.

In fact, control of extreme values and large variations in weights is desirable at each stage of adjusting the weights - after non-response adjustment, and then again after calibration. It is important to ensure that no step in the weighting procedure results in extreme values of the weights.

For a more technical discussion of the issue of trimming extreme values, see Potter (1990). The fact remains that there is no rigorous procedure for general use for determining the limits for trimming. While more sophisticated approaches are possible, it is desirable to have a simple and practical approach. Such an approach may be quite adequate for the purpose, at least in situations where the main problem is caused by a limited number of extreme values assigned during the weighting process.

After calculation of non-response weights, we have recommended and used the following simple procedure.

‘Normalise’ the non-response weight adjustment, meaning rescale these weights to average 1.0 over the sample cases. Then any computed non-response weights outside the limits 1/L to L are recoded to the boundary of these limits, i.e. weight adjustments smaller than 1/L are made to equal 1/L, and similarly any computed values greater than L are made equal to L. As a rule-of-thumb, a value not exceeding the limits L = 2.0-3.0 is suggested as reasonable for this parameter. Note that the limits imposed are in terms of the ratio of the adjusted to the original design weights, i.e. the factor by which the design weights are being modified by non-response and other adjustments.

After calibration, we can follow the same form of check and correction for any extreme values introduced as a result of adjustments of the weights to meet the imposed calibration controls on the structure of the sample. Limits on the range of weight adjustment can also be incorporated within the calibration procedure itself, but of course at the cost of increased technical complexity.
E. Scaling of weights

This refers to the overall inflation factor required to inflate the sample results to the corresponding population aggregates. The factor is meant to isolate the effect of the common, overall sampling rate from the other unit-specific factors discussed above. The latter can be arbitrarily scaled as convenient, such as to average 1.0 per unit.

For many uses, sample weights maybe scaled arbitrarily. The scale chosen for weights does not affect the survey estimates of proportions, means or other ratios. Scaling affects the numerator and the denominator of such statistics in exactly the same way, so that the effect cancels out in the result.

However, proper scaling is essential for certain purposes, such as the following.

1. It is common and convenient to scale the weights to average 1.0 of the sample units. This applies when no pooling of the data over domains or samples is involved, and the purpose of the survey is to estimate proportions, means and other ratios, rather than population aggregates.

2. If the results from different domains have to be put together, the weights for such an aggregation should be scaled such that for each domain the sum of weights for the sample cases is proportional to the population size of the domain. This ensures that the contribution of a domain is in proportion to its population size.

3. If results from two surveys covering the same population are to be put together, then normally the weights are scaled such that the sum of weights for any domain is in proportion to the sample size of that domain. This ensures that the importance given to a sample is directly proportional to its expected precision (since variance is approximately inversely proportional to sample size).

   Note that this is the same thing as making the average weight value in different samples to be the same, such as equal to 1.0.

4. For directly estimating population totals or aggregates from the sample, the weights have to be scaled to equal the inverse of the actual selection probabilities of the units. This is elaborated in the next subsection.

F. Final unit weights

The final unit weight is given by the product of the weighting adjustments at different steps in the procedure: design weights (A), times the non-response adjustment (B), times the adjustment resulting from calibration (C), corrected for trimming (D), and scaling (E), as necessary.

3.4.3 Estimation from sample data

A. Estimating proportions, means, ratios with weighted data

Attaching weights to the data for each analysis unit in the survey makes the process of estimating proportions, means, ratios and even more complex statistics very straightforward, without the need to explicitly refer to complexities of the sampling design.
The most common type of estimator encountered in surveys takes the form of a ratio of two sample aggregates, say $y$ and $x$:

$$y = \sum_i y_i = \sum_i \sum_j w_{ij} \cdot y_{ij}; \quad x = \sum_i x_i = \sum_i \sum_j w_{ij} \cdot x_{ij}; \quad r = \frac{y}{x}. \quad (3.10)$$

Both the numerator ($y$) and denominator ($x$) may be substantive variables – as, for example, in the estimation of income per capita from a household survey, where $y$ is the total income and $x$ the total number of persons estimated from the survey. For each household $j$ in PSU $i$, $y_{ij}$ may refer to its income and $x_{ij}$ to its size (= number of persons, in this example). Quantity $w_{ij}$ is the weight associated with the unit. In estimating ratios such as in Equation (3.10), the weights $w_{ij}$ can be scaled arbitrarily.

Ordinary means, percentages and proportions are just special cases of the ratio estimator. In a mean, the denominator is a count variable, that is, $x_{ij}$ is identically equal to 1 for all elements in the sample. In a proportion, in addition $y_{ij}$ is a (0,1) dichotomy.

In stratified samples, the normal practice is to use combined ratio estimates computed from results aggregated across strata ($h$) to achieve greater stability (lower mean square error):

$$r = \frac{\sum_h y_h}{\sum_h x_h} = \frac{\sum_h \sum_i y_{hi}}{\sum_h \sum_i x_{hi}} = \frac{\sum_h \sum_i \sum_j w_{hij} \cdot y_{hij}}{\sum_h \sum_i \sum_j w_{hij} \cdot x_{hij}}, \quad (3.11)$$

in which, despite the subscripts (denoting the element $j$ in PSU $i$ in stratum $h$), both the numerator and the denominator simply involve appropriately weighted aggregates across the strata.

It is important to re-emphasise that once appropriate sample weights have been attached to each sample case in the data, estimating statistics such as the above requires no explicit reference to the structure of the sample, so long as all ultimate units involved in the statistic along with their weights are accounted for. The reference to $h$ or $i$, respectively the stratum and PSU identifier in the above equation, is simply to explain the set of units included, rather than to indicate a dependence of the estimation on any feature of the sample structure other than the individual weights.

In a multi-stage sample, the probability of selection of an ultimate unit is the product of probabilities at the various stages of selection. In estimating proportions, means and other ratios as above, it is only the ultimate sampling probabilities and not the details at various stages that matter. In fact, apart from the weights, no other complexities of the sample appear in this estimation. For this reason, statistics like ratios are called ‘first order statistics’. These are distinguished from variances and other ‘second order statistics’, the estimation procedures for which must take into account the structure of the sample in addition to the sample weights of individual units.

B. Estimating totals

When the actual (not merely relative) selection probabilities of the units are known, then unit weights as inverse of the actual selection probabilities can be used to obtain
an estimate of the total population size from the sample data itself. For instance, if \( p_j \) are the actual selection probabilities of \( n \) units in the sample from a population of \( N \) units, estimates of this population size and of the total and mean of some variable \( y \) are:

\[
\hat{N} = \sum_{j=1}^{n} \left( \frac{j}{p_j} \right) = \sum_{j=1}^{n} W_j ; \hat{Y} = \sum_{j=1}^{n} W_j y_j ; \bar{y} = \left( \frac{\sum_{j=1}^{n} W_j y_j}{\sum_{j=1}^{n} W_j} \right) = \left( \frac{\hat{N} \hat{y}}{\hat{N}} \right).
\tag{3.12}
\]

As distinct from the arbitrarily scaled weights \( w \), we have used \( W \) to indicate weights which are the inverse of actual selection probabilities. In estimating \( \bar{y} \), the scale of the weights is immaterial.

The factor for inflating the sample size \( (n) \) to population size \( (N) \) is:

\[
\hat{N}/n = \frac{1}{n} \sum_{j=1}^{n} \left( \frac{1}{p_j} \right) = \frac{1}{n} \sum_{j=1}^{n} W_j = \bar{W},
\tag{3.13}
\]

that is, the average value of the weights \( W_j \) in the sample.

Estimate (3.12) of the population size can be used to estimate other aggregates from the corresponding sample mean \( \bar{y} \) as \( \hat{Y} = \hat{N} \bar{y} \), or equivalently, from sample total \( y \) as \( \hat{Y} = \bar{W} y \).

There are situations when the values of the sampling probabilities are known only in relative terms, and not in terms of their actual values. In that case, \( N \) cannot be estimated from the sample data.

However, even when it is possible to use Equation (3.12) to estimate \( N \), the results are often unreliable, especially with small samples subject to high rates of under-coverage and non-response. For example, the estimate is directly biased in proportion to the magnitude of the coverage error. Generally aggregates estimated from sample surveys are subject to underestimation. This is particularly true for a population which is difficult to survey. The population of labouring children is an obvious example.

Often the population size \( N \) can be estimated more reliably from external information. The procedure is in fact used for estimating any aggregates - as distinct from proportions, means, rates and ratios – from the sample. The procedure is as follows.

In place of simple inflation of the estimated sample aggregate \( y \) by a constant scaling factor \( \bar{W} \) (as defined above), we use the sample to estimate the required ratio, say \( r \):

\[
r = \frac{\hat{Y}}{\hat{X}} = \frac{\bar{y}}{\bar{x}} \text{ with } \hat{Y} = \hat{N} \bar{y}, \hat{X} = \hat{N} \bar{x},
\]

and use a ratio-type estimator for population total of \( y \)

\[
\hat{Y}_r = r \hat{X}.
\]

Here \( (\hat{Y}, \bar{y}) \) and \( (\hat{X}, \bar{x}) \) are estimated totals and means from the sample using equations like (3.12), with \( y \) as the variable of interest and \( x \) an auxiliary variable for which a more reliable population aggregate value \( X \) is available from some external source. A special case is when \( X \) refers to \( N \), the size of the population from a more reliable external source; in this case \( \bar{x} \equiv 1 \) for each unit in the sample.
3.4.4 Estimating from a sample: numerical illustration

A. Data for numerical illustrations

In this subsection, a small data set is presented which will be used to illustrate some sampling procedures in subsequent chapters. The data set will be used to compare several sample designs for estimating the number of working children in the population and their total earnings. Using a common data set for different sampling procedures facilitates their comparison. The numerical illustration has been implemented in an Excel folder prepared by Mehran (2012). The sample designs illustrated here include the following.

(1) Simple random sampling of households (illustrated in this section).

(2) Separating out and excluding areas empty (or nearly empty) of the target population of working children (Section 6.4).

(3) Adaptive cluster sampling of neighbouring households as working children are found in the sample (Section 9.8).

Let us consider a population of 80 households spread over 20 geographical areas. Some households have no children and some have more than one. There are altogether 215 children, 20 of whom are working – 5 are unpaid and 15 are paid workers, with total earnings of 45 currency units. The list of households and their main characteristics is given in Table 3.3.
### Table 3.3. The illustrative data set

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>31</td>
<td>8</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>61</td>
<td>16</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>32</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>62</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>33</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>63</td>
<td>16</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>34</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>64</td>
<td>16</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>35</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>65</td>
<td>17</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>36</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>66</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>37</td>
<td>10</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>67</td>
<td>17</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>38</td>
<td>10</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>68</td>
<td>17</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>10</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>69</td>
<td>18</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>10</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>70</td>
<td>18</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>41</td>
<td>11</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>71</td>
<td>18</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>42</td>
<td>11</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>72</td>
<td>18</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>43</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>73</td>
<td>19</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>44</td>
<td>11</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>74</td>
<td>19</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>45</td>
<td>12</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>19</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>46</td>
<td>12</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>76</td>
<td>19</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>47</td>
<td>12</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>77</td>
<td>20</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>48</td>
<td>12</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>78</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>49</td>
<td>13</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>79</td>
<td>20</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>50</td>
<td>13</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>80</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>10</td>
<td>51</td>
<td>13</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>22</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>23</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>52</td>
<td>13</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>53</td>
<td>14</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>7</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>54</td>
<td>14</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>25</td>
<td>7</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>26</td>
<td>7</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>55</td>
<td>14</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>26</td>
<td>7</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>27</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>56</td>
<td>14</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>27</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>7</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>57</td>
<td>15</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>28</td>
<td>7</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>29</td>
<td>8</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>58</td>
<td>15</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>29</td>
<td>8</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>8</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>59</td>
<td>15</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>8</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Column headings**

1. Household identification number
2. Area identification number
3. Number of children in household
4. No. of working children in household
5. Total wages of working children

**B. Sampling and target populations**

It is always useful to keep in view the distinction between the *sampling population* (of sampling units), and the *target population* (of analysis units). The relationship between sampling units and analysis units is elaborated in Chapter 4, in particular Sections 4.3 and 4.5.

The sampling population is the population for which there is a sampling frame and samples can be drawn with pre-assigned probabilities of selection. As seen from Table
3.3, in the present example the sampling population consists of 80 households almost evenly spread over 20 geographical areas. There are exactly 4 households per area except for area 3 and 5 where there are three and five households respectively. Sixty-eight households have children, and 12 have no children. There are altogether 215 children. Summary statistics of the sampling population are shown in Table 3.4.

<table>
<thead>
<tr>
<th>Table 3.4. Sampling population: Summary statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geographical areas</td>
</tr>
<tr>
<td>Number of households</td>
</tr>
<tr>
<td>Number of households with children</td>
</tr>
<tr>
<td>Total number of children</td>
</tr>
</tbody>
</table>

The target population is the population for which the parameters of interest are to be estimated. For certain purposes, households and all children within households constitute the target population. In this illustration, however, the main target population consists of the households with working children. There are 18 of them; each of these households has exactly one working child, except two which have two working children each. Therefore, there are altogether 20 working children in the population. Five are unpaid family workers. The other 15 working children are paid, with total earnings of 45 currency units. Summary statistics of the target population of households with working children are shown in Table 3.5.

<table>
<thead>
<tr>
<th>Table 3.5. Target population: Summary statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Households with working children</td>
</tr>
<tr>
<td>Number of working children</td>
</tr>
<tr>
<td>Of these, number receiving wages</td>
</tr>
<tr>
<td>Total earnings of these children</td>
</tr>
</tbody>
</table>

C. Simple random sample of households

Table 3.6 shows the results from a simple random sample of households. The sample size is \( n = 13 \) from a population of \( N = 80 \) households, giving a sampling rate of \( f = 13/80 = 0.1625 \). The weight of each household in the sample is the same, given by the inverse of the probability of selection, \( w = 1/0.1625 = 6.1538 \).

We have households as the sampling units, and analysis units of three different types: households, children and working children. With different designs, the relationship between units of the two types can be complex, but in the present case it is straightforward. Any analysis unit is linked to (comes into the sample through) one and only one sampling unit. The relationship between sampling-to-analysis units is of the form one-to-(none/one/many) in all cases. In this situation, the probability of an analysis unit appearing in the sample is exactly the same as the selection probability of the sampling unit to which it is linked. The same applies to the sample weights, of course.
### Table 3.6. Estimates from a simple random sample of households

<table>
<thead>
<tr>
<th>Sampling unit</th>
<th>Simple random sample (SRS)</th>
<th>Population estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SRS</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Weighting</td>
<td></td>
</tr>
<tr>
<td></td>
<td>of sample data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>and estimation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>from the sample</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>hh</th>
<th>area</th>
<th>Selection probability</th>
<th>Sample weight</th>
<th>Whether has children</th>
<th>No. of children</th>
<th>Whether working children</th>
<th>No. of working children</th>
<th>Total child wages</th>
<th>Whether has children</th>
<th>No. of children</th>
<th>Whether working children</th>
<th>No. of working children</th>
<th>Total child wages</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>2</td>
<td>0.16</td>
<td>6.2</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6.2</td>
<td>18.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>0.16</td>
<td>6.2</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6.2</td>
<td>30.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>0.16</td>
<td>6.2</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6.2</td>
<td>30.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>21</td>
<td>6</td>
<td>0.16</td>
<td>6.2</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>10</td>
<td>6.2</td>
<td>30.8</td>
<td>6.2</td>
<td>18.5</td>
<td>61.5</td>
</tr>
<tr>
<td>32</td>
<td>8</td>
<td>0.16</td>
<td>6.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>34</td>
<td>9</td>
<td>0.16</td>
<td>6.2</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6.2</td>
<td>18.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>38</td>
<td>10</td>
<td>0.16</td>
<td>6.2</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6.2</td>
<td>24.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>43</td>
<td>11</td>
<td>0.16</td>
<td>6.2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6.2</td>
<td>6.2</td>
<td>6.2</td>
<td>6.2</td>
<td>0.0</td>
</tr>
<tr>
<td>48</td>
<td>12</td>
<td>0.16</td>
<td>6.2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>6.2</td>
<td>18.5</td>
<td>6.2</td>
<td>6.2</td>
<td>12.3</td>
</tr>
<tr>
<td>53</td>
<td>14</td>
<td>0.16</td>
<td>6.2</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6.2</td>
<td>30.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>54</td>
<td>14</td>
<td>0.16</td>
<td>6.2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6.2</td>
<td>18.5</td>
<td>6.2</td>
<td>6.2</td>
<td>0.0</td>
</tr>
<tr>
<td>68</td>
<td>17</td>
<td>0.16</td>
<td>6.2</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6.2</td>
<td>18.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>79</td>
<td>20</td>
<td>0.16</td>
<td>6.2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6.2</td>
<td>6.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>80.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>73.8</td>
<td>252.3</td>
<td>24.6</td>
<td>36.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>68</td>
<td>215</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(2)</td>
<td>(3)</td>
<td>(5)</td>
</tr>
</tbody>
</table>

Using the household weights in Equation (3.12), the table shows, for the particular sample drawn, estimates of:

1. the number of households in the population,
2. number of households having children,
3. total number of children in the households,
4. number of households having working children,
5. total number of working children in the households, and
6. the total of the wages received by all children.

These estimates are also compared with the actual values for the population as given in Table 3.3. (Using the fuller data not reproduced in Table 3.3, we can also estimate the number of child workers who are paid – it is 37 compared to the actual population value of 25.)

Using the Excel folder repeatedly, Table 3.7 shows the results for an (arbitrary) set of 10 simple random samples. The table gives an idea of the variability we get between sample estimates with such a small sample. Note that the variability in estimates of the number of working children is higher than the variability in estimates of the total number of (working or not) children; the variability in estimating the total wages is higher still. If we were to consider all possible simple random samples of size \( n = 13 \), the average values of the estimates would be equal to the true population values because the sampling procedure is unbiased. The variability between samples will of course decline proportionately with increasing sample size.
3.5 Practical variance estimation procedures

There are two classes of variance estimation procedure of practical interest in the context of social surveys: estimating variance from ‘comparison among primary selections’, and from ‘comparison among sample replications’. The first method is simpler and computationally faster; it is normally used when applicable. There can, however be more complex situations – more complex sample designs, more complex statistics – which may require the second type of method, comparison among sample replications.

These procedures, as well as the analysis of design effects outlined later in this section, have been discussed widely in the literature. A simplified but practical and detailed exposition is available in Verma (1993).

3.5.1 Estimating variance from comparison among primary selections

This is perhaps the simplest approach for computing sampling errors of common statistics such as proportions, means, rates and other ratios, and the method can be easily extended to more complex functions of ratios such as differences of ratios, double ratios, indices etc. The method applies to multi-stage stratified designs with variable unit selection probabilities. This covers most practical applications in surveys of child labour. The limitation of the method is the difficulty in extending it to cover more complex statistics and/or more complex sampling designs.

This method is based on the comparison among estimates for independent primary selections within each stratum of a multistage design. The variance computation algorithm is derived using Taylor linearization; therefore the procedure is normally referred to as the linearization method.
The basic equations are as follows.

Consider a population total \( Y \) obtained by summing up individual values \( Y_{hij} \) for elements \( j \) over PSU \( i \), and then over all PSUs and strata \( h \) in the population:

\[
Y = \sum_h Y_h = \sum_i \sum_{ij} Y_{hij} = \sum_h \sum_i \sum_{ij} Y_{hij} \quad .
\]  

(3.14a)

The above is estimated by summing appropriately weighted values over the units in the sample:

\[
y = \sum_h y_h = \sum_i \sum_{ij} y_{hij} = \sum_h \sum_i \sum_{ij} w_{hij} \cdot y_{hij} \quad .
\]  

(3.14b)

The summation in (3.14a) is over all units in the population, and in (3.14b) over all units in the sample.

For the combined ratio estimator of two aggregates \( y \) and \( x \)

\[
r = \frac{y}{x} = \frac{\sum_h y_h}{\sum_h x_h} = \frac{\sum_i \sum_{ij} y_{hij}}{\sum_i \sum_{ij} x_{hij}} = \frac{\sum_h \sum_i \sum_{ij} w_{hij} \cdot y_{hij}}{\sum_h \sum_i \sum_{ij} w_{hij} \cdot x_{hij}} \quad ,
\]  

(3.15)

the general expression for variance is

\[
\text{var}(r) = \sum_h \left[ \left( 1 - f_h \right) \frac{a_h}{a_h - f_h} \sum_i \left( z_{hi} - \frac{z_h}{a_h} \right)^2 \right] .
\]  

(3.16)

where \( a_h \) is the number of primary selections in stratum \( h \), \( f_h \) the sampling rate in it, and the computational variable \( z \) is defined as

\[
z_{hi} = \frac{1}{x} \left( y_{hij} - r \cdot x_{hij} \right); \quad z_h = \sum_i z_{hi} .
\]  

(3.17)

The approach is based on the following assumptions about the sample design.

1. The sample selection is independent between strata.
2. Two or more primary selections are drawn from each stratum (\( a_h > 1 \)).
3. These primary selections are drawn at random, independently and with replacement.
4. The number of primary selections is large enough for the valid use of the ratio estimator and the approximation involved in the expression for its variance.
5. The quantities \( x_{hij} \) in the denominator (which often correspond to sample sizes per PSU) are reasonably uniform in size within strata.

The above variance estimation formulae are very simple despite the complexity of the design, being based only on weighted aggregations for the primary selections, and identification of the strata. A most important point is that the complexity of sampling within PSUs does not appear explicitly to complicate the variance estimation procedure. No separate computation of variance components is required.

Expressions very similar to (3.17) can be easily written for more complex statistics than ordinary ratios, for instance for differences of two ratios, weighted sums or products of ratios, or ratios of two ratios (‘double ratios’), weighted sums of double ratios, ratios
of double ratios or similar index numbers, and even for regression coefficients. (For a convenient listing of these expressions, see for instance, Verma, 1991.) Over the years, the range of application of the procedure has been greatly extended to complex statistics (see, for instance, Binder, 1983; Binder and Patak, 1994; Deville, 1999; Verma and Betti, 2011).

### 3.5.2 Estimating variance from comparison among replications of the full sample based on repeated re-sampling

The procedures of estimating variance from comparison among replications of the full sample based on repeated re-sampling are more complex and computer intensive, but their great advantage is that they can be applied to statistics of any complexity. They can also be extended more readily to complex sample designs (such as panels or overlapping samples), compared to the linearization approach described above.

The basic idea is that of ‘repeated re-sampling’. This approach refers to the class of procedures for computing sampling errors for complex designs and statistics in which the replications to be compared are generated through repeated re-sampling of the same parent sample. (See, for instance Rao, Wu and Yue, 1992.) Each replication is designed to reflect the full complexity of the parent sample.

The replication-based procedures are also more versatile than the linearization approach: by repeating the entire estimation procedure independently for each replication, the effect of various complexities, such as each step of a complex weighting procedure, can be incorporated into the variance estimates produced.

The important technical requirement is the construction of the appropriate set of replications in terms of which the required variance can be estimated. Once the replications have been created, no further explicit reference is required even to the structure of the sample. Descriptions of the procedure are widely available in the literature, and need not be repeated here. (For a simple and clear exposition, see: Rust, 1985; Verma 1993.)

The various re-sampling procedures available differ in the manner in which replications are generated from the parent sample and the corresponding variance estimation formulae evoked. There are three general procedures, known as the balanced repeated replication (BRR), the jackknife repeated replication (JRR), and the ‘bootstrap’. The JRR and bootstrap are quite versatile and convenient methods, particularly the former. The JRR procedure is outlined below.

### 3.5.3 Jackknife Repeated Replication (JRR)

Jackknife Repeated Replication (JRR) is a versatile and quite straightforward application of the above technique.

The basic model of the JRR may be summarised as follows.

Consider a design in which two or more primary selection units (PSUs) have been selected independently from each stratum of the population. Within each PSU, subsampling of any complexity may be involved, including weighting of the ultimate
3.5 Practical variance estimation procedures

In the “standard” version, each JRR replication is formed by eliminating one PSU from a particular stratum at a time, and increasing the weight of the remaining PSUs in that stratum appropriately, so as to obtain an alternative but equally valid estimate to that obtained from the full sample.

Let \( z \) be a full-sample estimate of any complexity, and \( z_{(hi)} \) be the estimate produced using the same procedure after eliminating primary unit \( i \) in stratum \( h \) and increasing the weight of the remaining \( (a_h - 1) \) units in the stratum by the factor \( g_h = w_h / (w_h - w_{hi}) \). With \( w_{hi} \) as the weights of sample units, \( w_h \) is the sum over sample units in PSU \( i \), and \( w_h \) is the sum over the PSUs in stratum \( h \).

Let \( z_{(h)} \) be the simple average of the \( z_{(hi)} \) over the \( a_h \) sample units in \( h \). The variance of \( z \) is estimated as:

\[
\text{var}(z) = \sum_{h} \left( 1 - f_h \right) \left( \frac{a_h - 1}{a_h} \right) \sum_{i} \left( z_{(hi)} - z_{(h)} \right)^2.
\]

The same relatively simple variance estimation formula holds for \( z \) of any complexity. Furthermore, apart from variance estimation of ordinary cross-sectional measures, application of the JRR methodology can be readily extended to indicators based on more complex designs, rotational or longitudinal samples for instance.

Defining sample structure: ‘computational’ strata and PSUs

In many practical situations some aspects of sample structure need to be redefined to make variance computation possible, efficient and stable. Of course, any such redefinition is appropriate only if it does not introduce significant bias in the variance estimation. Such redefinition is often necessary because practical variance estimation methods require the sample design to satisfy certain conditions noted earlier (conditions (1)-(5) in Section 3.5.1). Though these basic assumptions regarding the structure of the sample for application of the variance estimation methods are met reasonably well in many surveys, often the assumptions are not met exactly.

A very convenient approach in practice is to summarise the most essential information about the sampling design in the form of two variables, coded for each unit in the micro-data file: the ‘computational stratum’ and the ‘computational PSU’ to which the unit belongs.

The computation stratum has to incorporate all information about the stratification of the PSUs, including both explicit stratification and, where applicable, implicit stratification resulting from systematic sampling of the PSUs. It has also to ensure that each computational stratum contains at least two computational PSUs (which are then assumed to have been selected at random with replacement).

Starting from the actual PSUs, the variable computational PSU should seek to create units reasonably large and uniform in size, and small enough in number so as to avoid excessive computational burden. Redefinition for the above purpose usually involves some ‘collapsing’ of the sample structure. Some technical procedures for this purpose include: reducing the number of replications formed by deleting units in groups rather than singly as assumed in the basic model; dropping some of the replications from the computation; random grouping of PSUs within strata so as to reduce the number of
3. Basic sampling and estimation procedures

3.6 Analysis of design effects

3.6.1 The concept

A most useful concept for the computation, analysis and interpretation of sampling errors concerns the design effect (Kish, 1992a). Design effect is the ratio of the variance ($\nu$) under the given sample design, to the variance ($\nu_0$) under a simple random sample of the same size:

$$d^2 = \frac{\nu}{\nu_0}, \quad d = \frac{se}{se_0}. \quad (3.18)$$

Computing design effects requires the additional step of estimating sampling (or standard) error under simple random sampling ($se_0$), apart from its estimate under the actual design ($se$).

Proceeding from standard errors to design effects is essential for understanding the patterns of variation and determinants of the magnitude of the error, for smoothing and extrapolating the results for diverse statistics and population subclasses, and for evaluating the performance of the sampling design. Analysing design effects into components helps to better understand from where inefficiencies of the sample arise, to identify patterns of variation. Decomposition of variances and design effects identifies more ‘portable’ components, which may be more easily imputed (carried over) from a situation where they can be computed with the given information, to another situation where such direct computations are not possible. On this basis, valid estimates of variances can be produced for a range of statistics wider than that for which variance computations have been actually performed.

All the above reasons apply even more strongly for statistics based on small samples, such as at the regional or local level. Smaller sample sizes and less available information make the computation of sampling errors more difficult, sometimes impossible. The results of individual computations also tend to be less stable, thereby increasing the need for averaging over them. Disaggregation of the results to lower levels or subpopulations also increases the amount of computations involved. The practical solution is to seek procedures to extrapolate the results from a limited set of computations to other statistics and situations. These procedures require ‘portable’ measures such as individual components of the design effect.

3.6.2 Components of design effect

We may decompose the design effect into components as follows:

$$v = v_0 d^2 = v_0 (d_w d_u d_x)^2. \quad (3.19)$$
Here \(\nu_0\) is the variance (for the statistic concerned) in a simple random sample (SRS) of the analysis units (e.g. working children); \(d_w\) is the effect of sample weights; if applicable; \(d_u\) is the effect of clustering of analysis units into the ultimate sampling units (e.g. clustering of working children in establishments where they work); and \(d_x\) is the effect of other complexities of the design, such as clustering, stratification and other features of the sample structure. Factors other than \(d_x\) do not involve those features of the sample structure, but essentially depend only on the more numerous ultimate sampling and analysis units, and the sample weights associated with those units. Hence normally factors other than \(d_x\) are well estimated even with quite small samples.

Normally the effect of complexities of the design such as multiple stages and stratification is to make \(d_x > 1\), because the loss in efficiency of the sample due to clustering tends to be larger than any gain from stratification. We can expect \(d_x < 1\) in stratified random samples of analysis units.

### 3.6.3 Estimating the effect of weighting \((d_w)\)

As to the effect of weights \(d_w\), weighting generally inflates variance (weighting is primarily introduced to reduce bias). Large variation in weights can substantially inflate the design effect. In principle, the factor can be \(d_w < 1\), for example with suitable calibration weights.

A very simple expression for estimating \(d_w\) is the following from Kish (1965):

\[
d^2_w = \left[ \frac{n \sum (w_i^2)}{\left( \sum w_i \right)^2} \right] = 1 + cv^2 (w_i).
\]

This provides a very good approximation when the sample weights are ‘external’, not correlated with survey variables. Generally it over-estimates the effect of weighting. The effect can be estimated more precisely as:

\[
d^2_w = \left[ \frac{n}{\sum w_i} \right] \frac{\sum (w_i^2 z_i^2)}{\sum (w_i z_i^2)}.
\]

For a ratio, \(r = \sum w_i y_i / \sum w_i x_i\), we have \(z_i = (y_i - r x_i)\). In fact Equation (3.21) is found to apply also to statistics more complex than ratios (Verma and Betti, 2011).
3.6.4 Estimating the effect of clustering within ultimate sampling units \((d_u)\)

The effect of clustering of analysis units into the ultimate sampling units, \(d_u\), depends on the average number of analysis units (say working children) taken into the sample per ultimate sampling unit (for example, per establishment).

It is useful to distinguish three types of situation.

(1) The ultimate sampling units are identical to the analysis units. For instance, we may list all working children in establishments in an area, and then directly take a sample of working children (without the intermediate step of selecting establishments). Obviously, in such a situation, there is no clustering of analysis units within ultimate sampling units and, by definition, \(d_u = 1\).

Exactly the same would apply to the situation when we select exactly one analysis unit (child) from each ultimate unit (say an establishment or household) selected into the sample.

(2) The survey variable is measured at the level of the ultimate sampling unit (rather than at the level of the analysis unit), and then ascribed to each selected analysis unit within the given sampling unit.\(^{11}\) For instance, the survey may involve measurement of some characteristic of establishments (e.g. sector of activity, size or type of unit) or of households (size, composition, income level, characteristics of head), and then the value measured may be ascribed as an analysis variable to each child worker in that unit. Clearly, \((i)\) the number of ultimate units in the sample (establishments or households in the above example) constitute the effective sample size, while \((ii)\) the number of analysis units (working children in the above example) form the apparent sample size, and their ratio gives the effect of clustering of analysis units within ultimate sampling units: \(d_u^2 = (ii)/(i) \geq 1\).

It is important to be precise about how numbers \((i)\) and \((ii)\) are defined. The former, \((i)\), should only include ultimate sampling units containing at least one analysis unit selected into the sample – e.g. establishments employing children from which at least one is included in the sample, disregarding establishments not employing children or those from which no working child has been taken into the sample. The latter, \((ii)\), should only include analysis units selected into the sample – e.g. working children taken into the sample from the establishments selected. The ratio \((ii)/(i)\) is the average number of selected analysis units (e.g. working children), per ultimate unit (e.g. an establishment or household) containing a selected analysis unit (one or more working children in the sample). This obviously gives \(d_u \geq 1\).

(3) Ultimate sampling units are generally (and usually small) clusters of the analysis units, but the survey variable is measured at the level of the analysis units directly.

\(^{11}\) We use the term ‘measured’ in the general sense. It may refer to direct measurement for each (say) establishment or household in the survey, or to a variable constructed from (e.g by averaging or aggregating) values measured for analysis units it contains (e.g. income of household members aggregated to obtain the household’s total income), or by a combination of these two forms. By ‘ascribing’ this value to each analysis unit it contains, we mean that each such analysis unit receives the same value (for example: the income of its household, or the size, sector or type of ‘its’ establishment).
3.6 Analysis of design effects

This is a very common case – e.g. selecting a sample of establishments, followed by selecting one or more children within each sample establishment and measuring the survey variables directly for each child in the sample. The value of \( d_U \) would depend mainly on two factors: the intracluster correlation of analysis units within ultimate sampling units (e.g. between children in the same establishment), and the average number of analysis units per ultimate unit in the sample (e.g. the average number of children selected per establishment containing working children). Normally we still find \( d_U \geq 1 \), but smaller than that in case (2) above.

For completeness, we should note that it is possible, in principle, to have \( d_U < 1 \) for variables which are negatively correlated - for instance, certain demographic characteristics among members of the same household - but this situation is rare.

Estimating \( d_U \) in the more complex case (3) can be done using the ‘random grouping of elements’ method as described below.

### 3.6.5 Estimating the effect of clustering, stratification and other design complexities (\( d_X \))

The estimation of design effect arising from design complexity with a replication method such as JRR requires an indirect approach, involving what may be termed as ‘random grouping of elements’.

Consider variance computed under the following two assumptions about structure of the design:

i. Variance (\( \nu \)) under the actual design.

ii. Using the same procedure, variance (\( \nu_R \)) computed by assuming the design to be (weighted) simple random sampling of elements. This can be estimated from a ‘randomised sample’ created from the actual sample simply by completely disregarding its structure other than the weights attached to individual elements.

For computation (ii), the JRR replications are constructed as in the normal application of the JRR, but in place of the actual strata and primary selections, random grouping of the sample elements are used for this purpose. This provides a variance estimate corresponding to a sample of elements, i.e. without the effect of stratification, clustering or other complexities, but which still differs from the SRS estimate due to the effect of sample weights on variance.

Actually, the result \( \nu_R \) depends on the type of elements used for constructing the random groupings. Consider for instance a sample with households as the ultimate sampling units, within which individual members form the analysis units. When we have random groupings of individual persons (that is, without regard to whether they come from the same or from different households), the variance estimate obtained is:

\[
\nu_R = \nu_0 d_W^2 ,
\]

where, as defined earlier, \( \nu_0 \) is the variance under the assumption of simple random sampling of elements (individual persons in our example).
In a sample of households, we may use random groupings of households instead (that is, keeping all members of a household together in the same group), the variance estimate obtained is:

\[ v' = (v_0 d_w^2) d_u^2. \]  

(3.23)

Now we have three computations of variance, obtained for instance by the application of standard JRR procedure, but assuming different sample structures:

- \( v \) for the actual design, with all its features including clustering, stratification and weighting, i.e. incorporating the full design effect (\( d \));
- \( v' \) assuming an SRS of ultimate sampling units, i.e. retaining the effect only of clustering of analysis units within ultimate sampling units, (\( d_u \)), and the effect of weighting, (\( d_w \));
- \( v_R \) assuming an SRS of analysis units, i.e. retaining the effect only of weighting, (\( d_w \)).

With \( v, v' \) and \( v_R \) computed by using the standard JRR procedure but on the assumption of different sample structures as described above,

- application of (3.22) with \( d_w \) estimated from (3.21) gives \( v_0 \);
- then (3.18) gives the overall design effect \( d^2 = v / v_0 \), without the need to separate out other components;
- and (3.23) can be used to obtain \( d_u \) separately;
- finally, (3.19) gives \( d_x \), the effect of stratification, clustering and other design complexities.

### 3.7 Brief description of errors in survey data

#### 3.7.1 A typology of errors

Knowledge about data quality is required for their proper use and interpretation. Also, measures of data quality are important for the evaluation and improvement of survey design and procedures.

There are diverse forms and many different sources of errors in surveys, and various frameworks have been proposed for their classification. Different frameworks emphasise different aspects of the problem. None may be considered as ‘the best’, though some frameworks are more illuminating than others. The following framework is drawn from Verma (1981), further elaborated in Hussmanns et al. (1990). This framework distinguishes between the following two groups of errors affecting the survey process.

**A. Errors in measurement**

These arise from the fact that what is measured on the units included in the survey can depart from the actual (true) values for those units. These errors concern the accuracy of measurement at the level of individual units enumerated in the survey, and centre on substantive content of the survey: definition of the survey objectives and questions;
ability and willingness of the respondent to provide the information sought; and the quality of data collection, recording and processing. This group of errors can be studied in relation to various stages of the survey operation.

B. Errors in estimation

These are errors in the process of extrapolation from the particular units enumerated in the survey to the entire study population for which estimates or inferences are required. These centre on the process of sample design and implementation, and include errors of coverage, sample selection and implementation, non-response, and also sampling errors and estimation bias.

The above categorisation, in terms of errors in measurement and errors in estimation, is more fundamental than the distinction usually made between sampling and non-sampling errors.

In Figure 3.1 a third category, namely item non-response, has been added as an intermediate category between measurement and estimation errors.

Each group of errors may be further classified in more detail in order to identify specific sources of error, so as to facilitate their assessment and control. However, it is important to note that the various phases of a survey are closely related. While it is useful to classify the total survey error into components, errors cannot always be attributed to a particular type or source. The same or similar methods of assessment and control may be suited for measuring more than one type of error, and some of the indicators obtained may provide no more than a general or overall measure of data accuracy without being able to identify specific sources and types of error.

3.7.2 Errors in measurement

As noted, the broad range of ‘errors in measurement’ may be classified by source, for example as conceptual, response (‘data collection’) and processing errors. Conceptual errors concern the scope, concepts, definitions and classifications adopted in relation to the survey objectives, and are the most fundamental ones. The distinction between response errors concerning the process of data collection, and processing errors concerning the subsequent process of transforming the information into a micro database, is a useful one from the point of survey operations and methods of assessing and controlling these errors. Despite this operational distinction, however, the two classes of error are conceptually quite similar.

Various components of measurement error may be distinguished. Further operational classification within each category may be introduced. Each type of error may be decomposed into bias and variance components. These distinctions are useful in so far as the components differ in nature and in methods of assessment and control.

A. Measurement Bias

A part of the error is common to the work of all interviewers (or coders, etc.); this gives rise to response bias, i.e. more or less systematic errors in obtaining the required information. Bias arises from shortcomings affecting the whole survey operation: basic conceptual errors in defining and implementing the survey content; incorrect
instructions affecting all the survey workers; errors in the coding frame or programs for processing the data, etc. Errors also arise from inherent difficulties in collecting certain types of information, more or less independently of the specific technical design and procedures of the survey, given the general social situation and the type of respondent involved.

The first step in identifying bias is through logical and substantive analysis of the internal consistency of the data. Beyond that, the assessment requires comparison with more accurate information: data from external sources and data collected with special, improved methods. There are several possibilities in connection with bias assessment. For instance, the study of response bias may involve two interviews on a subsample following the original interview: a re-interview, which is an independent replication of the original interview and is aimed at measuring response variance, followed in discrepant cases by a reconciliation interview aimed at establishing correct responses and identifying biases and their sources.

B. Measurement Variance

This refers to variable errors in data collection and processing. In addition to biases common to the whole operation, each interviewer has his/her own particular bias, which affects the interviewer’s whole workload. This gives rise to correlated response variance, which indicates a lack of uniformity and standardisation in the interviewers’ work. By contrast, simple response variance is random, not correlated with any particular interviewer. It is an indicator of the inherent instability of particular items in the questionnaire. Its high value indicates the need for better training and supervision of survey work.

Assessment of measurement errors requires comparisons between independent repetitions of the survey under the same general conditions - there is no way, in a single survey, to distinguish between variation among the true values of units (which gives rise to sampling error), and the additional variability arising from random factors affecting individual responses.
### Errors in measurement

1. **Conceptual errors**
   - errors in basic concepts, definitions and classifications
   - errors in putting the above into practice (questionnaire design, preparation of survey manuals, training and supervision of interviewers and other survey workers)

2. **Response (or ‘data collection’) errors**
   - response bias
   - simple response variance
   - correlated response variance

3. **Processing errors**
   - recording, data entry and coding errors
   - editing errors
   - errors in constructing target variables
   - other programming errors

### Mixed category

4. **Item non-response**
   - only approximate or partial information sought in the survey
   - respondents unable to provide the information sought (“don’t knows”)
   - respondents not willing to provide the information (“refusals”)
   - information suppressed (for confidentiality or whatever reason)

### Errors in estimation

5. **Coverage and related errors**
   - under-coverage
   - over-coverage
   - sample selection errors

6. **Unit non-response**
   - unit not found or inaccessible
   - not-at-home
   - unable to respond
   - refusal (potentially ‘convertible’)
   - ‘hard core’ refusal

7. **Sampling error**
   - sampling variance
   - estimation bias

### Non-sampling errors = 1 to 6

*Adapted from Hussmanns et al. (1990).*
3.7.3 Errors in estimation

A. Coverage and related errors

Coverage errors arise from discrepancies between the target and the frame populations, and also from errors in selecting the sample from the frame. The condition of ‘probability sampling’ is violated if: (a) the survey population is not fully and correctly represented in the sampling frame; (b) the selection of units from the frame into the sample is not random with known non-zero probabilities for all units; or (c) not all the units selected into the sample are successfully enumerated. Coverage error concerns primarily (a), but also (b); (c) concerns non-response.

B. Non-response errors

Non-response refers to the failure to obtain a measurement on one or more study variables for one or more sample units. When a whole unit is missed, we have unit non-response. When a unit is included but information on some items for it is missed, we have item non-response (see Section 3.7.4). Non-response causes an increase in variance due to decreased effective sample size and/or due to weighting and imputation introduced to control its impact. More importantly, it causes bias in so far as non-respondents are selective with respect to the characteristic being measured. For instance, one might expect persons with high incomes to be more reluctant to give information on their income; similarly, poorer, unemployed and socially excluded persons are more likely to be missed in surveys. Classification of unit non-response according to the reasons or circumstances giving rise to it can be very helpful for identifying and controlling the extent of non-response and assessing its impact. It is most effective when the non-response categories are designed to capture the most important factors in the particular survey, are not too numerous, and are clear and non-overlapping. Examples are units not found or not accessible, not-at-home, unable or refusing to respond. In a repeat survey, it can be very useful to distinguish between ‘potentially convertible’ refusals, and ‘hard core’ refusals which have to be dropped from future rounds.

For composite units (e.g. a multi-adult household), any of the above reasons may result in ‘partial unit non-response’.

C. Sampling error

Sampling error is a measure of the variability between estimates from different samples, disregarding any variable errors and biases resulting from the process of measurement and sample implementation. Of course, sampling error represents only one component of the total survey error. For estimates based on small samples, this component may be the dominant one. In other situations, non-sampling errors, in particular sample selection, non-response and measurement biases, may be much more important.

---

12 “The non-response rates can be measured well if accurate accounts are kept of all eligible elements that fall into the sample. These are necessary for understanding the sources of non-response, for its control and reduction, for predicting it in future surveys, and for estimating its possible effect on the surveys. … These aims can be better served if the many possible sources of non-response are sorted into a few meaningful classes. A good classification of non-response depends on the survey situation ….” Kish (1965), Section 13.4A.
However, even in these cases, sampling error increases progressively as the estimates are produced for smaller and smaller subgroups of the population, such as for social classes or regions of a country. In a small enough subgroup, sampling error may well outweigh non-sampling errors.

### 3.7.4 Item non-response

Item non-response can be seen as an intermediate category between errors in measurement and errors in estimation. Like any other error in measurement, item non-response is subject-matter specific. At the same time, it can be viewed simply as an addition to the existing unit non-response in analysis involving the particular item affected, thereby amounting to an error in estimation. Item non-response is particularly important in surveys collecting complex and detailed information.

Information on an item may be incomplete simply because it is not feasible to seek it exactly or in full detail in an interview survey; these errors are akin to ‘conceptual errors’. The impact on the results may differ depending on the respondent’s characteristics and circumstances. Often information is missing because the respondent is unable to provide it, or the respondent may be unwilling to provide information which is considered too sensitive or personal. There can be an added, special reason for item non-response in surveys providing micro-data to researchers and other users: this is deliberate suppression of some information by the agency providing the micro-data, presumably based on confidentiality and similar considerations.

As in the case of unit non-response, classification of item non-response according to the reasons or circumstances giving rise to it can be very helpful in identifying and controlling the extent of non-response and assessing its impact.
II. Imperfect frames
Chapter 4
Sampling from imperfect frames

This chapter addresses two related themes: the content and quality of the sampling frame (Sections 4.1-4.4); and the different sampling situations defined in terms of the relationship between sampling units and analysis units in a survey (Sections 4.5-4.7).

We begin with clarifying the basic concepts of the survey population and the sampling frame for single-stage and for multi-stage sampling (Section 4.1). Next, we identify basic requirements of and common problems with area frames, and summarise desirable properties of area frames concerning quality, efficiency and cost (Section 4.2).

As to list frames, in practice lists are subject to imperfections, such as the presence of blanks, clustering of elements, duplications, under-coverage, units not found in the field, and changes in the identity, boundaries or characteristics of the units. In the worst case, we may have no frame at all. The chapter discusses these problems from a practical perspective. When these problems are limited and not systemic, we may still obtain an acceptable probability sample using a ‘conventional’ design, despite imperfections in the frame. The problems are illustrated and discussed in Section 4.3.

A few examples are given in Section 4.4 to illustrate and discuss some of the concepts and common aspects concerning sampling frames. The examples are mostly from establishment-based baseline surveys covering selected sectors of child labour. Technical issues concerning establishment surveys will be elaborated in Chapter 5.

The second theme of the chapter is to develop and explain the following fundamental concepts. In Section 3.4.2 we referred to the distinction between analysis units and sampling units: analysis units refer to the units for which survey estimates are produced from a sample of those units; the sample selection process is actually applied to sampling units. Notwithstanding the frame imperfections noted above, in most applications of conventional sampling we have one-to-one or one-to-many correspondence between sampling and analysis units: any analysis unit is associated with at most one sampling unit. We may term this as direct sampling. By implication, its complement is ‘indirect sampling’.

There are two forms of indirect sampling. First, a widespread situation encountered in surveying elusive populations is the presence of many-to-one and many-to-many links between sampling and analysis units - an analysis unit being associated with more than one sampling unit. Different types of link between sampling and analysis units give rise to sampling with multiplicity, as described in Section 4.5. Sampling with multiplicity is the common link between many of the sampling techniques discussed in subsequent chapters. Examples are multiplicity sampling (Chapter 7), multi-frame sampling (Chapter 8), adaptive cluster sampling (Chapter 9), and time-location sampling (Chapter 10). The multiplicity estimator of the form described in Section 4.6 provides the basis for estimation in these different sampling schemes. We also note in
this section the associated concept ‘weight-share method’ of estimation in the context of ‘indirect sampling’.

The second form of departure from conventional direct sampling involves situations when the sample has to be obtained by exploiting links between analysis units themselves, rather than primarily between analysis units and sampling units. This is link-trace sampling. Types we discuss include snowball sampling (Chapter 13), and respondent-driven sampling (Chapter 14).

4.1 Basic concepts

4.1.1 The target population

The definition of the population to which the sample results are to be generalised is a fundamental aspect of survey planning and design. It is important to specify the target population as precisely as possible. This specification is in terms of the content and extent of the population.

A. The population content

The population content is defined by the type and characteristics of the elementary units comprising it. The elements of a population are the units for which information is collected and analysed. For this reason we refer to them generally as the ‘analysis units’. The goal of the survey is to produce valid estimates for the population composed of these units.

The same survey may in fact involve analysis units of several types – such as households, individual household members, children, and working children - for each of which data are to be collected and analysed. The aggregate of each type of analysis units in a sense defines a distinct survey population. When analysis units of different types are involved in the same survey, they are normally connected to each other and often share the same data; but they are nevertheless distinct in terms both of data collection and data analysis.

Analysis units are included in or excluded from the target population on the basis of certain eligibility criteria, determined by the survey objectives. For example, a child labour survey would include children only within some specified age range. Only households containing children, or only those containing labouring children, may be the target population. A survey of establishments may include only establishments below a certain size, and/or establishments operating in particular sectors, which employ or use child labourers, and so on. Quality of the information for the identification of analysis units eligible for the survey affects the quality of survey implementation. The cost and difficulty of collecting such information with sufficient accuracy is a critical practical consideration.

B. The population extent

This is normally defined in terms of two criteria: (i) the population extent in space, i.e. the boundaries of its geographical coverage; and (ii) the population extent in
time, i.e. the time period to which it refers. Many household-based surveys of child labour, for instance, are national in coverage. By contrast, intensive studies such as rapid assessments of the child labour situation are often confined to local areas or communities. There have been many surveys of child labour in particular sectors which are confined to some geographical or administrative parts of the country. Such limits of the coverage may be determined by substantive objectives of the survey; but often they can also be dictated by limitations in the available time, resources and capacity for the survey.

Clear specification of the time period to which the target population refers is often given less attention than clear specification of its spatial boundaries. In part this practice may be justified, in so far as variations over time are smaller compared to geographical variations in the survey variables. However, specification of the population extent in time deserves attention. The socio-political situation can be subject to sharp changes, and hence also the child labour situation. Perhaps more pervasive is the effect of seasonality on the demand and supply of child labour. The implication is that the survey results cannot be automatically extended to times and seasons other than the ones covered in the survey.

C. The survey population

Both in terms of the content and the extent of the population, some units may be excluded from the survey, not for being outside the intended coverage but because they are too costly or difficult to cover. Examples are small population groups living in inaccessible areas, areas with a very low incidence of child labour, or areas where safety of the survey workers cannot be guaranteed, or where there are units missing from the frame used for selecting the sample. The term survey population, as distinct from the target population, is sometimes used to describe that part of the target population remaining after making such exclusions for practical reasons.

In short, the target population is what the survey intended to cover; the survey population corresponds to the population actually covered.

Some general points in relation to the choice of the survey population need emphasis.

- In any survey, rules of population inclusion and exclusion must be defined in clear operational terms. Otherwise confusion and errors result at the implementation stage.
- The limitations in the population covered must be kept in view in drawing inferences from the survey results, and in comparing results from different sources.
- Apart from deliberate and explicit exclusions, surveys also suffer from coverage errors which are less easily identified and measured. Painstaking work is required to control these errors and to assess their magnitude and impact on the survey results. Their magnitude depends on the quality of the sampling frame and sample implementation.
4. Sampling from imperfect frames

4.1.2 The sampling frame

A. Basic requirements

Probability sampling requires that the selection of units into the sample be made by a randomisation process that assigns to the units the desired probabilities, probabilities which are non-zero for all units in the population, and are known at least for all units finally selected into the sample. The randomisation process requires the application of an objective procedure for the actual selection of units into the sample. The sampling frame is the instrument for achieving this. A sampling frame is a representation of the population to be surveyed in a physical form. The frame comprises lists of units from which survey samples are selected, and normally also provides information for making or improving estimates from sample data.

In practice, the required frame is defined in relation to the required structure of the sample and the procedure for selecting it. The nature of the available frame, or the frame which can be constructed, is an important consideration in sample design. Relevant factors include the completeness and accuracy of coverage of the population in the frame, the type of sampling unit, and the amount and quality of the information on characteristics of the sampling units relevant for the efficiency of the sample selection and estimation procedures.

The cost of developing entirely new frames can be high. Where possible, one should use a design for which the sampling units involved are covered by existing frame materials, at least for the first stage of sampling. A frame may be constructed from a single source, or may have to be compiled by combining information from a number of sources. Different types and sources of frames may be used for different parts of the population. It is also possible to use more than one frame in combination to represent the same population more adequately. The use of multiple frames raises special issues in accounting for the units’ selection probabilities (Chapter 8).

B. Listings, sampling units and elements; rules of association

The definition of units of different types, the rules of association between them, and the empirical data needed for the implementation of those rules define the structure of the sampling frame.

Listings refer to the units to which the sample selection procedure can be applied in practice. Associated with them are sampling units to which the selection procedure is intended to apply. For instance, we may wish to obtain a sample of establishments (the sampling units), and select a sample from a frame which is meant to be a list of establishments. However, not all the listings in this frame may have a one-to-one correspondence with actual establishments. Some listings may correspond to establishments which no longer exist. Some others may correspond to establishments or other entities not in the target population of interest. Some listings may represent enterprises or other grouping of establishments, rather than individual establishments. Similarly in a household-based survey, listings may correspond to structural units such as dwellings or residential addresses, while the intended sampling units may be households, which are social units representing groupings of persons meeting certain specified criteria depending on the household definition adopted. Rules of association
link listings and sampling units. These rules can be more complex than the simple examples given above.

In addition to the formal rules of association, empirical information is needed to implement those rules in the field – to identify the actual sampling unit(s) associated with each listing selected into the sample. Such information may, for instance, be in the form of addresses, maps and verbal descriptions, informing on the location and identity of the units. In real frames, such information may be missing, incomplete, doubtful or erroneous.

In a multistage sample, the above requirements of correspondence and linkage between listing and sampling units apply separately at each stage of sampling, such as selecting areas from a list of area units at the first stage, and selecting households from a list of households within each selected area at the second stage.

At the last stage of sampling (and in any case in a single stage design) we need, in addition to rules and data associating listings and sampling units, rules associating the ultimate sampling units and the analysis units of final interest. Units of the two types may be the same, or may be different such as individual persons or children (as the analysis units) within households (as the ultimate sampling units).

C. Frame for multi-stage sampling

For a survey with multi-stage sample design, a frame is needed for each stage of selection; with such designs, the representation in the frame may be partly implicit, but it should still account for all the units. The sampling units used at the first stage of sampling are called primary sampling units (PSUs). For the first stage of selection, a frame of PSUs is needed, ideally covering the entire population exhaustively and without overlaps. Such a frame is called the primary sampling frame (PSF). Once a sample of PSUs has been selected, a frame for the next (second) stage is required only within the selected PSUs. This applies at any stage after the first: a frame is required only for the units selected at the preceding higher stage. The set of frames below the PSF are termed secondary sampling frames (SSFs). Units used at the last (ultimate) stage of sampling are called the ultimate sampling units (USUs). The creation of a PSF is normally a major and expensive operation since it must cover the entire population, and there are strong practical reasons to seek to compile it from pre-existing sources. By contrast, SSFs often have to be created or updated specifically for the survey.

In a frame involving sampling units of different types, it is useful to distinguish between two categories of units, which we may term respectively as macro units and micro units. This distinction is useful since, as concerns the sampling frame, the two categories of units tend to involve different requirements and problems. Macro units refer to relatively large collections of analysis units – examples being areas or large establishments. Consequently, macro units are relatively few in number and often also more stable over time. Micro units may be the same, or at least are of the same scale, as the analysis units: they are nearly as numerous and as small as the latter – such as households or small establishments (micro units in the frame), each containing a few individuals or working children (analysis units).

The above concepts are illustrated in Figure 4.1. The diagram on the right represents the situation for a two-stage design. The PSF contains a list and necessary auxiliary
CHAPTER 4

4. Sampling from imperfect frames

information for primary sampling units covering the whole population. These are macro units, which may be of different types. Commonly, these are area-based units. Within the selected PSUs, lists and necessary auxiliary information are compiled for the next stage units. After the last stage of selection, normally a sample of some sort of micro units is obtained.

The diagram on the left in Figure 4.1 depicts a single-stage design. The sampling units may vary in size, possibly from very large to very small. In the illustration we have shown them divided into two categories for convenience – large (macro) units, and small (micro) units. Often the two parts may be sampled using different frames and designs. Large units may have lists available from which they can be selected directly; all or a part of them may be included in the survey automatically. Selecting small units in a single stage requires detailed lists of such units covering the whole population. Alternatively, small units may be selected using multi-stage sampling as depicted in the diagram on the right.

**Figure 4.1. Representation of single- and multi-stage sampling**

---

D. Essential requirements for a frame

(1) The sampling frame contains a record for each frame unit. An item that is indispensable is unique identification of each unit. To ensure uniqueness and to control and facilitate sample implementation, a well-designed system of numerical identification is needed.
4.1 Basic concepts

(2) The frame requires auxiliary material and information for locating and demarcating the units in the field, such as area maps, descriptions and landmarks, especially near boundaries of the units.

(3) The effectiveness and efficiency of the frame is improved when it contains information on the unit’s characteristics beyond simply a unique identifier – such as information on the size and type of the unit, the stratum or sector it belongs to, and unit selection probability.

Table 4.1, based on United Nations (1986), lists the categories of information that may be included in records for frame units.

<table>
<thead>
<tr>
<th>Information categories</th>
<th>Purpose(s)</th>
<th>Examples, remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Type of unit code</td>
<td>Distinguish different types of frame units</td>
<td>Necessary when the frame includes more than one type of unit</td>
</tr>
<tr>
<td>2 Primary identifier</td>
<td>Uniquely identifies frame unit</td>
<td>Preferably numeric</td>
</tr>
<tr>
<td>3 Secondary identifiers</td>
<td>Aid in locating units</td>
<td>Names, addresses, keys to maps</td>
</tr>
<tr>
<td>4 Links to higher-level units</td>
<td>Identifies higher frame units and administrative divisions of which the unit is a part</td>
<td>Necessary for sub-national estimates. Some of this information may be incorporated in primary identifiers</td>
</tr>
<tr>
<td>5 Stratification variables</td>
<td>For grouping units prior to selection</td>
<td>Urban/rural, city size, economic activity, population density, etc.</td>
</tr>
<tr>
<td>6 Measure of size</td>
<td>For use in PPS selection, stratification, estimation</td>
<td>Census counts of population or household, etc.</td>
</tr>
</tbody>
</table>


Normally, sampling frames have to be used over a period of time following their initial creation. Durability of the frame is therefore a very desirable characteristic. The durability of the frame declines as we move down the hierarchy of units. The primary sampling frame (and to a lesser extent, frames of intermediate level units) usually represents a major investment for long-term use. By contrast, in most surveys, it is necessary to prepare fresh lists of the ultimate sampling units shortly before the survey enumeration. It is a major advantage in a survey if it can utilise lists which were originally prepared for some other survey or census. Lists of structural units such as dwellings are usually more durable than lists of social units such as households; pre-existing lists of individual persons are hardly ever useful, at least in most countries lacking good population registers.

This concept of differences in ‘durability’ among units of different types is particularly important in the context of linked surveys, when a second survey based on a sub-sample of the first survey, as in the case of a child labour survey based on a larger labour force survey.

For example, a list of area units may be provided by the LFS; a sub-sample of those units may be selected for the survey of child labour. A variety of sampling designs are possible. For instance, the CLS sample may be drawn from

(i) all areas in the LFS sample; or

(ii) only the areas containing establishments or households with children; or
(iii) only areas containing establishments or households with working children; or
(iv) actual lists of households selected in the LFS; or
(v) only the households successfully enumerated in the LFS; or
(vi) among those, only households which contain children; or
(vii) only households containing labouring children; or
(viii) actual lists of children listed in LFS households; or
(ix) lists only of children identified as workers; or finally
(x) one may even carry forward information on characteristics of such working children.

Clearly the durability of the units involved, as in the LFS-CLS links in the above example, declines as we move down the above list, and consequently also the acceptable time lag between the two operations.

4.2 Area frames

Often a distinction is made between ‘area’ samples or frames, and ‘list’ samples or frames. However, in a formal sense, generally all samples we use are drawn from lists of some sort. For example, even when we require a sample of areas, institutions, locations etc., we start with a list of the units to be selected. Listings refer to the individual units to which the sample selection procedure can be actually applied.

Nevertheless, the distinction between area frames and list frames is a useful one from a practical point of view. In this section we note some features of area frames, in particular some commonly encountered problems in using them. In subsequent sections, we will discuss in greater detail problems encountered in frames based on lists.

4.2.1 Basic requirements

Ideally, area units should cover specified land areas with defined boundaries, and the boundaries should be delimited in terms of physical features (streets, roads, railway lines, rivers, landmarks, etc.). However, sometimes area boundaries have to be created ‘artificially’ and defined in terms of verbal descriptions. Sometimes it is impossible to avoid using units which do not correspond to strictly delimited land areas – such as villages or other communities lacking sharp or officially defined boundaries. In such situations, special care and extra steps are required to ensure that fieldworkers apply uniform procedures in delimiting area boundaries defining the sampling units.

Strictly speaking, the frame units at each stage consist of the set of units of the type to be selected at that stage. In practice, sometimes units in the available frame material are of a different type than those required. These then have to serve as ‘building blocks’ to construct the type of units required by the sample design. For instance, the available area units may have to be segmented to form more suitable smaller sampling units (e.g. more compact blocks formed from census enumeration areas); or conversely, the
available units may have to be aggregated to form larger sampling units (e.g. larger
neighbourhoods formed from individual census enumeration areas).

In many surveys the area sampling frame has been based on census enumeration
areas, and the design has involved only a single area stage. Normally, fresh lists of
households, establishments or similar units are prepared within each selected area.

### 4.2.2 Common problems with area frames

Area frames are more stable than list frames. Nevertheless, area-based frames also
suffer from coverage and related errors. These usually arise from the failure to define
and identify physical boundaries of the area units correctly, and from poor quality of
the lists of the ultimate units such as establishments, dwellings or households within
sample areas. Common imperfections of area frames include the following.

1. The failure to cover the population of interest exhaustively. For instance, in a
   number of developing countries with inadequate cartographic work, the available
   frames are actually composed of lists of localities rather than of proper aerial
   units; scattered populations outside the listed localities may not be covered in such
   frames. Under-coverage also increases as the frame becomes outdated with time.

2. Errors and changes in area boundaries. These may arise from errors in identification
   of the boundaries and of boundary changes after the frame was prepared. The unit
   boundaries as defined in maps or descriptions may differ from boundaries of the
   actual sampling units. These in turn may differ from the boundaries of units on
   which other relevant information is available in the frame (information such as on
   size and density of the population).

3. Inappropriate type and size of units. The available units may be too large, too
   small, or too variable in size to serve as efficient sampling units. As noted, the
   available frame units may need amalgamation or segmentation before they can be
   used for sample selection.

4. Lack of auxiliary information. Information on size and other characteristics of
   the units, required for efficient sample selection, may be inaccurate or simply
   unavailable.

5. High cost. Area frames are generally expensive to create and maintain, unless they
   already exist for other, administrative purpose. Usually the investment is justifiable
   only when the frame can be used repeatedly, for many surveys and survey rounds.

### 4.2.3 Desirable frame properties

Table 4.2, from United Nations (1986), provides a checklist of desirable frame
properties. The properties shown in the list have been grouped into three categories:
properties related to quality, those related to efficiency, and those related to cost of the
frame.
A. Quality-related properties

These are the properties which help to control non-sampling errors, especially coverage errors.

The first desirable quality-related property is that the frame consists of well-defined units. Some area units are administrative units. Higher-level administrative units are usually well-defined in the sense that they have recognised boundaries that are usually clearly delineated in maps and descriptions, correspond to physical features, and are known to officials and people living in the area. Lower-level administrative units are not always well-defined in the same sense. It is also possible to use non-administrative area units, the most common example being enumeration areas (EAs) created for the population census. Sometimes small area units are established specifically for the use of sample surveys: for example, EAs may be divided into more compact segments or blocks using internal physical boundaries.

Table 4.2. Desirable frame properties

<table>
<thead>
<tr>
<th>Quality-related properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Well-defined units</td>
</tr>
<tr>
<td>Adequate identifiers</td>
</tr>
<tr>
<td>Complete</td>
</tr>
<tr>
<td>Up-to-date</td>
</tr>
<tr>
<td>Stable units</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Efficiency-related properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inclusion of accurate and up-to-date supplementary information</td>
</tr>
<tr>
<td>Flexibility - choice of sampling units available</td>
</tr>
<tr>
<td>Good quality maps of units available</td>
</tr>
<tr>
<td>Easy to manipulate and process</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cost-related properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low cost of acquisition or preparation</td>
</tr>
<tr>
<td>Low cost of use</td>
</tr>
<tr>
<td>Low cost of maintenance</td>
</tr>
</tbody>
</table>


The second desirable quality-related property is that all units have adequate identifiers. Of course, a unique identifier is indispensable for each frame unit. In addition to such a ‘primary’ identifier, there can be ‘secondary’ identifiers - such as maps, addresses, descriptions and other supplementary information - which aid in locating and demarcating the units in the field. For example, sometimes lower level administrative units (such as villages) serving as PSUs for the sample do not have unique names, and proper identification requires reference to the name of the higher level unit (such as the district) in which they lie. The latter serve as the secondary identifiers in this example.

The next desirable frame property is its completeness. Completeness has two aspects: the extent to which the intended coverage is actually achieved; and the extent to which the desired information (e.g. on stratification variables and measures of unit size) for each frame unit is included in the frame. It is necessary to devise appropriate checks to
verify both these aspects of coverage. The type and efficacy of the checks would differ depending on type of the frame.

Changes occur over time that affect both the number and characteristics of frame units. Smaller or lower-level units (villages, city wards) are likely to change more rapidly than larger higher-level units, but the latter are also by no mean immune to change. The frame being up-to-date means that, since the time of its creation, the changes in the number and characteristics of units in it are relatively small. This property may be different for units of different types in the same frame. As already noted, units of some types are more stable over time than others.

B. Efficiency-related properties

These are frame characteristics that facilitate efficient sampling design, meaning the achievement of lower sampling error for a given cost.

Among the most important of these properties is the inclusion in the frame of accurate, current supplementary data for each frame unit. Measures of unit size, such as the population, households or establishments in the area, are especially useful. Measures of unit size are useful for determining stratification, sample allocation, the sampling rates to be applied, the numbers of units to be selected at the various sampling stages, etc. These can also be useful at the estimation stage as auxiliary variables.

Flexibility of a frame refers to the possibility of choosing units of different types for different purposes from the same frame, such as for different surveys.

For area units, good definition depends to a large extent on the availability of accurate, detailed, large-scale maps showing the boundaries of each unit. Good maps permit the use of smaller more compact areas, which can be more efficient (cost effective).

Another frame property that facilitates the design of efficient samples is the ease of manipulation of the records (listings) of the frame units, such as through proper structuring, ordering and computerisation of the frame.

C. Cost-related properties

Properties that favour quality and efficiency in the use of the sampling frame have costs attached to them. A careful weighing of costs and benefits is usually needed, but is not always easy. Furthermore, different costs associated with the frame are interrelated. For instance, resources invested in the initial development of a frame can reduce subsequent costs of its use, maintenance and updating. The balance of these factors depends on how extensive the use of a frame is in terms of different surveys and extended time period.
4.3 Imperfect list frames

4.3.1 Basic consideration: the relationship between sampling and analysis units

In Section 4.1.2B, we referred to three types of units: listings in the frame, sampling units, and analysis units. Listings refer to the units to which the sample selection procedure can be applied in practice, and associated with them are sampling units to which the selection procedure is intended to apply. The fundamental distinction is, however, between listing or sampling units on the one hand, and on the other hand, the analysis units defining the target population on which the information is to be collected and analysed in the survey. The process of sample selection is applied to listings in the frame and through them to the sampling units. Then, on the basis of existing links between sampling units and analysis units, a sample of the analysis units is obtained.

Conventional sampling designs assume the situation where each analysis unit in the study population is linked uniquely to one and only one sampling unit. This is depicted in the diagram below. Symbol ‘A’ stands for a listing or sampling unit, and ‘B’ for an analysis unit in the target population of interest.

In practice, the relationship between units of the two types can be more complex. Listing or sampling units may be linked with analysis units in a variety of ways other than having a one-to-one correspondence. Different types of situation may be distinguished. The departure from one-to-one correspondence between sampling and analysis units may be: (1) exceptional, affecting a relatively small proportion of the cases; or it may be (2) quite widespread affecting a significant proportion of the frame; or it may be (3) general or systemic, a rule rather than an exception, affecting most of the frame.

In terms of its effect, the departure from one-to-one correspondence between sampling and analysis units may: (i) adversely affect efficiency of the sample, without introducing a bias; or (ii) it may introduce bias unless it is corrected for at the estimation stage, for instance by modifying sample weights. In addition there are also situations when (iii) the departure from one-to-one correspondence is in fact useful, such as for increasing the achieved sample size as in multiplicity sampling (Chapter 7), or improving coverage of difficult-to-reach subpopulations, or producing substantively more meaningful statistics (e.g. statistics of visitors on the basis of a sample of visits to a location or facility -
Chapter 10). In such situations, one may have introduced departures from one-to-one correspondence between sampling and analysis units deliberately. Several examples of such procedures – which can be very useful in sampling elusive populations of labouring children – will be developed and illustrated in subsequent chapters.

We will discuss in Section 4.5 various possible scenarios when departure from one-to-one correspondence between sampling and analysis units are systemic, affecting most of the frame. Here we confine our attention to the conventional design situation in which one-to-one correspondence between sampling and analysis units is the characteristic feature. Frame problems arise in this situation as well, because in real surveys the correspondence between sampling and analysis units is almost always less than perfect. In a small or large part, the linkage may involve one sampling unit linked to none, one or several analysis units, or conversely, one analysis unit linked to none, one or several sampling units. Such imperfections in a list frame are described below. The discussion is largely based on Kish (1965).

### 4.3.2 Imperfections which increase variance, but without introducing bias

**A. Blanks**

A blank means that a listing represents no real unit, but is merely a blank entry. The presence of blanks in the list does not affect the selection probabilities of the units, but the number of units selected becomes a random variable. If that number is kept fixed, the probabilities of selection become subject to random variation, and will become unknown if the number of actual units present in the list is not known. This scenario is depicted in Figure 4.3A.

*Figure 4.3A. Blanks*

Some listings (A) are associated with no analysis units (X), or only with units not in the target population, or the original unit no longer exists.

Symbol ‘X’ indicates that a listing or sampling unit (A) has no corresponding analysis unit. The listing does not result in the selection of any real unit from the target population.

*Blanks in the list do not represent non-response.* A common error is to treat blanks as non-response, and substitute for them or adjust the sample weights to compensate for them. Substituting for blanks increases the probability of selection of a unit in proportion to the number of blanks associated with it – for instance with systematic sampling the number of blanks that precede that unit in the list – which is clearly an incorrect procedure.
If the proportion of blanks in the list is small, the easiest option may be simply to ignore the problem – i.e. to select the sample from the existing list as it is, and accept blanks, if selected, as adding nothing to the sample obtained.

The proportion of blanks tends to increase as the time interval between the listing and the actual survey operation increases. The proportion of blanks also depends on the type of unit: some units are less stable than others, for instance listings of individual persons compared to listing of occupied addresses. Another determining factor is the quality of the listing operation.

Factors such as the above can not only increase the proportion of blanks in the list, but can also introduce uncertainty as to whether a blank found in the list implies a genuine absence of a population element, or has simply resulted from a failure to establish correspondence with an eligible element which in fact exists in the population. This uncertainty introduces bias in the survey estimates.

A different type of problem exists with lists containing a very high proportion of blanks as a result of the presence of ‘foreign elements’, i.e. elements not eligible for inclusion in the target population for the particular survey under consideration. This is a very common problem in child labour surveys, where the primary interest is in specific populations, such as households containing children, among the population of all households, or establishments employing children among the population of all establishments. In such situations, the important question is the relative cost of listing and screening the units. Some issues concerning these operations are considered in more detail in Chapter 6. Some of the important parameters are the following.

- What is the relative per unit cost of listing, and what proportion of the listed units belong to the target population of interest?
- How costly and difficult is the screening operation required to separate out units of the target population from other, foreign, units in the list?
- How accurate is this separation – both in terms of false inclusions and false exclusions?
- How costly and inconvenient is it to encounter foreign elements later during the more detailed survey interview?
- What is the (monetary, statistical, operational) cost of lack of control over sample size as a result of the presence of blanks in the selected sample?

It is not uncommon to have a design in which we wish to select a fixed number of eligible ultimate units from each area (or from some other type of unit) selected at the preceding stage, irrespective of the presence of blanks in the list from which these selections are made. For such a sample to be a probability sample, it is necessary to keep a record of the number (say n) of selections which had to be made in order to obtain the required number (say m) of actual units. If the frame contains (say) N listings, including blanks, then the ratio (n/N) is the sampling fraction applied. Note that the numerator is the number of listings selected, and not the number of units actually obtained in the sample.
B. Clustering of analysis units

This refers to single listings representing multiple elements or analysis units. For example, one may select a sample of establishments or of households from a list of such units, while the elements (survey units) of interest are labouring children within establishments or households. This scenario is depicted in Figure 4.3B. Symbol ‘A’ represents a listing or sampling unit, and ‘B’ the analysis units associated with it.

Figure 4.3B. Clustering of elements

More than one analysis unit (B) associated with the same listing (A); selecting the listing selects the whole cluster of analysis units it represents (is linked to).

Clustering of elements within listings available in the frame represent a very common situation. Different scenarios are possible.

(1) In certain situations, keeping in the sample all elements in each selected cluster is the appropriate design.

- Often the clusters involved are small, each listing containing no more than a few elements, and a high proportion containing none or only one element. In such situations, often the simplest solution is to include all the elements which occur within each selected listing – for example taking all the working children in each selected establishment or household.
- When the interviewing of the final elements in the clusters is relatively quick and cheap, compared to the cost of accessing and enumerating the cluster itself, it may not be worthwhile to subsample within clusters. A ‘take-all’ design may be the most economical and convenient.
- Sometimes subsampling may require a lot of information, which is complex or expensive to collect, or is error prone or burdensome for the respondent.

As such, the procedure involving ‘take-all’ sampling does not distort the selection probabilities, since each element receives the selection probability of the listing representing it. When the cost of clustering in terms of increased variance is likely to be small, lower costs and practical convenience can be useful advantages of take-all sampling.

(2) Secondly, there are situations when ‘take-all’ sampling is employed in a major part of the population, and any subsampling within clusters is confined to a part only.
• For example, in many surveys using residential addresses as sampling units in order to obtain a sample of households as analysis units, generally all households found at each address selected are taken into the sample, but a subsample is taken if the number of households at an address exceeds a certain specified limit, such as 3 or 4 households at the address.

• Sometimes, small and large sampling units are neatly divided into different strata, for instance strata of small and large establishments. In selecting children within establishments in this case, we may employ ‘take-all’ sampling for small establishments, and subsampling of children within large establishments in the sample.

(3) There are situations when the clustering of the final elements of interest is common and widespread in the population, and it necessary to adopt some special procedure to deal with it.

• When the clusters involved are common and relatively large, often the appropriate design is to introduce full-fledged multi-stage sampling, for instance sampling of clusters followed by subsampling within clusters using an appropriate procedure.

• When the clusters involved are common and quite variable in size, a more convenient (but possibly more expensive) procedure may be to prepare a complete list of all the ultimate elements of interest, in all clusters in the population, and then from this more perfect list frame, select a sample of the ultimate units directly, in a single stage.

(4) Often selecting one unit at random from the clustering is unnecessary. However, in certain situations, it is necessary, desirable or economical to avoid selecting more than one element from the same cluster in the sample. Consider a sample of individual persons in a household. There can be several reasons such as the following for confining the sample to selecting only one person from each sample household for inclusion in the survey.

• In some surveys involving sensitive topics, including more than one individual from the same household can result in ‘contamination’ of the results, i.e. responses of different individuals influencing each other.

• When the final interview involved is lengthy or burdensome, it may be too much to subject any household to more than one such interview.

• The information from different individuals in the same household may be highly correlated, making it inefficient to take more than one person from the same household. Very similar considerations can apply to sampling working children within establishments.

With a design involving the selection of one element per sample cluster, the selection probability of a unit (e.g. persons) varies - in comparison with the selection probability of its cluster (e.g. household) – in inverse proportion to the number of units in the cluster. This must be taken into account in determining the sample weights to be used at the analysis stage.
C. Change in unit characteristics

Efficiency of the sample design, selection and estimation depends on auxiliary information concerning characteristics of units in the population. Such information has to be available prior to sample selection, i.e. it has to be present in the sampling frame. Non-availability or inaccuracy in such information, including due to changes in unit characteristics subsequent to the preparation of the frame, tend to reduce efficiency of the sample. This scenario is depicted in Figure 4.3C.

Figure 4.3C. Changed unit characteristics

Identity, or eligibility, of the unit (B) to be included in the target population has changed (to B') between listing and sample selection or between sample selection and survey enumeration.

Here are a few examples.

- Information on unit characteristics and distribution is used for stratification. The effectiveness of stratification is reduced if units have been misclassified or have subsequently changed to move across strata boundaries. We may note that when units move across estimation domains in this way, their selection probabilities and hence design weights continue to be determined by the stratum from which they were selected, but their contribution to the survey estimate is to the domain they are in at the time of the survey.

- A common procedure is to select units with probability proportional to some measure of size – the actual size of the unit or some function of it. The effectiveness of this procedure depends on the correlation between the assumed unit size measure at the time of sample selection, and its actual size measure at the time of the survey data collection.

- Estimation procedures often use auxiliary information on unit characteristics coming from the frame or other sources external to the survey.

- In child labour surveys, or other such surveys targeting special subpopulations, often an important item of information is whether or not some larger units used for sample selection contain an element of the target population – for example whether a household or an establishment in the frame contains any working children. This characteristic of sampling units can be subject to relatively rapid change.
4. Sampling from imperfect frames

4.3.3 Imperfections resulting in potential bias

A. Under-coverage

Under-coverage means some units are not represented in the frame. This is the most serious and difficult problem, and it biases the results of many surveys. This is because non-representation in the frame is usually not random, but is selective in terms of characteristics of the units. This scenario is depicted in Figure 4.4A.

**Figure 4.4A. Under-coverage**

Some analysis units (B) are not associated with any listing or sampling unit (X); the list covers only a part of the target population.

Symbol ‘X’ indicates that an analysis unit (B) in the target population has no corresponding sampling unit in the frame. The analysis unit has no chance of being selected.

Under-coverage is a major problem particularly in surveying an elusive population of labouring children because often no lists of units employing child labour, and of course of labouring children individually, are available ready-made. Under-coverage may occur due to non-representation in the frame of units employing child labour (this is more commonly the case with establishments than with households containing working children), or due to the failure to identify working children within such units when selected.

There is no simple or cheap solution to the problem of under-coverage. The only advice that can be given is to increase the effort spent on the preparation and compilation of lists.

B. Non-response

Non-response arises from the failure to enumerate units selected into the sample. As such, it may not appear to be a problem resulting from shortcomings of the sampling frame. But in reality, the two types of problem can be related. This is because missing or imprecise information on the identity, location or other defining characteristics of units can result in increased non-response.

Furthermore, while factors causing non-response may be different, the effect of non-response can be similar – or appear to be similar – to other problems such as under-coverage, distortions in unit selection probabilities, or presence of blanks in the frame.

Therefore, it is useful to compare the problem of non-response with other problems resulting from frame imperfections. Non-response may be represented as in Figure 4.4B.
4.3 Imperfect list frames

Analysis unit (B) is selected into the sample through its association with a sampling unit (A), but is not successfully enumerated (X).

Analysis unit (B) is represented by sampling unit (A) in the frame, and is taken into the sample following the selection of (A). Symbol ‘X’ represents the situation when this analysis unit selected into the sample is not successfully enumerated to be included in the achieved sample. The relationship between the end points of the above representation, (A)—(X), appears similar to that for blanks (Section 4.3.2A). However, unlike blanks, non-response reduces the effective selection probabilities of units, and therefore requires adjustment of unit weights to reduce that effect. It can introduce bias in the results in any case. In fact, non-response affects representativeness of the sample in a way similar to the effect of under-coverage in the frame.

A technical discussion of issues concerning estimation under non-response and frame imperfections can be found in Särndal and Lundstrom (2005, Chapter 14).

C. Failure to locate units

Failure to locate units, including the failure to identify which unit(s) a selected listing represents, can be a problem when the lists used for sample selection are out-of-date or have been prepared with insufficient care. This scenario is depicted in Figure 4.4C.

The dotted line indicates uncertainty about the link between the sampling unit (A) and the analysis unit (B).
This is a common problem in the absence of a clear and complete description in the frame for identifying units in the field (such as difficulty in identifying children simply from their names). The problem can also be caused by insufficient effort by the field workers. The failure to locate selected units which actually exist constitutes non-response. This problem is often confused with that of blanks, which concerns units that actually do not exist (see above). Units not located are often indiscriminately reported as non-existent.

On the other hand, failure to locate units selected from the frame may also be the result of the fact that the units have actually ceased to exist by the time of survey data collection. In this case they indeed are blanks in the list. The only way to reduce this dilemma is to make a special effort to recheck the situation when units are reported as “not found”.

D. Duplications (multiple listings)

By ‘duplications’ we mean that the same unit is represented by more than one listing. This scenario is depicted in Figure 4.4D.

Sometimes the problem arises from the nature of the frame – as unavoidably occurs, for example, in the selection of households from an electoral roll (listing all eligible voters in each household), in the selection of a parent from a list of children at school, or in the selection of clients or service receivers from records of visits to a service facility. Taking these examples in turn, a household is selected if any of the registered voters in it is selected; a parent is selected if any of his/her children in the school list is selected; and a client is selected from any of his/her visits. Generally, selecting units subject to multiple events on the basis of a listing of individual events gives each unit a probability of selection proportional to the number of events it is associated with. There are situations when such ‘PPS’ selection is what is actually desired for the survey. But more generally, steps are required to compensate for or avoid such variations in selection probabilities.

Figure 4.4D. Duplications

More than one listing or sampling unit (A) associated with the same analysis unit (B); selecting any of those listings selects the associated analysis unit.
Eliminating all duplications in the frame is one solution, but it is not always necessary to do so. One may for instance use some system which identifies only one of the listings (such as the first or the last listing) as representing the unit, the other listings of the same unit being treated as blanks (i.e. the selection of one of those listings does not result in the selection of the unit concerned). Alternatively, each selection of a unit may be weighted inversely proportional to the number of listings in the frame representing that unit (including the listings not selected). Simply eliminating the duplications which happen to appear in the sample does not solve the problem.

Much more difficult is the problem of unsystematic duplications in the list, usually resulting from the failure to identify the fact that different listings actually represent the same unit. This can happen, for example, if the same unit is recorded in the list several times with slight differences in name, address or description. Such problems occur often in surveys of small establishments. In such cases painstaking work to eliminate all duplications in the list may be the only solution.

4.4 Illustrations from national practice

The following examples are given to illustrate some of the concepts concerning sampling frames defined in the preceding sections. The surveys in the examples are mostly establishment-based baseline surveys covering selected sectors of child labour. (Design of establishment surveys will be discussed more fully in Chapter 5.)

The following illustrations come from a series of baseline surveys on child labour conducted in Bangladesh. The first is a set of five closely related but still separate surveys in different sectors of activity. Their frames are based on a common listing operation covering the five sectors, and they use a common sample of primary sampling units. Samples within the selected PSUs are, however, independent and are comprised of establishments of different types as the final units.

The second illustration is of a multi-sectoral survey covering establishments of all types in metropolitan areas in Bangladesh. The example illustrates difficulties in obtaining lists of small units with good coverage.

The survey descriptions are extracted from published national reports on the surveys. Some critical observations are made following the description of the surveys in each illustration. Our primary interest in these comments is to point out shortcomings in past practice so as to contribute towards improving the sampling frames and designs in future surveys of child labour.

4.4.1 Set of five baseline surveys in Bangladesh, 2002-2003

In Section 2.3.2B, an example was given from the 2005 Baseline Survey on Child Domestic Labour in Bangladesh (Bangladesh, 2006a). That was a household-based survey. In this section, example is provided of a set of five earlier (2002-2003) baseline surveys in the country covering worst forms of child labour. These were all non-household based surveys. Three of the surveys were surveys of children working in small or informal sector establishments, a significant proportion of which employed
child labour; the remaining two involved sampling of children at certain sites or locations.

The surveys were a part of a broader undertaking, the National Child Labour Survey (NCLS) 2002-2003, conducted by Bangladesh Bureau of Statistics. The NCLS comprised the following components aimed at the development, testing and application of different methodologies:

1. a household-based survey of child labour (CLS) at the national level;

2. baseline surveys in five selected worst forms of child labour, namely: (A) battery recharging/recycling; (B) welding, (C) automobile workshops; (D) road transportation; and (E) street children (Bangladesh, 2004a, 2003b, 2003c, 2004b, and 2003d, respectively); and

3. an establishment-based child labour survey in selected metropolitan cities of the country (Bangladesh, 2003a).

Set (2) surveys are discussed in the next two subsections, and comments on (3) appear in Section 4.4.3. There was another major baseline survey in Bangladesh later (in 2005) which covered a wide range of hazardous child labour sectors, comments on which appear in the next chapter (Section 5.11.3) in the context of issues concerning sampling of establishments.

**Common sampling frame and primary units for the baseline surveys**

Due to lack of basic information from existing – whether primary or secondary - sources about the incidence and distribution of establishments or numbers of children working in these sectors, it was not possible to develop the sampling design for the surveys without a prior operation.

In order to construct a frame, a complete listing (census) of establishments in the five target sectors (sectors (A)-(E) listed above) was carried out through a so-called Quick Count Survey (QCS) conducted in all areas of the country. The listing obtained information on names and addresses of the establishments in sectors (A)-(C), and on the numbers of children aged 5-17 and adults employed in these establishments. For sectors (D) and (E) the units listed were locations of various types where children tended to congregate. The combined listing was a large operation, covering urban and rural areas and involving 483 district (Thana/Upazila) statistical offices (covering a total of 502 districts in the country), and 23 regional statistical offices throughout the 6 divisions into which the country is divided. The operation engaged 2,400 enumerators over November-December, 2002.

The listing census not only produced the frame for selecting samples of establishments for more detailed surveys, it also directly served as a source for a great deal of substantive information.

For the selection of the sample for the detailed survey, the districts (Thana/Upazila) covering the whole country formed the PSUs. Essentially all districts contained some units of the target population.
It is an unusual feature of the design that the selection of PSUs (districts) for the surveys was done only after the establishment listing operation covering all areas in the country. A common sample of districts was then used for all the five sectoral baseline surveys. For sample selection, the PSUs were divided into 3 strata:

- Urban Stratum I: 6 metropolitan/divisional cities of the country (a total of 47 districts)
- Urban Stratum II: 58 district towns/cities
- Rural Stratum III: all areas except the above urban strata (a total of 397 districts)

PSUs were selected with equal probability within each stratum, the probabilities being $f_1 = 0.5$, $0.5$ and $0.125$ respectively for the three strata. At the second stage, establishments were selected at a rate $f_2$ which was constant within each stratum and was proportional to $(1/f_1)$ for the stratum concerned, so that the overall sampling rate $f = f_1 f_2$ is uniform throughout. Hence the allocation is proportionate to the number of establishments. (This applies whether we consider all establishments, or only establishments employing children.)

The target sectors were concentrated in urban areas, though they had been expanding outside those areas. In order to obtain a similar sample size per PSU in terms of the number of establishments selected, a higher proportion of establishments needed to be selected in rural PSUs, compared to urban PSUs. For this reason, a smaller value of $f_1$, and hence a larger value of $f_2$, was taken in Stratum III.

A. Battery recharging/recycling establishments

**Listing ‘census’ of battery recharging/recycling establishments**

The listing census estimated a total of 12,200 battery recharging/recycling establishments in the country, employing 22,500 workers, of whom 5,500 (around 25 per cent) were child workers aged 5-17 (see Table 4.3A). The census revealed that indicators such as the average number of child workers per establishment and the ratio of child workers to adult workers were fairly uniform across the 23 regions of the country. All child workers were male, of which around 60 per cent were aged 15-17, 35 per cent aged 12-14, and 5 per cent as young as 5-11. Most (99 per cent) establishments were owned by individuals or families as single proprietors. Among the working children, 20 per cent were unpaid family workers; the rest being mostly regular paid workers and paid or unpaid apprentices.
Table 4.3A. Some results from listing for baseline survey covering battery recharging/recycling establishments

<table>
<thead>
<tr>
<th>Region</th>
<th>Establishments number</th>
<th>All workers number</th>
<th>workers per establishment</th>
<th>Child workers number</th>
<th>children per establishment as proportion of all workers</th>
<th>Sample size number (*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dhaka</td>
<td>4,170 34</td>
<td>7,145 32</td>
<td>1.71</td>
<td>1,781 32</td>
<td>0.43 0.25</td>
<td>137 33</td>
</tr>
<tr>
<td>Chittagong</td>
<td>1,972 16</td>
<td>4,108 18</td>
<td>2.08</td>
<td>942 17</td>
<td>0.48 0.23</td>
<td>69 17</td>
</tr>
<tr>
<td>Rajshahi</td>
<td>2,888 24</td>
<td>5,574 25</td>
<td>1.93</td>
<td>1,375 25</td>
<td>0.48 0.25</td>
<td>100 24</td>
</tr>
<tr>
<td>Khulna</td>
<td>1,659 14</td>
<td>3,009 13</td>
<td>1.81</td>
<td>711 13</td>
<td>0.43 0.24</td>
<td>55 13</td>
</tr>
<tr>
<td>Sylhet</td>
<td>725 6</td>
<td>1,461 7</td>
<td>2.02</td>
<td>318 6</td>
<td>0.44 0.22</td>
<td>27 6</td>
</tr>
<tr>
<td>Barisal</td>
<td>792 7</td>
<td>1,183 5</td>
<td>1.49</td>
<td>386 7</td>
<td>0.49 0.33</td>
<td>28 7</td>
</tr>
<tr>
<td>Total</td>
<td>12,207 100</td>
<td>22,480 100</td>
<td>1.84</td>
<td>5,513 100</td>
<td>0.45 0.25</td>
<td>416 100</td>
</tr>
</tbody>
</table>

(*) Sample size: number of establishments or number of children for more detailed interviewing. Generally one child per establishment was selected.

Source: Bangladesh, 2004a.

Sample for the detailed survey

The detailed survey was conducted during September-October 2003, i.e. with a gap of around 10 months from the listing operation. This is quite a long gap for the type of units (small establishments) being surveyed. The survey was conducted to obtain reliable regional estimates of the incidence and distribution of children aged 5-17 years currently working in battery recharging/recycling establishments. At the national level, the survey was designed to measure as many variables as possible: the number of working children, hours worked, health and safety, access to services, etc., with special emphasis on child abuses that are inherent in the worst forms of child labour.

At the third stage of sampling (i.e. selection of working children within establishments in the sample), it appears that mostly one working child actually appears in the sample per establishment employing child labour, though in principle up to 3 working children could be selected subject to availability. Since the average number of child workers per establishment is quite uniform across regions (in the narrow range of 0.43-0.49 in Table 4.3A), the proportionate allocation of the sample in terms of establishments gives nearly proportionate allocation in terms of working children in the population.

B. Welding establishments

From the listing operation, geographical distribution of the target population of welding establishments is found to be very close to that of the battery establishments described above. The main difference is that of size of the two populations of establishments and number of working children. The main results from the listing census are as in Table 4.3B.

As shown in the table, compared to the battery sector, the welding sector has more establishments, the establishments are larger on average in terms of the number of persons employed per establishment, and among the employees a higher proportion are children. In view of the similarity between the nature and distribution of the target populations in the two sectors, the sampling arrangements were also the same. The
field enumeration for the detailed survey of welding establishments was done during April 2003, i.e. with a gap of around 5 months from the listing operation.

The listing census estimated a total of 28,300 welding establishments in the country, employing 118,000 workers, of whom nearly 39,000 (33 per cent) were child workers aged 5-17. All child workers were male, of which 53 per cent were aged 15-17, 40 per cent aged 12-14, and 7 per cent as young as 5-11. Most (94 per cent) establishments were owned by individuals or families as single proprietors, the remaining 6 per cent were in partnership. Among the working children, only 1 per cent were unpaid family workers, the children being mostly regular paid workers or paid or unpaid apprentices.

There is more variation across regions in the average number of child workers per establishment – in the range 1.23 to 2.01.

### Table 4.3B. Some results from listing for baseline survey covering welding establishments

<table>
<thead>
<tr>
<th>Region</th>
<th>Establishments</th>
<th>All workers</th>
<th>Child workers</th>
<th>Sample size (*)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>number %</td>
<td>number %</td>
<td>number %</td>
<td>number %</td>
</tr>
<tr>
<td>Dhaka</td>
<td>9,493 34</td>
<td>40,168 34</td>
<td>12,236 31</td>
<td>136 34</td>
</tr>
<tr>
<td>Chittagong</td>
<td>5,083 18</td>
<td>26,056 22</td>
<td>8,392 22</td>
<td>72 18</td>
</tr>
<tr>
<td>Rajshahi</td>
<td>6,527 23</td>
<td>23,051 20</td>
<td>8,048 21</td>
<td>92 23</td>
</tr>
<tr>
<td>Khulna</td>
<td>4,087 14</td>
<td>15,222 13</td>
<td>5,691 15</td>
<td>58 14</td>
</tr>
<tr>
<td>Sylhet</td>
<td>2,062 7</td>
<td>10,200 9</td>
<td>2,582 7</td>
<td>31 8</td>
</tr>
<tr>
<td>Barisal</td>
<td>1,034 4</td>
<td>3,130 3</td>
<td>2,079 5</td>
<td>15 4</td>
</tr>
<tr>
<td>Total</td>
<td>28,286 100</td>
<td>117,827 100</td>
<td>39,028 100</td>
<td>404 100</td>
</tr>
</tbody>
</table>

(*) Sample size: number of establishments (total 404). Total number of working children interviewed was 434, just over one child per establishment.

Source: Bangladesh, 2003b.

### C. Automobile establishments

This survey used the same sampling frame and design, and in fact the same sample of PSUs as the Baseline Survey on Child Workers in Battery Recharging/Recycling and Welding sectors, but of course different samples of establishments and working children. The three sectors are similar in size and composition. An important parameter in the quality of the frame is the time lag between its creation and its use. In the automobile establishment survey, field enumeration was done during November 2002-January 2003, i.e. almost simultaneously with the listing operation. The main results from the listing census are as in Table 4.3C.
The listing census estimated a total of 9,600 automobile workshops/establishments in the country, employing 38,000 workers, of whom nearly 16,000 (42 per cent) were child workers aged 5-17. All child workers were male, of which 50 per cent were aged 15-17, 45 per cent aged 12-14, and 5 per cent as young as 5-11. Most (around 95 per cent) establishments were owned by individuals or families as single proprietors, the remaining 5-6 per cent establishments were owned in partnership. Among working children, practically none were unpaid family workers, being mostly regular paid workers or paid or unpaid apprentices. Table 4.4 gives comparison for selected items between the three baseline surveys (A)-(C).
D. Child workers in the road transport sector

This survey also obtained its sampling frame from the common listing operation described above. The sample of districts (PSUs) was common as well. The difference is the type of unit listed within areas. In this survey these units are not establishments but locations where road transport activities employing children are carried out. This gives rise to some fundamental differences from the other baseline surveys in the set.

The most important difference is that the lists obtained are subject to high level of under-coverage: locations carrying out certain activity are more difficult to identify and list than establishments with fixed premises. This is highlighted in the report of the survey as follows. The commentary in the report is preceded by a table of numbers of road transport activity sites and the numbers of children employed by region of the country.

“The above list of sites/locations should not be treated as complete because these only include the conspicuous ones which are well-known in the localities concerned. There may be many more scattered places / sites or road points where child workers are found engaged in various types of transport such as rickshaws, push carts, rickshaw vans, tempos and related road transports as helpers or pullers. So the total number of sites/places as well as the number of child workers may be much higher”.

The operation listed 9,873 road transport establishments, employing a total of 85,619 children (an average of 8.7 child labourers per site).

The second difference is that in the road transport sector, the population is more heterogeneous than in other sectors. This is because there are different types of road transport activity, and locations tend to specialise in the type of activity carried out there. Selection of the sample (within selected districts or PSUs) therefore required stratification of the frame according to the type of vehicle and the main type of activity at the location/site, in this case tempo/leguna, bus/minibus, rickshaw, and truck.

Thirdly, unlike establishments, sites do not present integrated activity with an identifiable owner or person in-charge. Hence the sample selection of children has to proceed immediately following the identification of working children found at the site. This can potentially increase the degree of under-coverage of the target population of working children.

At a sample of 350 sites/locations, a total of 442 children were interviewed in the detailed survey (an average of around 1.25 per site, i.e. on the average around 1 in 7 working children found at each site).

E. Baseline survey on street children

Obtaining a satisfactory sample of street children is a difficult task and requires special procedures. Several of these are discussed in subsequent chapters – for instance, capture-recapture and respondent-driven sampling procedures. Nevertheless, we give the example of the Bangladesh baseline survey of 2002-2003 on street children because it tried to use an ordinary list frame obtained from a combined operation of
listing establishments and sites/locations where working children are found. This in fact is the same frame as used for the other four baseline surveys described above.

The survey was designed to cover street children, defined as children: (i) aged 5-17 who are living (sleeping, eating, working, etc.) on the street; (ii) who sleep at night on the roadside, rail or bus stations, abandoned buildings, parks or other public spaces; and (iii) who are not living with their parents (or other related adults), even if on the roadside, pavement or in the slum. Furthermore, children being on the street was defined to be essentially an urban phenomenon, and therefore the survey was apparently confined to non-rural localities, i.e. to towns, cities and metropolitan areas.

It is a rather unusual feature of the survey that its frame is derived from the same listing operation as the one described above for the baseline surveys of establishments. The listing operation provided a frame of locations where street children congregate and also a count of the number of street children to be found at each such location. In total in the whole country, the operation found 413 locations with 2,573 street children, an average of just over 6 children per location. Six of the 23 regions of the country were reported to contain none or only an insignificant number (10 or fewer) of street children.

The number of locations where street children congregate and the total number of street children in these locations as obtained from the listing operation appear to be serious under-estimates for a country of the size of Bangladesh. The underlying cause is the under-coverage in the list frame. Procedures which may provide a reasonable coverage of establishments with a clearly defined identity and a contactable address, as in surveys (A)-(C) above, are less successful in covering more loosely defined units such as locations where children perform certain types of labour, as in the case of example (D). The coverage achieved is even less adequate when the units in addition tend to be mobile and reclusive, as in the case of street children.

The coverage achieved by a frame depends very much on the listing procedures followed in relation to the type of unit to be listed in the population. In the example of street children, success would depend on factors such as the following.

- How are the sampling units (street locations) and analysis units (street children) defined? How clear are these definitions for practical implementation in the field?

In addition to this basic consideration, the following issues have been pointed out in the survey report.

- What mechanism was adopted to know about the locations where the street children generally sleep at nights?
- What techniques were adopted to reach the street children?
- How were they counted? How effective was the counting system, given the fact that the street children did not have a permanent place of sleeping at night?
- What were the chances of missing some locations, and hence street children living or sleeping there?
4.4.2 Observations on the baseline surveys

The above examples illustrate several common aspects concerning sampling frames.

(1) Firstly, they illustrate the problem of surveying a population in the absence of an existing sampling frame. The creation of a sampling frame ‘from scratch’ can be a very large-scale and expensive operation: in the present case the operation of listing establishments and locations throughout the 6 divisions of the country involved 483 district and 23 regional statistical offices, engaging 2,400 enumerators over a two month period. In order to construct a frame, a complete listing (census) of establishments/locations in the five target sectors, (A)-(E), was carried out throughout the country through what has been termed a ‘Quick Count Survey’.

(2) It is often possible to economise by sharing the cost of the listing operation between different surveys. The present example provides a good illustration. Here a common listing operation served five national level baseline surveys carried out in five different sectors: battery recharging/recycling, welding, road transportation, automobile workshops, and children working on the street.

(3) Another aspect which affects the size of the listing operation is the amount of information to be collected during the operation. The basic objective of collecting detailed information during listing is to target the subsequent selection of the sample. In the present case, the listing obtained information on names and addresses of the establishments and locations, and the number of children and adults employed in these establishments. This required contact and interview with each establishment listed. A practical requirement is to minimise the amount of information to be collected from individual units during a large listing operation, or at least to avoid collecting any unnecessary information.

(4) Apart from providing a list from which the required samples can be selected, listing also has the purpose of providing information on unit characteristics to serve as a basis for efficient sampling design. In the present example, the information collected is used to determine different sampling rates by region for the two stages of sample selection. This however utilises only a small part of the information collected during listing which could have been used for improving the sample design – this making the procedure rather wasteful in the present example.

(5) Tables 4.3A-C presented above show that quite a wide range of information was collected during listing, which could be used for making substantive estimates about the population. It needs to be evaluated whether it would have been more efficient to collect some of this information on a smaller sample during the main survey subsequent to the listing operation. Most likely, it is wasteful to conduct a 100 per cent census to collect substantive information of the type reported for the surveys in the present illustration.

(6) The example is also exceptional in that the listing amounts to a complete census of establishments in the five sectors. Normally such listing is done on a sample basis, after selecting a sample of the PSUs.

(7) As emphasised in relation to survey (D), and especially to survey (E) above, the quality of coverage achieved can vary greatly depending on the listing procedure.
followed and the type of unit being listed. Units which lack clear boundaries or structures, are mobile, lack visibility, are difficult to approach, shy away from contact, etc., require special methods to be captured in the frame.

**Some critical remarks**

It is useful to make some critical remarks on the approach followed in these particular illustrations, in order to bring out some further practical points.

(1) As noted, the listing operation in the example involved a complete census of establishments in the five sectors, even though it is characterised as “a quick count survey throughout the country”. Such a major operation goes beyond the requirements of creating a sampling frame, especially a frame for drawing a two-stage sample. The objectives of such a census operation can, and should, be more general than simply serving as a basis for a survey of child labour. And if a complete census of establishments is to be undertaken, it would be more cost-effective and useful to cover all sectors, not just a few selected ones. An integrated approach with wider objectives would be preferable if a complete census is to be undertaken.

(2) If the objective is simply to obtain estimates of the numbers of establishments, the proportions employing child labour, and the numbers of children working in particular sectors, then a large-scale sample survey is likely to be more efficient than a complete census. Sampling introduces sampling error, but often yields a lower level of non-sampling errors and is likely to be substantially cheaper. In the household sector, for instance, information on child labour is almost always obtained through sample surveys rather than complete censuses. The same applies to surveys of units of other types.

(3) The normal procedure for multi-stage frame construction involves the following sequence of steps:

(i) the construction of a primary sampling frame (PSF) of large units, e.g. area units, as the PSUs;

(ii) selection of a sample of PSUs;

(iii) listing of ultimate units, e.g. establishments, within the selected PSUs;

(iv) collection of substantive data on a sample of the units listed; and possibly

(v) conducting a more intensive or detailed survey on a subsample of the above.

In this illustration, listing, step (iii), has been done for the whole population, before sampling PSUs, step (ii). The detailed lists produced for PSUs not selected for the survey subsequently remain unused, which is wasteful. Generally, it is desirable that the PSF is based on existing information (administrative sources, past censuses, other surveys, etc., as available) and/or on information collected on relatively large units covering the population. A complete census can be considered when the population is composed of large, relatively few units. Detailed listing involving a very large number of small units is generally best confined to a sample by using a multistage design.
The information for constructing the PSF can be approximate; it may, for example, be put together from different administrative sources, use simpler but correlated surrogate variables, or be derived from small scale studies specially conducted for the purpose; some of the information may even be qualitative, for example based on local expertise or observation. It is usually not necessary to undertake complete listing of the ultimate stage units (establishments or working children in the present examples) in order to identify characteristics of higher stage units (e.g. sample areas) needed for sample design and selection.

The purpose of listing is to create a frame of the ultimate sampling units. The focus should be on complete coverage of these units (within the selected higher stage units when the design has multiple stages), rather than on collecting a lot of substantive information. These two objectives – completeness of coverage versus collection of substantive information - are often conflicting.

The information available in the PSF (and in intermediate-level sampling frames if any) and in the lists of ultimate units should be exploited to the maximum in determining the sample structure and design parameters – such as the system of stratification, allocation of the sample among the strata or domains, the numbers of units to be selected and the sampling rates at the various stages. The present illustration is a rather poor example in this respect. Except for varying the sampling rates $f_1$ and $f_2 \propto (1/f_1)$ at the two sampling stages for one of the three strata, little use has been made in the design of the rich information collected in the process of constructing the sampling frame.

More attention should be paid to varying the design and procedures according to particular characteristics of the units to be enumerated. Procedures which may be suitable, for instance, for listing economic establishments may perform poorly when applied to enumerate locations of work or congregation of children.

### 4.4.3 Establishment-based survey in metropolitan areas, Bangladesh (2003a)

The next example is from a similar survey conducted in Bangladesh. As noted earlier, this was the third component of the National Child Labour Survey, 2002-03. The establishment-based survey aimed to cover child labour in all sectors, but with two restrictions in its scope: it was designed to cover only a segment of the total child labour force, namely children aged 5-17 employed for wage or salary in various establishments; and geographic coverage of the survey was confined to the six metropolitan areas of the country. The survey covered both formal and informal sector establishments, particularly small and unincorporated establishments of different types; large-scale establishments were outside the purview of the survey, as presumably they employed little or no child labour.

The detailed survey was fairly large in size, involving interviews with around 1,500 establishments. Unusually for a survey of this scope and size, the survey was not based on a probability sample. This arose from the haphazard nature and incompleteness of the list frame used for selecting the sample. As noted in the survey report, the survey was “conducted on the basis of non-probability sampling as probability sampling was not feasible due to absence of basic information which could serve as a frame for
the survey, such as an up-to-date and exhaustive list of establishments in the six metropolitan areas.”

As noted in Section 2.12.3, there can be two types of establishment-based survey. The main type are establishment surveys based on a sample selected directly from a frame of establishments likely to be employing children, and are aimed at producing estimates for the population of such establishments.

The second type are surveys based on samples of establishments selected on the basis of their link with working children identified in the household-based child labour survey. Such surveys aim at providing additional variables on children with whom the selected establishments are linked, the objective being to enrich analysis of the population of children. A sample of establishments identified through links with working children from a household-based child labour survey does not necessarily yield a representative sample of the total population of establishments employing children. In order to represent that population more adequately, the sample identified through a household-based survey needs to be supplemented with establishments selected directly from a frame of establishments.

In the present illustration, an attempt was made to obtain a list of establishments/employers from children reported to be in paid employment in the household-based survey interviews conducted during the same period. In any case, the sample derived from the household-based child labour survey turned out to be very small and inadequate. Merely 102 establishments were listed on this basis, and only 43 of them located, identified and successfully interviewed.

The bulk of the sample of 1,500 establishments therefore came directly from a frame of establishments. The survey report describes the procedure as follows.

“A supplementary list had to be prepared or collected for each of the cities … the field staff of the concerned Regional Statistical Offices (RSOs) were asked to collect the list of establishments (names and addresses, number of workers/employees etc.) from the local Chamber of Commerce and Industries or the local administrative authority. The RSOs could not collect the list of establishments from any of these cities due to non-availability of such lists with the local Chamber of Commerce and Industries or the local government authority. As a result, the concerned field offices were instructed to prepare a list of around 3,000 to 5,000 establishments [per metropolitan city] using the prescribed listing form supplied from Dhaka head office. An establishment list for each city was prepared based on concentration by type of establishment, although there were many more such establishments lying scattered in these cities [emphasis added] … The basic information that was collected for listing included: names and addresses of establishments or employers, total number of workers/employees, number of child workers aged 5-17 years, and type of establishment”.
On this basis, establishments with at least one paid child worker could be separated out. The overall numbers in the frame and the sample are summarised below.

<table>
<thead>
<tr>
<th></th>
<th>Household list</th>
<th>Establishment list</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of establishments in the frame</td>
<td>102</td>
<td>25,225</td>
<td>25,327</td>
</tr>
<tr>
<td>Number in the interviewed sample</td>
<td>43</td>
<td>1,461</td>
<td>1,504</td>
</tr>
</tbody>
</table>

It appears that the instruction to field offices was essentially to prepare lists of establishments “as they could”, with the required size of the list for each city specified only within a broad range (3,000-5,000). Hence the lists comprising the frame were selective, and incomplete to unknown and varying degrees. Their distribution neither by geographic region nor by sector of economic activity could be considered representative of the population of establishments or child employees in the metropolitan areas. Such a frame of course does not provide a basis for drawing a probability sample.

Frequently, the quality of list frames can be improved by making the listing operation more thorough and intensive – which may be achievable through reducing the size of the operation, such as by using a multi-stage design and restricting the information collected during listing to that most essential.

### 4.5 Types of linkage between sampling and analysis units

In Section 4.3, we considered the situation when, by and large, each analysis unit is associated uniquely with only one sampling unit. Departures from the above type of linkage exist more or less as exceptions arising from frame imperfections, though of course they have to be dealt with using appropriate procedures. The sampling situations we encounter are, however, often more complex than having a one-to-one relationship between sampling and analysis units.

In this section, we consider the types of relationship (linkage) between sampling and analysis units in a broader framework. The problem of imperfect frames of the type described above apply to any type of linkage. Beyond that, however, sampling situations can differ in a more fundamental respect as follows. The distinction between sampling units and analysis units is a basic distinction, especially in the complex sampling situations of the type encountered in surveys of an elusive population of labouring children. It is useful, therefore, to examine the issue further. Disregarding frame imperfections and similar problems for the present, we may identify five types of link between (ultimate) sampling units and analysis units:

1. **One-to-one** (one sampling unit to one analysis unit)
2. **One-to-many** (one sampling unit to more than one analysis unit)
3. **Many-to-one** (more than one sampling unit to one analysis unit)
4. **Many-to-many** (more than one sampling unit to more than one analysis unit)
5. Analysis units linked only to other analysis units, but not necessarily to any sampling unit directly.
4.5.1 Direct sampling

Conventional sampling designs, by assuming a unique link of each population element to only one sampling unit, include types (1) and (2). For example, we may select individuals from a list of individuals (type 1); or we may select households and interview all members of each selected household (type 2). See Figure 4.5.

Since conventional sampling assumes direct links between sampling units in the frame and analysis units for which a probability sample is required, the selection probabilities applied to the sampling units directly apply also to the associated analysis units. The inverse of the known selection probability of a selected sampling unit provides the weight for the associated analysis unit(s) for estimation from the sample.

Numerous specific approaches and techniques have been developed for situations and applications where direct sampling is possible. ‘Direct sampling’ in itself is not a sampling method, but merely a broad term referring to a type of sampling situation.

In practice, the association between sampling and analysis units may not be complete and unambiguous in all cases as a result of imperfections, as described in the preceding sections, in the frame or in sample selection and implementation. Nevertheless, if the imperfections are relatively infrequent and not overwhelming, the procedure can still be regarded as ‘direct sampling’.

All probability sampling is not direct sampling in the above sense. In certain situations, direct links between sampling units and analysis units are lacking. By exclusion, sampling involving other types of linkage between sampling and analysis units may be referred to as indirect sampling. A number of approaches and techniques have been developed for specific situations and applications where the sampling has to be indirect. Some of these procedures have great potential for practical application to surveys of elusive populations, and are discussed in considerable detail in this book. A wide range of examples have also been given.

It is useful to distinguish between two types of indirect sampling situation: sampling involving multiplicity, and sampling through tracing links between analysis units.
4.5.2 Sampling with multiplicity

Situations where links between sampling and analysis units exist, but are not direct in the sense defined above, are best described under the broad rubric of sampling with multiplicity.

In sampling with multiplicity, an analysis unit can be associated with more than one sampling unit, and can therefore appear in the sample through the selection of any of the associated sampling units. By the same token, a sampling unit may be linked with more than one analysis unit. Rules (3) and (4) apply to designs with multiplicity. The number of sampling units associated with an analysis unit is the multiplicity of the analysis unit. With such linkages, selection probabilities are directly available only for the sampling units, but not necessarily for the associated analysis units in the final sample. Multiplicity sampling, sampling from multiple frames and adaptive cluster sampling are examples of sampling with multiplicity. Special estimation procedures are required to determine the probability of selection of an analysis unit, given the selection probabilities of the (possibly multiple) sampling units with which the analysis unit is associated.

Figure 4.6A depicts two basic forms of sampling involving multiplicity. The diagram on the left refers to multiple representation (duplications) of analysis units in the frame. It has been already noted in the context of the discussion of duplications as a frame problem in Section 4.3.3D. In addition, there are situations when such multiple representation is a characteristic rather than an exceptional feature of the sampling situation. A typical example is the selection of units from multiple overlapping lists, as discussed in Chapter 8. Another possibility is to select persons from multiple locations – for instance the selection of child domestic workers through their household of residence as well as through households where they are employed as domestic workers. Or individuals may be selected through their multiple visits to a facility or location. Such scenarios are discussed in Chapter 10.

The second one is the situation when analysis units are linked to sampling units in the frame not only directly, but also through links with each other (diagram on the right in Figure 4.6A). The direct links alone do not necessarily involve multiple linkages, though they might. This is the classical ‘multiplicity sampling’ (which is often also referred to as ‘network sampling’) and is the topic of discussion in Chapter 7. This situation also characterises adaptive cluster sampling (Chapter 9).

---

13 We have used the term “multiplicity sampling” to refer to the specific sampling procedures of the type developed originally by Sirken (1970), which have been elaborated in Chapter 7 below. The term “sampling with multiplicity” has been used here to refer to the more general sampling procedure for situations when individual analysis units are linked to multiple sampling units. Multiplicity sampling, sampling from multiple frames and adaptive cluster sampling are specific applications of this general procedure.
Even if explicitly there is one-to-one correspondence between sampling and analysis units, the analysis units are also associated with each other through specified shared characteristics; selecting an analysis unit brings into the sample all other such units associated with it.

In typical applications of sampling involving multiplicity, analysis units have links to each other to form clusters, such that if any of the units in a cluster is selected into the sample, all other analysis units in that cluster are automatically included in the sample. This means that analysis units forming a cluster share all the links any of them has with sampling units; the aggregate of all the links available to a cluster of analysis units is available to every analysis unit in the cluster. For example, we may have a design in which, for any individual selected into the sample, the individual’s siblings are also taken into the sample automatically. Hence the siblings form a cluster, and an individual is taken into the sample because he/she has been selected directly, or because any of his/her siblings have been selected.

This is depicted in Figure 4.6B. Analysis unit B1 is linked to sampling unit A1, and similarly B2 to A2. If the two analysis units are also linked to each other to form a cluster, then the situation is equivalent to each of those being linked to both the sampling units A1 and A2 (the middle diagram in the figure). Or we can treat (B1+B2) as a single analysis unit, with multiple links to sampling units A1 and A2 (the diagram on the right in the figure).
4.5 Types of linkage between sampling and analysis units

With this scheme, it becomes possible to have non-zero selection probability even for an analysis unit with no link to a sampling unit, provided it is in a cluster with analysis unit(s) linked with some sampling unit. However, the characteristic feature of typical applications of sampling with multiplicity is that most of the analysis units have a direct link with some sampling unit, rather than only indirectly through their links with other analysis units. The primary difference from conventional sampling is the augmentation of available links through clustering of the analysis units.

More generally, sampling with multiplicity may involve links of both types (3) and (4) simultaneously – many-to-one and many-to-many between sampling and analysis units. An analysis unit maybe linked to multiple sampling units, and at the same time a sampling unit may be linked to more than one analysis unit. This is depicted in Figure 4.7.

In the discussion of sampling procedures in the following chapters, we will frequently encounter situations of sampling with multiplicity. For instance, we may select the sample from two frames which overlap in part so that some units appear in both the frames. A unit appearing in the two frames can be selected from either of them, or indeed from both. Similarly, suppose a sample of children is selected from several locations; children visiting more than one location have a chance of being selected from any of the locations they visit. For this reason, we will discuss estimation procedures for sampling with multiplicity in a general framework in Section 4.6.

Figure 4.6B. Clusters of analysis units with multiplicity sampling
4.5.3 Link-trace sampling

Rule (5) - analysis units linked only to other analysis units but not necessarily to any sampling unit directly – applies to situations where members of the population are hidden from the sampling process to varying degrees. Indeed, the characteristic feature is that most of the analysis units have no direct link with any sampling unit, but only indirect links through other analysis units. In other words, for parts of the population of interest, there are no known and identifiable links between the available sampling units and the required analysis units. Such analysis units can be identified for the sample, if at all, only through their links with other analysis units already included in the sample - where such links between analysis unit not only exist but are also reported by the already included units. Often such indirect links are tenuous in the sense that, in the context of the practical sampling situation, their prevalence is not predetermined but is realised (even created) only through the application of appropriate data collection procedures.

Situations where no tangible or identifiable links exist between sampling units and analysis units, and the sample relies essentially on links among analysis units themselves, may be referred to as link-trace sampling.

First we may distinguish the situation when no sampling frame whatever is available, from the situation where a usable frame is available for a part of the population (Figure 4.8).
4.5 Types of linkage between sampling and analysis units

**Figure 4.8. Lack of complete frame**

No frame
No lists exist or can be created [X] to cover the target population; this may happen, e.g. for altogether hidden populations (B).

Part frame
Usable frame ($A_1$) exists for a part of the target population ($B_1$).
But no lists exist or can be created [X] to cover the rest of the target population ($B_2$). For example, this may be the situation for mobile or hidden populations.

In the former case, we may be able to do no better than aim at a non-probability sample which reflects the target population in at least some respects. In the latter case, link-trace sampling may be able to provide a more representative sample of the target population.

The scheme is depicted in Figure 4.9. Examples of application of such sampling schemes include snowball sampling and respondent-driven sampling, discussed in detail in Chapters 13 and 14.

**Figure 4.9. Link-trace sampling**

A set of sampling units ($A_1$) gives the initial sample ($B_1$) of linked analysis units, which may be representative of the whole or only a part of the population, or may be largely non-representative.

Additional analysis units ($B_2$), associated with directly selected analysis units ($B_1$), are included in the sample. Further analysis units ($B_3$) linked with previously included analysis units ($B_2$) may be included. Similarly, further units linked with ($B_3$) etc. may be included.

The additional units included (in fact most of them) may not be associated with any listings in the frame.
4.6 The multiplicity estimator

4.6.1 Commonly used estimation procedures

The following are the commonly used estimation procedures in situations involving sampling with multiplicity. Different methods of estimation are possible, all with equal statistical validity but normally giving different numerical results (in terms of the estimates and their variances) in a particular application. We will make reference to the following main ones.

A. Horvitz-Thompson (H-T) estimator

The Horvitz-Thompson estimator is based on estimating the overall probability of the unit appearing in the sample. As to the notation used, let subscript $j$ indicate a unit, and subscript $k$ a particular ‘trial’, way, or source of its selection into the sample. Thus $f_{jk}$ denotes the probability of selection (i.e. of appearing in the sample) of unit $j$ from source $k$. The probability that the unit $j$ is not selected from source $k$ is $(1 - f_{jk})$. The probability that it is not selected from any of the $m_j$ sources, assuming independence, is the product of the above values for $k = 1$ to $m_j$, that is, \( \prod_{k=1}^{m_j} (1 - f_{jk}) \).

The probability that it appears in the sample is the complement of the above:

\[
\pi_j = 1 - \prod_{k=1}^{m_j} (1 - f_{jk}),
\]  

(4.1)

with weight $w_j = 1/\pi_j$ if unit $j$ is selected, and $w_j = 0$ if the unit is not selected into the sample.

The inverse of the selection probability so computed is used as the unit’s weight in estimation if the unit appears in the combined sample putting together the results from all the sources. If the unit has appeared in the sample, it does not matter (i) from which source it has been selected, or (ii) whether it has been selected more than once from that source, or (iii) whether it has been selected from more than one source. Each distinct unit in the sample is utilised once. The weight is of course zero if the unit has not been selected from any of the sources. The procedure requires knowing the units’ joint selection probabilities from different sources. The estimation procedure, especially variance estimation, is more complex than the estimation procedures outlined below.

B. Hansen-Hurwitz (H-H) estimator

The Hansen-Hurwitz estimator uses the expected number of times a unit is selected into the sample, taking into account its selection (and possibly multiple selections) from any of the sources. In this procedure, $f_{jk}$ stands for the expected number of selections of unit $j$ from source $k$. This number is essentially the same as the unit’s selection probability when that probability is small. However, when repeated selections of a unit are permitted, the expected number of selections ($f_{jk}$ as defined above, and of course also $p_j$ defined below) can exceed 1.0. The inverse of this expected number of selections is used as the weight applied to every appearance of the unit in the sample. Assuming independence between sources (and between multiple selections), the expected number of times a unit appears in the sample is:

\[
p_j = \sum_{k=1}^{m_j} f_{jk},
\]  

(4.2)
with \( w_j = (1/p_j) \) being the weight applied to every appearance of the unit in the combined sample putting together the results from all the sources and selections. If the unit has appeared in the sample, it does not matter from which source(s) it has been selected, but it matters how many times it appears in the sample. Note that each appearance of an analysis unit is uniquely related to one selection of an associated sampling unit.

Estimator \((4.2)\) is simpler than \((4.1)\) above. It is more practical and useful for our purpose since the probabilities and the corresponding weights are more easily calculated, and the variance estimation formulae are generally also simpler - these can in fact be quite straightforward, even for complex stratified multi-stage designs.

C. Multiplicity estimator

Sampling with multiplicity arises when an analysis unit (i.e. a unit in the target population of interest) can be selected through association with more than one sampling units. The unit may be associated with sampling units in more than one frame, and/or with multiple sampling units in any particular frame. There are several forms of the multiplicity estimator for such cases. A most practical and general one for our present purpose is of the following form.

Consider the selection of sampling units from a set of overlapping frames. An appearance of an analysis unit in the sample is associated with the selection of a particular sampling unit from a particular frame. Thus concerning units in the sample, we can take \( k \) to refer to a specific selection of a given analysis unit \( j \), i.e. the selection through association with a particular listing in any of the sampling frames. Each appearance \((k)\) of analysis unit \((j)\) in the sample is weighted inversely proportional to the probability of selection \((f_{jk})\) of the particular listing in a frame of the particular sampling unit from which the analysis unit has come into the sample, taking into account the number of possible sources \((m_j)\) from which that analysis unit could have been selected. Quantity \((m_j)\) is the sum over sampling units linked to the analysis unit in question, of the total number of times each of those sampling units is included in any of the frames.

**Example**

Suppose that there are two overlapping sampling frames from which the sample is selected. Consider that an analysis unit \( j \) is linked to 2 sampling units. The first linked sampling unit is present once in each of the two frames, giving it a total of 2 listings. The second linked sampling unit appears twice in the first frame and once in the second frame, giving it a total of 3 listings. Multiplicity \((m_j)\) of the analysis unit being considered is \( 2 + 3 = 5 \). In a probability sample, each of these listing has a probability of selection associated with it. Let these probabilities for the 5 listings be \( f_{j1} \) to \( f_{j5} \). Suppose that two of these listings happen to be selected into the sample, say listings 1 and 4, with probabilities be \( f_{j1} \) and \( f_{j4} \), respectively. Each of these selections results in an appearance of the linked analysis unit in the sample, with probability the same as that of the listing from which it comes, namely \( f_{j1} \) for the analysis unit’s first selection, and \( f_{j4} \) of its second selection.
The multiplicity probabilities and weights are:

\[ p_{jk} = m_j f_{jk}, \]

with \[ w_{jk} = \frac{1}{p_{jk}} = \frac{1}{m_j} \left( \frac{1}{f_{jk}} \right) \tag{4.3} \]

if the unit \( j \) is selected from listing \( k \), and \( w_{jk} = 0 \) if the unit is not selected from listing \( k \).

Note that Equation (4.2) replaces \( f_{jk} \) in (4.3) by its average value over the \( m_j \) listings representing unit \( j \) in the frame

\[ \bar{f}_j = \frac{1}{m_j} \sum_{k=1}^{m_j} f_{jk}. \]

The advantage of approach (4.3) is that the probabilities and corresponding weights applied to sample data are determined separately for each selection (or more precisely, for each appearance of the analysis unit into the sample), and do not depend on any other selections of the unit. In fact it is not necessary to know whether different selections refer to the same analysis unit or to different analysis units. That is, no matching of different appearances from different sources or selections of the same analysis unit is required, so long as the multiplicity \( m_j \) is known.

The procedure of sampling with multiplicity and the associated multiplicity estimator (4.3) provide a common link between a wide variety of sampling schemes, some of which have been described in detail in this book. These include for instance multiplicity (often also called ‘network’) sampling, multi-frame sampling, and adaptive cluster sampling – techniques widely used to improve the size and spread of samples of rare populations, and discussed in Chapters 7 to 9 in turn. Another area of widespread application concerns time-location sampling of mobile populations, discussed in Chapter 10.

D. Weight-share estimator

This commonly referred to estimator is the same as the multiplicity estimator (4.3), simply expressed at the level of the analysis unit by aggregating over different appearances \( (k) \) of that unit \( (j) \) in the sample. This is possible because the multiplicity \( (m_j) \) is common to different appearances of the same unit.

\[ w_j = \sum_k w_{jk} = \frac{1}{m_j} \left( \sum_k \frac{1}{f_{jk}} \right) = \left( \frac{m_j'}{m_j} \right), \tag{4.4} \]

where \( m_j' \) is the appropriately weighted number of times analysis unit \( j \) has been actually selected into the sample.

Note that computation of the estimation weights \( w_j \) in the weight-share form (4.4) is somewhat more restrictive and demanding in terms of the information required, than the weights in the multiplicity estimator (4.3). Form (4.4) requires that in a real survey:

- different selections \( (k) \) of a unit are recognised as referring to the same unit,
- that different enumerations of a unit in the sample yield the same value for its parameter \( m_j \).
4.6 The multiplicity estimator

- and that if the substantive variables to be estimated are being measured repeatedly at each appearance \( k \) of the unit \( j \) in the sample, then those measurements give identical values.

For instance, the contribution of a unit in the sample to the estimate of an aggregate with weights (4.3) is of the form \( \sum_k w_{jk} y_{jk} \), where the sum is over all appearances of the unit in the sample. That contribution using weights in the form (4.4) is \( w_j y_j \) where \( y_j \) stands for a single value replacing the different measurements \( y_{jk} \).

4.6.2 Technical details of the multiplicity estimator

The commonly used multiplicity estimator, of the form (4.3), is a Hanson-Hurwitz type of estimator. The estimator may be expressed as a summation over the sampling units selected into the sample, or as a summation over the analysis units associated with the selected sampling units.

A. The estimator in the form of summation over sampling units

Conceptually, each analysis unit may be seen as divided into \( m_j \) parts, where \( m_j \) is its multiplicity i.e. is the number of sampling units to which it is linked. One part is then associated with each sampling unit linked to that analysis unit. The sample weight of each part is \( \frac{1}{m_j} \) times the sample weight of the associated sampling unit. This reduces the sampling-to-analysis unit relationship to the one-to-many or one-to-one relationships of the conventional sampling, thus simplifying the parameter and variance estimation formulae.

In the following description, we will use subscript \( i \) for a sampling unit (e.g. a household), and subscript \( j \) for an analysis unit (such as a child) in the set of such units linked to it.

Let \( f_i \) be the selection probability of sampling unit \( i \) ; \( b_i \) the number of analysis units linked to it under the multiplicity rules used; \( m_j \) the multiplicity of analysis unit \( j \); and \( y_i = \sum_j \left( y_{ij} / m_j \right) \), the sum taken over the \( b_i \) analysis units linked to sampling unit \( i \). With a sample \( S^* \) of the sampling units, total \( Y \) is estimated as:

\[
\hat{Y} = \sum_{i \in S^*} \left( \frac{y_i}{f_i} \right), \quad \text{where} \quad y_i = \sum_{j \in b_i} \left( \frac{y_{ij}}{m_j} \right). \tag{4.5}
\]

Note that for all analysis units, we must have \( m_j > 0 \). A unit with zero multiplicity means that it is not linked with any sampling unit, and hence can never be selected. It remains unrepresented in the population.

Estimator (4.5) requires information on the multiplicity \( m_j \) of every analysis unit which is in the sample by virtue of being linked with a selected sampling unit. Information on selection probabilities \( f_i \) is required only for the sampling units which have been selected into the sample.

General form of the multiplicity estimator

Recalling that we use subscript \( i \) for a sampling (i.e. reporting) unit, and subscript \( j \) for the set of analysis (reported) units, let \( \delta_{ij} = 1 \) if sampling unit \( i \) is eligible to report on analysis unit \( j \), \( \delta_{ij} = 0 \) otherwise.
By definition, the multiplicity of analysis unit $j$ is the number of sampling units reporting on it, that is, $m_{ij} = \sum_i \delta_{ij}$, where the sum is over all sampling units $i$ in the population.

The general form of Equation (4.5) is:

$$\hat{Y} = \sum_{i \in S} \left( \frac{y_i}{f_i} \right) \text{ with } y_i = \sum_{j \in b_i} \left( \frac{y_{ij}}{m_{ij}} \right).$$

(4.6)

This multiplicity estimator is statistically unbiased (i.e. unbiased ignoring the effect of non-sampling errors) when, for any analysis unit $j$, the multiplicity weights $(1/m_{ij})$ satisfy the condition:

$$\sum_i \delta_{ij} \left(1/m_{ij}\right) = 1.$$

(4.7)

Several kinds of counting rule weights have been proposed in the literature (e.g. Sirken and Royston, 1976; Sirken, 1972). The standard ‘multiplicity estimator’ takes $m_{ij} = m_j$, depending only on the multiplicity of the analysis unit $j$ but not on the particular sampling units $i$ involved. This choice is simple and satisfies condition (4.7):

$$\sum_i \delta_{ij} \left(1/m_{ij}\right) = \left(1/m_j\right) \sum_i \delta_{ij} = 1,$$

since $m_j = \sum_i \delta_{ij}$ by definition.

The above model, $m_{ij} = m_j$, applies for instance to the situation when a person can be reported upon by $m_j$ households, including his/her own household. As an alternative, it is also possible to base the choice on numbers of reporting persons rather than on numbers of reporting households. Suppose person $j$ is eligible to be reported by $m_{ij}$ members of his/her own household, and a total of $m_j$ persons in all the different households in the population. It can be seen that condition (4.7) is satisfied in this case with the following choice of weights:

$$m_{ij} = \left(m_j/m_{1j}\right).$$

B. The multiplicity estimator in the form of summation over analysis units

It is useful to examine the multiplicity estimator (Equation 4.5) in further detail. This is a general formula applicable to the situation where the sampling units ($i$) and the analysis units ($j$) have many-to-many linkage. This situation is depicted in Figure 4.10.

The survey population can be seen as consisting of units of either of two types: sampling units (A), or analysis units (B). A single sampling unit may be linked to one or several analysis units. Similarly, a single analysis unit may be linked to one or several sampling units. A sample $S^A$ of sampling units is selected and all (say $b_{ij}$) analysis units associated with the selected sampling units are taken into the resulting sample $S^B$ of analysis units. In Figure 4.10, three units have been selected into $S^A$ ($i=1-3$) with probabilities $f_i$ and among them, they have four associated analysis units $S^B$ ($j=1-4$). This resulting sample is defined by all the links emanating from the selected sampling units to the population of analysis units. These links have been shown as solid lines in the diagram.

---

14 If an analysis unit is not linked to any sampling unit, the analysis unit can never be selected and is therefore not covered in the population surveyed, as noted earlier.
Now examine the links emanating from the analysis units in sample $S^B$ back to the sampling units in the population. Let $m_j$ be the number of links from analysis unit $i$, one link to each of $m_j$ sampling units to which it is linked. As noted above, the multiple estimator (4.5) may be viewed in the following terms. A measurement $y_j$ on analysis unit $j$ is seen as ascribed in equal part ($= y_j/m_j$) to the $m_j$ sampling units with which it is linked. A sampling unit $i$ which is linked to $b_i$ analysis units receives a total contribution, as already noted in Equation (4.5), of:

$$y_i = \sum_{j \in b_i} \left( \frac{y_j}{m_j} \right).$$

(4.8)

An estimate of total $Y$ from sample $S^A$ of these units is as given in (4.5), rewritten as:

$$\hat{Y} = \sum_{i \in S^A} \sum_{j \in b_i} \left( \frac{1}{f_i} \cdot \frac{1}{m_j} \right) y_j.$$

(4.9)
Note that the summation $\sum_{i \in S^A} \sum_{j \in S^B} m_{ij}$ is over all the links between samples $S^A$ and $S^B$, shown as solid lines in Figure 4.10. It can be seen that exactly the same set of links are covered by the summation $\sum_{j \in S^B} \sum_{i \in S^A} m_{ij}$ where $m_{ij}$ is the number of links between selected sampling unit $i$ and the associated analysis unit $j$. (For generality, we have taken this number as $m_{ij}$; in our present example, $m_{ij} = 1$.) Consequently, exactly the same estimator as (4.9) can be expressed in terms of summation over analysis units $j$ in sample $S^B$ as

$$\hat{Y} = \sum_{j \in S^B} \left( \sum_{i \in S^A} \left( m_{ij} / f_i \right) \right) \frac{y_j}{m_j} = \sum_{j \in S^B} \left( \frac{m_j'}{m_j} \right) y_j, \text{ say } . \quad (4.10)$$

Hence the estimate is a weighted sum of $y_j$ values over sample $S^B$ of analysis units, with the unit weights $w_j$ given by the ratio of the following two factors:

<table>
<thead>
<tr>
<th>$m_j'$ = $\sum_{i \in S^A} \left( m_{ij} / f_i \right)$</th>
<th>weighted sum of the links of analysis unit $j$ to sampling units in the sample, with each link weighted by $w_j' = (1/f_i)$, the inverse of the selection probability of the concerned sampling unit $i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_j$ = total number of links of analysis unit $j$ to all the sampling units in the population</td>
<td></td>
</tr>
</tbody>
</table>

It is instructive to note that the double summation over $j$ and $i$ in the above is the same as summation over different selections of the analysis units into sample $S^B$:

$$\hat{Y} = \sum_{j \in S^B} \left( \frac{(y_j / f_j)}{m_j} \right) \quad (4.13)$$

where index $j$ now stands for a particular selection of analysis unit $j$ in $S^B$ resulting from a particular selection of sampling unit $i$ in $S^A$. There is one-to-one correspondence between those selections, despite the presence of multiplicities in the relationship of analysis units to sampling units.

The equivalent estimators (4.9) and (4.13) are different in form because they differ in the order of summation over $i$ and $j$. In (4.9), quantities are constructed (from associated analysis units) and aggregated over sampling units in the directly selected sample $S^A$; in (4.13) quantities are constructed (from associated sampling units) and aggregated over appearance of analysis units in the resulting sample $S^B$.

### 4.6.3 Note on the weight-share method of estimation

As noted in Section 4.6.1D, this commonly referred to estimator is the same as the multiplicity estimator, simply expressed at the level of the analysis unit by aggregating over different appearances of that unit ($j$) in the sample. Equation (4.10) is in fact the multiplicity estimator, expressed in the weight-share form.
It is interesting to note the development and generalisation of the weight-share method of estimation.

In the weight-share method, the survey population is seen as consisting of units of either of two types: sampling units (A), and analysis units (B), with many-to-many relationship between units in the two sets. A sample $S^A$ of sampling units is selected and all analysis units linked to the selected sampling units are taken into the resulting sample $S^B$ of analysis units, using the relationships that exist between the two populations. In the generalised form (Deville and Lavallée, 2006), $S^B$ refers to clusters of units, the clusters defined such that if any one of the individual units in a cluster is selected into the sample, then all other units in the cluster are automatically included in the sample.

A major development of the weight-share method has been its application for computing sampling weights of households and persons in household panel surveys. In such surveys, a subset of individuals (‘longitudinal persons’) remain in the survey from one survey wave to another. Obviously, there is a one-to-one relationship between units which refer to the same person in different waves of the survey. However, the households in which these longitudinal persons reside may contain new members not initially selected into the survey. Procedures based on the weight-share method have been applied to household panel surveys in order to obtain cross-sectional weights applicable, for each wave of the survey, to the household and to all its members, including persons not in the original sample. Essentially, these weights are obtained by pooling and sharing among all current members the weights of individuals who came into the sample through their direct selection in the original sample.

Important contributions by Ernst and colleagues (in particular, Ernst, 1989) extended the method to deal with clusters as analysis units, specifically households as clusters of persons in longitudinal surveys. These clusters are such that the selection of any one of the units (e.g. of an individual) into the sample automatically involves the inclusion of all other units comprising the cluster (e.g. the household) of the selected unit.

Examples of panel surveys to which the weight-share method of estimation has been applied include Survey of Income and Program Participation (SIPP) in the United States (Ernst, Hubble and Judkins, 1984), Survey of Labour and Income Dynamics (SLID) in Canada (Lavallée, 1995), European Community Household Panel (ECHP) in 12 EU countries (Verma, 1995; Verma and Clémenceau, 1996), and EU Statistics on Income and Living Conditions (EU-SILC) survey in over 30 EU and associated countries (Verma, Betti and Ghellini, 2007).

Further generalisations of the weight-share method of estimation have been presented in detail by Deville and Lavallée (2006) and Lavallée (2007), who term it the ‘generalised weight-share method’ (GWSM). This procedure is described briefly below.

In the GWSM form, index $j$ in $S^B$ now refers to clusters of units (identified as, say, $jk$, meaning unit $k$ in cluster $j$), rather than to individual units. The clusters have to be such that if any one of the individual units in a cluster is selected into the sample, then all other units in the cluster are automatically included in the sample. We may select, for example, a sample of individual persons, identify households of the selected persons, and include in the final sample all members of the households so identified. Hence if any member of the household is originally selected, all household members
automatically come into the final sample. The same applies if we were to select a sample of children working in establishments of certain types (perhaps from some available lists), identify their establishments, and take into the sample all children working in those establishments even if they were not selected in the original sample.

Let \( l_{ik} = 1 \) if individual unit \( k \) in cluster \( j \) is linked to sampling unit \( i \), and \( l_{ik} = 0 \) if there is no such link. As before, the existence of a link means that if sampling unit \( i \) is selected in sample \( S^A \), then individual \((jk)\) is taken into sample \( S^B \).

Given the nature of the clusters being considered, the above selection means that all other units in the cluster of the selected unit are also included in the sample. An important feature of this scheme is that units can be included in the sample even if they have no direct representation in the sampling frame, so long as at least one unit in ‘their’ cluster has such a link and hence a non-zero chance of being selected. With this scheme, a sample \( S^A \) of sampling units \((i)\) gives a sample \( S^B \) of clusters \((j)\), and also of all units \((jk)\) in those clusters. Let

\[
   m_{ij} = \sum_k l_{i,jk}
\]

be the total number of links between sampling unit \((i)\) and all the units \(k\) in cluster \(j\). Define

\[
   m_j = \sum_i m_{i,j} \mid i \in U
\]

where the values are summed over all \(i\) (sampling units) in the population, and

\[
   m'_j = \sum_i \left( m_{i,j}/f_i \right) \mid i \in S^A
\]

where the values are summed over units \(i\) selected into sample \(S^A\) (with probabilities \(f_i\)). In estimating from sample \(S^B\) of analysis units, the weights are given by

\[
   w_{jk} = w_j = \left( m'_j/m_j \right)
\]

This is the same as Equation (4.4), the latter being a particular and somewhat restrictive form of the multiplicity estimator (4.3).

4.6.4 Note concerning ‘indirect sampling’

Just as for ‘direct sampling’, ‘indirect sampling’ in itself is not a sampling method, but merely a general and descriptive term referring to the type of sampling situation addressed by the specific methodologies developed for the purpose.

It should be pointed out that often the term ‘indirect sampling’ has been used, as an alternative term, to refer only to sampling with multiplicity, as defined above, rather than more generally as a complement of ‘direct sampling’ covering both sampling with multiplicity and link-trace sampling.

As to the weight-share method, it refers to an estimation procedure. The method is powerful and applies to a range of designs involving indirect sampling - specifically situations involving multiplicity sampling where links between sampling and analysis
units exist. The method provides a common framework for developing and describing estimation procedures for the sampling methods where it applies.

4.6.5 A formal framework

In this note, we will follow the convention of using the term ‘indirect sampling’ to refer to sampling with multiplicity, i.e. excluding the situation specifically requiring link-trace sampling. A useful framework has been developed for this type of sampling. It will be instructive to conclude by presenting a brief outline of it. The description is taken from Mehran (2012).

A. Direct sampling

In the conventional method of direct sampling, a sample of units is selected from a population according to some known probabilities; observations are made on the sample units and the sample results are extrapolated in order to make statements about the whole population. The full process involves three phases: sampling, observation, and estimation as shown schematically in Figure 4.11 (Mehran, 2011). The sampling phase connects the population to the sample. The population generally has a finite number of elements and is represented in practice by a sampling frame $U^A$. A high-quality sampling frame covers only and all units of the population, without duplication, and with sufficient information to access the units selected into the sample. Every unit of the population must have a chance of being selected into the sample. In probability sampling, each unit, $a$, in the sampling frame has a known, non-zero probability of selection ($\pi_a^A$). After selecting the sample, observations are made on every sample unit and the results are recorded according to a pre-determined procedure, generally based on a survey questionnaire and its accompanying manual, which explains the questions and provides instructions on filling in the questionnaire. Following the observation process, the sample results are used to calculate estimates of the parameters of interest regarding the original population from which the sample was drawn. One key feature of probability sampling is that the sample-to-population extrapolation weights can generally be derived directly from the probabilities of selection. The design weight of each unit is equal to the inverse of its probability of selection ($w_a^A = 1/\pi_a^A$).
CHAPTER 4

4. Sampling from imperfect frames

**Figure 4.11. Direct sampling**

### B. ‘Indirect sampling’

In indirect sampling, a random sample is drawn from a population $U^A$ and the result is used to produce an estimate for a target population $U^B$ taking advantage of the relationship that exists between the two populations (Lavallée, 2002). Indirect sampling is a natural method of sampling when no sampling frame exists for the target population, but one exists for a population that has appropriate link to it. The basic elements of indirect sampling are shown in Figure 4.12.

The central element of indirect sampling is the link function $\theta$ that links the sampling population to the target population. For every unit $a \in U^A$ and unit $b \in U^B$, the link function determines a non-negative value $\theta_{ab} \geq 0$, that may be regarded as the degree of the link between $a$ and $b$. A link exists if and only if $\theta_{ab} > 0$; and $\theta_{ab} = 0$ means there are no links between $a$ and $b$. In practice, the link function is usually defined as a dichotomous variable, taking the value 1 if a link exists and 0 if not. For indirect sampling to be well defined, it is necessary that each unit of the target population $U^B$ be related to at least one element of the sampling population $U^A$. Thus, $\theta_{+b} = \sum_{a \in U^A \cap b} \theta_{ab}$, the sum of all links to the sampling population $U^A$, should be strictly positive; no $\theta_{+b}$ should be zero. The relevant values of $\theta_{ab}$ are generally directly obtainable from the sample. For $\theta_{+b}$, it is often necessary to add one or more especially designed questions in the survey.

---

15 Here we are using the term ‘indirect sampling’ in the following more restricted sense, compared to the use in Section 4.5 and elsewhere in this book: the term is being used as a form of ‘sampling with multiplicity’ (Section 4.5.2), excluding link-trace types of sampling (Section 4.5.3).
Estimation of the variables of interest of the target population is obtained using extrapolation weights, $w_B^b$. These, calculated by the weight-share method, are:

$$w_B^b = \sum_{a \in S^A} \frac{\theta_{ab}}{\theta_{a+b}} w_A^a \quad b \in S^B,$$

where $S^A$ is the sample drawn from population $U^A$, $S^B$ the derived sample from population $U^B$, and $w_A^a$ is the weight of the sample unit $a$ in $S^A$, given by the inverse of the probability of selection, $\pi_A^a > 0$. (Equation (4.18) is implied by (4.4), written using a different notation.)

In finite populations, the link function may be represented by a matrix in which the size of the sampling population is the number of rows, the size of the target population, the number of columns and the cells values are $\theta_{ab}$. Different methods of sampling correspond to different forms of the link matrix. The optimal properties of the weight share method are studied in Deville and Lavallée (2006).
Chapter 5
Sampling establishments employing children

5.1 Introduction

5.1.1 Scope of the chapter

A major alternative to household-based surveys is to sample working children through establishments which employ child labour. In this chapter we discuss special considerations which arise in the design of samples for surveys of establishments. These are surveys of economic units, as distinct from population or household-based surveys. Such surveys are concerned with the study of employment and other characteristics of economic units like agricultural holdings, household enterprises, own-account businesses, or other establishments in different sectors of the economy. Our interest of course is more specific: establishments which employ, or are likely to employ, children.

From the point of view of sampling, and also of survey methodology generally, it is useful to distinguish between large and medium sized establishments on the one hand, and small and informal sector establishments on the other.

We begin this chapter with issues which are common to surveys of all sizes of establishments. These include stratification (Section 5.2), and sample allocation among domains or strata (Section 5.3). The next two sections address sampling of large and medium sized establishments. Usually such units are selected using list frames - lists of individual establishments with information on their location, size and other basic characteristics. These may be pre-existing lists, or lists specially created for the survey, or a combination of the two. They may cover the whole study area, or may be confined to selected areas for practical reasons. In fact, when the study area is geographically limited, sampling from list frames may also be practicable for small establishments. Technical aspects of two steps in the sample selection procedure are described in detail: selection of establishments from lists in a single-stage (Section 5.4); and selection of working children within establishments taken into the sample (Section 5.5). Issues relating to sampling within establishments are also relevant to surveys of small establishments.

While large and medium sized establishments do employ children, it is much more common to find child labourers in small and informal sector establishments. The focus of this chapter in subsequent sections will therefore be on issues concerning sampling of small, and in particular informal sector establishments – in other words, on sample surveys of economic units which, like households, are small-scale, numerous and widely dispersed in the population. Unlike medium-to-large establishments, usually small and informal sector units cannot be sampled on the basis of pre-existing lists; we discuss special considerations which arise in the design of samples for surveys of such establishments given this constraint. Specifically, the units and situations encountered include the following.
(1) Small-scale family farms
(2) Small-scale commercial crop production
(3) Self-employed professional and other small-scale economic units not in the informal sector
(4) Non-agricultural economic units, based within households
(5) Small non-agricultural economic units, including micro-enterprises, with separate fixed premises
(6) Mobile economic entities linked to a household
(7) Mobile workers employed by a small economic unit, which is household-based or has other fixed premises
(8) Mobile economic entities (household enterprises, establishments, individuals) with no fixed premises

The ‘informal sector’, as explained in Section 5.6, is normally taken to cover items (4)-(8) in the list. We take ‘small and informal sector establishments’ to indicate the wider coverage, (1)-(8). The nature of the place of location (premises) of the establishment has an important implication for the sampling procedures to be used. Items (1)-(5) have fixed premises from which they may be selected and where they may be contacted for the survey. Items (6)-(7) are mobile, but they are linked to some fixed location from which they can be selected in principle. Alternatively, it may be more convenient or efficient to sample and enumerate them from their mobile work locations. Such an arrangement has to apply in any case to units under item (8).

In this chapter we address only the situation when units can be sampled from known and fixed locations, as in (1)-(5), and also (6)-(7) if applicable. Sampling mobile units raises special issues which are addressed in Chapters 10 and 11.

A couple of other limitations defining the issues included for discussion in this chapter may be noted. While the small and informal sector economic units may be very unevenly distributed in the population, usually their patterns of distributions are not hidden, and they are usually also sufficiently numerous not to present special problems of sampling ‘rare populations’ considered in Chapters 6 to 9. Furthermore, a majority of informal sector workers in developing countries have little to conceal and therefore can be surveyed using the direct method of enquiry used in many other types of statistical survey; we assume here that in general they do not present special problems of surveying ‘reclusive populations’ of the type discussed in Chapters 13 and 14.

Outline of content of the remaining sections of the chapter is as follows.

Small and informal sector establishments usually form subpopulations which may be quite large but are unevenly distributed in the general population. For achieving an efficient sampling design it is necessary to understand the characteristic features of this type of establishment, and the implications of those features for sample design (Section 5.6).
A survey of small and informal sector establishments may take various forms. Options include an integrated multi-sectoral versus a separate single-sector survey, and a stand-alone survey versus a module attached to another survey (Section 5.7).

A design requirement can be to control the size of the sample by sector of activity. It is desirable to achieve this without making it necessary to sample establishments of different types at different rates in the same sample area. For this we define and make use of the concept of 'strata of concentration' (Section 5.8).

Sampling rates in the strata of concentration are adjusted so as to approximate the required sample allocation by sector. The more precisely we try to control the sample allocation by sector, the greater is the price to be paid in terms of sampling efficiency (Section 5.9).

In sampling small and informal sector establishments from sample areas, certain procedures requiring special care are involved (Section 5.10). This applies in particular to the listing operation. The considerations involved are rather different from those discussed in Section 5.4 for selecting large establishments in a single stage from list frames.

Some examples of country surveys of small and informal sector establishments are provided in Section 5.11. Finally, Section 5.12 notes some aspects of sample implementation for situations and units which require special procedures.

### 5.1.2 Design strategy

As to the design strategy, determination of the design involves a number of aspects such as the following.

A. Concerning sample structure
   - Definition of survey units; choice of the sampling frame.
   - Choice of overall sample size, determined by precision requirements, but also by cost, operational conditions and numerous other practical considerations.
   - Sample allocation, e.g. by geographical and administrative domains, and type of child labour activity.
   - Stratification of the population of establishments.

B. Concerning sample selection and estimation
   - Measures of size of the units (establishments), and unit selection probabilities.
   - Selection of units within strata.
   - Subsampling within units (selection of children within selected establishments).
   - Choice of appropriate estimation procedures, including for variance estimation.

Of course, these aspects are linked – but not deterministically or rigidly. Nevertheless, a good design strategy is to consider each of these aspects separately to the extent possible, so as to allow maximum flexibility in the choice of the various features of the design. A practical procedure is to scale the unit size measures which serve as unit selection probabilities such that they can serve as a basis for flexible approach to the sample design. We will elaborate on this strategy in relation to various aspects of the sample design and selection in the following sections.
C. Concerning sample size

The required sample size \( (n) \) depends on the sample design through efficiency of the latter ('design effect' – see Section 3.6). What matters for the purpose of meeting specified precision requirements is the 'effective sample size' = \( n/(\text{design effect}) \). In practice, sample size \( n \) can influence the sample structure (e.g. the size of clusters in multi-stage sampling), and hence influence the design effect.

The other main effect of \( n \) on the selection procedure is in the determination of the appropriate cut-off boundary between 'take-all' and 'take-some' strata (Section 5.2.3).

Apart from that, however, many features of the design are largely independent of the particular sample size chosen. In practice, to a considerable extent the design work can be carried out without reference to any particular sample size. The determination of the sample size is a complex question, dependent on various practical, cost and statistical considerations, largely independent of many other aspects of the design. For an extended discussion of the issue, see Verma (2008, Section 3.5).

5.1.3 Sampling stages in a sample of children working in establishments

The survey may involve sampling in several, often three, stages: sampling of area units, then of establishments within selected areas, and finally of working children within selected establishments. It is also possible that area units themselves are selected in multiple stages, each stage involving the selection of smaller area units within selected larger area units. In practice, however, most surveys on child labour (and on the labour force generally) use at most a single stage of area sampling, and we can assume this to be the case in the present discussion. In fact, it is quite common in practice to skip one or more of the three stages mentioned above. The various possible schemes for multi-stage sampling are summarised in Table 5.1.

Schemes (3), (4) and (7), involving direct sampling of establishments (or even of children), are the subject of Section 5.4. Schemes (5), (6) and (8), involving sampling within selected area units, are considered in subsequent sections. Scheme (2) can involve purposive selection of a few areas for intensive follow-up, or more appropriately, use the controlled selection (balanced sampling) procedure described in Chapter 12. This can also be the case sometimes with schemes (5), (6) or (8). Scheme (1) is likely to be used mainly for pilot or exploratory studies, or when the target population is small and compact.
5.2 Stratification

5.2.1 Important general considerations

The notion of stratification was briefly introduced in Section 3.2.1. It is useful to highlight some general points.

Stratification means dividing the units in the population into groups and then selecting a sample independently within each group. This permits separate control over design and selection of the sample within each stratum. This means that different parts of the population (strata) can be sampled differently, using different sampling rates and designs. The separation may also be retained at the stage of sample implementation and estimation and analysis, but this is not essential to the idea of stratification. It is common to pool the results from different strata to produce estimates for the whole population, or for major parts or domains of the population each of which is composed of a number of strata.

The advantages of stratification result from the separate control over sample design and selection within each stratum. The main advantages are as following.

(1) Firstly, in so far as the strata represent relatively homogeneous groupings of units, the resulting sample is made more efficient by ensuring that units from each grouping are appropriately represented in a controlled way.

There are many additional practical reasons for stratification.

(2) When data of specified precision are required separately for sub-divisions of the population, it is desirable to treat each subdivision as a ‘population’ in its own right, and select a sample of the required size and design from each independently.

---

### Table 5.1. Various schemes for multi-stage sampling

<table>
<thead>
<tr>
<th>No. of stages</th>
<th>Area units</th>
<th>Establishments</th>
<th>Child workers</th>
<th>Whether sampling involved (1=yes; 0=no)</th>
<th>Sampling scheme</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0 (all)</td>
<td>0 (all)</td>
<td>0 (all)</td>
<td>0</td>
<td>(1)</td>
<td>All working children in all the establishments in the study area (complete census)</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>(2)</td>
<td>All working children in sample of area units</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>(3)</td>
<td>Sample of establishments in the study area; all working children in each establishment selected</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>(4)</td>
<td>Sample of working children in every establishment in the study area</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>(5)</td>
<td>Sample of establishments in selected area units; all working children in each establishment selected</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>(6)</td>
<td>All establishments in selected area units; sample of working children in each selected establishment</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>(7)</td>
<td>Sample of establishments in the study area; sample of working children in each selected establishment</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>(8)</td>
<td>Sample of establishments in selected area units; sample of working children in each selected establishment</td>
</tr>
</tbody>
</table>
Stratification makes this possible.

(3) Sampling requirements and problems - as concerning sample size, design, availability of frame for sample selection, travel conditions, costs etc. - may differ markedly between different parts of the population. Stratification permits flexibility in the choice of the design separately within each part.

(4) A sample clearly controlled and distributed proportionately (or in accordance with some other specified criterion) across different parts of the population has the public relations advantage of appearing more representative and hence more acceptable to the users. In any case, control through stratification reduces the danger of getting a poorly distributed sample by chance.

(5) Stratification may also be introduced for administrative convenience; for instance, sample selection and implementation may be entrusted to different field offices on the basis of the strata.

5.2.2 Stratification criteria

A. Primary stratification: major geographical divisions; sectors of activity

Reporting by major geographical divisions is not always a primary objective of statistics, but often it is. Auxiliary information which may be available for improving the design is often based on administrative and geographical divisions. Primary stratification by geographical divisions can facilitate the expansion of the sample to additional areas at a later date.

The primary geographical/administrative strata are often further stratified in more detail using various criteria. These would almost always have to include stratification of establishments by sector or type of activity. Most of the non-household based child labour enquiries are sector-specific. Results are often required by sector. Furthermore, the data collection mode and methodology may also vary from one sector of activity to another.

B. Secondary stratification: geographical location; establishment size

Smaller strata, identifying the units' geographical location more precisely, often deserve attention, especially when the study area and the primary strata are large.

Next, the size of the establishment, measured in terms of turnover, the number of employees, or the number of children employed if available, can be a useful stratification criterion. However, given the primary requirements of stratification by geographic or administrative divisions and by sector/type of activity, the scope for explicit stratification by size becomes limited. Many establishment-based surveys use broad classes of unit size measure as a main criterion for stratification. With unit size so controlled, the sample within each size stratum can then be selected with equal probability, ignoring size variation within strata. This simplifies some sample selection and estimation procedures.

However, even such control by establishment size is unlikely to be available in a child labour survey when the sample size is relatively small, and where we still must control
for many other variables. Usually control by substantive variables, such as sector of activity, needs to be more explicit and precise than control by establishment size. Hence an alternative means of controlling the effect of unit size variations has to be invoked. Selection with probabilities proportional to unit size within explicit strata is often the appropriate solution. This does not preclude implicit stratification by size, for example by ordering units within strata by size and selecting the sample systematically.

### 5.2.3 Take-all strata

This refers to strata composed of units which are so large that all of them are to be included in the sample.

In surveys involving medium and large establishment as the sampling units, the distribution of a variable of interest is very often highly skewed. A limited number of big establishments contribute significantly to the population total. The remaining contribution is shared by more numerous less important units. The standard design strategy consists in defining a self-representing (‘take-all’) stratum gathering all units above a size threshold. Typical sample design involves a census among the biggest units, combined with a sample of smaller units.

For an informative study of the issue concerning complete coverage of large units, see Glasser (1962).

Procedures for defining optimal thresholds between these two components of the sample require micro data (sample or frame data), including information on target variables, which is generally not available in real surveys. We need a much simpler procedure, using only the unit size measures available in the sampling frame or some modified measure constructed from the given size measures.

Given that optimal allocation is obtained essentially by selecting units with probabilities proportional to an appropriately defined size measure, a simple and objective method for identifying the take-all stratum is as follows.

A particularly simple procedure would be to take units with size measures exceeding a limit $L_h$:

\[
X_i \geq \frac{\sum_{j \in h} X_j}{n_h} = L_h
\]  

(5.1)

into the sample with certainty. Here the sum is over units in the whole population of stratum $h$. Parameter $n_h$ is the required sample size for the stratum, so that $L_h$ is the sampling interval which would be applied in order to select the required number of units with systematic PPS sampling.

This procedure has some difficulty in dealing with large units with sizes just below $L_h$. This is because removing the large units exceeding $L_h$ in size reduces the remaining total size measure more than the reduction in the number of units still to be selected to reach the required sample size $n_h$. This requires increasing the sampling rate to be applied to the remaining units, which can make the required selection probabilities of the largest units among them to increase over 1.0. In order to deal with this problem, one has to repeatedly apply Equation (5.1), each time with a reduced value of $L_h$. 

The following, also quite straightforward, procedure is preferable for this reason (Verma, 1991).

Let the units be ordered in decreasing order of size (i.e. from the largest to the smallest), and the size measures cumulated as follows:

$$X_i \geq X_{i-1}; \quad S_i = \sum_{j=1}^{i} X_j.$$ 

Then with \(n\) units to be selected from \(N\) with PPS, the largest unit will be selected with certainty if \(X_i \geq S_N/n\). Beyond that, units up to and including the last unit \(i\) satisfying the following condition are taken into the sample with certainty:

$$X_i \geq \frac{S_N - S_{i-1}}{n - (i - 1)}.$$ \hspace{1cm} (5.2)

The remaining units \((j > i)\) in the ordered list are in principle subject to sampling with probabilities less than 1.0:

$$\text{Probability} \left( \frac{X_{j}}{I} < 1; \quad I = \frac{S_N - S_{j}}{n - i}, \quad j > i \right).$$

The above provides the basis. In practice, units with probabilities just below 1.0 may also be taken into the sample with certainty. Also, some small units with characteristics of special interest may be treated in the same way, even if only on the basis of judgement of the researcher.

The above procedure is stated in terms of a PPS design, i.e. when units are selected with probabilities proportional to size within strata. But it also provides the basis for the “simple random sampling within stratum” design, in so far as the sample allocation of each stratum depends on the sum of size measures of the units in it. The procedure identifies the set of strata with the largest units in which all the units are taken into the sample. The equal probability sampling may be seen as a special application of the PPS design where the same average measure of size is assigned to each unit in a stratum.

### 5.2.4 Limiting the number of strata and optimal stratification

The different combinations of the various dimensions of stratification associated with different levels of detail can be numerous. In practice, however, the number of strata should remain limited. Strata with a small number of units, especially in the presence of high rates of non-response, can cause estimation problems. Too many strata can also increase the burden of managing the survey implementation. Furthermore, the gain in statistical efficiency from stratification declines as the number of strata is increased. Indeed, it is more effective to use a multiplicity of stratification variables, each with a few categories, than to use many fine categories of a single variable.

Even more serious is the problem of strata with very small or even zero sample sizes, which can frequently occur when the total sample size itself is not large.

The issue of optimal stratification and optimal allocation to be described below are linked. We will comment on them further in Section 5.3.4.
5.3 Sample allocation

Allocation of the sample among strata or other study domains is determined by various factors, the main ones being the following:

1. Size of the stratum in terms of its relative contribution to the aggregate value of the variable to be estimated, or of some aspect related to that value. For instance, if the study objective is to estimate the number of working children in the population, then the sample size given to a stratum could be in proportion to the number of such children in the stratum.

2. Relative variability (variance) of the statistic being measured in the stratum. Efficiency of the sample is increased by taking more of the sample from the more variable strata.

3. Relative per unit cost of collecting and analysing the survey data. It is more economical to take larger samples in strata where these costs are lower.

4. Minimum sample size requirements for individual reporting domains, that is, for partitions of the population for which separate results need to be produced with some specified minimum precision. Examples of common reporting domains are individual sectors of economic activity, or classes of establishments employing children. This requirement tends to become more demanding as the number of reporting domain is increased.

5. Total available sample size is also a factor in sample allocation, especially in its interaction with point (4) above. Ensuring the required minimum sample size per domain can become an increasingly dominant requirement as the total sample size available is reduced.

5.3.1 Proportionate allocation

By this we mean allocation primarily in terms of objective (1) above, that is, allocation in proportion to the size of the stratum in terms of its relative contribution to the aggregate value of the variable to be estimated, or of some aspect related to that value. It is useful to distinguish between two sampling schemes: equal probability sampling of elements; and unit selection probabilities proportional to the value of some relevant characteristic.

A. Equal probability sampling of elements

Proportional allocation with the so-called ‘epsem’ - equal probability sampling of elements – is the simplest and the basic allocation scheme. In selecting a sample of size \( n \) from a population of size \( N \) units with uniform probability \( f = n/N \), each stratum \( h \) of size \( N_h \) is expected to receive a sample of size \( n_h = f N_h \), the allocation being:

\[
n_h \propto N_h .
\]  

(5.3)
B. Unit selection probabilities proportional to values of some characteristic

Scheme (5.3) requires information only on the number of units \((N_h)\) in the population of each stratum. But suppose that information is available on values \(Z_{hi}\) of some variable for each unit \(i\) in stratum \(h\) in the population and it is considered desirable (more efficient) to select units with probabilities proportional to those values. The number of units selected from a stratum will be proportional to the sum of those values in the stratum population, rather than simply to the count of units in the stratum:

\[
n_h \propto Z_h = N_h \bar{Z}_h, \text{ with } Z_h = \sum_{i=1}^{N_h} Z_{hi}, \quad \bar{Z}_h = Z_h / N_h.
\]  

(5.4)

The allocation is proportional to stratum total \(Z_h\). It is instructive to view this as allocation proportional to the product of the number of units in the stratum \((N_h)\), and the per unit average value \((\bar{Z}_h)\) of the chosen characteristic in the stratum.

Equation (5.4) represents a very versatile allocation scheme. We may note several possible options in the choice of variable \(Z\).

1. **Basic scheme**

   The basic scheme (5.3) corresponds simply to \(Z_{hi} \equiv 1\) for all units.

2. **Allocation proportional to unit size**

   Variable \(Z_{hi}\) may correspond to actual unit sizes (say \(X_{hi}\)) in terms of the numbers of elements in the unit, for instance the number of working children in establishment \(i\) in stratum \(h\).

   \[
n_h \propto Z_h = X_h = N_h \bar{X}_h.
\]  

(5.5)

3. **Allocation proportional to surrogate of unit size**

   Another common option is for \(Z_{hi}\) to be a surrogate measure of size, for instance the number of employees (of any age) in an establishment, when the actual elements of interest are working children, the numbers of which may be unknown at the time of sample selection.

4. **Allocation proportional to a function of unit size**

   Sometimes it is more efficient or convenient to take \(Z_{hi}\) as some function of the (actual or surrogate) size measure:

   \[
   Z_{hi} = f(X_{hi}), \quad n_h \propto Z_h = \sum_{i=1}^{N_h} f(X_{hi}).
   \]

An example is to select units with probability proportional to square root of unit size, \(Z_{hi} \propto \sqrt{X_{hi}}\). Suppose that this scheme is applied in a two-stage design. At the first stage, establishments are selected with probability proportional to square root of the number of children employed in the establishment, \(\sqrt{X_{hi}}\). At the second stage, within each selected establishment working children are selected with probability inversely proportional to \(\sqrt{X_{hi}}\). The overall selection probability of a working child would be a constant, while the number of working children selected from a sample establishment
would be proportional to \( \sqrt{X_{hi}} \) – bigger establishments giving more working children to the sample, but less than in proportion to the number of children they employ.

**An application of scheme (4): Model-based allocation**

An option is to base optimisation on an assumed (but hopefully realistic) regression model of the relationship between the auxiliary variable(s) \( X \) and the target variable(s) \( Y \). A commonly used model assumes that at the level of individual units, the \( Y_i \) values are proportional (within a random model error term) to \( X_i \), and that the (model) variance in the \( Y_i \) values, conditional on \( X_i \), is proportional to some power (say, \( g = 2 \)) of \( X_i \). It is reported that many populations conform to this variance structure, with values of parameter \( g \) in the range 0.5-1.0, and often in a narrower range 0.7-0.9.

This gives an optimal design as one in which units are selected with probabilities \( P_i \propto X_i^g \) and optimal allocation is given by \( n_h \) equal to the sum of \( P_i \) values of all units in the stratum population, i.e.

\[
n_h = \sum_{i \in h} P_{hi} \propto \sum_{i \in h} X_i^g.
\]

Note that this allocation is not the same as the ‘power allocation’ in (5) below. Here it is proportional to the sum of some power of the individual \( X_i \) values, rather than some power of the sum of those values as in scheme (5) below.

With \( g = 0 \) we have allocation simply proportional to the number of units in the stratum, irrespective of their size. Such allocation is often acceptable for household surveys, but is hardly worth considering in surveys of establishments with great differences in unit sizes. Parameter \( g = 0.5 \) gives unit selection probabilities proportional to square root of the size measures, and allocation proportional to the sum of those square roots.

Value \( g = 1.0 \) is often used, and believed to have been originally selected by the famous Indian statistician Mahalanobis. Parameter \( g = 1.0 \) simply means selection of units with probabilities in proportion to their size measures (the usual \( \pi \)PS sampling), and allocation in proportion to the sum of those size measures in the stratum population. For various practical reasons, this is a convenient and often the appropriate scheme. This is the case even if individually units are not selected with \( \pi \)PS in the stratum but, say, with uniform probabilities (stratified SRS). Nevertheless, the stratum allocation can be proportional to the sum of these size measures (just as in the \( \pi \)PS case), with the uniform selection rate to be applied in a stratum taken as proportional to the average of these size measures within the stratum.

In practice, sufficient information is often not available to estimate the appropriate value of \( g \) very precisely. Fortunately, the efficiency of the resulting allocation tends to be rather insensitive to the precise value used. This is an added reason for proposing the simple variation \( g = 1.0 \).

**(5) Allocation proportional to a function of stratum size**

A different version of scheme (4) above is to apply the functional relationship between the size measure and allocation at the aggregate (stratum) level, rather than at the level of individual units:
5. Sampling establishments employing children

\[ n_h \propto Z_h = f(X_h) = f(\sum_{i=1}^{N_h} X_{hi}). \]  

(5.6)

For instance, strata may be allocated sample sizes in proportion to square root of stratum total size measures \( n_h \propto Z_h = \sqrt{X_h} \).

It is common with such schemes to select units within each stratum at a uniform rate, proportional to per unit average of the transformed total \( f(X_h) \):

\[ f_h = n_h / N_h \propto f(X_h) / N_h. \]

An application of scheme (5): Power allocation

A particular form of scheme (5) is often referred to as ‘power allocation’. The idea is as follows.

When the objective is not only to produce the overall estimates most efficiently, but also to ensure that estimates for individual domains (e.g. regions, sectors) also meet certain minimum precision requirements, an option is to allocate the sample in proportion to some power \( \alpha < 1 \) of a stratum total \( Z_h \):

\[ n_h \propto Z_h^\alpha. \]

Ideally, one would like to have \( Z_h = Y_h \), stratum total of the target variable to be estimated. In any case, it is desirable that \( Z_h \) is correlated with \( Y_h \).

The parameter \( \alpha < 1 \) may have values in a range such as 1/3 to 2/3; a value =1/2 (allocation in proportion to the square root of size) is often used.

A problem with this formulation is that it does not ensure any minimum value for the allocation, which may be important for very small domains (such as important but small sectors). We have used a specification in the following form, which ensures a minimum allocation proportional to \( Z_0 \), even to the smallest domain:

\[ n_h \propto \left( Z_0^2 + Z_h^{2 \alpha} \right)^{1/2} \]

(5.7)
giving \( n_h \propto Z_0 \) for vanishing small domains \( Z_h \to 0 \).

(6) Allocation proportional to stratum total of the target variable

The efficiency of a sampling scheme depends on how closely \( Z_{hi} \) is correlated with the actual variable, say \( Y_{hi} \), to be estimated from the sample. Hence ideally \( Z_{hi} \) should be equal (or proportional) to \( Y_{hi} \), so that allocation (5.4) becomes

\[ n_h \propto Z_h = Y_h = N_h \bar{Y}_h, \text{ with } Y_h = \sum_{i=1}^{N_h} Y_{hi}, \bar{Y}_h = Y_h / N_h. \]  

(5.8)

As \( Y \) is the variable to be measured in the survey, it is generally unknown at the level of individual units at the time of sample selection prior to the survey. However, sometimes its stratum totals, \( Y_h \), may be available, at least in approximate and relative terms. These may then be used to determine the sample allocation according to (5.8). In the absence of information on unit-level values \( Y_{hi} \), units within each stratum may be selected at a uniform rate. This corresponds to taking \( Z_{hi} \propto \bar{Y}_h \), which still satisfies Equation (5.8). The above allocation means that strata with larger than average units are sampled at proportionately higher rates.
5.3 Sample allocation

(7) Allocation based on ratio of stratum totals of target to auxiliary variables

Finally, it is useful to note another possibility. Suppose that some stratum size measures $X_h$ are available. However, the $Y_h$ values corresponding to the variable to be measured in the survey are not available, but some values $(X'_h, Y'_h)$ correlated with corresponding $(X_h, Y_h)$ are available. Values $(X'_h, Y'_h)$ may for example be values of $(X_h, Y_h)$ from a previous survey, from an alternative source, or for a similar population. Then one may use the allocation

$$n_h \propto Z_h = \left(\frac{Y'_h}{X'_h}\right)X_h$$

as an approximation for $Y_h$ in (5.8).

One point common to all the above schemes (1)-(7) may be noted. The point also applies to other allocation schemes to be described below.

Often strata refer to detailed classification of the population, but the $Z_h$ total are available only for groups of strata. For instance these groups $(d)$ may be design or estimation domains, each divided into a number of finer strata $(h)$. In this situation, the sample allocation would have to be determined at the more aggregate domain level, such as $n_d \propto Z_d$.

5.3.2 Optimal allocation

Optimal allocation is traditionally discussed in terms of a simple variance-cost model of the following type:

$$V^2 = V_0^2 + \sum_{h} \frac{V_h^2}{n_h}, \quad J = J_0 + \sum_{h} J_h n_h.$$  

Variance $V^2$ is seen as decomposed into a constant term (independent of sample size), and a sum of stratum $(h)$ components, each inversely proportional to sample size $n_h$ of the stratum. Similarly, total cost $J$ is decomposed into a constant component independent of sample size, and a sum of stratum components, each directly proportional to the stratum sample size. It is assumed that only one type of unit is involved in each stratum. The expression is easily generalised to deal with units of different types in a multi-stage design, but that is not relevant in the present context of a single stage design of establishments.

The objective can be expressed as minimising total cost for given variance, or minimising variance for given cost. Usually it is in the latter form: i.e., optimal allocation is discussed in terms of the distribution of a given sample size among strata such that, for given cost, sampling variance is minimised. The well-known expression for optimal allocation is

$$n_h \propto N_h S_h / \sqrt{J_h}, \quad \text{or} \quad n_h = n \left( \frac{N_h S_h / \sqrt{J_h}}{\sum_{h} N_h S_h / \sqrt{J_h}} \right),$$

i.e., the total available sample size $n$ is allocated among strata:

- in direct proportion to stratum size (number of units in the stratum) $N_h$,
- and to standard deviation (the square root of unit variance), $S_h$, of the target variable $(Y)$ of interest,
but inversely proportional to square root of \( J_h \), the per unit survey cost in the
stratum (other than the constant cost component \( J_0 \)).

It is more instructive to view the above expression in terms of the coefficient of variation
\( C_h = (S_h/\bar{Y}_h) \) of \( Y \) among units in the population, rather than standard deviation \( (S_h) \)
which depends on its scale of measurement. The optimal allocation is

\[
n_h \propto N_h \bar{Y}_h \left( C_h / \sqrt{J_h} \right). \tag{5.9}
\]

Note that some additional constraints on the \( n_h \) values also need to be taken into
account. In order to estimate mean or total value for the stratum, \( n_h \) must be at least
1; to estimate stratum variance, \( n_h \) must at least equal 2; and of course, \( n_h \) has to
be an integer for the purpose of sample selection. These constraints may not be trivial
when numerous small strata are involved.

It is common in practice that information is lacking on \( J_h \) or it is suspected that any
variation in \( J_h \) values across the strata is not large. In such cases, a common practice
is to revert to ‘Neyman allocation’ described in Section 5.3.3 below, which disregards
differences in unit costs among the strata. Furthermore, it may also be the case
concerning within-stratum variability \( C_h \), for instance when information is lacking or
it is suspected that variations in \( C_h \) values across the strata are small. In such cases,
a common practice is to revert to simple ‘proportionate allocation’ in one of the forms
described in Section 5.3.1.

**X-optimal allocation**

In practice, the necessary information for the target variable(s) \( Y \) is normally not
available, and it is necessary to express the above in terms of characteristics of \( X_h \), the
auxiliary variable(s) available in the sampling frame. This is the so-called ‘x-optimal
allocation’

\[
n_h \propto N_h \bar{X}_h \left( C_h^{(X)} / \sqrt{J_h} \right). \tag{5.10}
\]

where the superscript in \( C_h^{(X)} \) indicates that the reference now is to the coefficient of
variation of the surrogate variable \( X \).

As to how efficient this allocation is, compared to the ‘true’ optimal in terms of the
target variable \( Y \), depends on the strength of the correlation between \( X \) and \( Y \). If this
correlation is perfect, then the x-allocation is in fact optimal. It should perform very well
with a correlation of 0.9 or higher, but the performance may significantly deteriorate
with a correlation below 0.8 or so.

**5.3.3 Neyman allocation**

In principle, unit variances can be estimated from external sources, including from past
surveys even if of moderate size. Reliable information on unit costs is, however, harder
to come by. Assuming per unit cost to be constant across the strata gives the Neyman
allocation:

\[
n_h \propto N_h \bar{Y}_h C_h. \tag{5.11}
\]
A practical point is worth noting in this context. In using $C_h$ from external sources, care should be taken to examine them for plausibility and consistency. In particular, where the information is obtained from diverse sources of uneven quality, it should not be taken at its face value, especially if large and implausible differences are found among the figures. It looks reasonable to assume that very large differences should not exist between population variances for similar variable, or among strata by sector of activity or establishment size. Sampling and response errors may make individual estimates of $C_h$ values unreliable. Where such problems are suspected, it is almost always preferable to work with suitably averaged quantities than with the original individual figures.

In any case, even though $C_h$ may be estimated more easily than costs $I_h$, it is quite common in practice that information is lacking on $C_h$ or it is suspected that any variation in $C_h$ values across the strata is not large. In such cases, a common practice is to revert to simple proportionate allocation in one of the forms described in Section 5.3.1.

### 5.3.4 Optimal stratification

The issues of optimal allocation and optimal stratification are linked. Apart from the unit size measures ($X$), most of the stratification variables are categorical (classificatory) rather than continuous. In any case, continuous variables can be grouped to form ordered classes. The following remarks apply to categorical variables or continuous variables so grouped.

(1) In relation to the Neyman allocation, the well-known “cumulative square root” ($\text{cum} \sqrt{f}$) rule of Dalenius and Hodges (1959) is often used. The procedure may be briefly summarised as follows. Suppose that units are arranged in ascending order of their size measures in the form of a histogram of class widths $U_x$ and frequencies $f_x$. The quantities $c_x$ are cumulated to $C_X$ as defined below

$$c_x = \sqrt{(U_x \cdot f_x)}; \quad C_X = \sum_{x=1}^{X} c_x.$$ 

Optimal strata (assuming Neyman allocation and unbiased estimator of a stratum total) are given by equal partitions of the cumulated $C_X$ values. The above has been stated in terms of the auxiliary variable $X$ because it is that rather than $Y$, the target variable to be measured in the survey, that is normally available in practice. The optimisation should ideally be in terms of the latter.

(2) The model-based optimal allocation (Section 5.3.1B, scheme (4)) corresponds to model-based optimal stratification in a similar way. Optimal strata (assuming the model-based allocation and, say, a ratio estimator) are given by equal partitions of the cumulative $X_i^g$ values of units arranged according to increasing $X_i$ values.

The real problem is that in procedure (1) we often have to rely on auxiliary $X$ rather than target $Y$ values, which lowers efficiency when the two variables are not very highly correlated. However, it is sometimes possible to have a good surrogate for $Y$, meaning some correlated but more easily available variable. Similarly in procedure (2), it may not be easy to obtain good estimates of the parameter $g$. Generally the variance tends to be rather insensitive to the $g$ value.
5.3.5 Scaling unit size measures to facilitate sample allocation and selection

For reporting, administrative or other reasons, minimum and/or maximum sample size constraints may have to be applied to various types of domain, such as regions, sectors according to type of economic activity, nature of establishments, type of workplace, etc.

The domains \((d)\) to which such constraints apply often differ from strata groups \((g)\) used for optimisation and from the final strata \((h)\) actually used for sample selection. These domains may be defined in terms of single variables, or as cross-classification of multiple variables, or as overlapping set of controls in terms of the marginal distribution by each of a number of variables. The allocation requirements may include both statistical and practical considerations. Furthermore, the minimum sample size requirements may be modified to incorporate information useful for optimising sample allocation, and also to meet fieldwork constraints. Individual units or sectors of high importance for other reasons may also be given inflated size measures.

As practical advice in applying the required allocation, stratification and sample selection procedures in a flexible and simple manner, we recommend that the original unit size measures be rescaled to take into account sample allocation requirements. A very useful and convenient procedure is to take the appropriate size measures in the frame, and use them as the basis for determining unit selection probabilities after transforming them to satisfy the sample allocation constraints by domain. In a sense, units 'carry with themselves' their determined selection probabilities, irrespective of the details of the stratification and of the exact procedure used for sample selection. This helps to separate out the issues relating to sample allocation from the procedures relating to stratification and sample selection.

The basic procedure is as follows.

Let \(Z_{di}\) be the unit size measure defined to be proportional to the required unit selection probability, in the same way as in Section 5.3.1B above. (The subscripts indicates unit \(i\) in sample allocation domain \(d\).) Let its aggregate over units in population of domain \(d\) be

\[
Z_d = \sum_{i \in d} Z_{di}.
\]

Note that we have replaced in \(Z_{di}\) the hitherto used reference to stratum \(h\) by reference to allocation domain \(d\), since strata may cut-across the allocation domains, though often strata are finer partitions of the allocation domains.

Just as in Equation (5.4), the expected sample size for a domain is proportional to its aggregated \(Z\) values: \(n_d \propto Z_d\).

It is common in practice to make adjustments to the sample allocation originally determined (for instance on the basis of optimisation criteria) so as to take into consideration additional factors not included in the original allocation. Let \(n'_d\) be the adjusted required sample size for the domain.

Defining rescaled size measures as

\[
Z'_{di} = \left(\frac{n'_d}{Z_d}\right)Z_{di},
\]

(5.12)
5.3 Sample allocation

It can be seen that

\[ \sum_{i \in d} Z_{di}' = n_d'. \]  

(5.13)

It is important to note that (5.13) implies that the size measures \( Z_{di}' \) rescaled as in (5.12) are not merely proportional to the required unit selection probabilities in order to obtain the target sample size \( n_d' \) for the domain, but are in fact numerically equal to those probabilities. This is true for all the domains. Hence once the size measures have been scaled in this manner, it is no longer necessary to refer to specific allocation domains in stratifying and actually selecting the sample.

The objective of rescaling unit size measures is to absorb differences in the sample selection requirements among the strata so that a uniform procedure can be used throughout as far as possible, removing the need for separate treatment for different domains or strata. Furthermore, with this approach the actual strata used for sample selection can be different (and can be more numerous) than the domains used for sample allocation and optimisation. The selection probabilities proportional to the modified size measures would tend to ensure that the required sample allocation by domain is obtained independently of the stratification used for the purpose of actual selection. Therefore, stratification can be determined by considerations of statistical efficiency, cost, administrative control, operational convenience, etc., separately from the sample allocation requirements. The size measures (and the associated unit selection probabilities) form an automatic basis of allocation of the sample among strata – independently of how exactly the strata are defined – meeting at the same time the sample allocation constraints. The stratum sample sizes are essentially proportional to the sum over the stratum population of unit selection probabilities as represented by their rescaled measures of size. The unit size measures also give a logical and simple method of separating out the ‘take-all’ stratum from the rest of the ‘take-some’ population to be sampled to obtain a given total sample size. The take-all stratum simply comprises all units with rescaled size measure \( Z_{di}' \geq 1.0 \). We may note that for PPS systematic sampling applied to rescaled size measures, the selection interval would equal 1.0 in all domains; the unit lists can be simply put together if desired and a uniform selection procedure used throughout.

5.3.6 Application in practice

Taking note of the procedures and problems for sample allocation summarised above, the following approach is proposed in practical application.

In the frame used for the survey, generally information is available only on unit size measures \( X_{hi} \) (e.g. number of employees), and little or no information is directly available on unit costs and variances. Hence it may be necessary to begin from the simple ‘proportional to size’ allocation using Equation (5.5):

\[ n_h \propto Z_h = X_h = N_h \bar{X}_h; \]

or, when information on unit costs and population variances of \( X \) within strata is available, with ‘X-optimal’ allocation using Equation (5.10):

\[ n_h \propto N_h \bar{X}_h \left( C_{h}(X)/\sqrt{L}_h \right). \]
The efficiency of this approach depends on the strength of the correlation between $X$ the auxiliary variable, and $Y$ the variable to be measured in the survey. The closer the relationship between the size measures available and the target variables to be measured and estimated, the more efficient generally will be the sample under any given design. It is desirable therefore to adjust, as far as possible, the given size measures to more closely reflect the target variables.

Of course, the target variables represent the statistics to be measured in the survey and hence, by definition, are not available in the sampling frame at the level of individual units in the population. But this does not preclude that at the aggregate (macro) level, at least approximate information is available on the average relationship between the size measures $X$ available in the frame and the target variable(s). For instance, if the survey objective is to measure the total number of child workers, and it is known from alternative or past sources that the proportion of children among employees varies in a certain way by establishment type, size or location, or other characteristics available in the frame, then it may be possible to incorporate such information as average factors for adjusting of the size measures within individual categories of the involved classification, thus reducing the average disparity between $X$ and $Y$ across the categories.

Furthermore, any known information on the relative variability of the relationship between $X$ and $Y$ (such as in the overall variation of the individual ratios $(Y_{hi}/X_{hi})$ for units in stratum $h$ may also be incorporated into the size measures used to determine the selection probabilities, so as to move closer to the optimal sample allocation. For example, we may know from alternative sources that the proportion of children among all employees is more variables in certain sectors, so that it would be efficient to increase the share of the sample allocated to them.

Model used for adjustment is of the form

$$Z_{hi} \propto X_{hi} \left( \frac{Y_h}{X_h} \right) \left( C \left\{ \frac{Y_{hi}}{X_{hi}} \right\}_h \right),$$

where the first factor modifying the given size measures $X_{hi}$ is the ratio between average $Y$ and average $X$ values in stratum ($h$) or similar domain, assumed known, and the second factor is the coefficient of variation of this ratio among units in $h$.

A point of practical importance to note is that the ‘strata’ $h$ referred to here need not be the final strata actually used for selecting the sample. Normally the latter will be much more detailed than the aggregations for which useful external information for improved allocation may be available. The two systems of ‘stratification’ may not overlap, but generally they are likely to be related.

The unit size measures may then be rescaled to automatically meet specified domain sample size requirements as in (5.12).
5.4 Single-stage sampling of medium and large establishments employing children

In this section we address issues concerning sampling of large and medium sized establishments. As noted, usually such units are selected using frames which list individual establishments with information on their location and basic characteristics. Furthermore, when the study area is geographically limited, sampling from list frames may also be practicable for small establishments. More precisely, therefore, this section is concerned with situations when the sample of establishments (including small establishments) can be selected directly from lists, in a single stage. Often a second stage of sampling is involved in the selection of working children within selected establishments, unless the establishments are small or employ only a few children.

In addition to household-based surveys of child labour, a large number of establishment-based surveys focussed on particular sectors of child labour activity have been conducted in developing countries. Our general impression from a review of this experience is, however, that there is a considerable scope for improving the sampling design and methodology of such surveys. Undoubtedly, the survey have to be carried out in difficult conditions, and with limited resources and insufficient prior information to optimise survey design. This makes it all the more important that a special effort is made to make the samples more representative and more efficient through using better technical design. With this in view, this section addresses a number of aspects in statistical terms, such as how to make better use of information on unit size measures, improve stratification, optimise sample allocation, and apply more efficient schemes for selecting establishments and child workers within establishments.

5.4.1 Basic considerations concerning the sampling approach

It is useful to note some basic recommendations concerning the sampling approach. It is assumed in this section that the appropriate design is a single-stage design in which units (establishments) are selected directly from list frames.

- The statistical units best suited for the collection of the data may vary depending on the nature of the child labour activity to be captured.
- As a rule, full use should be made of lists of establishments available from all sources, not only official sources.
- Allocation of the sample among geographical or administrative domains and also by type and conditions of child labour activity is normally dictated by the survey objectives, forming external constraints.
- Stratification by size is introduced where possible without making the strata too small. Since the scope for this is limited by other stratification criteria of higher priority (geographical location, type of activity, etc.), stratified variable probability rather than stratified equal probability design is often used to ensure control over sample size variations.
- Normally, establishments are selected with probabilities related to some measure of establishment size (PPS sampling). Ideally, the number of children working in the establishment should form the basis of the establishment size
measure. When this is not available, approximate and surrogate size measures should be sought.

- In any case, whether PPS or constant probability sampling is used within strata, sample allocation among the strata will be essentially in proportion to the sum of size measures of units in the stratum. In strata composed of large establishments, we may take all of them into the sample (‘take-all’ sampling), the other strata being subject to sampling (the ‘take-some’ strata). This division can be determined more precisely on the basis of unit size measures, once the sample size is determined.

- Often there is conflict between the requirement of controlling unit selection probabilities and sample size. A compromise between the two objectives is the appropriate choice in most situations.

The above technical points will be elaborated in the following.

### 5.4.2 Selection of units within strata: πPS versus stratified-SRS

Two important design choices to consider concern: selection probabilities of units within strata; and fixed probabilities versus fixed sample size.

#### A. Selection probabilities of units within strata

The first important design option concerns the variation of selection probabilities of units within strata. There are two common classes of design as concerns the procedure of selection within strata. These are:

(A1) the selection of units with probability proportional to size (πPS) within strata,

\[ P_i \propto X_i; \quad n_h \propto \sum_{i \in h} X_i; \text{ versus} \]

(A2) equal probability selection of units within each stratum,

\[ P_{i \in h} \propto \bar{X}_h = \frac{\sum_{i \in h} X_i}{N_h}; \quad n_h \propto \sum_{i \in h} X_i, \text{ the same as in design (A1)}. \]

Option (A1) is probably more complex, but the major drawback of option (A2) is that it requires fine stratification by unit size, so as to control for the effect of variation in that size. This is in competition with other stratification criteria – criteria which often have a higher, substantively determined priority - such as major reporting domains, or geographical/administrative divisions. For this as well as other reasons noted elsewhere, our preference is for a probability proportional to size (πPS) design.

In short, generally an efficient design is to select establishments with probabilities \((P)\) proportional to some measure of establishment size \((X)\) - or some function of size, \(P \propto Z = f(X)\), as discussed in Sections 5.2-5.3. To this, appropriate modifications are usually made, such as the following.

- Selection of large units \((X \geq X_L)\) with \(P = 1\).

- A minimum limit of the selection probability assigned to small units \((X \leq X_S)\), \(P = P_0\).
Varying the probabilities in proportion to some function of size, for instance 
\[ P_i \propto X_i^q, \quad 0 \leq q \leq 1. \]

Or alternatively, uniform selection probabilities within strata, for instance proportional to some function of the average unit size measure in the stratum.

Figure 5.1 illustrates the various types of relationship between size measures \( X \) and the unit selection probabilities \( P \). The patterns are shown below.

1. The unit selection probability \( P \) varies strictly in proportion to unit size.
2. The above is modified to put minimum \((P \geq P_0)\) and maximum \((P \leq 1)\) limits on the selection probabilities, whatever the unit size.
3. The variation in selection probabilities in (2) is smoothed, for instance to an “S” shaped curve within the minimum and maximum limits.
4. Unit size measures are grouped into size classes. Within any class, all units receive the same selection probability, proportional to the average size measure in the class.

A large number of schemes have been developed for sampling with unequal probabilities without replacement; for a wide-ranging review, see Hanif and Brewer (1980).

### B. Fixed probabilities versus fixed sample size

The objective of design is to control two aspects: unit selection probabilities and the numbers of units selected. The two objectives can conflict, and a choice has to be made as to which of these is given higher priority.

The two extreme options are:

- (B1) fixed unit selection probabilities \( (P_i) \), and the stratum sample sizes allowed to vary, the sample size \( (n_h) \) becoming a random variable, with expected value equal to the (fixed) target sample size, \( E(n_h) = n_h \);
- (B2) versus fixed stratum sample sizes \( (n_h) \), and the unit selection probabilities allowed to vary, the unit selection probabilities \( (P'_i) \) becoming a random variable, with expected value equal to the (fixed) target value, \( E(P'_i) = P_i \).
It is important to keep good control over the sample size obtained, especially when many strata with small sample sizes are involved. However, the control over sample size does not have to be absolute. Some reasonable variation can be accommodated in practice in most circumstances.

At the same time, it is also important to keep control over the sampling probabilities applied. Unnecessary variation reduces the efficiency of the design.

Our recommendation is for a compromise solution where possible, with somewhat higher priority given to control over sampling rates. Keep the design selection probabilities unchanged if the resulting sample size is within certain specified limits around its expected value under the design. Otherwise, make the minimum necessary adjustment to the selection probabilities to ensure that the sample size is not outside those limits.

These choices may also be considered largely independently of other aspects of the design, such as the choice of unit size measures and sample allocation. The difference between the two alternatives (A1) and (A2) reduces as stratification, especially size stratification, becomes more detailed. By contrast, the difference between the two alternatives (B1) and (B2) increases as stratification becomes more detailed.

5.4.3 Some specific sampling schemes

There are different methods available for selecting samples of establishments within strata. Some of the common ones are described below. A useful review of issues in sampling establishments from list frames is provided by Sigman and Monsour (1995).

A. Poisson sampling

Poisson sampling is a simple and flexible procedure which we can emulate in choosing an appropriate sampling method, keeping its merits to the extent possible, and minimising its limitations. The procedure is as follows.

The method subjects each unit \( i \) independently to selection with its assigned probability \( P_i \). This can be achieved by assigning to each unit in the frame a random number \( r_i \) from uniform distribution (0,1). The unit is included in the sample if \( r_i \leq P_i \), and not included otherwise.

The selection procedure has been represented in Figure 5.2 by the plot of given probabilities \( P_i \) against assigned random numbers \( r_i \) (0,1) to units in the population. Units in the shaded area are selected because for them the relationship \( r_i \leq P_i \) is satisfied. A line parallel to the horizontal axis represents selections with some constant probability \( P \). It can be seen that proportion \( P \) of this line is in the shaded area (representing selected units), and proportion \( (1 - P) \) is outside (representing units not in the sample).

Clearly, the method is extremely simple to apply. Its major drawback in this simplest form is the lack of control over the sample size obtained. This is because each unit is subject to selection independently, irrespective of how many other units have already been selected in the same way.
The expected sample size is $E(n') = n = \sum P_i$, summed over all units $i$ in the population, but the achieved sample size $(n')$ is a random variable with large variance. Its variance $V(n') = \sum P_i (1 - P_i)$ is approximately equal to $n$ for small $P_i$, which is a serious problem when we are dealing with small samples such as for individual strata.

Indeed, the probability of getting no sample at all is $Pr(n' = 0) = \prod (1 - P_i)$; the method is worth considering only when this probability is negligible.

The unbiased (Horvitz-Thompson) estimator of a total $Y$ is 

$$\hat{Y}_{HT} = \sum (Y_i / P_i),$$

where the sum is taken over the $n'$ units in the sample. Its variance is estimated by

$$\hat{V}(\hat{Y}_{HT}) = \sum \left(1 - P_i \right) \left( \frac{Y_i}{P_i} \right)^2.$$

This variance tends to be large, and much more precise is a ratio estimator which takes into account the sample size actually obtained. The estimator is:

$$\hat{Y}_R = \left(\frac{n}{n'}\right) \sum \left(\frac{Y_i}{P_i} \right),$$

with variance estimated as

$$\hat{V}(\hat{Y}_R) = \sum \left(1 - P_i \right) \left( \frac{Y_i}{P_i} - \frac{\hat{Y}_R}{n} \right)^2.$$

This shows that despite the lack of control over sample size, the procedure can be efficient (actually as efficient as simple random sampling when the $P_i$ values are constant), provided that the stratum sample sizes are not too small and the probability of getting a zero sample size is negligible.
5. Sampling establishments employing children

B. Collocated sampling

A special case of Poisson sampling is when the probability of selection is a constant \( P \) for all units in the population. (This case is commonly referred to as Bernoulli sampling.) One way to reduce the variability in sample size in the case of constant probability sampling is to replace the random numbers \( r_i \) by equidistant numbers \( x_i \) in terms of the rank of \( r_i \), ordered from the smallest to the largest value \((i = 1 \ldots N)\). That is,

\[
x_i = \frac{\text{rank}(r_i) - x}{N},
\]

with \( x \) as a random number \((0,1)\). A unit is selected into the sample if \( x_i \leq P \) (Brewer, Early and Hanif, 1984). Selection probability is strictly controlled, and also the sample size nearly equals \( PN \).

The method may not work with variable \( P_i \), however, because in that case it would be necessary to make the numbers equidistant at all levels of \( P_i \), or at least for sufficiently fine groupings of \( P_i \). The same would be required for individual strata when we have to control the stratum sample sizes. It may not be possible to ensure this fully when we have many strata with small sample sizes.

C. Sequential sampling

This is another variation on the Poisson sampling procedure aimed at controlling sample size with variable selection probabilities (Ohlsson, 1998). Consider the basic Poisson sampling procedure in which a unit is in the sample if the random number selected for it is at or below its probability of selection, \( r_i \leq P_i \). In terms of a transformed variable \( x_i = (r_i / P_i) \), the same result is obtained by taking a unit into the sample if \( x_i \leq 1 \).

Hence a straightforward method is to arrange the units in order of increasing \( x_i \) and take a certain number of units from the top of the ordered list into the sample. The specified selection probabilities are applied if we select up to the last unit for which \( x_i \leq 1 \).

But instead, in Sequential Poisson sampling, exactly \( n \) units from the top of the ordered list are taken into the sample, where \( n \) is the fixed required sample size. That is, the sample is defined by rank \( (x_i) \leq n \).

This has to be done separately within each stratum, or within groups of strata where it suffices to control the sample size within groups only. In the presence of many small strata, appropriate grouping of the strata may in fact be the only practical way of keeping the system manageable.

D. Sequential-collocated sampling

The attraction of the collocated sampling scheme is that it applies strictly the required selection probabilities. But in practice, it is not always easy or possible to ensure that equidistant random numbers are assigned separately to each different type of units.

By contrast, the sequential sampling method does not require the random numbers used for sampling to be made equidistant, though having the numbers bunched together can cause problems in obtaining representative samples. The essential point of the
sequential sampling scheme is that it controls the sample size. However, in practice, it is hardly ever necessary to impose this control over sample size rigidly. Sample sizes tend to vary in any case due to non-response and the presence of ineligible listings in the frame.

To maximise the positive points of each scheme, and minimise their limitations, we can combine them as follows (Verma, 2002).

1. The original random numbers \( r_i \) are first made equidistant as in the collocated sampling scheme. And then they are transformed as in sequential sampling.

2. Secondly, rather than insisting on an absolutely fixed sample size \( n \), we can allow variation within certain practically acceptable limits, say \( n_{\text{min}} \) to \( n_{\text{max}} \). Within that range, we retain the original sampling probabilities unchanged.

For instance, for small stratum sample sizes of 5-10 units, one may accept up to 20 per cent variation in the sample size actually obtained.

In other words, if the collocated scheme, which retains the selection probabilities unchanged, gives a sample size \( n' \) within the limits \( n_{\text{min}} \) to \( n_{\text{max}} \), then it is accepted. Otherwise the sequential scheme is used to obtain a sample of size lying at the boundary of this range nearest to \( n' \), that is, \( n = n_{\text{min}} \) if \( n' \leq n_{\text{min}} \), and \( n = n_{\text{max}} \) if \( n' \geq n_{\text{max}} \). Starting with the original random number \( r_i \) for the unit, inclusion into the sample is determined by the following procedure.

\[
x_i = \left( \frac{\text{rank}(r_i) - x}{N} \right) \cdot \frac{1}{P_i}.
\]

With \( n' = \text{count}(x_i \leq 1) \), the sample is defined by

\[
\text{rank}(x_i) \leq \max(n_{\text{min}}, \text{min}(n', n_{\text{max}})).
\]

### 5.4.4 Design criteria

A design may be evaluated in terms of how well it compromises on a series of criteria (Särndal, 1996):

1. Strictly \( \pi \text{PS} \) sampling, with unit selection probabilities \( \pi_i \)
2. Fixed sample size(s), \( n \)
3. Exact ‘design unbiasedness’ for the estimator of population total and its variance
4. Simple calculation of joint (second order) selection probabilities \( (\pi_{ij}) \) of units used in the variance estimation
5. Single-sum variance calculation

**Practicality of the design**

Criterion (1) refers to selection of units with probabilities strictly proportional to their size measures and without replacement.

Criterion (2) refers to constant sample sizes, not only overall but also separately for each stratum from which an independent sample is selected.
Criterion (3) refers to technical bias in the estimators used (and not to non-response or measurement bias in the results). It normally involves the Horvitz-Thompson estimator.

A lot of research has concerned the production of strategies satisfying (1)-(3). In principle, these together give a highly efficient design.

However, a strict adherence to (1)-(3) conflicts with (4)-(5).

Criterion (4) refers to the fact that when units are not selected independently and with uniform probabilities, the joint probabilities for pairs of units of both appearing in the sample become difficult to calculate. While for producing estimates of population total and means etc. only the probabilities of selection of individual units are required, estimation of variance generally requires information on joint probabilities as well. Criterion (5) refers to the fact that, in addition, in the estimation of variance these joint probabilities appear in double sums involving all pairs in the sample.\(^\text{16}\)

Strict adherence to (1)-(3) conflicts with (4)-(5) in the sense that the former results in complex or intractable joint selection probabilities and laborious sums over all pairs of units (rather than sums simply over individual units) in the computation of variances.

The conflicts and compromises between different design criteria are illustrated by the following examples of designs.

**Design (A). Unbiased design with fixed probabilities and sample size**

An example is the Horvitz-Thompson estimator and its variance under a design satisfying criteria (1)-(3) exactly, but violating (4)-(5):

\[
\hat{Y} = \sum_i \left( \frac{y_i}{\pi_i} \right) ; \quad \hat{V}(\hat{Y}) = \sum_i \Sigma_j \left( \frac{1}{\pi_{ij}} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j \right).
\]

In order to satisfy criteria (4)-(5), it is necessary to abandon at least some of (1)-(3). The best practical approach is to seek to satisfy (1)-(3) closely but not exactly, in such a way that computational simplicity of (4)-(5) can be achieved. Alternative variance estimators of the Horvitz-Thompson estimator, based on a simple sum of squared deviations, have been investigated by several authors; see Brewer and Hanif (1983) for a review.

**Design (B). Simple random sampling within strata**

As a counter example, stratified-SRS (equal probability simple random sampling without replacement within strata) designs are commonly used in business surveys. These abandon (1) and (3), so as to satisfy (2), (4) and (5). By using sufficiently fine stratification by size, and (uniform) selection probability proportional to the average unit size measure within each stratum, an approximation to the highly efficient \(\pi\)PS design may be achieved. But this is not possible when, for operational and reporting reasons, stratification by other more important criteria (such as region or sector of activity) must have priority over stratification by size.

\(^{16}\) In other words, in place of \(n\) quantities for the same number of units in the sample, one has to deal with \(n(n+1)/2\) quantities for all units and all pairs of units in the sample.
Design (C). Poisson sampling

Another rather different approach is Poisson sampling. Each unit is subject to independent selection with its probability $\pi_i$. Criterion (1) is thus satisfied exactly, as are (4) and (5). Criterion (3), requiring strictly unbiased estimators, gives unacceptably large variance, and hence this criterion has to be abandoned in favour of some close approximation, such as using ratio or regression estimators which are often nearly unbiased but much more efficient. The real problem with Poisson sampling is the lack of control over sample size – criterion (2) – which can be unacceptable particularly for strata or domain with small sample sizes. Even zero stratum sample sizes may occur.

As noted above, the best practical approach is to seek to satisfy (1)-(3) closely but not exactly, in such a way that the computational simplicity of (4)-(5) can be achieved. Better solutions lie in the direction of retaining as far as possible the advantages of designs (B) and (C) above, namely stratified-SRS and Poisson sampling, and avoiding the disadvantages of these designs. Satisfaction of criteria (4)-(5), which is common to both designs, should be retained; criterion (3) may be abandoned in favour of more efficient but nearly unbiased estimators.

It is worthwhile to seek a design in which Criterion (1) concerning $\pi$ selection is satisfied as it is in design (C) (i.e. Poisson sampling), and criterion (2) concerning constant sample size is satisfied as it is in design (B). It is not necessary to meet both these requirements exactly, but it is desirable that both are met with good approximation. There are two commonly used options which meet the above requirements. The choice between them depends upon which of the two requirements - (1) or (2), i.e. fixed sampling probabilities or fixed sample size – are satisfied exactly. The options, already described, are:

Design (D) Collocated Poisson sampling

Design (E) Sequential Poisson sampling

to which we may add,

Design (F) Sequential-collocated sampling.

5.5 Selecting child workers within establishments

5.5.1 Selecting establishments and child workers: some two-stage design options

In Section 3.3 we described several variations of probability proportional to size (PPS) sampling in a two-stage design. Here we consider them further, specifically in the context of selecting samples of establishments and child workers in them.
The following notation is used.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_i$, $\bar{S}$</td>
<td>Estimated size of establishment $i$, as available in the sampling frame prior to the survey; $\bar{S}$ is their mean value. This information is available for all units in the population.</td>
</tr>
<tr>
<td>$S'_i$</td>
<td>Actual size of establishment $i$, as obtained from the survey. Normally this refers to the number of working children in the establishment. This information is normally available only for units in the sample.</td>
</tr>
<tr>
<td>$X_i, X'_i$</td>
<td>Transformed size measures defining the selection probabilities, respectively at the first stage (selection of establishments) and the second stage (selection of working children). Normally $X_i$ is a function of $S'_i$, and $X'_i$ is a function of $S_i$ or $S'_i$.</td>
</tr>
<tr>
<td>$p_{1i}, p_{2i}, p_i$</td>
<td>Unit selection probabilities at the first and the second stages, and the overall selection probabilities of the ultimate units (working children).</td>
</tr>
<tr>
<td>$b_i$</td>
<td>‘Sample take’ (number of working children selected) from establishment $i$.</td>
</tr>
<tr>
<td>$a, n$</td>
<td>Sample size, respectively the total number of establishments and the number of working children selected; $n = \sum_{S} b_i$ where $\sum_{S}()$ is the sum over the $a$ sample units.</td>
</tr>
<tr>
<td>$l, b, p$</td>
<td>Constant parameters related to the population and sample sizes. $l = \sum_{P} X_i/a$ where $\sum_{P}()$ is the sum over all units (establishments) in the population. It is the average amount of $X_i$ in the population which contributes one establishment to the sample. Constant $b$ equals or is related to the average sample take per establishment, $(n/a)$; constant $p = b/l$ equals or is related to the average overall sampling rate. Details will be specified below.</td>
</tr>
</tbody>
</table>

**General scheme**

The overall sampling rate varies in proportion to the ratio $(X_i/X'_i)$ of the chosen design measures, and the sample take (number of working children selected) varies as $(S'_i/X'_i)$ for establishment $i$.

$$l = \sum_{P} X_i/a \quad n = \sum_{S} b_i.$$

$$p_{1i} = \frac{X_i}{l} \quad p_{2i} = \frac{b}{X'_i} \quad p_i = p_{1i} \cdot p_{2i} = \left(\frac{b}{l}\right) \frac{X_i}{X'_i} = p \left(\frac{X_i}{X'_i}\right) \quad b_i = p_{2i} S'_i = b \left(\frac{S'_i}{X'_i}\right) \quad (5.15)$$

**A. PPS with exact measures of size**

With the actual unit sizes $S'_i$ known exactly before sample selection, the selection probabilities at the two stages can be adjusted so as to obtain for the sample of working children a constant overall selection rate and also a constant sample take from each selected establishment.

$$X_i = X'_i = S_i = S'_i.$$
5.5 Selecting child workers within establishments

B. Constant probability at both stages

The probability is constant at each of the two stages, and hence is constant overall. Sample take per establishment varies in proportion to $S_i'$, the actual size of the establishment.

$$X_i = X_i' = S_i.$$  \[ (5.17) \]

\[
\begin{array}{|c|c|c|c|}
\hline
p_{1i} = \frac{S_i}{l} & p_{2i} = \frac{b}{S_i} & p_i = \frac{b}{l} = p \text{ (const.)} & b_i = p_{2i}S_i = b \text{ (const.)} \\
\hline
\end{array}
\]  \[ (5.16) \]

C. Fixed-take design

The sample size (number of children) to be selected from each sample establishment is fixed, but the overall selection probability of child workers varies in proportion to the ratio $(X_i'/S_i')$ of the chosen design measures.

$$X_i' = S_i'.$$

\[
\begin{array}{|c|c|c|c|}
\hline
p_{1i} = \frac{X_i}{l} & p_{2i} = \frac{b}{S_i'} & p_i = \left(\frac{b}{l}\right) \left(\frac{X_i}{S_i'}\right) = p \left(\frac{X_i}{S_i'}\right) & b_i = p_{2i}S_i' = b \text{ (const.)} \\
\hline
\end{array}
\]  \[ (5.18) \]

D. Measure of size $X_i$ is some more general function of unit size $S_i$

The measure determining selection probability of an establishment is

$$X_i = X_i' = f(S_i).$$

By appropriately choosing the probabilities at the two stages, the overall probability for selecting child workers can be made uniform, as follows:

\[
\begin{array}{|c|c|c|c|}
\hline
p_{1i} = \frac{f(S_i)}{l} & p_{2i} = \frac{b}{f(S_i)} & p_i = \frac{b}{l} = p \text{ (const.)} & b_i = p_{2i}S_i' = b \left(\frac{S_i'}{f(S_i)}\right) \\
\hline
\end{array}
\]

A useful scheme is to vary $X_i$ as the square root of $S_i$, i.e. with $f(S_i) = \sqrt{S_i}$. Assuming $S_i \approx S_i'$:

\[
\begin{array}{|c|c|c|c|}
\hline
p_{1i} = \frac{\sqrt{S_i}}{l} & p_{2i} = \frac{b}{\sqrt{S_i}} & p_i = \frac{b}{l} = p \text{ (const.)} & b_i = p_{2i}S_i' \approx b\sqrt{S_i} \\
\hline
\end{array}
\]  \[ (5.19) \]
In this case the overall probability of selecting children is constant, but the sample take varies in proportion to square root of establishment size.

E. Other variations, e.g. upper limit on sample take per establishment

Often it is desirable to take all child workers found in a selected establishment, but limit the number to be taken to some specified maximum. If the number of child workers found in an establishment exceeds that limit, only that maximum number are included in the sample, \( b_i' = \min(S_i', b_{\text{max}}) \). The following scheme gives this design.

If \( S_i' \leq b_{\text{max}} \):

\[
\begin{align*}
p_{1i} &= p \\
p_{2i} &= 1 \\
p_i &= p \text{ (const.)} \\
b_i &= p_{2i}S_i' = S_i'
\end{align*}
\]  

(5.20a)

If \( S_i' > b_{\text{max}} \):

\[
\begin{align*}
p_{1i} &= p \\
p_{2i} &= \frac{b_{\text{max}}}{S_i'} \\
p_i &= \left(\frac{b_{\text{max}}}{S_i'}\right)p \\
b_i &= p_{2i}S_i' = b_{\text{max}}
\end{align*}
\]  

(5.20b)

5.5.2 Numerical illustration of two-stage designs: selection of establishments and of working children within establishments

A. Numerical data for the illustration

In order to illustrate various schemes for selecting a two-stage sample of establishments and child workers engaged in them, Table 5.2 takes a number of populations of establishments, showing their distribution according to size, size taken as the number (0-9) of child workers employed. The number of establishments in each population has been taken as 100, so that the cells of the table can be seen as the number of establishments or as their per cent distribution by size in the population concerned. The illustrative populations vary from uniform distribution ‘1’, to distribution ‘7’ where most of the establishments are very small (employing very few children), and 80 per cent have no child workers.

Distribution ‘4’ in the table (with nearly 40 per cent of establishments not employing children, and another 25 per cent each employing exactly one child worker) has been highlighted. This is to indicate that details of the various sampling schemes will be discussed below with reference to this population.
### Table 5.2. Illustrative data

Distribution of establishments according to the number of children employed

<table>
<thead>
<tr>
<th>Population distribution</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Total number of children</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>100.0</td>
</tr>
<tr>
<td>2</td>
<td>14.1</td>
<td>13.4</td>
<td>12.6</td>
<td>11.8</td>
<td>10.9</td>
<td>10.0</td>
<td>8.9</td>
<td>7.7</td>
<td>6.3</td>
<td>4.5</td>
<td>100.0</td>
</tr>
<tr>
<td>3</td>
<td>26.0</td>
<td>21.0</td>
<td>16.6</td>
<td>12.7</td>
<td>9.4</td>
<td>6.5</td>
<td>4.2</td>
<td>2.3</td>
<td>1.0</td>
<td>0.3</td>
<td>100.0</td>
</tr>
<tr>
<td>4</td>
<td>39.5</td>
<td>25.9</td>
<td>16.2</td>
<td>9.5</td>
<td>5.1</td>
<td>2.5</td>
<td>1.0</td>
<td>0.3</td>
<td>0.1</td>
<td>0.0</td>
<td>100.0</td>
</tr>
<tr>
<td>5</td>
<td>50.5</td>
<td>26.9</td>
<td>13.3</td>
<td>5.9</td>
<td>2.4</td>
<td>0.8</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>100.0</td>
</tr>
<tr>
<td>6</td>
<td>67.0</td>
<td>23.4</td>
<td>7.2</td>
<td>1.9</td>
<td>0.4</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>100.0</td>
</tr>
<tr>
<td>7</td>
<td>80.0</td>
<td>15.0</td>
<td>4.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Columns (1) and (2) of Table 5.3 below show some statistics computed from each of the population distributions of Table 5.2. (Other figures in Table 5.3 will be commented upon subsequently.)

Column (1) shows the mean number of child workers per establishment for each distribution. Establishments not employing any children are included in computing the mean, which therefore falls sharply from 4.5 for distribution ‘1’ to only 0.3 for distribution ‘7’, reflecting a very high proportion of establishments not employing any children in the latter.

Column (2) shows the mean number of child workers per establishment for each distribution, but excluding establishments not employing child workers. The mean declines from 5.0 for distribution ‘1’, to 1.3 for distribution ‘7’, reflecting the concentration of small establishments in the latter.

### B. Simple random sampling of establishments

Consider that we begin from a simple random sample of establishments.

The objective is to obtain a sample of 20 working children. (This is out of, for example, 450 working children in population ‘1’, or only 26 working children in population ‘7’.)

Let us consider the following three variations of the design starting from a SRS of establishments.

**Design A.** SRS of establishments, with all working children included in selected establishments (‘take-all sampling’).

**Design B.** SRS of establishments, followed by selecting 1 child per establishment employing children.

**Design C.** SRS of establishments followed by taking 1-in-2 subsample of establishments employing 1 child, and then by selecting 1 child per establishment employing 1 or more children.

As noted, columns (1) and (2) of Table 5.3 give mean values for each population. Figures in the remaining columns are with reference to obtaining a sample of 20
working children, computed for each of the distributions ‘1’ to ‘7’, for each variation of
the design in turn. Footnotes to the table provide a description of the statistics shown.

Note that these are not results from one particular sample, but are expected values, i.e. averages of all possible samples under a given design.

<table>
<thead>
<tr>
<th>Table 5.3. Statistics for illustrative data for a SRS of establishments yielding 20 working children</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computed for Table 5.2 distributions according to the number of children employed</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistic (see description below)</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Design A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4.50</td>
<td>5.00</td>
<td>4.4</td>
<td>4.0</td>
<td>22.2</td>
<td>23.5</td>
</tr>
<tr>
<td>2</td>
<td>3.65</td>
<td>4.25</td>
<td>5.5</td>
<td>4.7</td>
<td>23.3</td>
<td>25.2</td>
</tr>
<tr>
<td>3</td>
<td>2.14</td>
<td>2.89</td>
<td>9.3</td>
<td>6.9</td>
<td>27.0</td>
<td>31.5</td>
</tr>
<tr>
<td>4</td>
<td>1.28</td>
<td>2.12</td>
<td>15.6</td>
<td>9.4</td>
<td>33.0</td>
<td>42.0</td>
</tr>
<tr>
<td>5</td>
<td>0.86</td>
<td>1.74</td>
<td>23.2</td>
<td>11.5</td>
<td>40.4</td>
<td>55.5</td>
</tr>
<tr>
<td>6</td>
<td>0.45</td>
<td>1.38</td>
<td>44.0</td>
<td>14.5</td>
<td>60.7</td>
<td>94.1</td>
</tr>
<tr>
<td>7</td>
<td>0.26</td>
<td>1.30</td>
<td>76.9</td>
<td>15.4</td>
<td>100.0</td>
<td>*</td>
</tr>
</tbody>
</table>

Notes to columns:

(1) Mean number of children employed per establishment
(2) Mean number considering only establishments employing one or more children

_in order to obtain a sample size of 20 working children:

Simple random sample of establishments, and taking all working children in selected establishments

(3) Expected number of establishments to be selected in a SRS of establishments, =20/(1)
(4) Of those, expected number employing children, =20/(2)

Sampling rate (%) for a SRS of establishments, with subsampling of working children in selected establishments ...

(5) if followed by selecting 1 child per establishment employing children, =0.20*(2)/(1)
(6) if followed by taking 1-in-2 subsample of establishments employing 1 child and then by selecting 1 child per establishment employing more than 1 child (computed using the two columns headed ‘0’ and ‘1’ in Table 5.2).

* Not possible to obtain the required sample size of 20 working children

Comments on Table 5.3 for the three designs A-C follow.

*Design A. SRS of establishments; all children in a selected establishment*

Column (3) gives the expected number of establishments to be selected with simple random sampling (SRS) to obtain a sample of _n_ = 20 working children, when all children in each selected establishment are included in the sample. In this case, this simply equals _n_ (=20), divided by the average number of working children per establishment, = (20/col(1)).

Column (4) is the expected number among these establishments which contain at least one working child; it equals (20/col(2)).

*Design B. SRS of establishments; 1 child per selected establishment employing children*

Column (5) is for a design in which establishments are selected with SRS at the first stage, and then one working child is selected from each establishment
5.5 Selecting child workers within establishments

employing children in the sample. If all establishments employed child workers, we would need a sample of 20 establishments in order to obtain the required number 20 of working children, i.e. a sampling rate of 20/100 = 0.20 for selecting establishments. The first stage sampling rate has to be increased because some of the selected establishments will contribute no working children to the sample. This is shown in column (5), = 0.20 * col(2)/col(1).

**Design C. As above except that, of the originally selected establishments employing 1 child, only 50 per cent are retained in the sample**

Column (6) is for a similar design, but this time dropping 50 per cent of the initially selected establishments which employ only one working child. The objective is to reduce the difference in the final selection probabilities of working children from establishments of different sizes, as explained below in the comments to Table 5.4, Design C. In any case, the initial sampling rate for selecting establishments in column (6) is higher than in column (5) because half of the selected establishments with only 1 working child are subsequently dropped from the sample. The difference between the two columns increases as the proportion of establishments employing only 1 child, among establishments employing any children, increases in the population, as from distribution ‘1’ to distribution ‘7’.

Computation of the figures requires reference to the data in Table 5.2. If $P_0$ is the proportion of establishments with no working children, and $P_1$ the proportion with exactly 1 working child each, then

\[
\text{col}(5) \text{ of Table 5.2(B)} = \frac{0.20}{(1 - P_0)},
\]

\[
\text{col}(6) \text{ of Table 5.2(B)} = \frac{0.20}{(1 - P_0 - 0.5P_1)}.
\]

Table 5.4 shows details of the sample with different designs applied to one of the distributions by establishment size, namely distribution ‘4’ of Table 5.2. For each design, the figures shown include: (a) the number of establishments to be selected; and (b) the expected number of working children to be selected, according to establishment size, in order to obtain a total sample of 20 working children. Additional figures are shown for some of the designs, as noted below.

---

17 Among establishments employing children, the proportion employing exactly 1 child increases from 11% (10/(100-10)) for Population ‘1’, to 75% (15/(100-80)) for Population ‘7’. See Table 5.2. This proportion equals $P_i/(1 - P_0)$, where $P_i$ is the proportion employing exactly $i$ children.
Table 5.4. SRS of establishments yielding 20 working children:
Expected sample size distributions by establishment size, for designs A-C.

| Distribution of establishments according to the number of children employed |
|-----------------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| Children employed           | 0                 | 1                 | 2                 | 3                 | 4                 | 5                 | 6                 | 7                 | 8                 | 9                 | total             |
| % distribution             | 39.5              | 25.9              | 16.2              | 9.5               | 5.1               | 2.5               | 1.0               | 0.3               | 0.1               | 0.0               | 100.0             |

Design A. SRS of establishments, with all working children included in selected establishments ("take-all sampling")

<table>
<thead>
<tr>
<th>(3) Number of establishments to be selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(4) Number of establishments to be selected (excluding those not employing children)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
</tr>
</tbody>
</table>

Expected number of working children selected

| 0.0  | 4.0  | 5.0  | 4.4  | 3.2  | 1.9  | 0.9  | 0.3  | 0.1  | 0.0  | 20.0  |

Design B. SRS of establishments, followed by selecting 1 child per establishment employing children

<table>
<thead>
<tr>
<th>(5) Number of establishments to be selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.0</td>
</tr>
</tbody>
</table>

Expected number of working children selected

| 0.0  | 8.6  | 5.3  | 3.1  | 1.7  | 0.8  | 0.3  | 0.1  | 0.0  | 0.0  | 20.0  |

Design C. SRS of establishments followed by taking 1-in-2 subsample of establishments employing 1 child and then by selecting 1 child per establishment employing more than 1 child

<table>
<thead>
<tr>
<th>(6) Initial number of establishments to be selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.6</td>
</tr>
</tbody>
</table>

Number of establishments retained in the sample

| 0.0  | 5.4  | 6.8  | 4.0  | 2.2  | 1.0  | 0.4  | 0.1  | 0.0  | 0.0  | 20.0  |

Comments on Table 5.4 for the three designs A-C follow.

Design A. SRS of establishments; all children in a selected establishment

The total number of establishments selected corresponds to column (3) of Table 5.3, broken down here according to the number of children employed in the establishment. Since establishments are selected at a uniform rate irrespective of establishment size, the total number selected is distributed proportionately across establishment size categories. The expected number of child workers selected from each size category is simply the number of establishments selected times the establishment size.

Design B. SRS of establishments; 1 child per selected establishment employing children

The total number of establishments selected corresponds to column (5) of Table 5.3. The number of child workers selected in each category is the same as the number of establishments selected in the category, since in all cases 1 child per establishment is taken into the sample. Other details are as in Design A.

Design C. As above, with only 50 per cent of establishments employing 1 child retained in the sample

The total number of establishments selected corresponds to column (6) of Table 5.3. The only other difference from Design B is that, following the initial selection, all establishments not employing children and 50 per cent of those employing...
only 1 child are dropped from the sample. The final selection probability of a working child appearing in the sample depends on the size of the establishment employing the child. In Design B, it equalled the selection probability, say \( p \), of an establishment (which is uniform irrespective of the size in the design being considered), divided by the establishment size \( S_i \) (since only 1 out of \( S_i \) working children in the establishment is selected). For sizes \( S_i = 1 \) and 2 in Design B, this probability equalled \( p \) and \( p/2 \), respectively. This difference is removed in Design C by reducing the final selection probability to \( p/2 \) also for establishments of size \( S_i = 1 \). Consequently, working children in the resulting sample with Design C have more uniform selection probabilities than in Design B.

C. Selecting establishments with probability proportional to some measure of size

Hitherto we have considered simple random samples, in which establishments are selected with a uniform probability. Now we consider probability proportional to size (PPS) sampling of establishments, where an establishment’s size is taken as the number of working children it employs, which is assumed to be known exactly prior to sample selection. (This assumption is likely to be somewhat unrealistic in practice, though approximate information on numbers of children employed may be available.) The design corresponds to Equation (5.16) above. We will refer to this as Design D. Table 5.5 shows some statistics for the design.

### Table 5.5. Statistics for illustrative data for a PPS of establishments yielding 20 working children

<table>
<thead>
<tr>
<th>Statistic (see description below)</th>
<th>(1)</th>
<th>(2)</th>
<th>(7)</th>
<th>(8)</th>
<th>(9)</th>
<th>(10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population distribution</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population</td>
<td>4.50</td>
<td>5.00</td>
<td>450</td>
<td>22.5</td>
<td>23</td>
<td>46</td>
</tr>
<tr>
<td>Expected sample values (Design D)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3.65</td>
<td>4.25</td>
<td>365</td>
<td>18.3</td>
<td>19</td>
<td>37</td>
</tr>
<tr>
<td>2</td>
<td>2.14</td>
<td>2.89</td>
<td>214</td>
<td>10.7</td>
<td>11</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.28</td>
<td>2.12</td>
<td>128</td>
<td>6.4</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>0.86</td>
<td>1.74</td>
<td>86</td>
<td>4.3</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>0.45</td>
<td>1.38</td>
<td>45</td>
<td>2.3</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>0.26</td>
<td>1.30</td>
<td>26</td>
<td>1.3</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

**Notes to columns:**
1. (1), (2) See Table 5.3
2. (7) Total measure of size of the 100 establishments, \( = 100^*(1) \).
3. (8) Sampling interval to be applied for systematic PPS sampling of 20 establishments employing children, followed by selecting 1 child per selected establishment, \( = (7)/20 \).
4. (9) The size at or above which an establishment could be selected more than once with the PPS scheme, \( = \text{int} (1 + [8]) \).
5. (10) The size at or above which an establishment will be selected more than once with the PPS scheme, \( = \text{int} (1 + 2 * [8]) \).

Columns (1) and (2) are population means as before. Column (7) shows the total measure of size, \( \sum S_i \), for each distribution. With the PPS scheme, only establishments with size \( S_i > 0 \) get a chance to be selected; establishments containing no working children \( (S_i = 0) \) are automatically excluded. Since we take 1 working child from each
selected establishment, \(a=20\) establishments have to be selected in order to obtain a sample of \(n=20\) working children.

Column (8) shows the sampling interval to be applied if the selection is made using systematic sampling, \(I = \sum S_i/a\).

With this procedure, some large establishments may be selected more than once. This may happen if the establishment size is larger than the systematic selection interval, \(S_i > I\). This will happen with certainty if \(S_i > 2I\). These two limits are shown in columns (9) and (10), respectively. Note that in this design, one working child is taken into the sample every time an establishment is selected, i.e. the number of working children taken from an establishment equals the number of times the establishment has been selected. It is the total number of selections, rather than the number of distinct establishments, which equals the required sample size of 20.

Table 5.6 shows details of the sample with different designs applied to one of the distributions by establishment size, namely distribution ‘4’ of Table 5.2. Designs E and F are variations on Design D, all described below. For each design, the figures shown are similar to those in Table 5.4.

**Design D. PPS selection of establishments, followed by selecting 1 child per selected establishment**

The total number of establishments selected corresponds to column (7) of Table 5.5. We need to select \(a=20\) establishments in order to obtain a sample of 20 working children (1 from each establishment selected). The systematic sampling interval to be applied is \(I = \sum S_i/a\), that is, \(128.3/20 = 6.4\). The number of establishments (and the number of working children) expected to be selected from each size category is its total measure of size divided by the common sampling interval \(I\).

**Design E. PPS sampling of establishments, ‘size’ being square root of the number of child workers in the establishment, followed by inverse-PPS sampling of children within selected establishments**

This design corresponds to Equation (5.19).

For an establishment with size \(S_c\) in size category \(c\), the measure of size and the sample-take per establishment both equal \(\sqrt{S_c}\). The total measure of size for that category is \(M_c\sqrt{S_c}\), where \(M_c\) is the number of establishments in that category in the population, given in the top panel of Table 5.4. As always with PPS sampling, the number, say \(m_c\), of establishments expected to be selected from a category is proportional to its total size measure, \(m_c = M_c\sqrt{S_c}/I\). Since each selected establishment gives \(\sqrt{S_c}\) working children to the sample, the total number of children from that category is \(n_c = m_c\sqrt{S_c} = M_cS_c/I\). Looking at this equation in terms of individual establishments and summing over all establishments, we get \(20 = \sum S_i/I\), giving the sampling interval to be applied for systematic sampling as \(I = \sum S_i/20 = 6.4\), which is exactly the same as that in Design D. The sum of the \(n_c\) values over all \(c\) equals the required sample size of 20 working children.
**Design F. Same as Design E, but with sample-take per establishment rounded to nearest integer, and establishment measure of size taken as (actual size)/(sample-take) of the unit**

Sample-take per establishment has been rounded to the nearest integral to make Design E more realistic, since in practice the sample size must be an integer. The establishment size measures have been modified such that the product of sample-take and size measure equals actual size (number of child workers) of the establishment, as it does also in Design E.

**D. Restricting the maximum number of children selected from any establishment**

There can be practical or statistical reasons to limit the maximum number of child workers to be selected from any establishment. Consider a simple random sample of establishments, followed by selection of working children from each selected establishment. Designs A and B above represent two ends of possible designs. In Design A, all working children in a selected establishment are taken into the sample. In Design B we take a maximum of one working child into the sample from any establishment, irrespective of its size. In an intermediate design, we may take all working children in a selected establishment into the sample if the number of children involved does not exceed a certain limit, say $m$. If an establishment has more than $m$ child workers, then a random sample of exactly $m$ children is selected from them.
5. Sampling establishments employing children

Table 5.6. PPS of establishments yielding 20 working children: expected sample size distributions by establishment size, for various sample designs

| Distribution of establishments according to the number of children employed |
|-----------------------------|----------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| Children employed           | 0              | 1       | 2       | 3       | 4       | 5       | 6       | 7       | 8       | 9       | total   |
| % distribution              | 39.5           | 25.9    | 16.2    | 9.5     | 5.1     | 2.5     | 1.0     | 0.3     | 0.1     | 0.0     | 100.0   |

(For Distribution ‘4’ - see Table 5.2)

**Design D.** PPS selection of establishments, followed by selecting 1 child per selected establishment

<table>
<thead>
<tr>
<th>Measure of size of the 100 establishments</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>25.9</td>
</tr>
<tr>
<td>32.3</td>
</tr>
<tr>
<td>28.4</td>
</tr>
<tr>
<td>20.5</td>
</tr>
<tr>
<td>12.3</td>
</tr>
<tr>
<td>6.1</td>
</tr>
<tr>
<td>2.2</td>
</tr>
<tr>
<td>0.5</td>
</tr>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>128.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Expected number of establishments selected (required PPS sampling interval = 6.4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>4.0</td>
</tr>
<tr>
<td>5.0</td>
</tr>
<tr>
<td>4.4</td>
</tr>
<tr>
<td>3.2</td>
</tr>
<tr>
<td>1.9</td>
</tr>
<tr>
<td>0.9</td>
</tr>
<tr>
<td>0.3</td>
</tr>
<tr>
<td>0.1</td>
</tr>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>20.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Expected number of working children selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>4.0</td>
</tr>
<tr>
<td>5.0</td>
</tr>
<tr>
<td>4.4</td>
</tr>
<tr>
<td>3.2</td>
</tr>
<tr>
<td>1.9</td>
</tr>
<tr>
<td>0.9</td>
</tr>
<tr>
<td>0.3</td>
</tr>
<tr>
<td>0.1</td>
</tr>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>20.0</td>
</tr>
</tbody>
</table>

**Design E.** PPS sampling of establishments, followed by inverse PPS sampling of working children within selected establishments. “Size” = √(number of working children in establishment)

<table>
<thead>
<tr>
<th>Measure of size (MoS), and also the sample take (number of children selected) per establishment</th>
</tr>
</thead>
<tbody>
<tr>
<td>- per establishment</td>
</tr>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>1.0</td>
</tr>
<tr>
<td>1.4</td>
</tr>
<tr>
<td>1.7</td>
</tr>
<tr>
<td>2.0</td>
</tr>
<tr>
<td>2.2</td>
</tr>
<tr>
<td>2.4</td>
</tr>
<tr>
<td>2.6</td>
</tr>
<tr>
<td>2.8</td>
</tr>
<tr>
<td>3.0</td>
</tr>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>25.9</td>
</tr>
<tr>
<td>22.9</td>
</tr>
<tr>
<td>16.4</td>
</tr>
<tr>
<td>10.2</td>
</tr>
<tr>
<td>5.5</td>
</tr>
<tr>
<td>2.5</td>
</tr>
<tr>
<td>0.8</td>
</tr>
<tr>
<td>0.2</td>
</tr>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>84.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Expected number of establishments selected (with PPS sampling interval 6.4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>4.0</td>
</tr>
<tr>
<td>3.6</td>
</tr>
<tr>
<td>2.6</td>
</tr>
<tr>
<td>1.6</td>
</tr>
<tr>
<td>0.9</td>
</tr>
<tr>
<td>0.4</td>
</tr>
<tr>
<td>0.1</td>
</tr>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>13.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Expected number of working children selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>4.0</td>
</tr>
<tr>
<td>5.0</td>
</tr>
<tr>
<td>4.4</td>
</tr>
<tr>
<td>3.2</td>
</tr>
<tr>
<td>1.9</td>
</tr>
<tr>
<td>0.9</td>
</tr>
<tr>
<td>0.3</td>
</tr>
<tr>
<td>0.1</td>
</tr>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>20.0</td>
</tr>
</tbody>
</table>

**Design F.** Same as Design E, but with sample take per establishment rounded to nearest integer and taking establishment measure of size (MoS) = (actual size)/(sample take)

<table>
<thead>
<tr>
<th>Sample-take</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>2</th>
<th>2</th>
<th>2</th>
<th>2</th>
<th>3</th>
<th>3</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure of size (MoS)</td>
<td>0.0</td>
<td>1.0</td>
<td>2.0</td>
<td>1.5</td>
<td>2.0</td>
<td>2.5</td>
<td>3.0</td>
<td>2.3</td>
<td>2.7</td>
<td>3.0</td>
</tr>
<tr>
<td>- per establishment</td>
<td>0.0</td>
<td>25.9</td>
<td>32.3</td>
<td>14.2</td>
<td>10.2</td>
<td>6.2</td>
<td>3.0</td>
<td>0.7</td>
<td>0.2</td>
<td>0.0</td>
</tr>
<tr>
<td>- total in group</td>
<td>0.0</td>
<td>4.0</td>
<td>5.0</td>
<td>2.2</td>
<td>1.6</td>
<td>1.0</td>
<td>0.5</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Expected number of establishments selected (with PPS sampling interval 6.4)</td>
<td>0.0</td>
<td>4.0</td>
<td>5.0</td>
<td>4.4</td>
<td>3.2</td>
<td>1.9</td>
<td>0.9</td>
<td>0.3</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Expected number of working children selected</td>
<td>0.0</td>
<td>4.0</td>
<td>5.0</td>
<td>4.4</td>
<td>3.2</td>
<td>1.9</td>
<td>0.9</td>
<td>0.3</td>
<td>0.1</td>
<td>0.0</td>
</tr>
</tbody>
</table>

As in Design E, the product of the MoS and sample-take in Design F equals the actual size of the unit.

**Design G** in Table 5.7 is for a simple random sample of establishments with limit \( m = 3 \) on the working children to be taken from any establishment selected.

**Design H** is for a simple random sample of establishments with a more restricted limit \( m = 2 \) on the working children to be taken from any establishment selected.

In order to complete the picture, we may recall that Design B in Table 5.4 corresponds to \( m = 1 \), while all working children are included in Design A.
5.6 Small and informal sector establishments employing children

5.6.1 Child labour activity in the informal sector

Much of child labour activity takes place in the ‘informal sector’. It is important to understand the characteristic features of informal sector activity because of their direct bearing on the choice of methodology for the measurement of child labour. The exact boundary defining whether particular types of activity are included or excluded from the definition of the ‘informal sector’ may vary depending on the specific situation and data requirements. For the purpose of data collection, a broad and inclusive definition of the informal sector may be adopted. The definition may be refined or restricted for particular analytical purposes by introducing further criteria on the basis of the information collected. Typically in the situation of developing countries, the characteristic features identifying the informal sector include the following (Verma, 1992).

(1) The sector consists primarily of very small-scale units, established and operated by self-employed persons for the purpose of creating their own employment and generating income. The very small and diverse economic units are predominantly (though not exclusively) household-based but, nevertheless, as survey units are conceptually different from households or individual persons.

(2) They are mostly operated by persons working alone, with the help of unpaid family members, and possibly with a few hired workers or apprentices.
(3) Many informal sector units lack permanent and visible premises: they are located in the operator’s home, in temporary premises, or without a fixed place of business.

(4) They have very little capital and technical skills, and mostly work at low levels of organisation, productivity and income.

(5) They tend to have little or no access to organised markets and credit institutions. In fact, a major problem in many countries is the lack of public policies and resources for the support of this large and expanding sector of national economies.

(6) While some of the informal sector activities resemble traditional crafts or services, many other activities are new and arise from the conditions of modern, increasingly urbanised and marginalised existence. The existence of many such units depends on their ability to replace or complement the production of the formal sector. In some situations it is common for formal sector units to engage persons under conditions and for activities typical of the informal sector.

(7) While a vast majority of informal sector activities are legal in their objectives, they are often carried out without compliance with administrative and technical requirements for their operations. The public authorities often tolerate the existence of informal sector activities performed outside or on the fringe of laws and regulations. While informal sector units may engage in deliberately concealed activities, the majority of informal sector workers in developing countries have little to conceal and therefore can be surveyed using the direct method of enquiry used in statistical surveys of other types.

Hence the core of the informal sector consists of household enterprises, irrespective of the enterprise size, type of premise used or other similar characteristics. A household enterprise is taken to mean an unincorporated enterprise which is owned and operated by persons alone or in partnership with members of the same or other households, without employing regular employees.

In addition, according to draft international recommendations on the subject (MELS, 1992, especially Ch. 6 ‘Informal Sector Data Collection’ of the ILO report), statistical definition of the informal sector has the following elements.

(8) The informal sector is limited to unincorporated economic units, meaning economic units (i) which are not formally or legally separated from the household as an economic entity, and (ii) for which generally no separate accounts are maintained. In addition it may include private unincorporated micro-enterprises which, while employing one or more regular employees, operate on a scale below a specified level. The criterion of scale may be measured in various ways depending on the kind of economic activity performed, and the limit separating the informal from the formal sector may vary from one sector of activity to another.

(9) As a subset of unincorporated economic units, coverage of the informal sector may be limited to units which produce goods and services for exchange or sale on the market and which are exclusively or mainly engaged in sectors other than agriculture and related activities, such as hunting, forestry and fishing. From this, units which are owned and operated by self-employed professionals or technicians may be excluded, depending on national circumstances. Certain other activities not meeting the conditions of ‘informality’ may also be excluded.
The rather more ‘formal’ criteria (8) and (9) may be important in surveys focused on
the measurement of economic activity of informal sector units. For the purpose of
data collection specifically aimed at measuring the numbers and conditions of child
labour, it is more suitable to adopt, as noted above, a broad and inclusive definition of
activities in the informal sector.

5.6.2 Characteristic features of small and informal sector establishments

As noted, much of small and informal sector economic activity is household based.
While no such activity may be present in many households, where it does exist there
is often a one-to-one correspondence between economic units and households.
Consequently, the sample designs used in typical household surveys provide the point
of departure in this discussion of sampling of small-scale economic units. Given the
prevalence of one-to-one correspondence between units of the two types, households
rather than the economic units themselves may directly serve as the ultimate sampling
units.

Despite such basic similarities, however, the sampling requirements of a survey of
the small and informal sector differ, in some fundamental respects, from those of a
typical household survey of the general population. These differences arise from the
fact that we are dealing with a population of units which tend to be less stable than
typical households, are often unevenly distributed, and tend to be very heterogeneous.
Furthermore, precision requirements, and possibly even the measurement objectives,
may differ from one category of economic units to another.

(1) Heterogeneity

Household surveys are generally designed to cover the entire population uniformly.
Different subgroups (such as households by size and type, age and sex groups in the
population, social classes, etc.) are often important analysis and reporting categories
but, except possibly for geographical subdivisions, they rarely form distinct design
domains. By contrast, economic units are characterised by their heterogeneity and
by a much more uneven spatial distribution. The population is comprised of multiple
economic sectors, often with great differences in the number, distribution, size and
other characteristics of the units in different sectors. These represent diverse types of
economic activity to be captured in the survey, possibly using different questionnaires
and different data collection methodologies. Separate and detailed reporting by sector
tends to be a much more fundamental requirement than it is in the case for different
population subgroups in household surveys. The economic sectors can, and often
do, differ greatly in size (number of establishments in the population) and in sample
size (precision) requirements, and therefore in the required sampling rates. Hence
an important difference from household surveys is that economic surveys covering
establishments of different types often require major departures from a self-weighting
design.
(2) Uneven distribution

These aspects are accentuated by the uneven geographical distribution of establishments of different types. Small-scale economic activity of various types, though widely dispersed, is often far from being uniformly distributed across the population. To be reasonably efficient, the sample design must take into consideration the patterns of concentration of various types of economic activity. The patterns of concentration can be complex and differ from one type (sector) of activity to another, varying from (i) some sectors concentrated in a few areas, to (ii) some sectors widely dispersed throughout, but in between these two extremes are (iii) many unevenly distributed sectors, concentrated or dispersed to varying degrees. The pattern can also differ by size and other characteristics of the units. True, population subgroups of interest in household surveys can also differ in their distribution (such as that implied in the typology of ‘geographical, ‘mixed’ and ‘cross’ subclasses proposed by Kish, Groves and Krotky, 1976), but normally type (ii) rather than type (iii) predominates in household surveys. By contrast, often situation (iii) predominates in economic surveys. The sampling design must take into account the pattern of distribution in the population of units of different types.

(3) Increased complexity

These and other factors contribute to increased complexity of the survey design, including sample selection and estimation procedures. While such complexity in the design cannot be avoided, it remains highly desirable that it be kept to a minimum. Even more important is the requirement that the survey procedures remain as uncomplicated as possible concerning their implementation in the field. By their very nature, many operations in a survey of wide scope and coverage have to be decentralised.

5.6.3 Consequences for survey design

It has been a common practice to collect data on child labour in the informal sector largely through exploratory qualitative studies, aimed at in-depth analysis of particular aspects of the phenomenon in specific and restricted situations. Of course such studies can be valuable in understanding the conditions and nature of activities performed by children. However, their scope has been limited by factors such as: (i) coverage only of selected branches of economic activity; (ii) coverage confined to specific geographical areas; (iii) interest only in particular aspects of the activities, neglecting others; (iv) the use of restricted and non-representative sampling procedures; (v) small sample sizes; (vi) adoption of non-standard methodologies in response to particular conditions; and (vii) generally, by the one-time and ad-hoc nature of most of the enquiries. Better and more representative surveys are needed. It is necessary to develop and implement surveys capable of collecting quantitative information on the basis of representative, probability samples.

Several special requirements may be noted for surveys involving small and informal sector establishments, in particular in relation to differences from typical household-based surveys.
(1) Coverage
Good coverage is a more critical and difficult requirement in the case of non-household-based surveys, than for instance in typical household-based surveys. This is because results of the survey are usually more seriously and directly affected by coverage error, than in the case of a population-based survey for which more reliable external information is available to adjust (calibrate) the ratios and distributions obtained from the survey in estimating population totals. Such information is often lacking for economic variables.

(2) Area-based multi-stage sampling
It is necessary to begin from the observation that many surveys on child labour deal with a population of numerous and small units. For such units, no explicit lists are generally available from which a sample can be selected in a single stage. In any case, sampling in a single stage is hardly a practical option because the units are usually widely dispersed. Furthermore, in most cases, at least in the circumstances of developing countries, data have to be collected through face-to-face interviewing. Hence it is necessary to resort to multi-stage sampling which imposes a sufficient degree of clustering on the sample obtained. Typically, a multi-stage sample involves the selection of area units in one or more stages, then within areas the selection of households, establishments or other appropriate units, and finally of children at the last stage of sampling. Areas are usually selected after stratification by type of place, region, and other basic geographical or ecological variables, often with probabilities proportional to some measure of size. Areas are more stable units, and generally existing frames can be used for their selection, possibly with some updating. By contrast, the ultimate units involved are much less stable and it is therefore necessary to freshly prepare (or use very recent) lists for the last stage of selection. Common problems with area and list frames have been discussed in Chapter 4.

(3) The sampling frame
The frame of area units should provide information on the type and distribution of the sectors of economic activity, apart from information on the population and households. In most situations, no separate frames exist for sampling for surveys of establishments employing child labour. Nevertheless, large-scale censuses and surveys can provide a useful indication of the size and distribution of different sectors of child labour activity, in particular geographic strata and areas in which activities of interest may be concentrated, and strata and areas which are empty or nearly empty of them. Existing censuses and surveys often provide a basis for constructing a sampling frame for surveying child labour - they can at least provide a starting point for the process of constructing a frame.

(4) Stratification
An important objective is to isolate areas of concentration for different types of activity. This is in addition to the usual stratification of areas by urban-rural and region etc. For good stratification and selection of areas, it is necessary to have information on the number and distribution of economic units of various types in the areas. This is in
addition to the information on population size and characteristics of the type required for household surveys.

(5) Primary sampling units (PSUs) of different types

The most suitable units may differ in type and size from one survey population to another, and there may be differences in the methods of selection to ensure that the required sample sizes for different types of economic activity are achieved. For instance, while a majority of the surveys are based on area units, many involve the selection of places with concentrations of work, assembly, recreation, exchange of legal or illegal goods and services, other commercial activity, institutions providing services, etc. Also involved may be samples based on mobile locations or mobile units of various types.

(6) Listing

Several factors make the listing of households, establishments or other ultimate units in the selected areas a critical and relatively expensive operation. A large number of units may have to be listed to secure sufficient samples; information is required for screening, stratification, and differential sampling of units of different types; visible as well as hidden units must be listed to ensure good coverage; units of different types such as households and establishments may be involved in the same survey, requiring different arrangements, and so on.

(7) Last stage of sample selection

Elaborate stratification and sampling procedures and differing sampling rates may be required to deal with units of different types. An important aim of the design strategy developed in the following sections is to avoid or minimise differences in the sampling procedures required for units of different types within the same ultimate area units.

5.6.4 Most commonly used area-based design

A. Type of area unit

Often census enumeration areas (EAs) are the most suitable basic area units for the type of survey under discussion. The source may be the population or the economic census, depending on the type and quality of the information available.

What is the optimal size of the area units to be used for sampling? We can discuss this question at least in relative terms. Often it has been argued that, in comparison with normal household surveys, it is desirable to work with larger areas in a survey of small and informal sector establishments. Larger areas may have certain advantage: it is easier to find the required number of establishments of various types in larger areas; for the same reason, larger areas are more suitable where the same set of units are to be retained in an on-going survey; larger areas permit larger sample-takes per area, thereby reducing the number of areas to be taken into the sample and thereby reducing travel costs and improving supervision; and boundary errors are likely to be less problematic with larger areas thus improving coverage, which is a particularly important consideration in an establishment survey.
From a given frame of basic area units such as EAs, ‘larger’ areas can be created in two ways: by grouping geographically contiguous areas; or by introducing higher stages of selection, such as towns or villages as PSUs within which EAs are selected at the next stage. The balance of advantages can be different in the two situations.

The use of larger areas also has a number of disadvantages. Using fewer larger areas tends to reduce efficiency of the sample (larger design effects); this can be serious to the extent economic activities of the same type tend to be geographically concentrated. If the sample-take is not increased in proportion to the size of the area, then the listing cost would increase; this can be a serious matter where listing may already be a major and costly operation. A particularly serious problem with grouping areas (or introducing higher stages of sampling) is the resulting reduction in effectiveness of the primary stratification. In any case, a major part of the informal sector sample is expected to lie in urban areas, where travel cost is generally a less critical consideration, usually better maps are available, and often small pockets of high concentration of similar activities exist. All this argues for the use of small, compact areas for the purpose of sampling.

The balance between the sets of factors acting in opposite directions is likely to be very situation-specific.

B. Selection and subsampling of area units

Surveys of small establishments typically use a design with area units selected in a single stage. The common design is to select the sample separately within strata. Within each stratum, areas may be selected with probability proportional to some measure of their size, and within each selected area, establishments may be selected with probability inversely proportional to the size measure. As to what constitute suitable ‘unit size measures’ depends on specific requirements and characteristics of the survey. All working children in a selected establishment may be included in the survey, in which case they all receive the same probability of selection as their establishment; sometimes a subsample of working children may be taken from the establishment. A desirable feature is to obtain a self-weighting sample within each stratum, meaning that all analysis units (children or working children) in the stratum receive the same selection probability. This probability may differ from one sampling domain or stratum to another. In so far as the size measures used for PSU selection correspond to actual sizes of the units, the self-weighting sample also gives a constant number of ultimate units in the sample for each selected PSU. There are some common variations to this basic design, as was noted in Section 3.3, and further discussed and illustrated in Section 5.5.

5.6.5 Using existing sources of information

The following remarks are relevant not only for studying child labour in small and informal sector establishments, but apply to all sort of studies and surveys of child labour. These issues will be elaborated in Section 6.2. Here it should be emphasised that using existing sources of information can be particularly important in developing surveys of child labour in the small and informal sectors of activity because of the often very uneven geographical distribution of such activities.
Efficient design of child labour surveys requires a better utilisation of existing censuses and surveys. This involves using all available information for identifying (or at least beginning the process of identifying) pockets of concentration and patterns of distribution of the rare population of labouring children. This includes making use of available economic and agricultural censuses and surveys, as well of population census and surveys. Economic censuses can be very effective in identifying pockets of various economic activities, including those constituting rare populations. Sample surveys are less able to identify pockets of concentration, unless the samples are large and well-dispersed. Large-scale regular surveys, in particular the labour force survey, are often useful for the purpose. Household-based national child labour surveys, where available and are large enough, can also be of help.

Another potential source can be supplementary sampling frames covering working children. A supplementary list of working children, for instance from administrative or child welfare organisations, may be quite incomplete yet prove helpful in locating and surveying concentrations of working children (see Chapter 8).

5.7 Basic design issues

5.7.1 Separate single-sector versus integrated multi-sectoral surveys

In Section 5.6.2 we noted heterogeneity and uneven distribution of small and informal sector establishments among the reasons which make surveys involving such units more complex than regular household surveys.

There are a number of other factors which make the design of economic surveys more complex than that of household surveys. Complexity arises from the possibility that the ultimate units used in sample selection may not be of the same type as the units involved in data collection and analysis. Units of the two types may lack one-to-one correspondence. For instance, the ultimate sampling units may be (often are) households, each of which may represent none, one, or more than one establishment of interest. The same household may undertake different types of economic activity. Some activities may involve the participation of more than one household.

Hence, seen in terms of the ultimate sampling units (for example, households), different sectors (substantive domains) of establishments are not separate but are overlapping. This gives rise to two possible design strategies: separate sectoral surveys or multi-sectoral integrated surveys.

A. Separate sectoral designs

This refers to surveys where activity of the selected households and persons pertaining only to the sector concerned is enumerated. In using households which may be engaged in economic activities covering several sectors as sampling units, the samples for the different sectoral surveys may overlap.

Separate sectoral surveys tend to be more costly and difficult to implement. The overlap between the sectoral samples may be removed by characterising the sampling units (households) in terms of their predominating sector. This helps to make the sampling
process more manageable. However, this precludes separate sectoral surveys: if the sample for any particular sector is restricted only to the households in which that sector predominates over all other sectors, the coverage of the sector remains incomplete. An integrated design, covering different sectors simultaneously, is often preferable.

B. An integrated design

This refers to a design based on a common sample of units such as households and/or establishments, in which all sectors of activity undertaken by a selected unit would be covered simultaneously.

Sampling units of different types in an integrated design raises some important practical questions. An integrated multi-stage design implies the selection of a common sample of areas to cover units of all types in a single survey. The final sampling stage involves the selection of establishments within each selected area. In such a survey, in practice it is often costly, difficult and error-prone to identify and separate out the establishments into different sectors and apply different sampling procedures or rates by sector within the same area. Hence it is desirable, as far as possible, to absorb any differences in the sampling requirements by sector at preceding area stage(s) of sampling, so as to avoid having to treat units of different types differently at the ultimate stage of sampling. This means that, for instance, any differences in the required sampling rates for different sectors have to be achieved in the preceding stages, by distinguishing between areas according to their composition in terms of establishments of different types.

There is a target sample size to be met separately for each type or sector of units. The central requirement of such a design is to determine how the overall selection probability of establishments may be varied across area units such that the sample size requirements by economic sector are met. This has to be achieved under the constraint that the procedure for the final selection of establishments within any sample area is identical for establishments of different types. In the following sections, a practical procedure for achieving this is developed. We elaborate a useful technique which involves defining ‘strata of concentration’ classifying area units according to the predominant type(s) of economic units contained in the area, and use this structure to vary the selection probabilities in order to achieve the required sample sizes by type (sector) of the units. Various strategies for achieving this are developed, and are evaluated in terms of the efficiency of the design.

The cost of such – undoubtedly very desirable - operational simplification is the increased complexity of the design this may involve. However, as noted in Section 5.6.3, while such complexity in the design cannot be avoided, it is important that the survey procedures minimise complexity as concerns their implementation in the field since, by their very nature, many survey operations have to be decentralised and entrusted to less well-trained personnel in the field or local survey offices.

5.7.2 Stand-alone survey, versus a module attached to another survey

Another basic aspect of the design is the relationship of the survey of small and informal sector establishments to some other, larger and more regular, survey. The
most appropriate survey for this purpose is likely to be the national labour force survey, or sometimes a household-based national child labour survey.

We may identify three possible models of the relationship between the two types of survey.

Option (1):

The attachment of a small set of questions concerning the informal sector as a module onto, say, a labour force survey, making design and operational adjustments as possible but without altering the core design of the LFS in any fundamental way.

Option (2):

An integrated compromise design which incorporates the measurement of both the labour force and the small and informal sector as its basic objective.

Option (3):

A more-or-less independent or separate survey specifically designed for small and informal sector establishments potentially employing child labour, aimed at covering diverse sectors and units in an integrated manner.

Diverse factors will be involved in the choice among these models. The primary considerations are the data needs and the available resources and experience.

The modular approach, (1), is suitable when the range of data sought on the small and informal sector is relatively limited, or when resources are not available for undertaking a more comprehensive enquiry, and there is an established labour force (or similar) survey which can serve as the core.

In operational terms, the arrangements in the modular approach and a more integrated approach, (2), may be similar, but the design in the latter is more clearly a compromise to meet the requirements of both topics. The integrated approach can yield more comprehensive and better quality data. It is particularly appropriate when there is need to develop new data sources on both topics. If a labour force survey already exists, it would need to be redesigned to incorporate the measurement of small and informal sector activities. A good opportunity for that is when the LFS has to be redesigned in any case, for instance after a population census. Because of the close relationship between labour force and informal sector variables, and the compatibility between the design requirements for the two, an integrated survey can be an efficient instrument for meeting the combined data needs. In fact, some of the ‘adjustments’ required to the LFS in order to accommodate informal sector measurement are such that they could be beneficially incorporated in the LFS design in any case.

Since integration to varying degrees is possible, the distinction between the modular and integrated approaches is a matter of degree.

Limitations of the modular or integrated approach include the following:

- the risk of disrupting the design and operations of an established labour force survey;
the complexity of the procedures required to accommodate the two components;
- the difficulty in covering units of different types, possibly requiring different
  arrangements and modes of data collection;
- limitations in the extent to which a single design can be developed to meet
  multiple objectives efficiently; and
- the limitation on the detail of information which can be collected on each of
  the topics.

There are also a number of important advantages:
- firstly, the combined approach is less expensive;
- it facilitates the linkage of informal sector information with data on the wider
  labour force, at the macro as well as micro level; and
- on the basis of a continuing labour force survey, the approach is well-suited to
  yield a time-series of data on small and informal sector establishments.

A separate survey of small and informal sector establishments, option (3), can be
fully geared to the requirements of measurement of that sector. It can be designed to
cover more efficiently diverse units and different sectors of activity, and there is more
freedom to allocate the sample sizes appropriately. The information collected can be
more detailed than that in a combined operation.

In a comprehensive programme of statistics, the combined and separate approaches
can in fact be used in combination. The former can provide more regular data on
the small and informal sector, but in less detail; it may form a part of the continuing
operations based on the labour force survey. The latter may be used periodically at
longer intervals to obtain more detailed and structural data on the sector.

The drawback of stand-alone surveys is their high cost – both in terms of the budget
and technical resources.

### 5.8 Sample allocation through ‘strata of concentration’

The following sections develop a detailed solution to the basic design issue identified in
Section 5.7.1 concerning an integrated multi-sectoral survey. The issue is to design the
sample to meet sample size requirements separately for each sector, yet use a uniform
procedure for selecting establishments within each sample area, avoiding the need to
identify and distinguish between establishments of different sectors.

The procedure described below is based on Verma (2001) and Verma, Ghellini and
Betti (2010).

#### 5.8.1 Description of the basic design

Let us now consider some basic features of an area-based multi-stage sampling design
for an integrated survey covering small-scale economic units of different types. The
population comprises a number of sectors involving establishments of different types.
We assume that, on the basis of some criterion, each establishment can be assigned to one particular sector. Sample size requirements in terms of number of establishments \(n_{i,j}\) have been specified for each sector \(i\). The available sampling frame consists of area units (which form the primary sampling units in a two-stage design, or the ultimate area units in a design with multiple area stages). Information is available on size measures indicating the expected number of establishments \(N_{k,i}\) in each area unit \(k\) by sector \(i\), and hence on their total \(N_{i}\) for the area and total \(N_{k}\) for each sector. It is assumed that the above measures are scaled such that \(N_{i}\) approximates the actual number of establishments belonging to sector \(i\).

The required average overall probability for selecting establishments varies by sector:

\[
p_i = n_{i,j}/N_{i}\ .
\]  

(5.21)

However, the design requirement is that within each area \(k\) the overall selection probability, say \(g_k\), is the same for all establishments of different types (i.e. the sampling rate is uniform for establishments of different types in an area). This gives the expected number of units of sector \(i\) contributed to the sample by area \(k\) as

\[
n_{ki} = g_k N_{ki}\ .
\]  

(5.22)

Note that the reference here is to the expected value of the contribution of the area to the sample. In a multi-stage design, the actual number of units selected from any area will be zero if the area is not selected at the preceding stage(s), and generally much larger if the area has been selected.

The sample size constraints by sector require that the \(g_k\) are determined such that the relationship

\[
\sum_k g_k N_{ki} = n_{i}\ 
\]  

(5.23)

is satisfied simultaneously for all sectors \(i\) in the most efficient way. The criterion of efficiency also needs to be defined (see Section 5.9).

The sum in equation (5.23) is over all areas \(k\) in the population (and not merely the sample). As noted, the above formulation assumes that at the ultimate sampling stage, establishments within a sample area are selected at a uniform rate irrespective of the sector. This is a very desirable feature of the design in practice. It is often costly, difficult and error-prone to identify and separate out the ultimate survey units into different sectors and apply different sampling procedures or rates by sector. The preceding sampling stages are assumed to absorb any difference in the required sampling rates by sector through incorporating those in the area selection probabilities in some appropriate form.

---

18 Note that this information requirement by sector is more demanding than a single measure of population size normally required for probability proportional to size (PPS) sampling in a household survey. It is important that, especially in the context of developing countries with limited administrative sources, potential sources such as population, agricultural and economic censuses are designed to yield such information, required for efficient design of surveys of small-scale economic units during the post-censual period.
The most convenient (but also the most unlikely) situation in the application of (5.23) is when units of different types (sectors) are geographically completely segregated, i.e. when each area unit contains establishments belonging to only one particular sector. Indicating areas in the set containing establishments of sector \( i \) only (and of no other sector) by \( k(i) \), it can be seen that (5.22) is simply (5.21):

\[
g_{k(i)} = p_i = n_i / N_i. \tag{5.24}
\]

In reality the situation is more complex because area units generally contain a mixture of establishments of different types (sectors), and a simple equation like (5.24) cannot be applied. Clearly, we should inflate the selection probabilities for areas containing proportionately more establishments from sectors which need to be over-sampled, and vice versa. These considerations need to be quantified more precisely. We know of no exact or theoretical solutions to Equation (5.23), and have to seek empirical (numerical) solutions determining \( g_k \) for different areas depending on their composition in terms of establishments of different types, solutions which involve trial and error and defy strict optimisation.

### 5.8.2 Constructing and using ‘strata of concentration’ (StrCon)

#### A. The concept and notation

In order to apply variable sampling rates to achieve the required sample size by sector, it is useful to begin by classifying areas into groups on the basis of the particular sector which predominate in the area concerned. The basic idea is as follows.

For each sector, the corresponding ‘stratum of concentration’ is defined to consist of the set of area units in which that sector ‘predominate’ in the sense as defined below. One such stratum corresponds to each sector. The objective of constructing such strata is to separate out geographical areas according to an important aspect of their composition in terms of economic sectors. In order to distinguish these from ‘ordinary’ strata used for sample selection, we will henceforth refer to them as ‘StrCon’.

The following notation has been used:

\( i \) subscript identifying a sector, \( i = 1 \text{ to } l \)

\( k \) subscript identifying an area

\( N_{ki} \) number of ultimate units (establishments) of sector \( i \) in area \( k \)

\( N_k \) total number of establishments in area \( k \) (all sectors combined)

\( N_i \) total number of establishments of sector \( i \) in the population

\( A_i \) number of areas containing an establishment of sector \( i \) (areas with \( N_{ki} > 0 \))

\( B_i \) average number of establishments of sector \( i \) per area (counting only areas containing at least one such unit), \( = N_i / A_i \)

\( R_{ki} \) ‘index of relative concentration’ of sector \( i \) in area \( k \), \( = N_{ki} / B_i \)
5. Sampling establishments employing children

$j$ index identifying stratum of concentration (StrCon) of a sector, i.e. group of area units \(k\) in which sector \(i = j\) has the largest \(R_{ki}\) value - each area \(k\) belongs to one particular StrCon \(j\)

\(k(j)\) where necessary, \(k(j)\) is used to refer to any area \(k\) in a given StrCon \((j)\). Hence for instance, summed over areas in the frame for StrCon \((j)\):

\[
N_{ji} = \sum_{k(j)} N_{ki}, \quad N_{j} = \sum_{k(j)} N_{k}.
\]

\(N_{ji}\) Here \(N_{ji}\) is the total number of establishments of sector \(i\) in stratum of concentration \(j\). These amount to proportion \((N_{ji}/N_{i})\) of all establishments in sector \(i\); \((N_{ii}/N_{i})\) is the proportion of sector \(i\) establishments which are in ‘their own’ stratum of concentration, the remaining being scattered in StrCon ‘belonging’ to other sectors. The sector \((i)\) which has the largest value of \(R_{ki}\) in the area unit concerned \((k)\) is the stratum of concentration (StrCon \(j = i\) to which the area belongs. In this way, each area unit can be assigned to exactly one of the strata of concentration (apart from any ties, which may be resolved arbitrarily), the number of such strata being exactly the same as the number of sectors involved.\(^{19}\)

\(N_{j}\) is the total number of establishments in StrCon \(j\) (all sectors combined).

Note that the ‘index of relative concentration’ \(R_{ki}\) has been defined in relative terms: the number of units of a particular sector \(i\) in area \(k\) in relation to the average number of units of the same sector per area, excluding areas containing no establishments of the sector concerned. Defining this simply in terms of the absolute number of units would result in automatic over-domination by the largest sectors. Another point is that for a sector not large in overall size but concentrated within a small proportion of the areas, the average per area (including zeros) would tend to be small, making the sector likely to be the dominant one in too many areas. Hence we find it appropriate to exclude areas with \(N_{ki} = 0\) in computing the average \(B_{i}\).

B. Data by StrCon and sector (aggregated over areas)

Once the StrCon has been defined for each area unit, the basic information aggregated over areas on the numbers of units classified by sector and StrCon can be represented as in Table 5.8.

By definition, there is a one-to-one correspondence between the sectors and the StrCon. Subscript \(i\) (rows 1 to \(I\)) refers to the sector, and \(j\) (columns 1 to \(J\)) to StrCon. Generally, any sector is distributed over various StrCon. Any StrCon contains establishments from various sectors; in fact it contains all the establishments in the area units included in it. The diagonal elements \((i = j)\) predominate to the extent sectors are geographically segregated, and each tends to be concentrated within its ‘own’ StrCon.

\(^{19}\) Note that, in order to avoid unnecessary notational complexity, we have used a common notation \(N_{ki}\) to identify sector \(i\) and individual area \(k\), and \(N_{ji}\) to identify sector \(i\) and StrCon \(j\) (which is simply a grouping of areas). Subscripts \(k\) and \(j\) have been used consistently to identify, respectively, individual area units and StrCons. We use the notation \(k(j)\) only when it is necessary to identify the StrCon \((j)\) of an area \((k)\).
5.8 Sample allocation through ‘strata of concentration’

Table 5.8. Classification of units by sector and the ‘strata of concentration’

<table>
<thead>
<tr>
<th>StrCon</th>
<th>( j = 1 )</th>
<th>( \ldots )</th>
<th>( j )</th>
<th>( \ldots )</th>
<th>( \ldots )</th>
<th>( j = 1 )</th>
<th>( \text{Total} )</th>
<th>( \text{Target sample size} )</th>
<th>( \text{Achieved sample size} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sector</td>
<td>( i = 1 )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( i = j )</td>
<td>( \text{n}_1 )</td>
<td>( \text{n}_1 )</td>
<td>( \text{N}_i )</td>
<td>( \sum_j g_j N_i )</td>
</tr>
<tr>
<td>( i = j )</td>
<td>( N_{ji} )</td>
<td>( \text{n}_i )</td>
<td>( \text{n}_i )</td>
<td>( \text{n}_i )</td>
<td>( \text{n}_i )</td>
<td>( \text{n}_i )</td>
<td>( \text{n}_i )</td>
<td>( \text{N}_i )</td>
<td>( \sum_j g_j N_i )</td>
</tr>
<tr>
<td>( i = j = 1 )</td>
<td>( \text{n}_{11} )</td>
<td>( \text{n}_{11} )</td>
<td>( \text{n}_{11} )</td>
<td>( \text{n}_{11} )</td>
<td>( \text{n}_{11} )</td>
<td>( \text{n}_{11} )</td>
<td>( \text{n}_{11} )</td>
<td>( \text{N}_{11} )</td>
<td>( \sum_j g_j N_{11} )</td>
</tr>
</tbody>
</table>

Sampling rate applied: \( g_1 \), \( g_j \), \( g_l \)
Achieved sample size: \( n_1 = g_1 N_1 \), \( n_i = g_i N_i \), \( n_l = g_l N_l \), \( \sum_j g_j N_j \)

* In the ‘basic model’, described and illustrated in the following sections, it is assumed that a uniform sampling rate is applied to all area units in a given StrCon. Some more elaborate versions of the model are noted in Section 5.9.4.

C. Using StrCon for determining the sampling rates: a basic model

The use of StrCon as defined above is the fundamental aspect of the strategy for determining the selection probabilities to be applied to area units for the purpose of achieving the required sample size by sector in terms of the number of establishments, under the constraint that the procedure for sampling establishments within a selected area units is the same for establishments of different types.

We will first consider a simple model using a uniform sampling rate within each StrCon, but varied across StrCon with the objective of obtaining the required allocation by sector. This basic model will be illustrated numerically in Section 5.9.3 and some possible refinements noted.

The basic model is to apply a constant overall sampling probability \( g_j \) to all establishments in all areas \( k(j) \) belonging to StrCon \( j \). The expected sample sizes obtained with this procedure are

\[
n_{ji} = g_j N_{ji}, \quad n_j = \sum_i n_{ji} = g_j N_j.
\]  

(5.25)

The \( g_j \) values have to be determined such that the obtained sample sizes by sector \( i \)

\[
\sum_j g_j N_{ji} = \sum_j n_{ji} \approx n_i
\]

agree with the required sizes \( n_i \) simultaneously for all sectors, at least approximately.

The solution to (5.26) is trivial if all establishments are concentrated along the diagonal of Table 5.8, i.e. when each area contains establishments from only one sector: as noted in Equation (5.24), the solution is simply \( g_j = p_i = (n_i/N_i) \) for StrCon \( j = i \).
At the other extreme, when establishments of all types are uniformly dispersed across the population areas, no solution of the above type is possible: we cannot select establishments of different types at different rates simply by varying the selection probabilities at the area level, without distinguishing between establishments of different types in the same area.

In fact the simple procedure (5.26) gradually ceases to give useful results as establishments of different types become increasingly uniformly dispersed across the population. This is because satisfying it requires increasingly large differences among StrCon sampling rates, thus reducing the efficiency of the resulting sample.

Nevertheless, the basic procedure has been found to be quite useful in practice. It is common in real situations for establishments of different types to be fairly concentrated or at least unevenly distributed in the population.

### 5.9 Efficiency of the allocation procedure; possible improvements

The evaluation criterion is the effect of weights on sampling precision.

Equation (5.21), if it could be applied, gives an equal probability or self-weighting sample for each sector \(i\) meeting the sample size requirements in terms of number of establishments \(n_i\) to be included.

The constraint is that within each area \(k\) establishments of different types are to be sampled at a common area-specific rate \(g_k\) determined by (5.23). This means that the establishment probabilities of selection have to vary within the same sector depending on the area units from which those establishments come. This design is generally less efficient than a self-weighting sample within each sector.

Following some numerical illustrations in Section 5.9.3 of the basic model described below, some more flexible empirical refinements will be suggested for situations where that model appears insufficient.

#### 5.9.1 The effect of ‘random’ weights

The design effect, which measures the efficiency of a sample design compared to a simple random sample of the same size, can be decomposed under certain assumptions into two factors – respectively \(d_W\) and \((d_U d_X)\) in the notation of Section 3.6:

\[
\begin{align*}
(d_W), & \text{ the effect of sample weights, and} \\
(d_U d_X), & \text{ the effect of all other aspects of the sample design.}
\end{align*}
\]

We are concerned here with the first component – the effect of sample weights on precision which is generally to inflate variances and reduce the overall efficiency of the design. The increase in variance depends on the variability in the selection probabilities \(g_k\) or their inverse, the resulting design weights \((1/g_k)\). Given that in the present context, the weights are essentially ‘external’ or ‘arbitrary’, arising only from sample
allocation requirements, their effect on variance is estimated by equation (3.20), written here in terms of weights of the analysis units (say \( w_u \)):

\[
d^2_w = \left(1 + cv^2 (w_u)\right), \tag{5.27}
\]

where \( cv(w_u) \) is the coefficient of variation of the weights of the analysis units in the sample. The expression approximates the factor by which sampling variances are inflated, i.e. the effective sample size is reduced.

With weights \( w_u \) for individual units \( u \) in a sample of size \( n \), the above can be written as follows, the sum being over units in the sample:

\[
d^2_w = \left(\frac{\Sigma w^2_u}{n}\right) / \left(\frac{\Sigma w^2_u}{n}\right)^2, \tag{5.27a}
\]

or, for sets of \( n_k \) units with the same uniform weight \( w_k \) the above becomes:

\[
d^2_w = \left(\frac{\Sigma n_k w^2_k}{\Sigma n_k w^2_k}\right) / \left(\frac{\Sigma n_k w^2_k}{\Sigma n_k w^2_k}\right)^2. \tag{5.27b}
\]

The sums in the above equations are over units in the sample.

**Computation of \( d^2_w \) from the frame**

It is useful to write the above equations in terms of the weights of units in the population, so that different design strategies can be evaluated without actually having to draw different samples. The sums below are over units in the population:

\[
d^2_w = \left(\frac{\Sigma (1/w_u)}{N}\right) / \left(\frac{\Sigma (w_u)}{N}\right), \tag{5.28}
\]

or, for sets of \( N_{ki} \) units from a given sector \( i \) with the same uniform weight \( w_k = 1/g_k \), the above becomes:

\[
d^2_w = \left(\frac{\Sigma_k (N_{ki}/w_k)}{\Sigma_k N_{ki}}\right) / \left(\frac{\Sigma_k (N_{ki}/w_k)}{\Sigma_k N_{ki}}\right). \tag{5.29}
\]

### 5.9.2 Meeting sample size requirements

The above equations can be applied to the total population as in (5.28), or separately to each sector \( i \) as in (5.29). As noted, it is assumed that sub-sampling within any area \( k \) is identical for all sectors \( i \), implying uniform weights \( w_k \) for units of all types in the area. The average of \( d^2_w \) values over \( I \) sectors

\[
\bar{d}^2_w = \frac{\sum_i d^2_w}{I} \tag{5.30}
\]

may be taken as an overall indicator of the inflation in variance due to weighting. The objective is to minimise this indicator by appropriately choosing the \( g_k \) values satisfying the required sample size constraints (5.23).
The loss $\bar{d}_w^2$ due to variation in weights has to be balanced against how closely we are able to obtain the target sample sizes by sector. We may measure the latter by, for instance, the mean absolute deviation

$$E = \frac{\sum |\Delta n_i|}{\sum n_i},$$  \hspace{1cm} (5.31)

where $n_{i}$ is the target sample size for sector $i$, and $\Delta n_i$ is the discrepancy between this and the sample size achieved with the selection procedure being considered. Usually, as we try to satisfy the sample size constraints more precisely, the loss in precision due to weighting tends to increase, which may be more marked in certain sectors than others. Often sample size constraints have to be relaxed in order to limit the loss in efficiency due to non-uniform weights.

Incidentally, it may be noted that the loss due to weighting also has an effect on the effective sample size actually achieved in each sector. The effective sample size $n'_i$ in the presence of arbitrary weights is smaller than actual sample size $n_i$ by the factor $(1/d_{w,i}^2)$. The sectoral sample size constraints should be seen as applying to $n'_i$ rather than to $n_i$.

### 5.9.3 Numerical illustrations

Table 5.9 shows three simulated populations of establishments. The distribution of the establishments by sector is identical in the three populations – varying linearly from 5,000 in Sector 1 to 1,000 in Sector 5. The populations differ in the manner in which the establishments are assumed to cluster geographically. In Population (1) different sectors are geographically completely separated – each area unit containing establishments from only one sector.

We can construct five StrCon corresponding to the five sectors as described in preceding sections. In Population (1) these completely coincide with the sectors: StrCon1 for instance containing all Sector 1 establishments and none from any other sector; StrCon2 containing all Sector 2 establishments and none from any other sector, etc. In this case, all the cases in the sector-by-StrCon cross-tabulation lie on the diagonal. The last row of the panel shows the percentage of establishments in each StrCon which are in the diagonal cell, that is, are from the sector corresponding to the StrCon. The last column shows the percentage of establishments in each sector which are in the diagonal cell, that is are from the StrCon corresponding to the sector. In the special case of Population (1) all these figures are 100 per cent, of course.
### Table 5.9. Number of establishments, by economic sector and ‘stratum of concentration’ (three simulated populations)

<table>
<thead>
<tr>
<th>Population (1)</th>
<th>Stratum of concentration (StrCon)</th>
<th>Sector</th>
<th>StrCon1</th>
<th>StrCon2</th>
<th>StrCon3</th>
<th>StrCon4</th>
<th>StrCon5</th>
<th>Total</th>
<th>% diagonal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>5,000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5,000</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>4,000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4,000</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3,000</td>
<td>0</td>
<td>0</td>
<td>3,000</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2,000</td>
<td>0</td>
<td>2,000</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1,000</td>
<td>1,000</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td></td>
<td>5,000</td>
<td>4,000</td>
<td>3,000</td>
<td>2,000</td>
<td>1,000</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>% diagonal</td>
<td></td>
<td></td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td></td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Population (2)</th>
<th>Stratum of concentration (StrCon)</th>
<th>Sector</th>
<th>StrCon1</th>
<th>StrCon2</th>
<th>StrCon3</th>
<th>StrCon4</th>
<th>StrCon5</th>
<th>Total</th>
<th>% diagonal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>3,443</td>
<td>509</td>
<td>431</td>
<td>394</td>
<td>223</td>
<td>5,000</td>
<td>69%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>512</td>
<td>2,400</td>
<td>437</td>
<td>420</td>
<td>231</td>
<td>4,000</td>
<td>60%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>439</td>
<td>478</td>
<td>1,460</td>
<td>393</td>
<td>230</td>
<td>3,000</td>
<td>49%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>354</td>
<td>373</td>
<td>326</td>
<td>785</td>
<td>162</td>
<td>2,000</td>
<td>39%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>202</td>
<td>215</td>
<td>196</td>
<td>158</td>
<td>229</td>
<td>1,000</td>
<td>23%</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td></td>
<td>4,950</td>
<td>3,975</td>
<td>2,850</td>
<td>2,150</td>
<td>1,075</td>
<td></td>
<td>55%</td>
</tr>
<tr>
<td>% diagonal</td>
<td></td>
<td></td>
<td>70%</td>
<td>60%</td>
<td>51%</td>
<td>37%</td>
<td>21%</td>
<td></td>
<td>37%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Population (3)</th>
<th>Stratum of concentration (StrCon)</th>
<th>Sector</th>
<th>StrCon1</th>
<th>StrCon2</th>
<th>StrCon3</th>
<th>StrCon4</th>
<th>StrCon5</th>
<th>Total</th>
<th>% diagonal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1,964</td>
<td>835</td>
<td>806</td>
<td>790</td>
<td>605</td>
<td>5,000</td>
<td>39%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>703</td>
<td>1,542</td>
<td>656</td>
<td>609</td>
<td>490</td>
<td>4,000</td>
<td>39%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>559</td>
<td>554</td>
<td>1,071</td>
<td>454</td>
<td>362</td>
<td>3,000</td>
<td>36%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>387</td>
<td>365</td>
<td>309</td>
<td>680</td>
<td>259</td>
<td>2,000</td>
<td>34%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>187</td>
<td>179</td>
<td>158</td>
<td>167</td>
<td>309</td>
<td>1,000</td>
<td>31%</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td></td>
<td>3,800</td>
<td>3,475</td>
<td>3,000</td>
<td>2,700</td>
<td>2,025</td>
<td></td>
<td>37%</td>
</tr>
<tr>
<td>% diagonal</td>
<td></td>
<td></td>
<td>52%</td>
<td>44%</td>
<td>36%</td>
<td>25%</td>
<td>15%</td>
<td></td>
<td>37%</td>
</tr>
</tbody>
</table>

The proportion of cases in the diagonal are lower for the other two populations: 55 per cent in Population (2) and 37 per cent in Population (3) in our simulation. Leaving aside the overall level, the pattern across StrCon in terms of the proportion in the diagonal cell happens to be very similar in these two populations, as can be seen from the last row in each panel. However, the pattern across sectors happens to differ markedly between these two populations, as can be seen from the last column of Table 5.9 for the two populations.

Table 5.10 shows the selection of a sample (n=1000) from Population (1), assuming the target sample sizes by sector to be as shown in column titled "Required n". This is a very simple case: it involves no more than applying the required overall sampling rate for each sector to its corresponding StrCon, since the two in fact are identical in this case. The sampling rate is simply the size of the sample divided by the known
5. Sampling establishments employing children

population size of the sector (or the corresponding StrCon): for instance, 267/5,000 for sector or StrCon 1; or 119/1,000 for sector or StrCon 5.

| Sector | Required n | Achieved n by StrCon | Achieved n | $d_{w,i}^2$ | $|\Delta n_i|$ |
|--------|------------|-----------------------|------------|--------------|----------------|
|        | StrCon1    | StrCon2               | StrCon3    | StrCon4      | StrCon5        |
| 1      | 267        | 267                   | 0          | 0            | 0              | 267 1.00 0 |
| 2      | 239        | 0                     | 239        | 0            | 0              | 239 1.00 0 |
| 3      | 207        | 0                     | 0          | 207          | 0              | 207 1.00 0 |
| 4      | 169        | 0                     | 0          | 0            | 169            | 169 1.00 0 |
| 5      | 119        | 0                     | 0          | 0            | 119            | 119 1.00 0 |
| All    | 1,000      | 267                   | 239        | 207          | 169            | 119 1,000 0 |

Sampling rate 0.05 0.06 0.07 0.08 0.12
Weight 1.25 1.12 0.97 0.79 0.56

The achieved sample size by sector, shown in column titled “Achieved n”, is identical to the target sample size. Hence the absolute discrepancy $|\Delta n_i| = 0$ for all sectors. The sample is self-weighting within each domain; there is no loss from weighting, $d_{w,i}^2 = 1$ throughout.

The sample weights shown in the last row are inversely proportional to the probability of selection of establishments in the StrCon or the corresponding sector. The weights have been scaled to average 1.0 per establishment in the sample.

Table 5.11 shows the results of sample selection for Population (2), with the same required sample allocation by sector as in the previous illustration. Columns in the table correspond to iterations (0-5). The iterations were performed using the following procedure.

Applying the required overall sampling rate ($p_i$) for the sector to the StrCon $j = i$ corresponding to it provides a starting point for the iterations.

$$g_j^{(0)} = c^{(0)} p_i \text{ with } p_i = n_i/N_i, \ j = i, \ i = 1 \ to \ i.$$ (5.32)

Superscript (0) refers to the starting point, “iteration 0”. Constant $c^{(0)}$ is determined from the constraint that the total sample size $n$ is given, that is,

$$\sum_j g_j^{(0)} N_j = n.$$ (5.33)

Application of the StrCon sampling rates in Equation (5.32) gives the achieved sample sizes by sector as

$$n_t^{(0)} = \sum_j g_j^{(0)} N_{ji}.$$ (5.34)

As against $n_{t,i}$, the target sample size by sector.
The next iteration adjusts the StrCon rates to move closer to the sectoral target sample sizes. In general, the $t^{th}$ iteration is

$$g_j^{(t)} = c^{(t)} g_j^{(t-1)} \left( \frac{n_i}{n_i^{(t-1)}} \right),$$

(5.35)

with $c^{(t)}$ determined such that $\sum_j g_j^{(t)} n_j = n$.

<table>
<thead>
<tr>
<th>Stratum of concentration (StrCon)</th>
<th>Sample weight (average=1.00)</th>
<th>StrCon1</th>
<th>1.26</th>
<th>1.48</th>
<th>1.65</th>
<th>1.79</th>
<th>1.87</th>
<th>1.93</th>
</tr>
</thead>
<tbody>
<tr>
<td>StrCon2</td>
<td>1.13</td>
<td>1.28</td>
<td>1.44</td>
<td>1.57</td>
<td>1.68</td>
<td>1.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>StrCon3</td>
<td>0.97</td>
<td>1.04</td>
<td>1.13</td>
<td>1.23</td>
<td>1.32</td>
<td>1.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>StrCon4</td>
<td>0.80</td>
<td>0.73</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
<td>0.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>StrCon5</td>
<td>0.56</td>
<td>0.38</td>
<td>0.29</td>
<td>0.25</td>
<td>0.22</td>
<td>0.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ratio (StrCon1/StrCon5)</td>
<td></td>
<td>2.24</td>
<td>3.90</td>
<td>5.67</td>
<td>7.22</td>
<td>8.41</td>
<td>9.26</td>
<td></td>
</tr>
</tbody>
</table>

In the results for each iteration in Table 5.11, the weights have been scaled to average 1.0 per unit in the sample. The first panel shows the sample weights for each StrCon, and how these change by iteration. The contrast among the sectors in these weights becomes more pronounced with each iteration.

<table>
<thead>
<tr>
<th>Sector $i$</th>
<th>$d_{W,i}^2$</th>
<th>Design effect due to weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.04</td>
<td>1.20</td>
</tr>
<tr>
<td>2</td>
<td>1.04</td>
<td>1.21</td>
</tr>
<tr>
<td>3</td>
<td>1.04</td>
<td>1.22</td>
</tr>
<tr>
<td>4</td>
<td>1.05</td>
<td>1.27</td>
</tr>
<tr>
<td>5</td>
<td>1.09</td>
<td>1.52</td>
</tr>
<tr>
<td>average</td>
<td>1.05</td>
<td>1.28</td>
</tr>
</tbody>
</table>

The second panel of Table 5.11 shows $d_{W,i}^2$, the design effect (factor by which effective sample size is reduced) for each sector as a result of variation in sample weights introduced to meet the target sample sizes in Population (2). The difference between the achieved and target sample sizes by sector are shown in the third panel; the sum of absolute values of the differences as a percentage of the total sample size appears in the last row of the panel.

<table>
<thead>
<tr>
<th>Sector $i$</th>
<th>Required $n$</th>
<th>$E_i$</th>
<th>% mean deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>267</td>
<td>35</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>239</td>
<td>23</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>207</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>169</td>
<td>-21</td>
<td>-12</td>
</tr>
<tr>
<td>5</td>
<td>119</td>
<td>-42</td>
<td>-32</td>
</tr>
<tr>
<td>All</td>
<td>1000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% mean absolute deviation</td>
<td>12.5</td>
<td>8.7</td>
<td>5.9</td>
</tr>
</tbody>
</table>
Columns of the table show the results obtained by sequential application of a simple iterative procedure aimed at making the achieved sample sizes closer to the target sizes by sector. The mean absolute deviation $\bar{\delta}$ declined from over 12 per cent initially to only under 2 per cent after five iterations. On the other hand, mean design effect $\overline{d^2}$ increases from nearly 1.0 initially to around 1.5 after five iterations. This is the price to be paid for meeting the target sample sizes more closely. These variations are shown in Figure 5.3.

**Figure 5.3.** Design effect due to weights versus percentage mean absolute deviation from target sample sizes by sector. Population (2)

The corresponding results for Population (3) are shown Table 5.12 and Figure 5.4. In this population, establishments in each sector are more widely dispersed across the population, and consequently the basic procedure described above gives less satisfactory results than the previous illustration. For example, the mean absolute deviation, which is high (16 per cent) to begin with, remains quite high (8 per cent) even after five iterations. At the same time, the loss due to weighting increases more and more rapidly with iteration; starting with 1.1, it increases to 2.5 after five iterations.

Note also how the difference between StrCon sampling rates and hence the associated weights becomes more marked with iterations, the StrCom1-to-StrCom5 ratio becoming nearly 20:1 after five iterations from the initial value close to 2:1. This points to the need for seeking solutions more flexible than the basic procedure developed above. Some promising directions are outlined in the next subsection.
### Table 5.12. Sample weights, design effect due to weights, and deviation from target sample size. Results with five iterations. Population (3)

<table>
<thead>
<tr>
<th>Stratum of concentration (StrCon)</th>
<th>Sample weight (average=1.00)</th>
<th>Design effect due to weighting</th>
<th>% mean deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>StrCon1</td>
<td>1.36 1.84 2.46 3.20 4.08 5.10</td>
<td>1.07 1.26 1.52 1.83 2.20 2.61</td>
<td>15.6</td>
</tr>
<tr>
<td>StrCon2</td>
<td>1.21 1.50 1.84 2.22 2.61 3.01</td>
<td>1.06 1.22 1.44 1.69 1.98 2.28</td>
<td>13.1</td>
</tr>
<tr>
<td>StrCon3</td>
<td>1.05 1.15 1.27 1.38 1.47 1.55</td>
<td>1.06 1.21 1.40 1.63 1.89 2.16</td>
<td>11.3</td>
</tr>
<tr>
<td>StrCon4</td>
<td>0.86 0.80 0.78 0.77 0.76 0.75</td>
<td>1.07 1.23 1.45 1.72 2.01 2.35</td>
<td>10.0</td>
</tr>
<tr>
<td>StrCon5</td>
<td>0.61 0.44 0.35 0.31 0.28 0.27</td>
<td>1.10 1.36 1.71 2.11 2.57 3.06</td>
<td>9.1</td>
</tr>
<tr>
<td>ratio (StrCon1/StrCon5)</td>
<td>2.24 4.23 6.96 10.36 14.37 18.99</td>
<td>1.07 1.25 1.50 1.80 2.13 2.49</td>
<td>8.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sector</th>
<th>Required n</th>
<th>$\varepsilon_i$</th>
<th>% mean deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>267</td>
<td>55</td>
<td>41</td>
</tr>
<tr>
<td>2</td>
<td>239</td>
<td>23</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>207</td>
<td>-6</td>
<td>-7</td>
</tr>
<tr>
<td>4</td>
<td>169</td>
<td>-29</td>
<td>-24</td>
</tr>
<tr>
<td>5</td>
<td>119</td>
<td>-43</td>
<td>-34</td>
</tr>
<tr>
<td>All</td>
<td>1000</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

% mean absolute deviation 15.6 13.1 11.3 10.0 9.1 8.4
5.9.4 More flexible models: an empirical approach

As clear from the above illustrations, depending on the numbers and distribution of units of different types and on the extent to which the required sampling rates by sector differ, a basic model like equation (5.26) may be too inflexible, and may result in large variations in design weights and hence in large losses in efficiency of the design. It may even prove impossible to satisfy the sample allocation requirements in a reasonable way.

Lacking a general theoretical solution, we have tried more flexible empirical approaches for defining the sampling rates to be applied to area units in order to meet the required sample allocation, and for achieving this more efficiently.

Basically, the approach has involved supplementing (5.26) by further modifying the area selection probabilities in a more targeted fashion within individual StrCon. In place of taking the overall sampling rate $g_j$ as a constant for all area units $k(j)$ in StrCon $j$, we may introduce different rates $g_{k(j)}$ which may vary among groups of area units within $j$ or, in principle, even by individual area. The models we have tried and found empirically useful are of the form:

$$g_{k(j)} = g_j \left(1 - p_{k(j)}\right) + \left(1 + c_j h(p_{k(j)})\right)p_{k(j)}.$$

In the above, $p_{k(j)} = N_{k(i)}/N_k$, is the proportion, for a given area $k$, of establishments belonging to the sector ($i = j$) which corresponds to StrCon ($j$) of the area. This is an indicator of how ‘purely’ the area represents the StrCon it belongs to. For instance, if an area is completely homogeneous in the sense of containing establishments only
of the sector corresponding to its StrCon \( j \), then \( P_{k(j)} = 1 \). At the other extreme, if establishments are distributed over areas entirely at random, expected value of \( P_{k(j)} \) equals \( \left( \frac{N_i}{N} \right) \), \( i = j \).

Parameters \( g_j \) and \( c_j \) are constants for StrCon \( j \), and \( h(\cdot) \) indicates a function.

Model (5.36) allows the sampling rate to be varied according to the extent to which establishments in the sector corresponding to the area’s StrCon predominate in the area concerned. It allows greater flexibility in achieving the target sample sizes by sector, with a better control on the increase in variance resulting from variations in sample weights (factor \( \alpha_{w,i}^2 \), Equation 5.29). The essential motivation of model (5.32) is that by taking \( c_j h(P_{k(j)}) > 0 \) we can have, within the same StrCon, higher sampling rate \( g_{k(j)} \) in areas which have higher concentrations \( P_{k(j)} \) of the sector corresponding to that StrCon. Function \( h(\cdot) \) determines the rate at which \( g_{k(j)} \) increases with increasing \( P_{k(j)} \).

Empirical choice of parameters of a function of the type (5.36) is made on the basis of a balance between (i) meeting the target sample sizes by sector as closely as possible, and (ii) limiting the loss due to variations in sample weights. An empirical approach could involve the following steps:

- choosing a form for the function \( h(P_{k(j)}) \) in equation (5.36);
- choosing parameter \( c_j \);
- using the above in (5.23) to iteratively determine \( g_j \) values by StrCon which meet the sample allocation requirements \( n_i \) by sector;
- computing the implied losses in efficiency due to weighting \( \alpha_{w,i}^2 \) from equation (5.29) for each sector, and their average over sectors \( \bar{\alpha}_{w,i}^2 \) from equation (5.30);
- computing the mean absolute deviation between the achieved and target sample sizes, \( \bar{E} \) from equation (5.31);
- making a choice of the combination \( (\bar{\alpha}_{w,i}^2, \bar{E}) \) from the range of possible combinations obtained with the given model and parameters;
- and finally, comparing this outcome against a range of other outcomes obtained with different models and parameters \((h, c)\), and choosing the most reasonable solution from among those computed.

The objective is to identify and choose the ‘best’ model, at least among those empirically evaluated. Comparing large numbers of numerical trials points at least to the direction we should be moving in concerning the choice of the design parameters.

Here are examples of some of the forms which we have investigated and compared. We may note that in the basic model (5.26), we take \( c_j \equiv 0 \), so that \( g_{k(j)} = g_j \) same for all area units in the StrCon. A simple and useful model is to take \( c_j = c > 0 \), a constant for all areas. Concerning parameter \( h_i \), in practice we have hitherto taken its functional form to be the same for all StrCon in a given trial of the procedure, i.e. independent of particular area \( j \). Forms we have found useful include the following.
The specific examples given above are meant to be merely illustrative of some useful solutions to the basic design problem in multi-stage sampling of heterogeneous and unevenly distributed small-scale economic units. The solution depends on the nature of the population at hand, and the approach sketched above has been used by the author in designing a number of samples, covering surveys of diverse types in different situations (countries). These have included surveys of agricultural and non-agricultural small-scale units, and even a survey of schools where the objective was to control the sample allocation by ethnic group. In this last-mentioned example, it was not ethically permissible to identify and sample differentially students of different ethnic (racial) groups: all variations in the required overall sampling rates by ethnic group had to be achieved by appropriately adjusting the school selection probabilities according to prior and approximate information on the schools’ ethnic composition.

### 5.9.5 Note on sample selection procedure

In the preceding, we have mainly addressed issues concerned with meeting the *sample allocation* requirements, which of course is a fundamental aspect of the design for surveys of the kind under discussion. In describing these procedures, we have not explicitly considered any aspects of the structure of the required sample except for two: (i) that the selection of establishments involves a multi-stage design, with the final selection of establishments preceded by one or more area stages; and (ii) certain target sample sizes in terms of the number of establishments in different sectors are to be met. However, any real sample would normally also involve other complexities and requirements, such as stratification, variation in sampling rates and even different selection procedures by geographical location or type of place, selection of units using systematic sampling with variable probabilities, etc. These determine the actual *sample selection* procedures to be used.

To the extent possible, the issue of sample allocation – meeting the externally determined target sample size requirements by sector – should be kept separate from the structure and process of actual sample selection.

One technique, already described in Section 5.3.5, for meeting the sample allocation requirements automatically without complicating the sample selection procedures determined by other aspects of the design, is to *incorporate the sample allocation requirements into the size measures of units* used for PPS selection. In other words, once the $g_j$ values are determined using Equation (5.26), all unit size measures, say $m^q_k$, may be replaced by modified size measures $m'_k = g_k m_k$. Or in the more general form using Equation (5.37), once we determine $g_{k(j)}$ values, all unit size measures, say $m_{k(j)}$, may be replaced by modified size measures $m'_{k(j)} = g_{k(j)} m_{k(j)}$.

With size measures adjusted in this way, we can apply a uniform selection procedure (such as systematic PPS sampling with a constant selection interval) to *all the areas*,

<table>
<thead>
<tr>
<th>$c_j$</th>
<th>$h(\cdot)$</th>
<th>$g_{k(j)}$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_j = 0$</td>
<td>$1$</td>
<td>$g_{k(j)} = g_j$</td>
<td>constant</td>
</tr>
<tr>
<td>$c_j = c &gt; 0$</td>
<td>$= 1$</td>
<td>$g_{k(j)} = g_j (1 + c P_{k(j)})$</td>
<td>linear</td>
</tr>
<tr>
<td>$c_j = c &gt; 0$</td>
<td>$= P_{k(j)}$</td>
<td>$g_{k(j)} = g_j (1 + c P_{k(j)}^2)$</td>
<td>quadratic</td>
</tr>
</tbody>
</table>
irrespective of differences among areas in the required overall selection probabilities \( g_k \). Once the size measures are so adjusted, the required differences in the sampling rates by sector are automatically ensured and it is no longer necessary to apply the sample selection operation separately by StrCon or even by sector. As noted, the units always ‘carry with themselves’ their assigned size measure - incorporating the relative selection probabilities required for them, irrespective of details of the selection process. This automatically ensures the required allocation.

The use of adjusting size measures to simplify the selection operation is a useful and convenient device, applicable widely in sampling practice. The advantage of such separation of allocation and selection aspects is that the structure and process of selection (stratification, multiple sampling stages, sub-sampling etc.) can be determined flexibly, purely by considerations of sampling efficiency, and need not be constrained by the essentially ‘external’ requirements of sample allocation (Verma, 2008; Section 4.4, “Rescaling of size measures to facilitate sample selection”).

5.10 Selection of establishments within sample areas

In sampling small and informal sector establishments from sample areas, certain procedures requiring special care are involved, as described below.

5.10.1 Segmentation and listing

The quality of estimates of aggregates (total number of economic units, employment, child workers, etc.) from the survey depends on the completeness of coverage, which in turn depends on the quality of listing.

General problems in sampling from list frames were discussed in Section 4.3. Below we note specific aspects concerning sampling of small and informal sector establishments.

In a multi-stage sample, the ultimate units need to be listed only within the last stage area units selected into the sample. However it is generally necessary to prepare fresh lists in these areas prior to each survey to ensure that the most up-to-date situation is reflected. The period between listing and main survey interviewing should be minimised to the extent possible. At the same time it is generally desirable that the listing operation be organised separately from the main survey, ideally using different enumerators for the two operations.

Two other operations also have to be accommodated: (i) possible segmentation of the areas considered too large to be completely listed, followed by selection of a sample of segments where created; and (ii) selection of the final sample after listing.

Segmentation is a difficult operation, and in many surveys a tendency has been noticed for the selected segments to be systematically under-sized, especially if the staff involved are the same as those responsible for subsequent listing. This makes it desirable to separate listing not only from the later interviewing, but also from the earlier segmentation. The cost of adding a new operation to the survey can be considerable. Another complication with ad hoc segmentation is that its introduction...
requires a revision of the last stage sampling rates if the overall rate has to be kept unchanged. All these considerations imply that a special operation for segmentation should be introduced only if it results in a substantial saving in the listing work.

To summarise, after the selection of sample areas, there may be up to four field operations - segmentation (perhaps on a selective basis), listing, sample selection, and enumeration of the final sample. Each step needs to be controlled and checked in turn, making it highly desirable that the steps are operationally separated to the extent possible. It is clearly desirable to design the survey such that the number of steps involved is minimised.

In a survey of establishments, the listing stage has several objectives:

- to identify and produce a complete list of survey units;
- to obtain information on characteristics of the units listed so that those within the scope of the survey as defined by the eligibility criteria can be identified precisely (eligibility may be determined by a combination of several criteria);
- to obtain information for secondary stratification of the ultimate units;
- to obtain other information required for sample selection;
- and possibly, sometimes the information collected during listing may be used for producing substantive estimates for the survey population.

### 5.10.2 Listing units

In principle, a representative sample of small and informal sector economic units can be obtained by listing households and household members, or by listing establishments and activities. However, a better alternative is to organise listing in the form of a dual system combining the household and establishment approaches. The alternatives are described below in turn.

#### A. Listing through households

In a survey of small and informal sector establishments, households form the core or basic units because a large proportion of the economic units of interest consist of own-account activity with one-to-one correspondence with households. Even when a number of activities exist within the same household, it may be unavoidable (sometimes even advantageous) to treat the household as a single integral economic unit, though in general it is desirable to list each activity separately. Completeness in the identification of economic units depends on the type and detail of questions asked at the listing stage. Ideally, this should involve the listing of all household members individually, and asking information on variables such as current activity status, type of activity, status in employment, secondary activity and location of work of each member, using a definite short reference period.

However, given that the number of households to be listed can be several folds larger than the final sample to be enumerated, the cost of listing involving detailed questioning can be high. Consequently many surveys in practice use a much simpler (hence cruder) approach. For example, only a minimal set of questions (e.g. “Do any persons in this household have their own business or other activity?”), followed by
5.10 Selection of establishments within sample areas

A listing of individual activities by type) may be asked in an attempt to identify self-employed economic activity. Unfortunately, with such a simplified approach, usually a heavy price has to be paid in the form of poor quality of coverage. Some surveys have used a more elaborate form involving explicit probing using a specified list of activities, though still at the level of the whole household rather than that of individual members. A useful improvement would be to precede the above with a question on the main source of income of the household. Such a question, requiring a specific answer from each household rather than a simple ‘yes-no’ response, can be effective in identifying economic activity of household members, and whether it involves child labour.

B. Listing of establishments

The identification of economic activity exclusively through households is obviously the appropriate approach for listing household-based own-account activity. The approach is less satisfactory in dealing with informal sector economic units which are located separately from the household. Such units are often fewer in number and may require special sampling and data collection procedures. Examples are micro-economic units employing one or more hired workers, or enterprises operated jointly by more than one household, which are frequently located outside the premises of any one household. Concentration of establishments such as in the market place also falls in this category. The quality and coverage of such units can be improved by identifying and listing them using the establishment approach.

Separate identification and enumeration of establishments located outside private households, in non-residential buildings in the sample area, also has a number of additional advantages: (i) such establishments can be sampled separately, possibly using different procedures, and usually at higher rates; (ii) subsequent interviewing at the location of work is facilitated; and (iii) certain estimation complexities such as those arising from an establishment being operated by more than one household can be avoided.

A purely establishment approach is of course not suited to identifying the small own-account activity units which have no fixed location or are otherwise not visible from the outside: hence the need for a combined (dual) household-establishment approach.

C. The dual approach

The idea of the dual approach is to divide the population of units into two categories, in principle non-overlapping and exhaustive: the bulk of smaller units which are best covered through a household listing operation; and units which require special treatment and are appropriately listed using the establishment approach. To define these categories clearly, the three situations A/C in Table 5.13 have to be distinguished.
### Table 5.13. Distinguishing situations in the dual approach

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type A</td>
<td>Establishments located within the sample area, in a building or structure other than an occupied residential dwelling. The owner(s) of the establishment may or may not reside within the sample area.</td>
</tr>
<tr>
<td>Type B</td>
<td>One or more small or informal sector activities carried out within the household premises, owned and operated by persons resident in the household. Hired workers may or may not be employed. In principle, a variation is possible (Type B'): as above, but activity owned and operated by a person not resident in the household, e.g. self-employment activity carried out in someone else’s household. This is likely to be a rare case, and may be treated as just a variation under Type B.</td>
</tr>
<tr>
<td>Type C</td>
<td>All other small and informal sector activities of persons residing in the sample area, carried out without a fixed or definite location, irrespective of whether the activity is conducted within or outside the sample area.</td>
</tr>
</tbody>
</table>

To keep the categories distinct, it is important to define in operational terms what is meant by a ‘building’, ‘structure’, ‘occupied dwelling unit’ etc. In particular, it should be made clear whether kiosks, stalls and other make-shift structures are to fall under type A (having a separate fixed premises) or type C.

Listing would involve the coverage of all structures in the area, whether residential or non-residential. All residential or mixed-use buildings are covered using the *household component* of the listing operation. It identifies all households, and all informal sector activity carried out by household members, obtaining information on the sector or type of activity, and its location and type of premise if any. Information on the size (the number of regular hired workers for example), and other characteristics for the identification of economic units and for determining whether they are in-scope of the survey, is also obtained. From the list of informal activities so obtained, those of type A, i.e. conducted at a fixed location outside the household, are separated out and eliminated from the list. Instead, they are added to the second list described below if they are located within the sample area, after eliminating any duplicates in that list, of course. In this way the household component of the list covers types B and C activities. (If needed, a question can be added to include here the residual category B’ as well.)

Through coverage of all structures in the sample area the second, *establishment component* of listing identifies all establishments located in buildings other than occupied residential dwellings (type A). To these are added for completeness any missed establishments of type A located within the area which have been discovered through the household listing operation as described above.

The two components of the list can be kept apart and sampled and enumerated separately as described in Section 5.10.4B.

The main drawback of the dual approach is its higher cost. Also, special care is needed to ensure that there is no double counting or gaps in the coverage. Despite this, there is much to recommend this approach.

### 5.10.3 Cost and quality of the listing operation

The high cost of the listing operation is a concern whatever the procedure used. A number of steps can be taken to reduce the size of the listing operation, for example: reducing the number of area units selected and increasing the last stage sampling rate, even to ‘take-all’ sampling; over-representing more dense areas (i.e. areas with more economic units for a given number of households) in the sample; and in some cases...
even dropping areas with no or little economic activity from the study population. Generally, these features make the sample less efficient, i.e. increase design effects. In most cases however, some reduction in sampling efficiency is a less worrying factor than the non-sampling error resulting from poor quality of coverage. Coverage error has a direct and proportionate impact on the estimation of aggregates from the survey. The improvement of the quality (coverage) of listing should therefore be a primary concern.

5.10.4 Selecting establishments

A. Basic practical requirements

Before discussing the specifics of sampling from the dual household-establishment lists, it is useful to note several desirable features of the design in relation to the stage of selecting households or economic units within the sample areas.

From practical point of view, it is highly desirable that the process is uncomplicated. Selecting units in different sectors of activity with different rates should be avoided if at all possible, as emphasised in the preceding section. It is better and often possible to absorb any required differences within the preceding stages of sampling, which involve much smaller and better controlled operations. In any case, too much reliance cannot be placed on the classification of units, such as according to sector of activity, on the basis of often approximate information obtained during listing. This by no means precludes the use of such information for stratification of the units by sector of activity prior to sample selection.20

However, special procedures and sampling rates may be necessary for certain special categories of units, particularly types of units occurring infrequently. Examples are micro-enterprises if covered in the survey, or more generally, units covered through the establishment approach as described above. When the procedures must be varied within the same area, such variation should be minimised. When an entirely uniform procedure is not possible, one should explore before introducing any further complications whether the objectives can be reasonably met by dividing the units into only two categories for the purpose of sampling: those which are sampled using the normal or uniform procedure; and others which can all be included in the survey without the need for sampling.

It is essential that the procedures adopted do not result in departures from probability sampling. There are examples in country surveys (and even in some international documents) where the desire to achieve certain specified sample sizes for various domains resulted in the adoption of procedures which do not yield a probability sample. Given that the procedures and rates of sampling may vary by type of unit, a particularly important requirement is to ensure that records are kept of the number of units of various types listed and the number selected in each sample area, so that the sampling rates (and sampling weights to be applied at the estimation stage) can be computed.

A desirable feature of the design is good control over the sample sizes and workloads. In household survey practice, sample areas are usually selected with PPS, and there

---

20 Indeed, experience shows that it is quite simple to design forms which divide the list into columns by sector, from which a stratified systematic sample is selected easily.
is often a debate as to whether at the last stage it is preferable (i) to aim at a self-weighting sample, or (ii) at a sample with fixed workloads. For theoretical as well as practical reasons the first option is generally favoured in normal household surveys, except in ‘heavy’ or repeated surveys where strict control of interview workloads is considered a critical requirement. The balance of argument is likely to be more in favour of the second option in the case of surveys of small establishments. Variations in sample sizes and workloads are likely to be a more serious problem in such surveys; consequently, the need to control these variations is typically greater. Samples tend to depart from self-weighting in any case because of the need to cater for different sectors and diversity of units.

B. Selecting the sample from a dual establishment-household list

The two (household-based and establishment-based) components of the list can be kept apart and sampled and enumerated separately. In the household component, the ultimate sampling units are households with one or more small or informal sector economic activities. If a household is selected, all its activities can be included in the sample, even if each activity is identified and enumerated separately at the subsequent stage. For operational simplicity, it is desirable that a uniform sampling procedure is applied to all households with one or more relevant activities, without distinguishing by sector or type of activity. Generally, the convenient arrangement would be to conduct the main interview at the location of the household, irrespective of where the activity is carried out.

In the establishment component, different sampling and enumeration procedures can be used. The appropriate unit of sampling will be the establishment, rather than the household. And generally, the most convenient arrangement may be to conduct the main interview at the location of the establishment. Establishments of several types are likely to fall within this component:

- Micro-units with one or more hired workers. These are usually few in number, and may be selected at a high rate or even all taken into the sample. If sampled, stratification by type and size can be useful.
- Concentrated establishments, such as in a market place. These can be stratified by type, and if there are too many in number, they can be sampled at a low rate.
- For establishments operated by more than one household, using the establishment rather than the household as the sampling unit avoids some of the estimation complexities which would otherwise be present in such cases.
- Concerning establishments of other types, it is generally desirable for simplicity that they are sampled in the same way as the household component, the only difference being that here the sampling unit will be the establishment.
5.11 Some examples of country surveys of small and informal sector establishments

Surveys of children working in small and informal sector establishments take many different forms. For instance, they may be focussed on one or a few particular sectors, or their scope may be more general such as covering the whole informal sector. They may be ‘stand-alone’ surveys collecting a lot of detailed information on child labour, or they may be organised simply as brief modules attached to other surveys. In this section a few example are given of such surveys in developing countries.

A number of examples from baseline surveys in Bangladesh were given in Chapter 4, primarily in the context of discussion of sampling frames and frame problems. The following illustrations address some technical aspects of sampling small and informal sector establishments for child labour enquiries. The earlier examples in Chapter 4 are also relevant in this context.

In the following three illustrations, the first one from Pakistan concerns a relatively small-scale survey confined to a single sector. The example from Uganda has a broader coverage of child labour in whole of the informal sector, and involves three independently selected components: surveys of children, of household heads, and of establishment owners. The example from Bangladesh is from an integrated survey covering establishments from 45 different sectors of activity.

5.11.1 Baseline survey on child labour in glass-bangles industry, Pakistan

The Baseline Survey (BLS) in Hyderabad District of Pakistan was a part of a set of four baseline surveys in the country, the other three covering: coal mines in Chirat (Noshera) and Chakwal; tanneries in Kasur; and surgical instruments manufacturing in Sialkot (Pakistan, 2004).

The research was primarily conducted through two separate surveys, named the Baseline Survey and the School Dropout Survey. Focus group discussions and in-depth interviews with key informants and other stakeholders were also used to provide qualitative data and corroborate the findings from BLS.

The main purpose of the baseline survey in Hyderabad was to establish reliable and verifiable data on the bangles manufacturing industry in terms of the nature, magnitude, causes and consequences of the worst form of child labour. The school dropout survey was conducted in order to understand the underlying causes of the high rate of dropouts and to determine the extent of linkage between school dropout and child labour. For this survey, a control group was selected from private and public schools within the immediate vicinity of the sampled glass bangle manufacturing establishments surveyed in the BLS.

Sampling for the baseline survey

For the baseline survey, the study population was divided into 3 strata. Each stratum was further divided into establishment blocks (EB’s) and each EB into establishments.
A two-stage stratified random sampling with probability proportional to size (PPS) was employed. At first stage, establishment blocks were selected at random from each stratum with PPS; these formed the primary sampling units (PSUs). At the second stage, individual establishments were selected at random from the selected establishment blocks (i.e. the PSUs) with inverse PPS; these formed the secondary sampling units (SSUs). The sample consisted of 258 establishments randomly selected from 9 establishment blocks. All establishments were enlisted and approached. The reported response rate was 100 per cent.

**Sampling for school dropout survey**

The control group interviews were conducted at fifteen schools in immediate vicinity of sample establishments covering nine primary and six middle schools (including 3 primary and 2 middle schools for girls). Two students from each class and two teachers from each school were interviewed.

Sample sizes for the two surveys are given in Table 5.14, taken from the survey report.

<table>
<thead>
<tr>
<th>Working children survey</th>
<th>School dropouts survey</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Category</strong></td>
<td><strong>Interviews conducted</strong></td>
</tr>
<tr>
<td>Working children</td>
<td>527</td>
</tr>
<tr>
<td>Parents</td>
<td>54</td>
</tr>
<tr>
<td>Employers</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>638</strong></td>
</tr>
</tbody>
</table>

*Source: Baseline survey on child labour in glass-bangles industry, Pakistan (2004).*

**Comments**

The above example illustrates the use of two-stage sampling for drastically reducing the amount of listing of micro-units to be done. At the first stage of selection of establishment blocks, only 9 of those were selected for the listing and sampling of individual establishments. However, a very important technical point should be noted concerning the choice of such a small sample of PSUs. This small number (only 9) of sample PSUs might be sufficient for a small-scale study such as the one being illustrated. Full-scale surveys covering an extensive population – a whole country, a region, or a large city, for instance – usually require a larger sample of PSUs, say a minimum of 25-30.

The second noteworthy point concerns the procedure used for selecting the sample of schools. In Section 2.12.2A we noted two types of sample commonly used for selecting schools in the context of child labour. One is to enumerate schools attended by a sample of working children identified in the main child labour survey. The primary objective of such a survey is to study the interaction between the children’s employment and educational performance. The second type is to seek a representative sample of schools themselves. Among other things, this requires a precise specification of the population of schools which is being sampled. In the present illustration, the procedure for sampling schools is different from both the above schemes: all or a sample of
5.11 Some examples of country surveys of small and informal sector establishments

5.11.2 Child Labour and the urban informal sector, Uganda

Informal sector enterprises constitute a distinct and a growing sector in Uganda as well as in many other African economies. It is noted in the survey report that informal sector enterprises provide 60-80 per cent of manufacturing employment, and about 20 per cent contribution to the GDP per-capita. As of 1999, there were an estimated 800,000 informal sector enterprises in Uganda, employing over 1.5 million people. The sector was thought to be providing over 90 per cent of Uganda’s non-agricultural employment and to account for over 80 per cent of Uganda’s productive units. The overall outcome of the study was to generate information on the dynamics of child labour in Uganda’s urban informal sector relevant for policy formulation and implementation at national and local levels. Some substantive information from this survey (Uganda, 2004) is presented in Annex A. Here we note, as an illustration for the topic under discussion in this chapter, some details on sampling methodology of the survey.

This study covered children aged 5-17 working in the urban informal sector in four districts of Uganda. The sector comprised micro enterprises (employing less than 5 persons) and small enterprises (employing 5-20 persons). Micro enterprises are largely dependent on family labour or own-account employees. The small enterprises usually operate from a fixed location and employ more wage labour.

The study employed a mix of quantitative and qualitative tools. For the quantitative part of the survey, three populations were studied, namely; (a) the working children; (b) the households (heads) employing or using child workers; and (c) establishments (owners) who were employers of child workers. For each of these, a simple random sample was drawn. As to the sample size, quantitative data were collected from 433 working children, 249 informal sector establishments (owners), and 172 households.

**Household sample**

A list of all residential zones (enumeration areas) in the urban areas of the districts selected was prepared, and from that list, a simple random sample of 2 residential zones was drawn in Kampala while in other districts only 1 zone for each district was drawn. Using systematic sampling, one out of every four households was selected.

**Informal sector establishments sample**

From the list of the zones, a classification was made of commercial zones according to concentration of particular subsectors of informal sector activity. From each ‘sub-sector concentration areas of informal sector establishments’ so created, a sample of 1 zone was randomly selected. In each of the selected zones, all metal work centres, repair works, wood works, food processing, restaurants/hotels and bars were listed. A simple random sample of the informal businesses was then drawn using systematic sampling.
CHAPTER 5

5. Sampling establishments employing children

The working children/child labour sample

The working children were identified from three segments, namely informal sector establishments, households, and the self-employed category. Working children from households and establishments were selected randomly if more than one was found in the household or establishment. Where only one working child would be found, he/she was automatically interviewed. The self-employed category was obtained from listed areas where children were observed to be working. The areas included, among others, the streets, markets, car washing bays, ‘boda-boda’ concentration areas, and any other areas in which working children could be observed. Their selection was based on convenience sampling owing to the fact that the child labourers were highly mobile and therefore a sampling frame could not be developed.

The distribution of the three samples by district is shown in Table 5.15, quoted from the survey report.

<table>
<thead>
<tr>
<th>District</th>
<th>Number of selected working children</th>
<th>Number of selected establishments</th>
<th>Number of selected households</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kampala</td>
<td>318</td>
<td>145</td>
<td>104</td>
</tr>
<tr>
<td>Kasese</td>
<td>32</td>
<td>24</td>
<td>22</td>
</tr>
<tr>
<td>Bushenyi</td>
<td>34</td>
<td>39</td>
<td>23</td>
</tr>
<tr>
<td>Tororo</td>
<td>49</td>
<td>41</td>
<td>23</td>
</tr>
<tr>
<td>Total</td>
<td>433</td>
<td>249</td>
<td>172</td>
</tr>
</tbody>
</table>


Data Collection Instruments

Questionnaires were used for each of the key sample categories (working children, owners of informal sector establishments, and household heads). Wherever necessary, the method of data collection through survey enumeration was complimented from observation, from informal interviews, and from photographs particularly of the children’s activities and of their surroundings. Secondary data were also extensively used, making use of existing studies, policy statements and legislation.

The instrument for working children targeted all working children aged 5-17 years whether self-employed or employed by someone else. Children were interviewed wherever they were located following a specified procedure (e.g. at home, at the establishment, or on the street).

The instrument for households targeted household heads. Information was also collected on all other members of households aged 5 years and above, including on their socio-demographic profiles, through the household head. Using this instrument, it was possible to get information on both working and non-working children. This made it possible to determine prevalence of child labour, which would not be possible if the instrument were administered only to working children.

The establishment instrument, like the household instrument, targeted establishment owners. Information was collected through the owner on children aged 5-17 years
working in the establishment on a range of issues including type of work done, payment, and hazards experienced at the place of work.

Comments

The study in this illustration involves three distinct samples. The samples cover the same population (of 4 districts), but are drawn independently using different sampling frames and different designs. The samples of establishments and of households are meant to be representative samples of their respective type of units. The two samples employ different systems of stratification and generally do not overlap in the PSUs (zones) selected.

The sample of working children is a complex one, and involves a number of features which will be elaborated in subsequent chapters. It is made up of three samples coming from three different frames: one frame (and a selected sample) of establishments, one of households, and one of locations where child labourers congregate. The frames may overlap, in the sense that the same child may be present in more than one frame. This complicates the selection probabilities of children into the overall sample, on the lines discussed in Chapter 8 on multiple frames. As already noted, a part of the sample – consisting of children in self-employment – lacked a probability basis, being based on convenience sampling in the absence of a proper sampling frame.

5.11.3 Baseline survey on hazardous child labour sectors, Bangladesh

This is another example from Bangladesh baseline surveys (Bangladesh, 2006b). The survey involved a multi-sectoral design, as distinct from single-sector surveys (see Section 5.7.1 on this distinction).

The objective of this baseline survey was to identify the child workers engaged in selected sectors, to quantify the incidence and distribution of worst forms of child labour, and to generate comprehensive data on children’s activities and occupations, and the effect of their work on their education, health and moral development. The survey covered 45 targeted sectors across the country. The variety of activities covered is shown by the list in Table 5.16.

It was a probability-based large-scale survey, and was conducted in two parts based on two sampling frames, namely: (i) a list frame; and (ii) an area frame.

A. Sampling from the list frame

The Economic Census 2001-03 was used as the list frame for establishments with 10 or more workers. The list frame contained over 85,000 establishments, of which around 12,000 were in the 45 sectors covered in the survey. The list of establishments was stratified by sector of activity and geographical location. From the list frame, a total of over 2,000 sample establishments from the 45 sectors covered in the survey were selected. However, of these 2,000 sample establishments, nearly 750 establishments were not found or were thought to have closed their activity, and around 100 had shifted to another location or refused to respond. Hence only around 1,250 establishment were successfully interviewed from the 2,000 selected from the list frame.
B. Sampling from the area frame

The second part of the survey consisted of sampling of small establishments in a two-stage design, with sampling from an area frame at the first stage, and listing and sampling of establishments at the second stage. The Population Census 2001 with supplementary information (based on quick count enumeration of establishments of certain types conducted in 2005) was used as the area frame. Geographical units called mouza and mahalla formed the PSUs. The area frame excluded all establishments covered under the list frame. The area frame was stratified into one ‘special’ stratum, and four ‘general’ strata. The special stratum comprised the PSUs having the highest concentration of establishments of one of the target sectors. All of a total of 350 PSUs from this stratum were included in the survey. The four general strata were formed based on area type: rural; municipality; other urban; and ‘statistical metropolitan area’. A total of 1,400 PSUs were included in the sample from the four general strata. This gave a total of 1,750 sample PSUs based on the area frame to cover establishments which had less than 10 workers/persons and were engaged in one of the 45 targeted sectors. The final sample consisted of around 11,000 establishments.

Table 5.16. The 45 targeted hazardous child labour sectors

<table>
<thead>
<tr>
<th>MANUFACTURING</th>
<th>OTHER</th>
<th>30</th>
<th>rickshaw, van puller</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 bidies</td>
<td>15</td>
<td>brick, stone breaking</td>
<td>31 laundry</td>
</tr>
<tr>
<td>2 matches</td>
<td>16</td>
<td>textile dyeing, bleaching</td>
<td>32 automobile workshop</td>
</tr>
<tr>
<td>3 glass, glass products</td>
<td>17</td>
<td>fishing, drying fish</td>
<td>33 saw mill</td>
</tr>
<tr>
<td>4 leather footwear</td>
<td>18</td>
<td>shrimp farming/processing</td>
<td>34 tokai</td>
</tr>
<tr>
<td>5 Plastic/ rubber products</td>
<td>19</td>
<td>salt mining, refining</td>
<td>35 limestone, chalk products</td>
</tr>
<tr>
<td>6 aluminium products</td>
<td>20</td>
<td>ship breaking</td>
<td>36 milling rice, spices</td>
</tr>
<tr>
<td>7 GI sheet products</td>
<td>21</td>
<td>leather tanning, dressing</td>
<td>37 spirit, alcohol blending</td>
</tr>
<tr>
<td>8 cigarettes</td>
<td>22</td>
<td>carpentry</td>
<td>38 jute textiles</td>
</tr>
<tr>
<td>9 jarda, quivam</td>
<td>23</td>
<td>garment waste</td>
<td>39 saw mill</td>
</tr>
<tr>
<td>10 plastics</td>
<td>24</td>
<td>restaurant, tea stall</td>
<td>40 iron and steel foundry</td>
</tr>
<tr>
<td>11 soap, detergent</td>
<td>25</td>
<td>vulcanizing</td>
<td>41 casting iron, steel</td>
</tr>
<tr>
<td>12 pharmaceutical prod.</td>
<td>26</td>
<td>steel furniture painting</td>
<td>42 battery re-charging</td>
</tr>
<tr>
<td>13 cement, lime, plaster</td>
<td>27</td>
<td>metal works</td>
<td>43 poultry dressing</td>
</tr>
<tr>
<td>14 jewellery, ornaments</td>
<td>28</td>
<td>engineering workshops</td>
<td>44 welding</td>
</tr>
<tr>
<td>29 steel re-rolling mills</td>
<td>45 fireworks</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Bangladesh, 2006b.

Each of the four general strata was divided into substrata, each substratum capturing in some way the concentration of a target sector. The procedure, applied separately within each of the 4 strata, was essentially as follows. Within each stratum, the sectors were ranked according to the number of PSUs (in the frame) in which the sector had one or more establishments. From this the sector having establishments in the smallest number of PSUs was identified. Then all the PSUs which contained any (one or more) establishments from that sector were assigned to the first substratum, say SS-01. These PSUs were removed from the frame, and the process was repeated for the remaining sectors to identify SS-02. The only difference from the previous step was that in comparing the remaining sectors, any establishments those sectors had
in the PSUs already included in SS-01 were disregarded. Similarly, at the third step, the sectors still remaining were compared (disregarding any establishments they may have in SS-01 and SS-02 already defined), and the one having establishments in the smallest number of the remaining PSUs identified. All PSUs which contained any establishments from that sector were assigned to substratum SS-03. And so on.

From each sample PSU, all the establishments in the target sectors were listed, “irrespective of whether they were household or non-household establishments, fixed premises or not, and home-based or not”. Among the listed establishments, only those containing at least one child worker were identified as eligible for the survey. The report describes sampling within sample areas as follows.

“All the establishments falling under a sector were treated separately for sampling of establishments. For each sector, 4 eligible establishments were surveyed per PSU, subject to their availability. Again, for each selected establishment, two children were selected for survey subject to availability. In case the number of establishments in a PSU was more than the required number, establishments were selected by simple random sampling without replacement (SRSWOR). Otherwise all of them were selected for the survey. Similarly, for child workers, if the number of child workers in a sample establishment was more than 2, a sample of 2 child workers was selected by SRSWOR”.

C. Comments

The above illustration provides an example of a survey where the sample is made up of two parts, using different frames and different sampling designs. The two frames (and the two samples) are independent, covering different segments of the target population.

A list frame covered large establishments (employing 10 or more workers). Possible shortcomings of this type of frame were discussed in Section 4.3. Such problems tend to increase with increasing time lag between the listing and data collection operations. In the present illustration, the list frame suffered from serious problems of under-coverage and from the failure to locate and identify many units in the field: over one-third of the selected establishments could not be located.

An area-based frame covered smaller establishments (employing fewer than 10 workers). The first stage of sampling consisted of selecting a sample of area units. Potential problems and shortcomings of area frames were discussed in Section 4.2.

The second stage consisted of listing and selection of establishments within area units selected at the first stage. General problems of list frames of the type discussed in Section 4.3 can be present in this case as well, but there is more scope for controlling them through minimising the time lag between updating of the lists and data collection for the survey. This time lag needs to be shorter with smaller (and hence less stable) survey units. More specific problems of lists used for selecting small and informal sector establishments in multi-stage sampling are discussed in Section 5.10.

The procedures used for stratification and for sampling of establishments and children within sample areas are quite complex. It needs to be empirically investigated whether stratification was effective in separating out different types of areas and sectors of
activity. As to the selection of establishments and children, it is essential that details are kept of the numbers of eligible units listed and the numbers selected. Otherwise, the sampling rates applied cannot be computed, and the probability basis of the sample can be lost.

### 5.12 Remarks concerning sample implementation

#### 5.12.1 Units requiring special treatment

There are several situations requiring special treatment when the ultimate units of sampling lack one-to-one correspondence with units of observation and analysis.

**Several economic units in the same household**

As a sampling issue, this presents no special problem. If a household is selected all economic units in it can be taken into the sample, and each unit receives the same selection probability as the household to which it belongs. The problem in such situations can be the complexity of the data collection operation.

**Establishments with several types of activity**

Again, this is not a problem in terms of sampling. If a unit has been selected into the sample, details of all its activities can be enumerated. All data on the unit are given the same weight, determined according the procedure for selecting the unit. At the stage of tabulation and analysis, however, different types of activity of the same unit may appear under different classifications such as different economic sectors, depending on their characteristics. It is not necessary for these sectors to correspond with the sector or stratum from which the unit was selected, though the sampling weight is always determined in accordance with the original selection.

**Changes in the type of activity and other characteristics of the units**

The above argument applies to this situation as well. The sampling weight of a unit is always determined by how it was selected on the basis of its characteristics as determined at the time of listing. Information pertaining to the unit is always tabulated according to its characteristics as determined at the time of enumeration.

**Units with partners in different households**

Small and informal sector units may be owned or operated in partnership. Even if the overall proportion of such enterprises is small, the proportion may be significant in certain sectors of activity or in certain areas. There is no problem if all owners of a partnership reside in the same household, since in that case the unit appears in the sample with the same probability as the household. The problem arises when the owners of a partnership unit reside in different households, and the units concerned are taken into the sample on the basis of the selection of associated households.

As a sampling issue, the last mentioned situation is in fact similar to that in an economic survey, where the units of sampling and observation are generally individual
establishments, but in a proportion of the cases information can be collected only at the level of enterprises each with multiple establishments. Hence the technical issue is of a wider relevance than the present context alone. They involve sampling with multiplicity, introduced in Section 4.5 above.

5.12.2 Departures from probability sampling

Unfortunately, it is not uncommon to find examples of more or less serious departure from the standards of probability sampling in the design and implementation of surveys of small and informal sector establishments.

Several factors may contribute to this state of affairs.

(1) Often samples are not designed to ensure good control over sample size, particularly on the sample-takes in individual areas. There are also difficulties in ensuring such control due to the lack of relevant and accurate information on the numbers and characteristics of units in the sampling frame.

(2) Another contributing factor is the need to ensure that minimum sample targets are achieved for each of a number of reporting domains: the problem is not only that of overall sample size but also of the sizes required separately for different types or sectors of activity.

(3) Large variation in the sample sizes from individual areas is a related problem. Finding large and unexpected concentrations of units contributes to this loss of control, as do large and variable rates of non-response.

(4) Often decisions to adjust survey procedures to deal with these problems have to be taken ad hoc in the field, at lower levels of the survey organisation. These adjustments can result in departures from the standards of probability sampling if the procedures are not correctly formulated and controlled.

The identification of some common errors should be helpful in reducing their incidence. Here are a few examples.

(1) Sometimes it is argued that the survey timing should coincide with peak periods of the activity or activities to be covered. This can be a biased procedure. It is better to spread out the enumeration over time so as to capture seasonal and other variations.

(2) The same argument applies to substitution for activities found to be dormant at the time of the survey. The substitution of inactive units by active units obviously results in over-estimation of the extent of activity at any particular time.

(3) Sometimes quotas are fixed in order to achieve a pre-specified, fixed sample size for different categories of units. For instance, a survey may be carried on until a certain sample size is achieved and discontinued thereafter. Since the sample areas are not covered in a random order, such a procedure would normally result in a non-probability sample.

(4) Another example with similar results is provided by certain procedures for the selection of units at the last stage. For instance, in a survey a specified number of
units of a certain type may be taken from each sample area, if the area contains a sufficient number of units of the required type. Only otherwise are units from other categories taken to achieve the required quota. Clearly this does not yield a probability sample for the last mentioned categories of units.

(5) Taking fixed sample sizes rather than fixed sampling rates increases the problem, though this is sometimes unavoidable if adequate control cannot be achieved otherwise.

(6) Uncontrolled substitution for non-responding units is a common source of bias in small and informal sector surveys.

(7) Many of the problems during sample implementation arise from undue importance being given to fix sample sizes and sample-takes from individual areas, when steps have not been taken to ensure such control in the design itself. In this connection, a few important points may be noted.

- It is neither necessary nor useful to aim at absolutely fixed sample sizes. A considerable amount of variation from the ‘target’ sample size can usually be accommodated.
- Of course, beyond a certain point it becomes necessary to control such variation. Variation in sample size is often a more serious problem in surveys of small and informal sector establishments, than in certain other types of survey, because of the diversity of the sectors and units to be covered and the specific requirements for each type. A balance is required between accepting these variations and making adjustments to the sampling process at the ultimate stage.
- Often it is possible to reduce the problems of variation in sample sizes by adopting an appropriate survey design. In a survey of small heterogeneous units it is desirable, as noted earlier, to introduce the necessary variations and controls at higher stages of sampling to the extent possible, so that the need to introduce special measures at the stage of final sample selection and enumeration can be minimised.
III. Rare populations
Chapter 6
Sampling rare populations

6.1 Introduction: making use of clustering and uneven distribution

A rare population refers to a subpopulation which forms a small proportion of the total population, such as 1 per cent-10 per cent. Children working in particular types or sectors of activity may form a rare population among the total population of all children.

However, from the point of view of discussing sampling issues, it is not sufficient for a population to be small for it to be termed ‘rare’. We reserve the term ‘rare’ for small subpopulations which also present certain special sampling problems.

The characteristic feature of a rare population is that sampling the whole population with normal procedures (such as equal probability sampling of elements) does not yield samples of adequate size for the subpopulations or domains of interest because of their small size. Procedures are required for more intensive and targeted sampling. These cover several different types of situation, including the following common ones.

Rare traits. The objective is to estimate small proportions possessing certain specified traits (e.g. children who are engaged in a certain type of labour). It is not the base population which is necessarily small, but the population possessing the characteristic of interest which forms numerator of the (small) proportion.

Rare populations. Here the objective is to estimate measures (proportions, mean values, etc.) for a small population (e.g. some characteristic of children engaged in a certain type of labour). It is the base population (the denominator of the statistic) which is small.

Small areas and other small domains. Here estimates are required not for some selected subpopulations but more generally for small partitions making up the whole. Samples of any reasonable size (or even of some maximum feasible size) cannot meet these reporting requirements. The characteristic feature is the need to use less complex or up-to-date but much larger data sets from alternative sources (registers, censuses) in conjunction with data from the more complex and up-to-date but smaller sample surveys.

This chapter addresses issues concerning rare traits and rare populations. Special procedures such as small area estimation methods will not be discussed.

If a small population is confined to a particular stratum or set of areas, and the extent and location of its concentration is known, it is likely that it can be adequately sampled using the normal sampling procedures of the type suitable for drawing samples from the total population. This would also be the case when the small subpopulation – even if not physically confined to a particular location – is concentrated in a sampling frame of its own, or in a separate and known part of the frame for the total population.
Rare populations may be uniformly or widely dispersed in the general population, or may be unevenly distributed or clustered to varying degrees. In the latter case, the extent and locations of clustering may not be known for sample selection. By rare populations in the present context we mean subpopulations which are both a small proportion of the total population, and are mixed up with the total population both in terms of physical location and in the sampling frame. In surveys of different types of child labour, the rare populations of interest (working children) are generally unevenly distributed among the general population of children. Particular types or sectors of work (e.g. work in bars, tourism and entertainment sectors, as porters, at garbage dumps, etc.) tend to be confined to particular locations and time periods.

Mostly, surveys of child labour use multi-stage sampling with area-based frames, while the ultimate sampling units are usually institutions, establishments, households, or even children within selected sample areas. The rare population is often diffused over these areas with unknown patterns. A successful sampling strategy involves identifying and making use of the pockets of concentration and patterns of distribution of the rare population of interest. This chapter discusses techniques and procedures through which this may be achieved.

There are six aspects of the strategy, which are distinct but are often used in combination for greater effectiveness. These will be discussed in turn in the following sections.

1. Locating concentrations of the rare population using existing large-scale sources

Most surveys of child labour deal with a population of numerous small units. Efficient design of child labour surveys requires utilisation of existing censuses and surveys to the maximum extent possible. The requirement is to use all available information for identifying (or at least beginning the process of identifying) pockets of concentration and patterns of distribution of the rare population of labouring children.

Various useful sources for the purpose are identified in Section 6.2. This covers making use of available economic and agricultural censuses and surveys, as well of population censuses and surveys. Censuses can be useful in identifying pockets of various economic activities. Sample surveys are less able to identify pockets of concentration, unless the samples are large and well-dispersed. Administrative and other sources, especially those providing lists of working children which can supplement the available sampling frames, can be valuable.

2. Separating out near-empty clusters

The next step is to collect additional information to further the process of locating concentrations of the rare population. Generally the procedures would involve some fieldwork, but to be useful, the additional fieldwork must be quite limited.

Firstly we discuss the task of separating out empty or near-empty clusters, i.e. areas containing few members of the rare population of interest. Two specific techniques are described in Section 6.3.

- Use of sequential tests to remove strata or areas which are empty or nearly empty of the rare population of interest.
Using lists where available to separate out areas containing rare population persons recorded in the lists, from other areas which do not contain such persons. As noted by Sudman (1985), even incomplete lists can be very useful in identifying areas where the target population is located.

Numerical illustrations of gains in efficiency from separating out empty (or nearly empty) clusters are provided in Section 6.4.

(3) Partitioning the frame according to degree of concentration of the rare population

As a refinement of (2), a procedure involving screening small samples in area segments with the objective of sorting them into strata with different concentrations of the rare population is described in Section 6.5. A detailed numerical illustration is provided to explain the steps involved.

(4) Oversampling strata with higher concentrations

This makes use of the patterns of concentration identified in (1)-(3) above. For oversampling to be effective, the over-sample strata should represent significantly above-average levels of concentration, and should also account for a significant part of the rare population. Specific techniques are discussed with detailed illustrations in Section 6.6.

(5) Listing, screening and two-phase sampling

Once the targeted, disproportionate sample of areas has been obtained, the next step is to refine it and proceed with the identification and sampling of the final elements of the rare population, such as households or children. The cost of this operation depends on the extent and per unit cost of the listing and screening operations required for separating out members of the rare population from the rest of the population. If screening involves the collection of a lot of complex information and therefore screening costs are too high, the process may be broken into two or more steps – introducing multi-phase sampling. The objective of steps (1)-(4) above is to limit the listing and screening costs. The issues involved in this step are discussed in Section 6.7.

(6) Special procedures for enlarging the rare population sample size

The above procedure still may not be sufficient to achieve the required sample sizes. Special procedures can be used to increase selection probabilities of units in the rare population and hence increase the achieved sample size. In Section 6.8 we summarise one particular technique, namely cumulating information over surveys. There are a number of other procedures to increase selection probabilities of rare population units, discussed in detail in subsequent chapters: multiplicity sampling (Chapter 7), multi-frame sampling (Chapter 8), adaptive cluster sampling (Chapter 9), location sampling (Chapter 10), snowball sampling (Chapter 13), and its development called respondent-driven sampling (Chapter 14).
6.2 Locating concentrations of the rare population using existing large-scale sources

This section concerns uses and limitations of existing sources for identifying pockets of concentration and patterns of distribution of the population of labouring children. These include:

- economic and agricultural censuses and surveys;
- population censuses and surveys and;
- administrative and other sources, including sources providing lists to supplement the sampling frame.

6.2.1 Use of economic censuses and surveys

A. Potential and limitations

Possible advantages of economic censuses and surveys include the relatively low marginal cost of including some information relevant to the sectors in which children work, and the reliability and detailed nature of the information which can be obtained on characteristics of the economic units concerned.

However, a number of serious limitations also exist. First of all, censuses are usually conducted infrequently, with a gap of several years between successive censuses. They cannot meet the requirement for more frequent data. The data increasingly become out-of-date as we move away from the time of the last census. Furthermore, economic censuses and surveys are not presently conducted in all countries, or conducted on a regular basis.

While often economic censuses are comprehensive in covering establishments of different sizes and in diverse sectors of activity, common limitations of such censuses include: (i) the exclusion of establishments below a certain size; (ii) exclusion of certain sectors of economic activity; and/or (iii) exclusion of certain geographical areas. In India, for instance, while economic censuses have generally covered establishments of all sizes, they have been restricted to non-agricultural activities. Excluding very small units is more common in situations where the data collection is undertaken frequently, such as annually. Even when the coverage is extended to smaller units, it may still remain confined to units with recognisable external features, excluding numerous informal sector units which exist in other forms. Often the scope may be limited to certain sectors of the economy, such as industry. There is little scope for supplementing employment and other economic information with information on personal and household characteristics.

Some improvements are of course possible. An economic census or survey may be extended to cover all sectors of the economy, and to include small, unlisted and less visible units as well. To the extent possible, steps may be taken to improve the possibility of using conventional economic censuses and surveys as a basis for subsequent enquiries specifically aimed at child labour. This includes the construction of sampling frames for such surveys, and the collection of information on the basis of which small economic units can be identified.
6.2 Locating concentrations of the rare population using existing large-scale sources

B. Area frame from the economic census

A recent economic census may be pertinent for surveying child labour in so far as it contains information on economic activity which can be used for stratification and selection of area units. In many countries, population census enumeration areas (EAs) are used for the economic census as well. However, unless based entirely on the population census areas, an economic census area frame is usually not as complete in coverage and as good in the demarcation and mapping of area units as a population census frame. This is because of differences in the resources typically available for the two types of census.

If the time lag between the population census and economic census is long, then there can be problems of proper identification of the population census EAs for use in a subsequent economic census. Sometimes special operations are undertaken to maintain an up-to-date frame between population censuses, but this practice is not common in developing countries. There are exceptions, of course. In India for instance, a better frame for urban areas – known as the urban frame survey (UFS) blocks – is available, but still the economic census does not document fresh demarcation and mapping of areas for future use in follow-up surveys.

In some countries, economic censuses have been conducted independently of the population census in the sense that the two do not use the same set of area units; consequently, the quality of the area frame resulting from the economic census is not as good as it could be. We have already noted above a number of other common limitations in the scope and coverage of economic censuses. These factors tend to limit the usefulness of economic censuses as a source of frame for sample surveys. The way to overcome these limitations is to link the economic census with the house-listing operation of the population census and widen its scope and coverage.

6.2.2 Use of population-based censuses and surveys

A. Potential and limitations

Population-based censuses and surveys can have two major advantages. They can in principle cover all types of activity, irrespective of the sector of activity, size and location of the units, and whether it is primary or secondary activity. Secondly, economic information can be related to many other characteristics of the population and households.

However, limits on the type of information which can be included and the extent to which the design of an existing survey can be adapted to the measurement of child labour can present serious limitations to the use of these sources. Their real potential lies in the possibility of serving as the basis for the undertaking child labour surveys.

B. The population census frame

Many population censuses contain information on the numbers of persons by status in employment and often also by sector. With some idea of the proportion which children form of the total labour force in different types and sectors of activity, the above information may be useful in designing a child labour enquiry. To be usable,
such information needs to be tabulated at the level of individual area units. Much more useful is the cross-tabulation of status in employment against the sector or branch of activity at the level of individual areas. The usefulness of the population census frame is further enhanced if it contains information on the nature of economic units. Two desirable items may be mentioned in particular.

(1) Information on the number of workers (especially on the average number of hired or paid employees) in the establishment can help to separate out household-based own-account activities from micro-establishments, and these two from larger establishments. Such information is useful not only for a survey on child and adult labour, but also in the planning of economic censuses and surveys.

(2) Information on the legal status and type of enterprise can help in identifying the target population for the survey more clearly.

C. Large-scale household-based surveys

The following two are among the most important large-scale household-based surveys which can provide sampling frames for surveys of child labour:

- the labour force survey; and
- a national household-based survey of child labour.

National labour force surveys are often relatively large, established and continuing surveys. Increasingly, labour force surveys are being carried out more than once a year – twice a year, or even quarterly or monthly in some countries. This increased frequency is motivated by the need both to have up-to-date data and to monitor short-term and seasonal variations in the labour force. Regular labour force surveys often contain some information about child labour among children 14 years of age and older – though usually there is little information regarding the younger age groups. Given their size, wide coverage, and similarity in content, national labour force surveys provide a great deal of pertinent information for the design of child labour enquiries.

A national household-based survey of child labour can also provide a basis for the identification of where and to what extent child labour activities of interest are concentrated. Typically, national household-based child labour surveys are smaller and much less frequent than national labour force surveys, but their concepts, definitions, classifications, subject matter and methodology are likely to have a great deal in common with the more difficult and specialised child labour surveys addressed in this book.

6.2.3 Other sources

In addition to regular censuses and surveys, there are often other potential sources of supplementary sampling frames covering working children. For special purposes, it may be possible to use alternative sources for the frame, or at least to supplement the main census-based frames. For instance, sometimes use can be made of lists of electricity users, especially if domestic and business use can be distinguished. Certain activities may be concentrated in a few known locations which can serve as part of the frame. In certain urban areas, usable lists or registers of vendors or businesses may be available which can be aggregated over areas. A supplementary list of working
children, for instance from administrative or child welfare organisations, may be quite incomplete yet prove helpful in locating and surveying concentrations of working children.

Multiple sampling frames may overlap, meaning that some population elements appear in more than one frame. This affects the unit selection probabilities, and has to be taken into account in the estimation procedure. This is a form of sampling with multiplicity (Chapter 7). Use of multiple frames is discussed in Chapter 8.

### 6.3 Separating out near-empty clusters

#### 6.3.1 Introduction

Typically, little data are available on geographical distribution of the rare population of interest. It may be possible sometimes to identify from available census or administrative data areas which clearly do not contain any members of the rare population of interest. (An obvious – but easy – example is the identification from the census of areas which contain no private households.) This is unlikely to be sufficient, however. Additional screening is normally necessary in order to identify strata and areas which are empty or nearly empty of the rare population, so that they can be excluded from the frame for selecting a targeted sample of the population.

Sequential sampling and tests provide a useful device for the purpose of identifying empty or nearly empty strata and areas. The procedure is detailed and numerical illustrations are provided in the following subsections. Sequential sampling may be supplemented by other techniques in order to identify how the proportion of the rare class varies across areas.

#### 6.3.2 The sequential test

A procedure for identifying empty and near-empty areas which can be excluded from the survey of the rare population is based on sequential analysis. Its basics are as follows. (For a very clear exposition of the procedure, see Dixon and Massey, 1951.)

Let us take a survey of children working in private households as domestic servants as an example. Suppose we decide that $p = p_0$ is a low enough proportion of households in an area using child domestic work for the area to be excluded in a survey of such labour. The implication is that it is considered wasteful and/or unimportant to survey areas with such a rare incidence of the phenomenon being studied. And suppose that we decide that $p = p_1$ is a high enough proportion of households using child domestic work for the area to be included in a survey of such labour. For each area we have to make a decision, based on examining only a small sample of households in the area, concerning whether to keep the area in the child labour survey or to exclude it. We cannot afford to investigate all the households in the area.

Since the decision has to be based on a small random sample of households in the area, we have to accept a certain degree of uncertainty about its correctness. Let $\alpha$ be the probability (the proportion of times) that we would end up deciding to include
an area even though actually \( p = p_0 \) (i.e. even though \( p \) is too small for the area to be included). Let \( \beta \) be the probability (the proportion of times) that we would end up deciding to exclude an area even though actually \( p = p_1 \) (i.e. even though \( p \) is large enough for the area to be retained in the frame). If our hypothesis is that \( p = p_0 \) (so that the area should be excluded from the child labour survey frame), then \( \alpha \) and \( \beta \) are, respectively, the so-called ‘Type 1’ and ‘Type 2’ errors.

According to our hypothesis, \( p_0 \) is the proportion of households in the area having child domestic workers. We take a random sample of size \( m \) households and find that \( m_1 \) of those have child labourers. Given \( p_0 \), the likelihood of this happening is

\[
P_0(m,m_1) = p_0^{m_1} (1-p_0)^{m-m_1}.
\]

But if, according to the alternative hypothesis, the proportion of households in the area having child domestic workers is \( p_1 \), the likelihood of the above outcome is

\[
P_1(m,m_1) = p_1^{m_1} (1-p_1)^{m-m_1}.
\]

The hypothesis " \( p = p_0 \)" would be accepted if \( P_0(m,m_1) \) is much higher than \( P_1(m,m_1) \); hypothesis " \( p = p_1 \)" would be accepted if \( P_1(m,m_1) \) is much higher than \( P_0(m,m_1) \). It can be shown that with given levels of risk \( \alpha \) and \( \beta \), the “critical region” of the test is as follows.

1. If \( \frac{P_1(m,m_1)}{P_0(m,m_1)} \leq \left( \frac{\beta}{1-\alpha} \right) \),

the hypothesis " \( p = p_0 \)" is accepted, i.e. we decide to exclude the area from the child labour survey.

2. If \( \frac{P_1(m,m_1)}{P_0(m,m_1)} \geq \left( \frac{1-\beta}{\alpha} \right) \),

the alternative hypothesis " \( p = p_1 \)" is accepted, i.e. we decide to include the area in the child labour survey frame.

3. If \( \left( \frac{\beta}{1-\alpha} \right) < \frac{P_1(m,m_1)}{P_0(m,m_1)} < \left( \frac{1-\beta}{\alpha} \right) \),

we take another observation \( (m+1) \) and repeat the above checks. The process continues till either condition (1) or condition (2) above is satisfied, permitting us to make a definite decision about whether to exclude or include the area in the frame of the child labour survey.
6.3.3 Numerical illustrations of the sequential test

A. Example 1

Table 6.1 provides an illustration of the procedure. The hypothesis being tested is that “the area does not contain a sufficiently high proportion of households employing child domestic work, and consequently we can exclude it from the survey on the topic”. If the hypothesis is rejected, we retain the area in the child labour survey frame.

The following parameters have been used in the illustration in Table 6.1.

(1) \( p_0 = 0.10, \alpha = 0.20 \)

This means that an area would be included in the child labour sampling frame (i.e. our hypothesis would be rejected) \( \alpha = 20\% \) of the time if \( p_0 = 10\% \) of households in the area employ child domestic work. Such a high value of \( \alpha \) implies that if in doubt, we would keep an area in the survey rather than exclude it.

(2) \( p_1 = 0.20, \beta = 0.20 \)

This means that we would fail to include an area \( \beta = 20\% \) of the time if \( p_1 = 20\% \) of households in the area employ child domestic work. This is perhaps too high a value for \( \beta \), but it is useful for illustrating the procedure. In the final example we will use a smaller \( \beta \) value and see its effect.
CHAPTER 6

6. Sampling rare populations

Table 6.1. Identifying areas to retain or to leave out from the survey population (Example 1)

<table>
<thead>
<tr>
<th>Parameters</th>
<th>( p_0 )</th>
<th>( p_1 )</th>
<th>( \alpha )</th>
<th>( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.10</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
</tr>
</tbody>
</table>

\[ X(m, m_1) = \ln(p_1(m, m_1)/p_0(m, m_1)) \]  

Critical values: Lower \( \ln(\beta/(1 - \alpha)) = -1.39 \); Upper \( \ln((1 - \beta)/\alpha) = 1.39 \).

<table>
<thead>
<tr>
<th>m</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.12</td>
<td>0.69</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.24</td>
<td>0.58</td>
<td>1.39</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-0.35</td>
<td>0.46</td>
<td>1.27</td>
<td>2.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-0.47</td>
<td>0.34</td>
<td>1.15</td>
<td>1.96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-0.59</td>
<td>0.22</td>
<td>1.03</td>
<td>1.84</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-0.71</td>
<td>0.10</td>
<td>0.92</td>
<td>1.73</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-0.82</td>
<td>-0.01</td>
<td>0.80</td>
<td>1.61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-0.94</td>
<td>-0.13</td>
<td>0.68</td>
<td>1.49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>-1.06</td>
<td>-0.25</td>
<td>0.56</td>
<td>1.37</td>
<td>2.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>-1.18</td>
<td>-0.37</td>
<td>0.44</td>
<td>1.25</td>
<td>2.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>-1.30</td>
<td>-0.48</td>
<td>0.33</td>
<td>1.14</td>
<td>1.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>-1.41</td>
<td>-0.60</td>
<td>0.21</td>
<td>1.02</td>
<td>1.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>-0.72</td>
<td>-0.09</td>
<td>0.90</td>
<td>1.71</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>-0.84</td>
<td>-0.03</td>
<td>0.78</td>
<td>1.59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>-0.96</td>
<td>-0.14</td>
<td>0.67</td>
<td>1.48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>-1.07</td>
<td>-0.26</td>
<td>0.55</td>
<td>1.36</td>
<td>2.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>-1.19</td>
<td>-0.38</td>
<td>0.43</td>
<td>1.24</td>
<td>2.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>-1.31</td>
<td>-0.50</td>
<td>0.31</td>
<td>1.12</td>
<td>1.93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>-1.43</td>
<td>-0.62</td>
<td>0.19</td>
<td>1.01</td>
<td>1.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>-0.73</td>
<td>-0.08</td>
<td>0.89</td>
<td>1.70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>-0.85</td>
<td>-0.04</td>
<td>0.77</td>
<td>1.58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>-0.97</td>
<td>-0.16</td>
<td>0.65</td>
<td>1.46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>-1.09</td>
<td>-0.28</td>
<td>0.53</td>
<td>1.35</td>
<td>2.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>-1.20</td>
<td>-0.39</td>
<td>0.42</td>
<td>1.23</td>
<td>2.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>-1.32</td>
<td>-0.51</td>
<td>0.30</td>
<td>1.11</td>
<td>1.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>-1.44</td>
<td>-0.63</td>
<td>0.18</td>
<td>0.99</td>
<td>1.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>-0.75</td>
<td>-0.06</td>
<td>0.87</td>
<td>1.69</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>-0.87</td>
<td>-0.05</td>
<td>0.76</td>
<td>1.57</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>-0.98</td>
<td>-0.17</td>
<td>0.64</td>
<td>1.45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>-1.10</td>
<td>-0.29</td>
<td>0.52</td>
<td>1.33</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A row corresponds to the number of households examined so far in the area, and a column corresponds to the number employing child domestic workers among the examined households. Thus row \( m = 10 \) and column \( m_1 = 2 \) means that among the 10 households examined, 2 are found to employ child domestic labour. The value in the cells is \( X(m, m_1) \) defined above. Their critical values are functions of the chosen parameters \( (\alpha, \beta) \).

We decide to exclude the area if that value equals or falls below the lower critical limit. We retain the area in the frame if that value equals or exceeds the upper critical limit. If neither condition is met, we take more observations of households in the area, until a definite decision can be taken.
The procedure is to take a random sample of households in the area, one by one. Going
down the column marked “m” in Table 6.1 gives the number of households examined
so far in the area. Let us say we are on the 12th household (m = 12). Going along
the row from there shows the number of households (m₁) found so far who employ
child domestic workers. If we have found no such households so far (m₁ = 0), we
accept our hypothesis, i.e. decide to exclude the area from the child labour survey.
This is because \( P₁(m, m₁) = P₁(12,0) \) defined earlier (which is based on the alternative
hypothesis being true) turns out much lower than \( P₀(m, m₁) = P₀(12,0) \) (which is based on
the our hypothesis being true). To explain more precisely, the test is that

\[
X(m, m₁) = \ln(P₁(m, m₁)/P₀(m, m₁))
\]

equals or falls below the lower critical limit \( \ln(\beta/(1 - \alpha)) = -1.39 \).

The figure shown in cell \( m = 12, m₁ = 0 \) of the table is \( X(m, m₁) = -1.41 \). It falls below
the lower critical limit of -1.39.

After examining a sample of 12 households in the area and finding none employing
child domestic workers, we stop looking at more households and decide to exclude the
area from the survey.

Now suppose that among \( m = 12 \) households we had found \( m₁ = 1, 2 \text{ or } 3 \) households
employing child labour. Under our hypothesis and the particular choice of parameters,
the information is not sufficient to reach a decision and we go on to examine the next
(13th) household, and carry on to the 14th, 15th etc. until a decision can be reached.

But now suppose that among \( m = 12 \) households we had found \( m₁ = 4 \) households
employing child labour. This result is much more likely if the proportion of households
employing child labour is high \( (p₁ = 0.20) \) rather than low \( (p₀ = 0.10) \) : the ratio of
probabilities \( P₁(12,4)/P₀(12,4) = 1.83 \) is high, exceeding the upper critical limit
\( \ln((1 - \beta)/\alpha) = 1.39 \).

After examining a sample of 12 households in the area and finding as many as 4
employing child labour, we stop looking at more households, and decide to include the
area in the child labour survey.

The logs of the two critical limits are in fact two parallel straight lines in the plot of \( m₁ \)
versus \( m - m₁ \) or \( m₁ \) versus \( m \). They never meet in our illustrative Table 6.1. We have
to terminate the household searching process at some stage by making some arbitrary
choice. In the illustration the process is terminated after \( m = 30 \) households. The
choice made for this purpose was that if at this stage (after examining 30 households)
we have found fewer than 5 households employing child labour, we decide to exclude
the area – even though the probability ratio has not yet fallen below the lower critical
limit. If the number employing child labour has reached 5, we decide to include the
area – even though the probability ratio has not yet risen above the upper critical limit.

Note: “include” here means including in the frame, i.e. in the population covered in the child labour
survey. Following inclusion in the frame, whether the unit is included in the sample depends on
outcome of the random selection procedure.
B. Example 2 showing decision path

Table 6.2 provides another illustration, and Table 6.3 shows the decision path corresponding to this illustration. The choice of parameters is the same as before, except for a lower value for $p_0$:

1. $p_0 = 0.05, \alpha = 0.20,$

2. $p_1 = 0.20, \beta = 0.20.$

By lowering the $p_0$ value, we are making the process more decisive. We would include an area after examining only the first (randomly selected) household if that household is found to employ child labour. We would stop after 3 households if 2 of them (actually the 2nd and the 3rd) are found to employ child labour. We would exclude an area if after 9 households we still have not found any employing child labour. And so on.
### Table 6.2. Identifying areas to retain or to leave out from the survey population (Example 2)

<table>
<thead>
<tr>
<th>Parameters</th>
<th>(p_0)</th>
<th>(p_1)</th>
<th>(\alpha)</th>
<th>(\beta)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.05)</td>
<td>(0.20)</td>
<td>(0.20)</td>
<td>(0.20)</td>
<td></td>
</tr>
</tbody>
</table>

\[X(m, m_1) = \ln(P_1(m, m_1)/P_0(m, m_1)).\]

Critical values: Lower \(\ln(\beta/(1 - \alpha)) = -1.39\); Upper \(\ln((1 - \beta)/\alpha) = 1.39\).

<table>
<thead>
<tr>
<th>(m)</th>
<th>(m_1)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.17</td>
<td>1.39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.34</td>
<td>1.21</td>
<td>2.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-0.52</td>
<td>1.04</td>
<td>2.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-0.69</td>
<td>0.87</td>
<td>2.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-0.86</td>
<td>0.70</td>
<td>2.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-1.03</td>
<td>0.53</td>
<td>2.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-1.20</td>
<td>0.36</td>
<td>1.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-1.37</td>
<td>0.18</td>
<td>1.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>-1.55</td>
<td>0.01</td>
<td>1.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>-0.16</td>
<td>1.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>-0.33</td>
<td>1.23</td>
<td>2.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>-0.50</td>
<td>1.05</td>
<td>2.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>-0.68</td>
<td>0.88</td>
<td>2.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>-0.85</td>
<td>0.71</td>
<td>2.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>-1.02</td>
<td>0.54</td>
<td>2.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>-1.19</td>
<td>0.37</td>
<td>1.92</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>-1.36</td>
<td>0.19</td>
<td>1.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>-1.54</td>
<td>0.02</td>
<td>1.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>-0.15</td>
<td>1.41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>-0.32</td>
<td>1.24</td>
<td>2.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>-0.49</td>
<td>1.07</td>
<td>2.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>-0.66</td>
<td>0.89</td>
<td>2.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>-0.84</td>
<td>0.72</td>
<td>2.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>-1.01</td>
<td>0.55</td>
<td>2.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>-1.18</td>
<td>0.38</td>
<td>1.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>-1.35</td>
<td>0.21</td>
<td>1.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>-1.52</td>
<td>0.03</td>
<td>1.59</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>-0.14</td>
<td>1.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>-0.31</td>
<td>1.25</td>
<td>2.81</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>-0.48</td>
<td>1.08</td>
<td>2.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

See notes to Table 6.1.

Table 6.3 clarifies further the decision process corresponding to data in the previous table by taking examples of four areas (A)-(D). In the left hand panel the four columns show the result in each area of examining a sample of households one by one. Either we find children being employed in the household (1) or do not (0). In the panel on the right, the position of an area moves down one row after each household examined. If the household is found to have child labour, the position of the area is moved one cell to the right; otherwise, we continue down the column.
### Table 6.3. Classification of areas according to outcome in a sample of households (Example 2)

<table>
<thead>
<tr>
<th>Number of hhs examined in the area (m)</th>
<th>Whether employs child domestic labour (1=yes)</th>
<th>Number of households in area (A-D) found to employ child domestic labour (m₁)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>(B) (C) (D)</td>
<td>(B) 1.39</td>
</tr>
<tr>
<td>1</td>
<td>0 1 0 0 (A,C,D)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0 (Y) 0 0 (A,C,D)</td>
<td>2.77</td>
</tr>
<tr>
<td>3</td>
<td>0 1 1 (A) (C,D)</td>
<td>2.60</td>
</tr>
<tr>
<td>4</td>
<td>0 0 0 (A) (C,D)</td>
<td>2.43</td>
</tr>
<tr>
<td>5</td>
<td>0 0 0 (A) (C,D)</td>
<td>2.26</td>
</tr>
<tr>
<td>6</td>
<td>0 0 0 (A) (C,D)</td>
<td>2.09</td>
</tr>
<tr>
<td>7</td>
<td>0 0 0 (A) (C,D)</td>
<td>1.91</td>
</tr>
<tr>
<td>8</td>
<td>0 0 0 (A) (C,D)</td>
<td>1.74</td>
</tr>
<tr>
<td>9</td>
<td>0 0 0 (A) -1.55 (C,D)</td>
<td>1.57</td>
</tr>
<tr>
<td>10</td>
<td>(X) 0 0 (C,D)</td>
<td>1.40</td>
</tr>
<tr>
<td>11</td>
<td>0 0 (C,D)</td>
<td>2.78</td>
</tr>
<tr>
<td>12</td>
<td>0 1 (C) (D)</td>
<td>2.61</td>
</tr>
<tr>
<td>13</td>
<td>0 0 (C) (D)</td>
<td>2.44</td>
</tr>
<tr>
<td>14</td>
<td>0 0 (C) (D)</td>
<td>2.27</td>
</tr>
<tr>
<td>15</td>
<td>0 1 (C) (D)</td>
<td>2.10</td>
</tr>
<tr>
<td>16</td>
<td>0 (Y) (C)</td>
<td>1.92</td>
</tr>
<tr>
<td>17</td>
<td>0 (C)</td>
<td>1.75</td>
</tr>
<tr>
<td>18</td>
<td>0 (C) -1.54 (C,D)</td>
<td>1.58</td>
</tr>
<tr>
<td>19</td>
<td>(X)</td>
<td>1.41</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>2.80</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td>2.62</td>
</tr>
<tr>
<td>22</td>
<td></td>
<td>2.45</td>
</tr>
<tr>
<td>23</td>
<td></td>
<td>2.28</td>
</tr>
<tr>
<td>24</td>
<td></td>
<td>2.11</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>1.94</td>
</tr>
<tr>
<td>26</td>
<td></td>
<td>1.76</td>
</tr>
<tr>
<td>27</td>
<td></td>
<td>-1.52</td>
</tr>
<tr>
<td>28</td>
<td></td>
<td>1.59</td>
</tr>
<tr>
<td>29</td>
<td></td>
<td>1.42</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>-0.48</td>
</tr>
</tbody>
</table>

For example in area (A), we examined 9 households and found none with child labour (all zeros in the column for this area in the left panel). In the panel on the right, we continued to move down the same column until we hit the lower critical boundary after 9 households. We stop the search there and decide to exclude the area. (The exclusion is indicated by an ‘X’ at the end of the column concerned.)

In area (B), the very first household examined employed child labour. In the panel on the right we move one step to the right in view of this outcome, and hit the upper
CHAPTER 6

6.3 Separating out near-empty clusters

critical boundary. The area is retained. (‘Y’ at the end of the column indicates that the area is retained in the frame for sample selection.)

In both areas (C) and (D), we found the 3rd household to contain child labour. Corresponding to this we move for both the areas a step to the right in the right-hand panel. Thereafter no more households are found with child labour until the 12th household, where child labour is found in area (D) but not in area (C). For (D) we move a step to the right, and once again after the 15th household when another case with child labour is found. With this we hit the upper boundary, and the area is retained. By contrast in area (C), no more households employing child labour are found. We continue down the same column and hit the lower critical boundary after the 18th household. The area is excluded.

C. Simplification: examining households in batches

Although it is convenient to move by examining only one household at a time and deciding how to proceed at each step, this may be too complex a task for fieldworkers, also too disruptive, wasteful and error prone. It is more practical to examine a batch of households at a time, record the results as described above and make a decision on the basis of the results for the whole batch whether to (i) keep the area, (ii) exclude it, or (iii) continue and examine the next batch of households. In practice it may be preferable to simplify the decision criteria, even if the theoretical critical limits are not being met exactly. Table 6.4A provides a practical example.22

With the assumed values of parameters \( \alpha = 0.20 \) and \( \beta = 0.10 \), we have

Lower critical value \( \ln(\beta/(1 - \alpha)) = -2.08; \)

Upper critical value \( \ln((1 - \beta)/\alpha) = 1.50. \)

A darkly shaded cell in Table 6.4A is the first cell in a given column where the tabulated value \( X(m, m_1) \) equals or falls below the lower critical limit. In the normal application of the procedure, the search is terminated if we reach such a cell, and we decide to exclude the area from the frame. In lightly shaded cells, \( X(m, m_1) \) equals or exceeds the upper critical limit. The search is terminated if we reach any such cell, and we decide to retain the area from the frame. Reaching an unshaded cell means that sufficient information is still lacking to determine whether the area should be excluded or included, and the search has to continue.

An examination of the above results suggests that the decision procedure can be approximated by the fieldworkers selecting and dealing with a batch of 10 households at a time, and deciding after each batch on whether or not it is necessary to continue the search. Four batches are indicated by the boxed rows (\( m = 10, 20, 30 \) and 40) in Table 6.4A. The decision procedure is as follows. It is summarised in Table 6.4B below.

1. If in the first batch no households are found with child labour, the search is terminated and the area is dropped. (Note that the actual critical limit for dropping

---

22 If we follow the decision criteria strictly, working in batches would increase the average number of households we need to examine.
the area in this case is 13 households with no child labour, but we have changed it to 10 households in the above example in order to simplify the procedure.)

(2) If 2 or more households in the first batch of 10 are found to have child labour, the search is terminated and the area is retained for the child labour survey.

(3) If only one of the 10 households is found with child labour, we have to select and examine a second batch of 10 households.

(4) The results after the second batch are examined in similar fashion. The area is excluded if no more households are found with child labour in the second batch (i.e. in the two batches combined, only 1 household has been found with child labour). The area is retained if in the two batches combined we have found 3 or more households with child labour. But we must continue with the third batch of 10 households if the combined first two batches have yielded exactly 2 households employing child domestic workers.

(5) At some stage the process has to be terminated on the basis of some arbitrary criterion if necessary. In the illustration, we stop after four batches: dropping an area if fewer than 5 households out of 40 have been found to contain child labour, and keeping the area if 5 or more households with child labour have been found.
### Table 6.4A. Examining households in batches (Example 3)

<table>
<thead>
<tr>
<th>Parameters</th>
<th>$p_0$</th>
<th>$p_1$</th>
<th>$\alpha$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.05</td>
<td>0.20</td>
<td>0.20</td>
<td>0.10</td>
</tr>
</tbody>
</table>

$X(m, m_1) = \ln(P_1(m, m_2)/P_0(m, m_2))$.

Critical values: Lower $\ln(\beta/(1 - \alpha)) = -2.08$; Upper $\ln((1 - \beta)/\alpha) = 1.50$.

<table>
<thead>
<tr>
<th>$m_1$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.17</td>
<td>1.39</td>
<td>2.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.34</td>
<td>1.21</td>
<td>2.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-0.52</td>
<td>1.04</td>
<td>2.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-0.69</td>
<td>0.87</td>
<td>2.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-0.86</td>
<td>0.70</td>
<td>2.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-1.03</td>
<td>0.53</td>
<td>2.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-1.20</td>
<td>0.36</td>
<td>1.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-1.37</td>
<td>0.18</td>
<td>1.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>-1.55</td>
<td>0.01</td>
<td>1.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>-1.72</td>
<td>-0.16</td>
<td>1.40</td>
<td>2.96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>-1.89</td>
<td>-0.33</td>
<td>1.23</td>
<td>2.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>-2.06</td>
<td>-0.50</td>
<td>1.05</td>
<td>2.61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>-2.23</td>
<td>-0.68</td>
<td>0.88</td>
<td>2.44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>-0.85</td>
<td>0.71</td>
<td>2.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>-1.02</td>
<td>0.54</td>
<td>2.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>-1.19</td>
<td>0.37</td>
<td>1.92</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>-1.36</td>
<td>0.19</td>
<td>1.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>-1.54</td>
<td>0.02</td>
<td>1.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>-1.71</td>
<td>-0.15</td>
<td>1.41</td>
<td>2.97</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>-1.88</td>
<td>-0.32</td>
<td>1.24</td>
<td>2.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>-2.05</td>
<td>-0.49</td>
<td>1.07</td>
<td>2.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>-2.22</td>
<td>-0.66</td>
<td>0.89</td>
<td>2.45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>-0.84</td>
<td>0.72</td>
<td>2.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>-1.01</td>
<td>0.55</td>
<td>2.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>-1.18</td>
<td>0.38</td>
<td>1.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>-1.35</td>
<td>0.21</td>
<td>1.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>-1.52</td>
<td>0.03</td>
<td>1.59</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>-1.70</td>
<td>-0.14</td>
<td>1.42</td>
<td>2.98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>-1.87</td>
<td>-0.31</td>
<td>1.25</td>
<td>2.81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>-2.04</td>
<td>-0.48</td>
<td>1.08</td>
<td>2.64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>-2.21</td>
<td>-0.65</td>
<td>0.91</td>
<td>2.46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>-0.82</td>
<td>0.73</td>
<td>2.29</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>-1.00</td>
<td>0.56</td>
<td>2.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>-1.17</td>
<td>0.39</td>
<td>1.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>-1.34</td>
<td>0.22</td>
<td>1.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>-1.51</td>
<td>0.05</td>
<td>1.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>-1.68</td>
<td>-0.13</td>
<td>1.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>-1.86</td>
<td>-0.30</td>
<td>1.26</td>
<td>2.82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>-2.03</td>
<td>-0.47</td>
<td>1.09</td>
<td>2.65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>-2.20</td>
<td>-0.64</td>
<td>0.92</td>
<td>2.47</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 6.4B. Examining households in batches of 10 households from the area – summary

<table>
<thead>
<tr>
<th>Outcome $m_1 =$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5+</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st batch, $m = 1 – 10$</td>
<td>X</td>
<td>?</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>2nd batch, $m = 11 – 20$</td>
<td>X</td>
<td>?</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>3rd batch, $m = 21 – 30$</td>
<td>X</td>
<td>?</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th batch, $m = 31 – 40$</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“X” means terminate the search and exclude the area from the sampling frame
“?” means continue the search by examining the next batch of 10 households in the area
“Y” means terminate the search and retain the area in the sampling frame

D. Operating characteristic function of the procedure

Finally, Figure 6.1 shows the ‘operating characteristic function’ of the procedure used. What is the probability ($h_p$) that we would retain an area if proportion ($P$) of households in the area contain child labour? (Here $P$ refers to the actual proportion in the whole area in question, and not in our sample of households from the area.) If $P$ is small, there is a low chance ($h_p$) that the area would be retained; for large $P$ the chance ($h_p$) of retaining the area is also high. The following table gives the relationship between $P$ and ($h_p$).

<table>
<thead>
<tr>
<th>$P$</th>
<th>($h_p$)</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>If no child labour exists in an area, it will always be rejected from a child labour survey</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>If all households in an area contain child labour, the area will be always accepted</td>
</tr>
<tr>
<td>$p_o$</td>
<td>$\alpha$</td>
<td>$\alpha$ is the confidence level on the hypothesis $P = p_o$. If $P$ is in fact equal to $p_o$, we have $\alpha%$ chance of keeping the area rather than excluding it.</td>
</tr>
<tr>
<td>$p_i$</td>
<td>$(1 - \beta)$</td>
<td>Our intention is that an area should be retained if $P = p_i$. Given our hypothesis that $P = p_o$, we have a $\beta%$ chance of excluding an area when in fact $P = p_i$.</td>
</tr>
</tbody>
</table>
6.3 Separating out near-empty clusters

**Figure 6.1.** Operating characteristic: Probability of retaining area in the survey \((h_p)\) as function of the proportion \((P)\) of households with a child domestic worker in the area.

<table>
<thead>
<tr>
<th>(P)</th>
<th>(h_p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>(p_0)</td>
<td>0.05</td>
</tr>
<tr>
<td>(p')</td>
<td>0.11</td>
</tr>
<tr>
<td>(p_i)</td>
<td>0.20</td>
</tr>
<tr>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

6.3.4 Using incomplete lists for identifying areas containing the rare population

As noted, even incomplete lists can be very useful in identifying areas where the target population is located.

Suppose that we have put together a consolidated list of working children using all the information available from different organisations, informed persons, and other sources. We also assume that each listing contains information to identify the area (cluster) where the child concerned may be found.

Let \(L\) be the number of listings (working children) in the whole list, and \(L_k\) the number of listings from cluster \(k\). Suppose that one out of the \(L\) listings is selected at random, and we take into the sample the cluster \((k)\) to which this listing (child) belongs. With repeated application, the procedure selects a sample of clusters in which the selection probability of a cluster is proportional to the number of listings it contains. In fact, with one draw the actual probability is \((L_k / L)\). If a sample of \(n\) listings is drawn then
-- ignoring that more than one listing may come from the same cluster -- we have a sample of \( n \) clusters, with clusters selected with probability \( n \left( \frac{L_k}{L} \right) \).

A cluster is selected through any of its \( L_k \) listings in the frame. This is a particular application of multiplicity sampling, discussed in detail in Chapter 7 below.

This is not a complete sample of clusters, since clusters which contain no listings \( (L_k = 0) \) have no chance of being selected. In practice, partial lists tend to be very unevenly distributed over clusters: often there are many entries (large \( L_k \)) from a minority of clusters, and none or very few from a majority.

The above procedure can therefore help to divide clusters in the frame into two parts:

(1) Clusters which contain one or more listings (and hence one or more members of the target population), and which have been selected into the sample. In fact, since clusters in this set have been selected with probabilities proportional to the number of listings in the cluster, clusters containing many listings tend to be overrepresented in this part of the sample.

(2) Clusters which may not contain any listing. This group in fact contains areas of three types: (i) clusters which do contain one or more listings of rare population members, but did not get selected at the first stage; (ii) clusters which contain members of the rare class, but none of which are in the list, so that all of these clusters had a zero chance of being selected at the first stage; and (iii) clusters which contain no members of the rare class, and hence also no listings of those, and therefore had a zero chance of being selected at the first stage.

The size of group 2(ii) depends on the sample size \( (n) \) at the first stage. Recalling that the selection equation at the first stage is

\[
f_k = n \left( \frac{L_k}{L} \right) = \frac{L_k}{(L/n)},
\]

units with \( L_k \geq \left( L/n \right) \) listings are expected to be automatically selected into the sample. For \( L_k < \left( L/n \right) \), \( f_k \) is the proportion of units with size \( L_k \) which are expected to be selected. Obviously

\[
(1 - f_k) = \frac{L_k}{(L/n)}
\]

is the proportion of clusters with \( L_k \) listings which remains in group (2). We can adjust the relative sizes of groups (1) and (2) by changing the sample size \( n \).

### 6.4 Gain from separating out empty clusters: numerical illustration

#### A. The simulated data

Let us refer back to the small data set presented in Table 3.3, giving a list of households and their main characteristics, with summary statistics in Tables 3.4 and 3.5. As noted there, we use this data set to illustrate and compare some sampling procedures. The numerical illustration has been implemented by Mehran (2012) as an Excel folder containing programs to select samples according to different sample designs. The illustrative population consists of 80 households spread over 20 geographical areas. Some households have no children and some have more than one. There are altogether
215 children, 20 of whom are working – 5 are unpaid and 15 are paid workers, with total earnings of 45 currency units.

In Section 3.4.4C, results were presented for a simple random sample of size n=13 households, selected from the whole population of 80 households. Using the same data set, here we compare with those some results from a design involving separation and exclusion of areas which contain no working children. In the reduced frame, a stratified random sample is selected which applies different sampling rates in different strata. Each stratum is a group of areas containing child workers.

The example in this section assumes that information is available, before sample selection, for each area on whether or not it contains any child workers. Data on the geographical distribution of child workers may be available, for instance, from a pilot survey or from a preliminary screening process as in multiphase sampling. This is a major item of information, and it can be expected that its use in the design and selection of the sample would increase the efficiency substantially. In reality, such complete information – which permits identification and exclusion of the part of the frame containing no units of the target population - is rare. For instance, the procedure described in the previous section identifies areas which definitely contain elements of the target population. However, the status of the remaining areas is not known in this respect; they may or may not contain elements of the target population.

For the illustration, we assume that of the 20 geographical areas, it is known that households with child workers are confined to 7 areas. These areas contain 28 (of the original 80) households. The remaining 13 areas are devoid of child workers. In the same way as in Table 3.3, characteristics of the population in the seven areas of concentration are shown in Table 6.5. The household identifiers are the same in the two tables.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>11</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>11</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>11</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>12</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>12</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td>12</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>12</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>14</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>14</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>14</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>14</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>18</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>18</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>18</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>19</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>74</td>
<td>19</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>19</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>19</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>28</td>
<td>89</td>
<td>20</td>
<td>45</td>
<td></td>
</tr>
</tbody>
</table>

**Column headings**
(1) Household identification number
(2) Area identification number
(3) Number of children in household
(4) No. of working children in household
(5) Total wages of working children
The total sample size of 13 households is allocated to the seven areas containing child workers by the square root method (Section 6.6.5 below) based on data on the distribution of child workers among the areas. Accordingly, three sample households are to be drawn from area 5, two each from areas 6, 11, 12 and 18, and one each from areas 14 and 19, as shown in Table 6.6.

### Table 6.6. Estimates from a stratified random sample of areas containing child workers

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(2a)</th>
<th>(2b)</th>
<th>(3a)</th>
<th>(3)</th>
<th>(4a)</th>
<th>(4)</th>
<th>(5a)</th>
<th>(5)</th>
<th>(3a)</th>
<th>(3)</th>
<th>(4a)</th>
<th>(4)</th>
<th>(5a)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>5</td>
<td>0.60</td>
<td>1.67</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1.7</td>
<td>6.7</td>
<td>1.7</td>
<td>1.7</td>
<td>1.7</td>
<td>5.0</td>
</tr>
<tr>
<td>17</td>
<td>5</td>
<td>0.60</td>
<td>1.67</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1.7</td>
<td>6.7</td>
<td>1.7</td>
<td>1.7</td>
<td>1.7</td>
<td>3.3</td>
</tr>
<tr>
<td>18</td>
<td>5</td>
<td>0.60</td>
<td>1.67</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1.7</td>
<td>3.3</td>
<td>1.7</td>
<td>1.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>23</td>
<td>6</td>
<td>0.50</td>
<td>2.00</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.7</td>
<td>2.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>24</td>
<td>6</td>
<td>0.50</td>
<td>2.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>41</td>
<td>11</td>
<td>0.50</td>
<td>2.00</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2.0</td>
<td>8.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>8.0</td>
</tr>
<tr>
<td>44</td>
<td>11</td>
<td>0.50</td>
<td>2.00</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.0</td>
<td>6.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>46</td>
<td>12</td>
<td>0.50</td>
<td>2.00</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2.0</td>
<td>10.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>6.0</td>
</tr>
<tr>
<td>48</td>
<td>12</td>
<td>0.50</td>
<td>2.00</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2.0</td>
<td>6.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>4.0</td>
</tr>
<tr>
<td>56</td>
<td>14</td>
<td>0.25</td>
<td>4.00</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4.0</td>
<td>4.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>71</td>
<td>18</td>
<td>0.50</td>
<td>2.00</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2.0</td>
<td>8.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>4.0</td>
</tr>
<tr>
<td>72</td>
<td>18</td>
<td>0.50</td>
<td>2.00</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2.0</td>
<td>8.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>4.0</td>
</tr>
<tr>
<td>76</td>
<td>19</td>
<td>0.25</td>
<td>4.00</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4.0</td>
<td>8.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Total</td>
<td>29.0</td>
<td></td>
<td></td>
<td></td>
<td>76.7</td>
<td>15.0</td>
<td>15.0</td>
<td>13.3</td>
<td>34.4</td>
<td>27.0</td>
<td>89.0</td>
<td>18.0</td>
<td>20.0</td>
<td>14.0</td>
<td>45.0</td>
</tr>
</tbody>
</table>

**Column headings**

(1) Household identification number  
(2) Area identification number  
(2a) selection probability  
(2b) sample weight  
(3a) whether has children  
(3) Number of children in household  
(4a) whether has working children  
(4) No. of working children in household  
(5a) has a paid child  
(5) Total wages of working children  

**B. Estimation and comparison**

In the manner of Table 3.6, Table 6.6 presents the sample areas and estimates from the sample data. The structure of the table is the same as that of the earlier table.

Columns (1) and (2) identify the sample households and their geographical locations. Column (2a) gives the probability of selection and column (2b) gives its inverse, the sampling weight of each sample household. In this example, the probabilities of selection and the corresponding sampling weights are not all equal as was the case with simple random sampling in Table 3.6.

As in the case of the simple random sample of the example in Section 3.4.4C, we have here households as the sampling units, and analysis units of three different types: households, children and working children. The relationship between units of the two types is again straightforward. The relationship between sampling-to-analysis units is
of the form one-to-(none/one/many) in all cases; any analysis unit is linked to (comes into the sample through) one and only one sampling unit. Consequently, the probability of an analysis unit appearing in the sample, and hence its sample weight, is exactly the same as the selection probability of the sampling unit to which it is linked. Using the household weights in Equation (3.12), the table shows estimates, for the particular sample drawn, of:

- Col. (2b) the number of households in the population,
- Col. (3a) number of households having children,
- Col. (3) total number of children in the households,
- Col. (4a) number of households having working children,
- Col. (4) total number of working children in the households, and
- Col. (5) the total of the wages received by all children.\(^{23}\)

These estimates are also compared with the actual population values as given in the last row of the table.\(^{24}\)

Using the Excel folder repeatedly to draw samples, Table 6.7 shows the results for an (arbitrary) set of 10 stratified random samples drawn using the frame which excludes areas without child workers. The table gives an idea of the variability we get between sample estimates with such a small sample. Note that the variability in estimates of number of working children is higher than in estimates of total number of (all, whether working or not) children, the variability in estimating the total wages is higher still. If we were to consider all possible simple random samples of size \(n=13\), the average values of the estimates will be equal to the true population values because the sampling procedure is unbiased. The variability between samples will of course decline in proportion to the sample size.

The results are compared with the results from Table 3.7 for a set of 10 simple random samples drawn from the whole population. Clearly, the estimates for the number of child workers and the total wages received by child workers are less variable, and are generally closer to the actual population values, compared with the results from a SRS drawn without separating out and excluding ‘empty’ areas containing no child workers. The reduced variability simply results from the fact that with empty areas excluded in a situation when they constitute a majority of the areas, we have a much reduced proportion of blanks in the list from which the sample is drawn. The presence of blanks increases sampling variance.

This improvement does not apply to the estimate of the total number of children in the population. Excluding areas empty of child workers does not affect much the distribution of all children, when we consider both working and non-working children. In any case, most households in the illustrative population have children: (68/80) in the whole study population, and (27/28) after excluding areas with no child workers.

\(^{23}\) Columns (3a)-(5) refer to the right hand panel of Table 6.6.

\(^{24}\) Using the fuller data not reproduced in Table 6.5, we can also estimate the number of child workers who are paid – it is 37 compared to the actual population value 25.
Table 6.7. Survey estimates for a set of 10 stratified samples from frame excluding areas without child workers, compared with SRS from the whole population

<table>
<thead>
<tr>
<th>Sample</th>
<th>Number of children</th>
<th>Number of working children</th>
<th>Total child wages</th>
<th>Number of children</th>
<th>Number of working children</th>
<th>Total child wages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>252.3</td>
<td>36.9</td>
<td>73.8</td>
<td>94.0</td>
<td>17.0</td>
<td>43.0</td>
</tr>
<tr>
<td>2</td>
<td>196.9</td>
<td>30.8</td>
<td>67.7</td>
<td>106.0</td>
<td>15.0</td>
<td>39.0</td>
</tr>
<tr>
<td>3</td>
<td>184.6</td>
<td>18.5</td>
<td>12.3</td>
<td>110.0</td>
<td>31.0</td>
<td>73.0</td>
</tr>
<tr>
<td>4</td>
<td>215.4</td>
<td>18.5</td>
<td>30.8</td>
<td>108.7</td>
<td>27.0</td>
<td>60.0</td>
</tr>
<tr>
<td>5</td>
<td>240.0</td>
<td>36.9</td>
<td>86.2</td>
<td>92.0</td>
<td>21.0</td>
<td>43.0</td>
</tr>
<tr>
<td>6</td>
<td>221.5</td>
<td>24.6</td>
<td>61.5</td>
<td>98.0</td>
<td>25.0</td>
<td>67.7</td>
</tr>
<tr>
<td>7</td>
<td>258.5</td>
<td>30.8</td>
<td>73.8</td>
<td>102.7</td>
<td>21.0</td>
<td>42.0</td>
</tr>
<tr>
<td>8</td>
<td>227.7</td>
<td>30.8</td>
<td>98.5</td>
<td>92.7</td>
<td>17.0</td>
<td>38.7</td>
</tr>
<tr>
<td>9</td>
<td>190.8</td>
<td>24.6</td>
<td>86.2</td>
<td>92.0</td>
<td>25.0</td>
<td>57.7</td>
</tr>
<tr>
<td>10</td>
<td>184.6</td>
<td>6.2</td>
<td>12.3</td>
<td>104.7</td>
<td>17.0</td>
<td>48.3</td>
</tr>
</tbody>
</table>

Average of 10 samples

- Simple random sample (all areas; see Table 3.7)
  - Number of children: 217.2
  - Number of working children: 25.8
  - Total child wages: 60.3

- Stratified design excluding areas without child workers
  - Number of children: 100.1
  - Number of working children: 21.6
  - Total child wages: 51.2

Population value

- Simple random sample (all areas; see Table 3.7)
  - Number of children: 215.0
  - Number of working children: 20.0
  - Total child wages: 45.0

- Stratified design excluding areas without child workers
  - Number of children: 89.0
  - Number of working children: 20.0
  - Total child wages: 45.0

C. Variance and cost

Usually, the choice between different sampling procedures involves a balance between the magnitude of sampling variance and the associated survey cost. More precise designs are usually also more costly, and vice versa. We choose the design which gives the lowest variance for a given cost, or what amounts to the same thing, the design which has the lowest cost for the same variance. However, the comparison between the two designs in Table 6.7 – a simple random sample from the whole population, and a stratified random sample only from the ‘areas of concentration’ containing working children - is not of this type. Restricting the sample to areas of concentration, when such areas are known prior to the survey, reduces survey costs. The sample is geographically more confined. Consequently, costs of travel to sample areas, of preparing the sampling frame and lists, of control and supervision of fieldwork, are all likely to be lower compared to a sample spread throughout the population. At the same time, by excluding areas not containing any child workers, sampling error is reduced - as can be seen from Table 6.7.

Hence the design excluding areas devoid of working children is both cheaper and more efficient. This situation arises because this design is based on a better sampling frame incorporating additional information about characteristics and distribution of the target population. Such information may come at a great cost if it has to be collected for the whole population specially for the survey, or it may cost very little if it already exists in available sources. Hence it is not simply a matter of procedures, but also of the circumstance in which the survey is being conducted. Apart from this special factor, below are some hints concerning quantification of the information on costs and variances.
Costs

Let us measure the cost of a survey assuming the following per unit costs for the different operations involved.

(i) Travel to and supervision of work in an area: 4 currency units per sample area
(ii) Listing and selection of households in an area: 0.5 currency unit per household
(iii) Contacting and interviewing selected households: 2 currency units per household
(iv) Identifying and interviewing working children in households: 3 currency units per working child

Table 6.6A summarises information on numbers of sample units involved from Tables 6.6 and 3.6. These numbers affect the total survey costs.

<table>
<thead>
<tr>
<th>Number of:</th>
<th>(1) All areas</th>
<th>(2) Areas with working children</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Population</td>
<td>Sample</td>
</tr>
<tr>
<td>areas</td>
<td>20</td>
<td>11</td>
</tr>
<tr>
<td>households</td>
<td>80</td>
<td>13</td>
</tr>
<tr>
<td>households/area</td>
<td>4.0</td>
<td>1.2</td>
</tr>
<tr>
<td>children</td>
<td>215</td>
<td>41</td>
</tr>
<tr>
<td>children/household</td>
<td>2.7</td>
<td>3.2</td>
</tr>
<tr>
<td>working children</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>working children/household</td>
<td>0.25</td>
<td>0.46</td>
</tr>
</tbody>
</table>

(1) Simple random sample (Table 3.6)
(2) Stratified sample in areas with working children (Table 6.6)

The figures shown are for the same sample size n=13 in terms of the number of households; the number of working children in the sample would be different in the two designs.

In both cases, we are considering a single-stage sample of households. Therefore, if the number of areas in the population is very large, we would expect the number of areas in the sample to be similar to the household sample size – possibly a little lower as more than one household may be selected from some areas. The number of areas which come into the sample cannot exceed the total number of areas in the population, nor the number of households to be selected. These limits for the two designs in our example are: 13 for design (1), the simple random sample (13 being the number of households selected); and 7 for design (2), the stratified sample in areas with working children (7 being the total number of areas in the population). The actual number of areas in the two particular samples being shown are close to this (11 and 7, respectively). The number of households per area is practically identical (4.0) in the two populations. Hence the costs of (i) travel to and supervision of work and of (ii) listing and selection of households in sample areas are proportional to the number of
areas in each sample, namely in the ratio 11:7. On the other hand, the total cost of (iii) contacting and interviewing selected households is identical (ratio 1:1) in the two cases since the household sample size is taken to be the same. Sample (2) catches more working children per household than sample (1). The cost of (iv) identifying and interviewing working children in each household is therefore higher in sample (2) – by a factor nearly of 3:1 or at least 5:2. This last-mentioned cost is in fact a ‘good cost’ since it proportionately increases the number of working children included in the sample.25

With the above parameters and per unit costs assumed earlier, the costs and expected sample sizes for the two designs are as follows.

<table>
<thead>
<tr>
<th>Cost</th>
<th>(i)</th>
<th>(ii)</th>
<th>(iii)</th>
<th>(iv)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design (1) Simple random sample</td>
<td>11x(4.0)</td>
<td>11x4x(0.5)</td>
<td>13x(2.0)</td>
<td>13x0.25x(3.0)</td>
<td>102</td>
</tr>
<tr>
<td>Design (2) Stratified sample in areas with working children</td>
<td>7x(4.0)</td>
<td>7x4x(0.5)</td>
<td>13x(2.0)</td>
<td>13x0.71x(3.0)</td>
<td>96</td>
</tr>
</tbody>
</table>

Variance

Sampling errors can be estimated from the results of the one sample we have in each case. Our illustrative population is very small, but this is artificial in the sense that in practice we normally deal with larger populations and smaller sampling fractions than here. With small sampling rates, generally the variance estimation procedures can be greatly simplified. For instance, one can use, uniformly for different designs, approximate methods of extending the variance estimator for sampling with replacement to the case of sampling without replacement (Särndal, Swensson and Wretman, 1992).

In our present illustration, the sample size is very small, and therefore estimates of sampling errors are themselves subject to large variability and do not provide a reliable basis for comparing different sampling procedures. Instead, in the last row of Table 6.7 we have provided estimates of sampling error (in the form of coefficient of variation or relative error = standard error/mean) obtained directly from the (albeit very partial) sampling distribution. Standard deviation of the sampling distribution of a statistic is, by definition, its standard error. The substantially higher precision of Sample (2) is apparent.

6.5 Identifying strata with differing concentrations

6.5.1 The procedure

We used the procedure of Section 6.3.2 to separate out strata with no or very few clusters containing members of the rare class (group 0) from the rest (group 1). When some lists of the rare population elements are available, it may be possible to separate the last mentioned group into two parts, as described in Section 6.3.4:

25 Of course, these figures from the samples are subject to random variation determined by the sample size, and the variation can be large for the very small samples we have in the illustration.
clusters \textit{definitely} containing rare population members who are present in those lists (new group 1); and other clusters (group 2). The latter group may contain clusters with elements from the lists, but this is not certain; it may contain clusters with rare population elements which are not in the lists; or it may contain clusters altogether empty of such elements.

The next possible step is to decompose group (2) into strata according to the degree of concentration of the rare population. Efficiency in sampling a rare population can be improved by stratifying the sampling frame according to the degree of concentration of the rare population (discussed in this section), and then oversampling strata with higher concentrations (Section 6.6). A detailed numerical illustration will be provided in the Case Study in Annex B. The objective of that case study is to numerically illustrate the procedure, and provide a quantitative indication of the potential gain in efficiency.

The objective of the procedure described below is to separate out clusters according to their degree of concentration of the rare population. In terms of the total population, the clusters involved in the process should be fairly large in size (say on the average 200-300 households). Consider a design such as the following.

Each cluster is divided into segments (say into 6-12 segments per cluster, with an average size of 20-30 households per segment). A field trip is made to each cluster and, say, 2 or 3 households are selected at random, one by one, in each segment. Consider the first selection in each segment. If the selected household contains a member of the rare population, the segment is retained in group (2). If the selected household does not contain a member of the rare population, the segment is reallocated to group (3).

With the above procedure, segments have been retained in group (2) with probability proportional to the proportion \( p_k \) of households in segment \( k \) which contain a member of the rare population. The expected proportion of segments which are retained in group (2) is similar to the average proportion of households in the original group (2) which contain a member of the rare population, except that the latter is a weighted average with \( H_k \), the total number of households in segment \( k \), as the weights. The sums below are over \( A \) segments in the population.

Expected proportion of segments retained = \( \frac{\sum_{k=1}^{A} p_k}{A} \).

Average proportion of households which contain a rare population element = \( \frac{\sum_{k=1}^{A} H_k p_k}{\sum_{k=1}^{A} H_k} \).

Now we consider a second selection in each segment in group (3) and apply the same procedure as above. If the household from the second selection contains a member of the rare population, it is retained in group (3). The probability of a segment being selected in this way is \( (1 - p_k)p_k \), the first factor being the probability that the segment is not selected during the first selection, and the second factor the conditional probability that it is selected during the second selection. If the household from the second selection does not contain a member of the rare population, the segment is reallocated to group (4).

The procedure may be continued if it appears useful. But the point to note is that in the \( g^{th} \) selection, units are selected with probabilities proportional to \( (1 - p_k)^{g-1} p_k \), which increasingly become less sensitive to differences among segments in the \( (p_k) \) values.
6.5.2 Illustration of the procedure

Table 6.8 provides an illustration of the procedure. Suppose that we have already removed segments which definitely contain units from the list frame as group (1). There remain 100 segments in group (2) which are shown in the table. Let us assume that these segments are fairly uniform in population size, and column (2) in the table shows the proportion of the population in the segment which belongs to the rare population of interest. This proportion \((p_k)\) varies from 0 to 1, with an average of \(p = 0.47\) in our illustration. These proportions of course are not known, and the actual procedure followed in practice has to be as follows.

A field visit is made to the 100 segments and in each segment, 2 households are selected at random. Information is obtained from each of the two selected households on whether it contains a member of the rare population (e.g. a child domestic worker employed by the household). The two selected households in each segment are distinguished according to whether they come from the first or the second selection in the segment.

The procedure is to retain in group (2) only those areas where in the 1st selection, the selected household was found to contain a member of the rare population. The chance that this happens for an area \(k\) is the proportion \((p_k)\) of households in it which contain a member of the rare population. Hence \((p_k)\) is the selection probability of area \(k\) in the 1st selection. The objective of the illustration is to show how these probabilities work.

For brevity in the illustration, we do not reproduce listings of individual households in each segment. In fact, the application of the procedure in the field does not need such a listing in so far as a random sample of a couple of households in the area can be obtained without it. In the illustration, we simulated the expected results of the selection procedure as follows. We take a random number \(r_k\) (from uniform distribution) and consider area \(k\) selected if \(r_k \leq p_k\). This gives the required selection probability \((p_k)\). The range of values for \(r_k\) is zero to \(p_{\text{max}}\), the maximum value of \((p_k)\) among the segments in the frame. The expected proportion of segments selected equals the average value of \(p\) in the frame. In our illustration, \(p = 0.47\). (The proportion of areas selected turns out to be 0.48.) The above, of course, is not the procedure we can apply in practice since the \(p_k\) values are not known. The actual selection procedure involves selection of a couple of households in each area for determining whether the area is to be selected. For the purpose of constructing this illustration and discussing the sampling procedure we have replaced this actual random selection procedure by the statistically equivalent procedure of selecting areas on the basis of the \((p_k, r_k)\) values as described above. We have used this latter procedure for the illustration because it clarifies the basis of the method.
### Table 6.8. Numerical illustration of the procedure
Separation of segments into strata with different concentrations of the rare population

<table>
<thead>
<tr>
<th>(1) segment</th>
<th>(2) probability</th>
<th>(3) selected if $r \leq p$</th>
<th>(4) probability</th>
<th>(5) selected if $r \leq p(1-p)$</th>
<th>“Group (4)” remaining segments</th>
<th>Final group</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>p</td>
<td>r</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.61</td>
<td>0.70</td>
<td></td>
<td>0.24</td>
<td>0.23</td>
<td>0.61</td>
</tr>
<tr>
<td>2</td>
<td>0.49</td>
<td>0.40</td>
<td>0.49</td>
<td>0.24</td>
<td>0.00</td>
<td>0.41</td>
</tr>
<tr>
<td>3</td>
<td>0.41</td>
<td>0.60</td>
<td></td>
<td>0.24</td>
<td>0.08</td>
<td>0.35</td>
</tr>
<tr>
<td>4</td>
<td>0.46</td>
<td>0.16</td>
<td>0.46</td>
<td>0.23</td>
<td>0.08</td>
<td>0.35</td>
</tr>
<tr>
<td>5</td>
<td>0.35</td>
<td>0.82</td>
<td></td>
<td>0.23</td>
<td>0.08</td>
<td>0.35</td>
</tr>
<tr>
<td>6</td>
<td>0.36</td>
<td>0.10</td>
<td>0.36</td>
<td>0.17</td>
<td>0.06</td>
<td>0.22</td>
</tr>
<tr>
<td>7</td>
<td>0.22</td>
<td>0.70</td>
<td></td>
<td>0.25</td>
<td>0.11</td>
<td>0.44</td>
</tr>
<tr>
<td>8</td>
<td>0.44</td>
<td>0.60</td>
<td></td>
<td>0.25</td>
<td>0.11</td>
<td>0.44</td>
</tr>
<tr>
<td>9</td>
<td>0.43</td>
<td>0.22</td>
<td>0.43</td>
<td>0.07</td>
<td>0.06</td>
<td>0.08</td>
</tr>
<tr>
<td>10</td>
<td>0.51</td>
<td>0.49</td>
<td>0.51</td>
<td>0.11</td>
<td>0.23</td>
<td>0.00</td>
</tr>
<tr>
<td>11</td>
<td>0.62</td>
<td>0.70</td>
<td></td>
<td>0.23</td>
<td>0.19</td>
<td>0.62</td>
</tr>
<tr>
<td>12</td>
<td>1.00</td>
<td>0.95</td>
<td>1.00</td>
<td>0.05</td>
<td>0.11</td>
<td>0.05</td>
</tr>
<tr>
<td>13</td>
<td>0.45</td>
<td>0.70</td>
<td></td>
<td>0.25</td>
<td>0.24</td>
<td>0.45</td>
</tr>
<tr>
<td>14</td>
<td>0.53</td>
<td>0.04</td>
<td>0.53</td>
<td>0.07</td>
<td>0.06</td>
<td>0.08</td>
</tr>
<tr>
<td>15</td>
<td>0.08</td>
<td>0.97</td>
<td></td>
<td>0.00</td>
<td>0.23</td>
<td>0.00</td>
</tr>
<tr>
<td>16</td>
<td>0.63</td>
<td>0.58</td>
<td>0.63</td>
<td>0.05</td>
<td>0.11</td>
<td>0.05</td>
</tr>
<tr>
<td>17</td>
<td>0.36</td>
<td>0.04</td>
<td>0.36</td>
<td>0.05</td>
<td>0.11</td>
<td>0.05</td>
</tr>
<tr>
<td>18</td>
<td>0.11</td>
<td>0.02</td>
<td>0.11</td>
<td>0.05</td>
<td>0.11</td>
<td>0.05</td>
</tr>
<tr>
<td>19</td>
<td>0.00</td>
<td>0.34</td>
<td></td>
<td>0.00</td>
<td>0.23</td>
<td>0.00</td>
</tr>
<tr>
<td>20</td>
<td>0.74</td>
<td>0.18</td>
<td>0.74</td>
<td>0.00</td>
<td>0.23</td>
<td>0.00</td>
</tr>
<tr>
<td>21</td>
<td>0.91</td>
<td>0.38</td>
<td>0.91</td>
<td>0.00</td>
<td>0.23</td>
<td>0.00</td>
</tr>
<tr>
<td>22</td>
<td>0.05</td>
<td>0.32</td>
<td></td>
<td>0.00</td>
<td>0.23</td>
<td>0.00</td>
</tr>
<tr>
<td>23</td>
<td>0.97</td>
<td>0.31</td>
<td>0.97</td>
<td>0.00</td>
<td>0.23</td>
<td>0.00</td>
</tr>
<tr>
<td>24</td>
<td>0.68</td>
<td>0.08</td>
<td>0.68</td>
<td>0.00</td>
<td>0.23</td>
<td>0.00</td>
</tr>
<tr>
<td>25</td>
<td>0.11</td>
<td>0.70</td>
<td></td>
<td>0.00</td>
<td>0.23</td>
<td>0.00</td>
</tr>
<tr>
<td>26</td>
<td>0.07</td>
<td>0.63</td>
<td></td>
<td>0.07</td>
<td>0.21</td>
<td>0.07</td>
</tr>
<tr>
<td>27</td>
<td>0.70</td>
<td>0.34</td>
<td>0.70</td>
<td>0.00</td>
<td>0.23</td>
<td>0.00</td>
</tr>
<tr>
<td>28</td>
<td>0.98</td>
<td>0.59</td>
<td>0.98</td>
<td>0.00</td>
<td>0.23</td>
<td>0.00</td>
</tr>
<tr>
<td>29</td>
<td>0.09</td>
<td>0.50</td>
<td></td>
<td>0.00</td>
<td>0.23</td>
<td>0.00</td>
</tr>
<tr>
<td>30</td>
<td>0.01</td>
<td>0.02</td>
<td></td>
<td>0.00</td>
<td>0.23</td>
<td>0.00</td>
</tr>
</tbody>
</table>

**Notes:**
- $p$ = proportion of the rare population in the segment
- $r$ = random number
- Group (0), empty or nearly empty of members of the rare population, has already been separated out using the procedure of Section 6.3.2.
- Group (1), comprising clusters definitely containing members of the rare population from available lists, has been separated out using the procedure of Section 6.3.4.
### Table 6.8. Numerical illustration (cont.)

<table>
<thead>
<tr>
<th>(1) segment</th>
<th>(2) probability</th>
<th>(3)</th>
<th>(4) selected if ( r \leq p )</th>
<th>(5) probability</th>
<th>(6)</th>
<th>(7) selected if ( r \leq p(1-p) )</th>
<th>“Group (4)” remaining segments</th>
<th>Final group</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>p</td>
<td>r</td>
<td>p</td>
<td>p(1-p)</td>
<td>r</td>
<td>p</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>0.66</td>
<td>0.47</td>
<td>0.66</td>
<td>0.06</td>
<td>0.24</td>
<td>0.06</td>
<td>(4)</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>0.06</td>
<td>0.17</td>
<td></td>
<td>0.07</td>
<td>0.20</td>
<td>0.08</td>
<td>(4)</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>1.00</td>
<td>0.69</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>0.08</td>
<td>0.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>0.39</td>
<td>0.34</td>
<td>0.39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>0.00</td>
<td>0.46</td>
<td></td>
<td>0.00</td>
<td>0.15</td>
<td>0.00</td>
<td>(4)</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>0.61</td>
<td>0.39</td>
<td>0.61</td>
<td>0.01</td>
<td>0.11</td>
<td>0.01</td>
<td>(4)</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>0.01</td>
<td>0.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>0.56</td>
<td>0.49</td>
<td>0.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>0.48</td>
<td>0.06</td>
<td>0.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>0.61</td>
<td>0.53</td>
<td>0.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>0.17</td>
<td>0.72</td>
<td></td>
<td>0.14</td>
<td>0.07</td>
<td>0.14</td>
<td>(3)</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>0.25</td>
<td>0.59</td>
<td></td>
<td>0.19</td>
<td>0.14</td>
<td>0.25</td>
<td>(3)</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>0.62</td>
<td>0.28</td>
<td>0.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>0.35</td>
<td>0.24</td>
<td>0.35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>0.36</td>
<td>0.41</td>
<td></td>
<td>0.23</td>
<td>0.14</td>
<td>0.23</td>
<td>(3)</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>0.57</td>
<td>0.72</td>
<td></td>
<td>0.24</td>
<td>0.13</td>
<td>0.24</td>
<td>(3)</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>0.58</td>
<td>0.85</td>
<td></td>
<td>0.24</td>
<td>0.05</td>
<td>0.24</td>
<td>(3)</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>0.00</td>
<td>0.38</td>
<td></td>
<td>0.00</td>
<td>0.02</td>
<td>0.00</td>
<td>(4)</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.59</td>
<td>0.93</td>
<td></td>
<td>0.24</td>
<td>0.02</td>
<td>0.24</td>
<td>(3)</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>0.03</td>
<td>0.09</td>
<td></td>
<td>0.03</td>
<td>0.17</td>
<td>0.03</td>
<td>(4)</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>1.00</td>
<td>0.41</td>
<td>1.00</td>
<td>0.00</td>
<td>0.17</td>
<td>0.00</td>
<td>(4)</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>0.98</td>
<td>0.36</td>
<td>0.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>0.56</td>
<td>0.68</td>
<td></td>
<td>0.25</td>
<td>0.22</td>
<td>0.25</td>
<td>(3)</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>1.00</td>
<td>0.37</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>0.64</td>
<td>0.93</td>
<td></td>
<td>0.23</td>
<td>0.21</td>
<td>0.23</td>
<td>(3)</td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>0.78</td>
<td>0.79</td>
<td></td>
<td>0.17</td>
<td>0.25</td>
<td>0.17</td>
<td>(4)</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>0.01</td>
<td>0.13</td>
<td></td>
<td>0.01</td>
<td>0.15</td>
<td>0.01</td>
<td>(4)</td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>0.39</td>
<td>0.91</td>
<td></td>
<td>0.24</td>
<td>0.01</td>
<td>0.24</td>
<td>(3)</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>0.47</td>
<td>0.18</td>
<td>0.47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>0.55</td>
<td>0.27</td>
<td>0.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>0.88</td>
<td>0.18</td>
<td>0.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>0.02</td>
<td>0.61</td>
<td></td>
<td>0.02</td>
<td>0.13</td>
<td>0.02</td>
<td>(4)</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>0.71</td>
<td>0.95</td>
<td></td>
<td>0.20</td>
<td>0.11</td>
<td>0.20</td>
<td>(3)</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>0.43</td>
<td>0.97</td>
<td></td>
<td>0.25</td>
<td>0.20</td>
<td>0.25</td>
<td>(3)</td>
<td></td>
</tr>
</tbody>
</table>
### Table 6.8. Numerical illustration (cont.)

<table>
<thead>
<tr>
<th>(1) segment</th>
<th>Selection for “Group (2)”</th>
<th>Selection for “Group (3)”</th>
<th>“Group (4)”</th>
<th>Final group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(2) probability</td>
<td>(3) selected if $r \leq p$</td>
<td>(5) probability</td>
<td>(6) selected if $r \leq p(1-p)$</td>
</tr>
<tr>
<td>k</td>
<td>p</td>
<td>r</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>66</td>
<td>0.08</td>
<td>0.63</td>
<td>p</td>
<td>0.07</td>
</tr>
<tr>
<td>67</td>
<td>0.96</td>
<td>0.12</td>
<td>0.96</td>
<td>0.02</td>
</tr>
<tr>
<td>68</td>
<td>0.02</td>
<td>0.59</td>
<td>p</td>
<td>0.22</td>
</tr>
<tr>
<td>69</td>
<td>0.34</td>
<td>0.56</td>
<td>p</td>
<td>0.89</td>
</tr>
<tr>
<td>70</td>
<td>0.89</td>
<td>0.09</td>
<td>p</td>
<td>0.41</td>
</tr>
<tr>
<td>71</td>
<td>0.96</td>
<td>0.12</td>
<td>0.96</td>
<td>0.02</td>
</tr>
<tr>
<td>72</td>
<td>0.02</td>
<td>0.59</td>
<td>p</td>
<td>0.22</td>
</tr>
<tr>
<td>73</td>
<td>0.89</td>
<td>0.09</td>
<td>p</td>
<td>0.41</td>
</tr>
<tr>
<td>74</td>
<td>0.34</td>
<td>0.56</td>
<td>p</td>
<td>0.89</td>
</tr>
<tr>
<td>75</td>
<td>0.89</td>
<td>0.09</td>
<td>p</td>
<td>0.41</td>
</tr>
<tr>
<td>76</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>77</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>78</td>
<td>0.96</td>
<td>0.12</td>
<td>0.96</td>
<td>0.02</td>
</tr>
<tr>
<td>79</td>
<td>0.34</td>
<td>0.56</td>
<td>p</td>
<td>0.89</td>
</tr>
<tr>
<td>80</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>81</td>
<td>0.02</td>
<td>0.59</td>
<td>p</td>
<td>0.22</td>
</tr>
<tr>
<td>82</td>
<td>0.89</td>
<td>0.09</td>
<td>p</td>
<td>0.41</td>
</tr>
<tr>
<td>83</td>
<td>0.34</td>
<td>0.56</td>
<td>p</td>
<td>0.89</td>
</tr>
<tr>
<td>84</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>85</td>
<td>0.96</td>
<td>0.12</td>
<td>0.96</td>
<td>0.02</td>
</tr>
<tr>
<td>86</td>
<td>0.02</td>
<td>0.59</td>
<td>p</td>
<td>0.22</td>
</tr>
<tr>
<td>87</td>
<td>0.89</td>
<td>0.09</td>
<td>p</td>
<td>0.41</td>
</tr>
<tr>
<td>88</td>
<td>0.34</td>
<td>0.56</td>
<td>p</td>
<td>0.89</td>
</tr>
<tr>
<td>89</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>90</td>
<td>0.96</td>
<td>0.12</td>
<td>0.96</td>
<td>0.02</td>
</tr>
<tr>
<td>91</td>
<td>0.02</td>
<td>0.59</td>
<td>p</td>
<td>0.22</td>
</tr>
<tr>
<td>92</td>
<td>0.89</td>
<td>0.09</td>
<td>p</td>
<td>0.41</td>
</tr>
<tr>
<td>93</td>
<td>0.34</td>
<td>0.56</td>
<td>p</td>
<td>0.89</td>
</tr>
<tr>
<td>94</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>95</td>
<td>0.96</td>
<td>0.12</td>
<td>0.96</td>
<td>0.02</td>
</tr>
<tr>
<td>96</td>
<td>0.02</td>
<td>0.59</td>
<td>p</td>
<td>0.22</td>
</tr>
<tr>
<td>97</td>
<td>0.89</td>
<td>0.09</td>
<td>p</td>
<td>0.41</td>
</tr>
<tr>
<td>98</td>
<td>0.34</td>
<td>0.56</td>
<td>p</td>
<td>0.89</td>
</tr>
<tr>
<td>99</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>100</td>
<td>0.96</td>
<td>0.12</td>
<td>0.96</td>
<td>0.02</td>
</tr>
</tbody>
</table>

6.5 Identifying strata with differing concentrations
Table 6.8. Numerical illustration (Summary)

<table>
<thead>
<tr>
<th>Group</th>
<th>number of segments</th>
<th>proportion of rare class</th>
<th>% share of rare population in the group</th>
</tr>
</thead>
<tbody>
<tr>
<td>original</td>
<td>100</td>
<td>0.47</td>
<td>1.00</td>
</tr>
<tr>
<td>group (2)</td>
<td>48</td>
<td>0.67</td>
<td>0.68</td>
</tr>
<tr>
<td>group (3)</td>
<td>31</td>
<td>0.40</td>
<td>0.26</td>
</tr>
<tr>
<td>group (4)</td>
<td>21</td>
<td>0.09</td>
<td>0.04</td>
</tr>
</tbody>
</table>

The \((p_k)\) values of the units selected after the above-described first selection are shown in column (4) of the table. The 48 selected units are retained in group (2), and the remaining 52 are provisionally reallocated to group (3). Segments in group (2) are seen to have a higher average proportion of households containing a member of the rare population. The average value of \(p\) for the group is 0.67 (compared with the overall average of 0.47).

In a similar manner, group (3) identified above has been decomposed into two parts. Column (5) is the segment’s probability of selection in the 2\textsuperscript{nd} selection. It equals \((1 - p_k)p_k\) as noted earlier. The selection procedure has been simulated in the illustration by taking a random number \((r_k)\) in the range 0 to \((1 - p_k)p_k\) in this case, and retaining a segment in group (3) if \(r_k \leq (1 - p_k)p_k\).

Segments not selected are reallocated to group (4). The \(p_k\) values of the segments retained in group (3) and segments reallocated to group (4) are shown in columns (7)-(8) of the table.

The procedure has decomposed the original 100 segments into 3 groups with different degrees of concentration of members of the rare population, as shown summarised at the end of Table 6.8.

6.6 Oversampling strata with higher concentrations

6.6.1 Introduction: when is oversampling useful?

One way of obtaining a larger sample of a rare population is to identify strata where larger concentrations of the population exist, and to sample those strata at a higher rate.
A. Some basics

It will be useful here to note some basic points about sample allocation in order to appreciate what gains in precision, if any, we might expect from stratification and disproportionate allocation.

First consider the whole population which contains the rare population of special interest.

In stratified sampling the population is partitioned such that every element in the population appears in only one part. The sample is selected independently within each part or stratum \((h)\). Suppose that the proportions individual strata form of the total population \((U_h, \sum U_h = 1)\) are known constants. Normally these quantities are defined in terms of the number of elements in the stratum \((N_h)\) and in the total population \((N)\):

\[
U_h = \frac{N_h}{N}.
\]

The mean estimated from the sample in each stratum \((\bar{y}_h)\) relates to the estimate from the whole sample \((\bar{y})\) as

\[
\bar{y} = \sum U_h \bar{y}_h,
\]

and the variance as

\[
\text{var}(\bar{y}) = \sum U_h^2 \text{var}(\bar{y}_h).
\]

The gain in stratification comes from the fact that only variability within strata, but not any variability between strata, contributes to the resulting variance in a stratified sample.

Issues concerning sample allocation among strata have been discussed in Section 5.3. Here we reiterate a couple of important points. A commonly used design is to apply the same sampling rate to all strata – this is ‘proportionate allocation’. With \(n\) denoting the sample size, and \(f\) the common sampling rate, proportionate allocation means the stratum sampling rates

\[
f_h = \frac{n_h}{N_h} = \frac{n}{N} = f, \ \text{a constant}.
\]

Further gains in precision of \((\bar{y}_h)\) can be obtained with ‘optimum allocation’, in which the sampling rates are not uniform across strata, but are higher in more heterogeneous strata (i.e. in strata where there is more variability, measured in terms of \(S_h\) defined below), and in strata where per unit survey costs \((C_h)\) are lower. The actual expression for optimum strata sampling rates is

\[
f_h \propto \frac{S_h}{\sqrt{C_h}}.
\]

Parameters \(S_h\) and \(\sqrt{C_h}\) often do not differ much across strata. Usually we use the same survey method in different strata, and it is unlikely to find large differences in per unit costs among the strata. Furthermore, only the square root of the cost differences appears in the above expression – which is even less variable across strata. Differences in element variances (or standard deviations \(S_h\)) are often also small. For means and similar statistics assuming \(S_h\) to be the same in different strata is often a reasonable approximation. For proportions \((P_h)\), \(S_h = \sqrt{P_h(1-P_h)}\) is quite insensitive to variations in
values, unless those variations are extreme. For instance $S_a = 0.4$ for $P_a = 0.2$, and $S_a = 0.5$ for a very different stratum with $P_a = 0.5$.

This does not preclude that there are situations in which large differences in costs and variances exist across strata, so that the optimum sampling rates can differ considerably. The point is that in many situation differences are not large enough to merit departures from proportional allocation.

B. Objectives in surveying a subpopulation

Now turning from the total population to surveying a subpopulation, it is useful to distinguish objectives of several types.

(1) To estimate the size of the subpopulation and the proportion it forms of the total population.

For example: What is the number ($M$) of working children aged 7-17? What proportion ($P = M/N$) of children aged 7-17 ($N$) are working? Here $N$ is the size of a large demographic class, and is usually known reasonably accurately.

(2) To estimate parameters for the subpopulation, such as the mean value for a certain variable $Y$,

$$\bar{Y} = \frac{1}{M} \sum_{i=1}^{M} Y_i .$$

For example, what proportion of working children are able to read and write? What are their mean number of working hours per week?

(3) To obtain a sufficiently large sample size for the rare population for in-depth studies, such as on living conditions, conditions at work, treatment by the employer, attitudes towards working, etc., classified by various characteristics, for instance by the child’s type and location of work, or sex and age. Some of these indicators will be more descriptive than quantitative; others may be more clearly quantitative but it may suffice to have approximate figures in order to draw conclusions or define policy objectives. What is needed is to have a sufficiently large and representative sample in order to capture diverse aspects of the children’s work.

(4) To optimise sample allocation to achieve an appropriate compromise among different objectives. This refers to the situation when the rare population of interest comprises a number of domains (e.g. different geographical regions, demographic and other subclasses, different types and sectors of activity), and estimates are required for the whole rare population and for each domain separately. The sample allocation needs to be a compromise to meet the different objectives.

An example of objective (1) is the ‘child labour survey’ (CLS), and of (2) and (3) the ‘labouring children survey’ (LCS) as defined in Verma (2008). Often these two types of survey are combined into a single operation; such a combined survey provides an example of (4).

In the following subsections, each of these objectives is discussed in turn.
6.6.2 Estimation of size of the rare population

A. Optimum versus proportional allocation

The first objective is to estimate the size of the rare population and proportion \((P)\) it forms of the total population. For this purpose, the base population is of course the total population, and the discussion above in Section 6.6.1 (part A) applies.

Using the symbols introduced in that section, the general expressions for variance and cost in a stratified sample, with simple random sampling (SRS) within each stratum, are:

\[
\text{Variance} = \sum_h \left( U_h^2 S_h^2 / n_h \right), \quad \text{Cost} = \sum_h (C_h n_h).
\]

1) With proportional allocation \( n_h = n U_h \) (stratum sample size proportional to stratum population size), the above become:

\[
\text{Variance} = \frac{1}{n} \left( \sum_h U_h S_h^2 \right), \quad \text{Cost} = n \left( \sum_h U_h C_h \right), \quad \text{giving}\quad (\text{Variance} \times \text{Cost}) = \left( \sum_h U_h S_h^2 \right) \left( \sum_h U_h C_h \right). \tag{6.3}
\]

2) With optimum allocation \( n_h \propto U_h S_h / \sqrt{C_h} \), that is \( n_h = n \left( U_h S_h / \sqrt{C_h} \right) \sum_n \left( U_h S_h / \sqrt{C_h} \right) \), we have

\[
\text{Variance} = \frac{1}{n} \sum_h \left( U_h S_h / \sqrt{C_h} \right) \sum_h \left( U_h S_h / \sqrt{C_h} \right), \quad \text{Cost} = \frac{n}{\sum_h \left( U_h S_h / \sqrt{C_h} \right)} \sum_h \left( U_h S_h / \sqrt{C_h} \right),
\]

\[
\text{giving}\quad (\text{Variance} \times \text{Cost}) = \left( \sum_h U_h S_h / \sqrt{C_h} \right)^2. \tag{6.5}
\]

The ratio \((G)\) of \((\text{Variance} \times \text{Cost})\) with optimum allocation to that with proportional allocation gives the gain with optimum allocation:

\[
G = \frac{\left( \sum_h U_h S_h / \sqrt{C_h} \right)^2}{\left( \sum_h U_h S_h^2 \right) \left( \sum_h C_h U_h \right)}. \tag{6.6}
\]

Often per unit costs are similar across strata. With that assumption the above expression becomes:

\[
G = \left( \sum_h U_h S_h \right)^2 \left( \sum_h S_h^2 \right). \tag{6.7}
\]

The above assumes simple random sampling within strata (with finite population correction neglected), apart from assuming constant per unit costs throughout. As noted earlier, for proportions \((P)\), \( S_h = \sqrt{P_h(1-P_h)} \); when \( P_h \) is small, \( S_h \approx \sqrt{P_h} \).
B. Numerical illustrations

It is important to get an idea of the above expressions in terms of numbers. For this purpose we have constructed illustrative Tables 6.9-6.11.

Consider very rare populations comprising less than \( P = 10 \) per cent of the total population. We will give numerical illustrations for very small (tending to 0) and small (2.5%, 5.0% and 7.5%) values of \( P \). The illustrative population is assumed to be divided into two strata, with stratum 2 having a higher concentration of the rare population of interest. The objective is to estimate \( P \), the proportion the rare population forms of the total population. Since the total population forms the denominator for the statistic in both strata, it is reasonable to assume that per unit cost is the same in the two strata.

The numerical results can be presented in a variety of forms. From the range of parameters which appear in the equations, choosing values for a subset determines the values of the remaining parameters. The form of presentation depends on which subset of parameters is chosen to display the data. In the tables, we have chosen the following three parameters:

<table>
<thead>
<tr>
<th>( P )</th>
<th>The proportion of the total population belonging to the rare class. It tells us how rare the class is.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( U_2 )</td>
<td>Relative size of stratum 2, which has the higher concentration of the rare population. ( U_1 + U_2 = 1 ).</td>
</tr>
<tr>
<td>( P_2/P )</td>
<td>Relative concentration (proportion) of the rare class in stratum 2, compared with the average concentration ( (P) ) in the population.</td>
</tr>
</tbody>
</table>

Since by definition \( (P_1U_1 + P_2U_2) = P \), \( (U_1 + U_2) = 1 \), we have

\[
\frac{P_2}{P} = \frac{1 - U_2(P_2/P)}{1 - U_2}.
\]

Table 6.9(1) shows \( (P_1/P) \) values. Consider, for example, the case \( (P_2/P) = 7 \) and \( U_2 = 0.10 \) in the table. This means that the stratum of concentration of the rare class comprises 10 per cent of the population, and in it the rare class concentration is 7 times higher than the average concentration. In this case we find \( (P_1/P) = 0.33 \), implying \( (P_2/P_1) \approx 0.20 \). This corresponds to a very high level of concentration in stratum 2 – over 20 times higher than that in stratum 1.

It is also instructive to see these figures in terms of the share of the rare population which is found in its concentration stratum: \( C_2 = U_2(P_2/P) = 7 \times 10\% = 70\% \). This figure is high in the example chosen.

Table 6.9(2) shows the ratio \( (g) \) of sampling rates to be applied to the two strata in order to optimise for estimating \( P \). This is for the case when all \( P \) values are small, so that we can take the approximation

\[
S_h \approx \sqrt{P_h}, \text{ giving } f_h \propto S_h = \sqrt{P_h}, \text{ hence } g = f_2/f_1 = \sqrt{P_2/P_1}.
\]

For the same case as above, with \( (P_2/P) = 7 \) and \( U_2 = 0.10 \), we have \( g = f_2/f_1 = 4.6 \), meaning that it is optimum to sample stratum 2 at a rate four and a half times higher than stratum 1.
### 6.6 Oversampling strata with higher concentrations

#### Table 6.9. Optimum allocation for estimating $P$, the rare population as a proportion of total population: for very rare classes (small $P$)

<table>
<thead>
<tr>
<th>$P_2/P$</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>0.95</td>
<td>0.89</td>
<td>0.82</td>
<td>0.75</td>
<td>0.67</td>
<td>0.57</td>
<td>0.46</td>
<td>0.33</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.89</td>
<td>0.78</td>
<td>0.65</td>
<td>0.50</td>
<td>0.33</td>
<td>0.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.84</td>
<td>0.67</td>
<td>0.47</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.79</td>
<td>0.56</td>
<td>0.29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.74</td>
<td>0.44</td>
<td>0.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.68</td>
<td>0.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.63</td>
<td>0.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.58</td>
<td>0.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.53</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### (2) Values of $g$ when all $P$’s are small

<table>
<thead>
<tr>
<th>$P_2/P$</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>1.5</td>
<td>1.6</td>
<td>1.6</td>
<td>1.7</td>
<td>1.9</td>
<td>2.1</td>
<td>2.4</td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.8</td>
<td>2.0</td>
<td>2.2</td>
<td>2.4</td>
<td>3.0</td>
<td>4.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2.2</td>
<td>2.4</td>
<td>2.9</td>
<td>4.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2.5</td>
<td>3.0</td>
<td>4.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2.9</td>
<td>3.7</td>
<td>7.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>3.2</td>
<td>4.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3.6</td>
<td>6.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>3.9</td>
<td>9.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>4.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### (3) Values of $G$ when all $P$’s are small

<table>
<thead>
<tr>
<th>$P_2/P$</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>0.99</td>
<td>0.98</td>
<td>0.97</td>
<td>0.95</td>
<td>0.93</td>
<td>0.91</td>
<td>0.88</td>
<td>0.83</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.97</td>
<td>0.93</td>
<td>0.89</td>
<td>0.83</td>
<td>0.75</td>
<td>0.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.94</td>
<td>0.87</td>
<td>0.78</td>
<td>0.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.91</td>
<td>0.80</td>
<td>0.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.88</td>
<td>0.71</td>
<td>0.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.84</td>
<td>0.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>0.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.76</td>
<td>0.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Per unit costs are assumed to be the same in all strata.*

- $P$: The proportion of the total population belonging to the rare class.
- $U_2$: Relative size of stratum 2, with higher concentration of the rare population. $U_1 + U_2 = 1$. 
Relative concentration (proportion) of the rare class in stratum 2, compared with the average concentration ($P$) in the population.

$g$ Relative stratum sampling rates (stratum 2 : stratum 1).

$G$ Relative gain in precision (reduction in variance) over proportional allocation, for the same cost.

Table 6.9 (3) shows $(G)$, the gain in precision for the same cost (or the reduction in cost for the same precision), with optimal allocation compared to proportional allocation. Again, this is based on the simplifying assumption $S_h \approx \sqrt{P_h}$. Thus for $(P_2/P)=7$ and $U_2=0.10$, we find variance reduced by nearly 40 per cent (i.e. reduced to 0.60 of the original value). Of course with lower degrees of concentration (smaller $(P_2/P)$ values) the gains are smaller. Also, it is important to note that usually the gains achieved are smaller than the theoretical values in the table because of lack of precise information for making optimum choices.

In Tables 6.10 and 6.11, the full expression $S_h = \sqrt{P_h (1-P_h)}$ has been used. This makes the results dependent on the value of $P$ (in addition to the chosen parameter $U_2$, $P_2/P$).

Table 6.10 shows the ratio of stratum 2 to stratum 1 sampling rates ($g$) when optimised for estimating $P$. The smaller the rare class, the greater is the oversampling required for optimisation (i.e. for achieving the maximum possible reduction in variance for a given total sample size). Thus again considering the point $(P_2/P) = 7$ and $U_2=0.10$, the table shows that $g$ increases from 3.2 for a moderately large subclass ($P=7.5$ per cent), to 4.2 for a very small rare class ($P=2.5$ per cent), and to 4.6 for the limiting case of a vanishingly small rare class as seen in Table 6.9.

Table 6.11 shows the relative efficiency of optimum versus proportionate allocation $(G)$. The reduction in variance becomes a little less marked as the rare population size $(P)$ goes up. Thus for $(P_2/P) = 7$ and $U_2=0.10$, the table shows $G$ to be 0.77 (i.e. 23 per cent reduction in variance with optimisation) for a moderately large subclass ($P=7.5$ per cent), down to 0.66 for a very small rare class ($P=2.5$ per cent), and to 0.61 for the limiting case of a vanishingly small rare class.
### Table 6.10. Optimum allocation for estimating $P$, the rare population as a proportion of total population: relative stratum sampling rates ($g$)

<table>
<thead>
<tr>
<th>$P$</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1/P$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>1.4</td>
<td>1.5</td>
<td>1.5</td>
<td>1.6</td>
<td>1.7</td>
<td>1.8</td>
<td>2.0</td>
<td>2.4</td>
<td>3.2</td>
<td>3.2</td>
</tr>
<tr>
<td>3</td>
<td>1.8</td>
<td>1.9</td>
<td>2.1</td>
<td>2.4</td>
<td>2.9</td>
<td>4.4</td>
<td>4.4</td>
<td>4.4</td>
<td>4.4</td>
<td>4.4</td>
</tr>
<tr>
<td>4</td>
<td>2.1</td>
<td>2.3</td>
<td>2.8</td>
<td>3.8</td>
<td>3.8</td>
<td>3.8</td>
<td>3.8</td>
<td>3.8</td>
<td>3.8</td>
<td>3.8</td>
</tr>
<tr>
<td>5</td>
<td>2.4</td>
<td>2.8</td>
<td>3.9</td>
<td>3.9</td>
<td>3.9</td>
<td>3.9</td>
<td>3.9</td>
<td>3.9</td>
<td>3.9</td>
<td>3.9</td>
</tr>
<tr>
<td>6</td>
<td>2.7</td>
<td>3.4</td>
<td>6.6</td>
<td>6.6</td>
<td>6.6</td>
<td>6.6</td>
<td>6.6</td>
<td>6.6</td>
<td>6.6</td>
<td>6.6</td>
</tr>
<tr>
<td>7</td>
<td>2.9</td>
<td>4.2</td>
<td>4.2</td>
<td>4.2</td>
<td>4.2</td>
<td>4.2</td>
<td>4.2</td>
<td>4.2</td>
<td>4.2</td>
<td>4.2</td>
</tr>
<tr>
<td>8</td>
<td>3.2</td>
<td>5.4</td>
<td>5.4</td>
<td>5.4</td>
<td>5.4</td>
<td>5.4</td>
<td>5.4</td>
<td>5.4</td>
<td>5.4</td>
<td>5.4</td>
</tr>
<tr>
<td>9</td>
<td>3.5</td>
<td>7.9</td>
<td>7.9</td>
<td>7.9</td>
<td>7.9</td>
<td>7.9</td>
<td>7.9</td>
<td>7.9</td>
<td>7.9</td>
<td>7.9</td>
</tr>
<tr>
<td>10</td>
<td>3.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$P$</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1/P$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>1.4</td>
<td>1.5</td>
<td>1.5</td>
<td>1.6</td>
<td>1.7</td>
<td>1.8</td>
<td>2.0</td>
<td>2.3</td>
<td>3.2</td>
<td>3.2</td>
</tr>
<tr>
<td>3</td>
<td>1.7</td>
<td>1.8</td>
<td>2.0</td>
<td>2.3</td>
<td>2.8</td>
<td>4.2</td>
<td>4.2</td>
<td>4.2</td>
<td>4.2</td>
<td>4.2</td>
</tr>
<tr>
<td>4</td>
<td>2.0</td>
<td>2.2</td>
<td>2.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
</tr>
<tr>
<td>5</td>
<td>2.2</td>
<td>2.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
</tr>
<tr>
<td>6</td>
<td>2.4</td>
<td>3.1</td>
<td>6.0</td>
<td>6.0</td>
<td>6.0</td>
<td>6.0</td>
<td>6.0</td>
<td>6.0</td>
<td>6.0</td>
<td>6.0</td>
</tr>
<tr>
<td>7</td>
<td>2.6</td>
<td>3.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2.8</td>
<td>4.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>3.0</td>
<td>6.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>3.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$P$</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1/P$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>1.4</td>
<td>1.4</td>
<td>1.5</td>
<td>1.5</td>
<td>1.6</td>
<td>1.8</td>
<td>2.0</td>
<td>2.3</td>
<td>3.1</td>
<td>3.1</td>
</tr>
<tr>
<td>3</td>
<td>1.7</td>
<td>1.8</td>
<td>1.9</td>
<td>2.2</td>
<td>2.7</td>
<td>4.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.9</td>
<td>2.1</td>
<td>2.5</td>
<td>3.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2.1</td>
<td>2.4</td>
<td>3.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2.2</td>
<td>2.8</td>
<td>5.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2.3</td>
<td>3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2.3</td>
<td>3.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2.3</td>
<td>5.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Per unit costs are assumed to be the same in all strata.**

- $P$: The proportion of the total population belonging to the rare class. The table is shown for a specimen of different $P$ values.
- $U_2$: Relative size of stratum 2, with higher concentration of the rare population. $U_1 + U_2 = 1$.
- $(P_2/P)$: Relative concentration (proportion) of the rare class in stratum 2, compared with the average concentration ($P$) in the population.
- $g$: Relative stratum sampling rates (stratum 2 : stratum 1).
Table 6.11. Optimum allocation for estimating \( P \), the rare population as a proportion of total population: gain in precision with optimum allocation over proportional allocation (G)

<table>
<thead>
<tr>
<th>( P )</th>
<th>( 0.05 )</th>
<th>( 0.10 )</th>
<th>( 0.15 )</th>
<th>( 0.20 )</th>
<th>( 0.25 )</th>
<th>( 0.30 )</th>
<th>( 0.35 )</th>
<th>( 0.40 )</th>
<th>( 0.45 )</th>
<th>( 0.50 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P/P )</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>1</td>
<td>0.99</td>
<td>0.98</td>
<td>0.97</td>
<td>0.96</td>
<td>0.94</td>
<td>0.91</td>
<td>0.88</td>
<td>0.84</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.97</td>
<td>0.94</td>
<td>0.90</td>
<td>0.84</td>
<td>0.76</td>
<td>0.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.95</td>
<td>0.89</td>
<td>0.80</td>
<td>0.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.93</td>
<td>0.82</td>
<td>0.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.90</td>
<td>0.75</td>
<td>0.46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.87</td>
<td>0.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.84</td>
<td>0.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.81</td>
<td>0.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

See notes to the preceding tables.
The computations confirm that for proportions, \( S_h = \sqrt{P_h(1-P_h)} \) values are insensitive to variations in \( P_h \) values, unless those variations are large. In fact, differences in \( S_h \) values are often small. Actually the same applies in relation to differences in per unit costs \( C_h \). Large differences are needed in the ratio \( (P_h/C_h) \) and hence \( f_h \approx \sqrt{(P_h/C_h)} \) for optimum allocation to result in a significant improvement in estimation of \( P \).

It would be useful to extract a few figures from the preceding tables to appreciate the above in quantitative terms. For a concentration stratum comprising, say, \( U_2 = 0.1 \) of the population and having an average concentration \( P_2/P = 5 \) times the average concentration \( P \) in the population, optimal allocation reduces variance (or increases effective sample size) by a factor = 0.80. This is the amount of reduction when \( P \) is very small; the reduction is somewhat smaller with larger \( P \) values: for instance it is by a factor 0.82 with \( (P = 0.025) \), 0.85 with \( (P = 0.050) \), and 0.88 with \( (P = 0.075) \). In any case, these are rather modest levels of reduction in variance. The reduction in variance is more significant with larger relative sizes of the concentration stratum \( (U_2) \), and with increasing degree of concentration in that stratum \( (P_2/P) \). A few figures are summarised below from the detailed tables.

<table>
<thead>
<tr>
<th>( U_2 = 0.10 )</th>
<th>( P_2/P = 3 )</th>
<th>( P = 0 )</th>
<th>0.025</th>
<th>0.050</th>
<th>0.075</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_2/P = 5 )</td>
<td>0.93</td>
<td>0.94</td>
<td>0.95</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>( U_2 = 0.10 )</td>
<td>0.80</td>
<td>0.82</td>
<td>0.85</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>( P_2/P = 5 )</td>
<td>0.63</td>
<td>0.66</td>
<td>0.69</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>( U_2 = 0.15 )</td>
<td>0.50</td>
<td>0.54</td>
<td>0.61</td>
<td>0.70</td>
<td></td>
</tr>
</tbody>
</table>

These figures demonstrate that gains in efficiency of estimating the proportion or total number of a rare population from oversampling of strata with higher concentrations are often modest, but can become substantial in the presence of high degrees of concentration.

Any gain from optimal allocation is likely to be reduced when strata are defined - as they often are – in terms of areas units, even if the rare population is geographically unevenly distributed and occurs in areas of concentration. By contrast, it is possible sometimes that one or more strata are defined, not in terms of geographical areas but on the basis of lists of population elements of interest. In such cases large gains in precision can result, even if the lists are not complete or contain non-members. However, for this to happen it is also necessary that the list strata account for a reasonably large proportion of the rare population of interest.

### 6.6.3 Estimating means and similar statistics for the rare population

**A. Rare population rather than total population as the sample base**

Below we note some important differences in the estimation procedure when a rare population, rather than the total population, forms the sample base for estimation.
There are five main differences between estimating means and similar statistics for a rare population, and estimating the proportion it forms of the total population discussed in the previous subsection.

(1) Base population

The rare population rather than the total population forms the denominator in estimating means and similar statistics for it. Everywhere, the strata weights refer to the (relative) numbers \(X_h = U_h (P_h / P)\), \(\sum_h X_h = 1\) in the rare population, rather than the numbers \(U_h\) in the total population.

(2) Need to estimate strata weights

The usual situation is that strata are constructed with sizes known in terms of the total population: \(U_h = (N_h / N)\). We cannot separate out the rare population beforehand to define the strata in terms of its numbers. We may know that the rare population is more concentrated in some strata than others, but not the precise extent of that concentration. Therefore, the stratum sizes \(X_h\) in terms of the numbers of the rare population \(M_h, \sum M_h = M\),

\[X_h = (M_h / M) = U_h (P_h / P)\]

with \(P_h = M_h / N_h,\quad P = M / N\)

are usually not known but have to be estimated from the sample. Let \(f_h\) be the sampling rate applied to stratum \(h\) (applied to the whole population in the stratum, including the rare class which is not yet separated out), and \(m_h\) be the number of rare population elements found in the stratum sample. The estimate of the total number of such elements in the stratum is

\[\hat{M}_h = m_h w_h,\]

where \(w_h = (1 / f_h)\) are the weights to be applied in the estimation. The relative stratum sizes are

\[\hat{X}_h = \left( \frac{m_h w_h}{\sum_h m_h w_h} \right).\]

Since the number of rare population elements cannot be controlled, a proportionate sample in terms of the total population \((f_h = N_h / N)\) is not necessarily proportionate in terms of the rare population \((f_h \neq M_h / M)\). In fact, for small subpopulations, any gain in efficiency from proportional allocation in terms of total population tends to be lost. But the effects of optimal or other disproportionate allocation remain.

(3) More uniform element variance (or standard deviation \(S_h\))

Usually it is reasonable to assume that element variance is fairly uniform across strata. Recall that in estimating the rare class as proportion \(P_h\) of the total population, we have \(S_h = \sqrt{P_h (1 - P_h)}\) which is insensitive to variations in \(P_h\) values. This variation in \(S_h\) may even be smaller when we consider a rare population. Just because more rare elements are concentrated in certain strata (large differences in \(P_h\) values), does not mean that those elements are systematically and significantly different from other elements of the same rare population which may be well-dispersed across strata. In other words, strata
6.6 Oversampling strata with higher concentrations

designed for the total population are often less effective in separating out homogeneous classes of the rare population.

(4) But more significant differences in per unit cost ($C_h$)

In estimating the rare class as proportion $P_h$ of the total population, it is often reasonable to assume that the differences among strata in per unit costs $C_h$ are small, especially when the relevant quantity is $\sqrt{C_h}$. However, in estimation with the rare population as the base, per unit cost differences can be large – meaning here per unit of the rare population. This is because identification of elements of the rare population from the total population normally requires screening (discussed further in Section 6.7.2). For each unit of the rare population in a stratum, there are $(1 - P_h)/P_h$ units from the rest of the population, and the cost of screening the latter has to be 'loaded' on to the cost of the rare population unit.

If $C_1$ is the total per unit cost of interviewing a rare population element, and $C_0$ is the per unit cost of screening, then the effective per unit cost of surveying the rare population is

$$C_h = C_1 + C_0 \left(1 - \frac{P_h}{P_h}\right).$$

(For simplicity, we have assumed that the per unit costs above include the share of common travel and listing costs etc. associated with each sample area in a multi-stage design.)

As in estimating $P_h$, it is often reasonable to assume that differences among strata in per unit interviewing and screening costs ($C_1$ and $C_0$) are negligible. However, this does not apply to $C_h$ when $P_h$ values differ among strata, especially between the strata of concentration and the rest.

Since all costs are relative, we may write the above expression more simply by dividing it by $C_1$:

$$C_h = 1 + c \left(1 - \frac{P_h}{P_h}\right),$$

where $c = C_0/C_1$, per unit screening cost as a fraction of the total interview cost per unit of the rare population.

(5) Using large clusters

It is advisable to select a large sample per cluster when the interest is in a rare population, rather than in the total population. This is because for a given size of cluster in terms of the total population, the effective cluster size in terms of the rare population elements is much smaller (by factor $P_h$); hence for comparable values of the intra-cluster correlation, design effects are also much smaller. The optimum cluster size (meaning average sample size per cluster) in terms of the total population is likely to be much larger. The is discussed further in quantitative terms in Section 6.7.2 (part C).

For simplicity, in the following we will disregard the approximation due to (2) above (i.e. the variability of strata weights); we will also assume simple random sampling within strata, disregarding design effect related to (5).
B. Proportional and optimum allocation

Using the symbols introduced earlier,

(i) With proportional allocation we have \( m_h = m X_h, \quad f_h = m_h / M_h = m / M = f, \)

constant, and

\[
(Variance \times Cost) = \left(\sum_h X_h S_h^2\right) \left(\sum_h X_h C_h\right).
\]

(ii) With optimum allocation, we have \( m_h = m \left(\frac{X_h S_h / \sqrt{C_h}}{\sum_h \left(X_h S_h / \sqrt{C_h}\right)}\right), \quad f_h \propto \left(\frac{S_h / \sqrt{C_h}}{1}\right), \)

and

\[
(Variance \times Cost) = \left(\sum_h \left(X_h S_h \sqrt{C_h}\right)\right)^2.
\]

Assuming the per unit cost to vary as noted earlier,

\[
C_h = 1 + c \left(\frac{1 - P_h}{P_h}\right),
\]

and with \( S_h \) assumed constant, the ratio (optimum/proportional) of (Variance X Cost) becomes

\[
G = \left(\sum_h \left(X_h \sqrt{C_h}\right)\right)^2 / \left(\sum_h X_h C_h\right).
\]

C. Numerical illustrations

In our numerical illustrations, we take the same population as in Section 6.6.2. There are two strata, stratum 2 having a higher concentration of the rare population.

From the above, (stratum2/stratum1) ratios are as follows:

ratio of per unit cost, \( k = \frac{C_2}{C_1} = \frac{1 + c \left(\frac{1 - P_2}{P_2}\right)}{1 + c \left(\frac{1 - P_1}{P_1}\right)}; \)

ratio of sampling rates, \( g = \frac{f_2}{f_1} = \sqrt{\frac{C_1}{C_2}}; \) and

relative efficiency, \( G = \frac{\left(X_1 \sqrt{C_1} + X_2 \sqrt{C_2}\right)^2}{\left(X_1 C_1 + X_2 C_2\right)}. \)

The numerical results in the illustration are displayed as \( U_2 \) versus \( (P_2 / P) \) as before, with \( c \) and \( P \) as two additional parameters.
6.6 Oversampling strata with higher concentrations

Parameters, (1) \( \left( \frac{P_1}{P} \right) = \frac{1-U_2 \left( \frac{P_2}{P} \right)}{1-U_2} \), and (2) \( X_h = U_h \left( \frac{P_h}{P} \right) \) are shown in Table 6.12 for reference. With \( \left( \frac{P_2}{P} \right) = 7.0 \) for example, \( \left( \frac{P_1}{P} \right) = 0.68 \), i.e. the former is oversampled by over 10 times, this case representing a high degree of concentration of the rare population in stratum 2.

<table>
<thead>
<tr>
<th>( P_2/P )</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>0.95</td>
<td>0.89</td>
<td>0.82</td>
<td>0.75</td>
<td>0.67</td>
<td>0.57</td>
<td>0.46</td>
<td>0.33</td>
<td>0.18</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>0.89</td>
<td>0.78</td>
<td>0.65</td>
<td>0.50</td>
<td>0.33</td>
<td>0.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.84</td>
<td>0.67</td>
<td>0.47</td>
<td>0.25</td>
<td>0.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.79</td>
<td>0.56</td>
<td>0.29</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.74</td>
<td>0.44</td>
<td>0.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.68</td>
<td>0.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.63</td>
<td>0.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.58</td>
<td>0.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.53</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As shown in Tables 6.13-6.15, when screening cost is high and the size of the rare population is very small: (i) there is sharp difference in per unit cost between the strata – the cost being much higher in the stratum with low proportion of the rare class; (ii) the optimum sampling rates are more sharply different – the stratum of concentration being over-sampled; and (iii) the improvement in efficiency over proportional allocation is larger – though still rather modest. These differences between strata become less marked with lower screening cost and a larger size of the rare population.

The following notation is used in Tables 6.13-6.15:
- \( P \) The proportion of the total population belonging to the rare class.
- \( U_2 \) Relative size of stratum 2, with higher concentration of the rare population. \( U_1 + U_2 = 1 \).
- \( P_2/P \) Relative concentration (proportion) of the rare class in stratum 2, compared with the average concentration \( P \) in the population.
- \( c \) Per unit screening cost as a fraction of the total interview cost per unit of the rare population.
### 6. Sampling rare populations

Table 6.13. Optimum allocation for estimating means and similar statistics over a rare population: ratio of per unit costs (Stratum 2 / Stratum 1)

1. **High listing cost (c=0.25)**
   - **Very small subpopulation (P=0.025)**

<table>
<thead>
<tr>
<th>$P_2/P_0$</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>0.51</td>
<td>0.48</td>
<td>0.45</td>
<td>0.41</td>
<td>0.37</td>
<td>0.32</td>
<td>0.26</td>
<td>0.19</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.34</td>
<td>0.30</td>
<td>0.25</td>
<td>0.20</td>
<td>0.13</td>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.26</td>
<td>0.21</td>
<td>0.15</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.20</td>
<td>0.15</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.17</td>
<td>0.10</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.14</td>
<td>0.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.12</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.10</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. **High listing cost (c=0.25)**
   - **Not-so-small subpopulation (P=0.075)**

<table>
<thead>
<tr>
<th>$P_2/P_0$</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>0.57</td>
<td>0.54</td>
<td>0.50</td>
<td>0.47</td>
<td>0.42</td>
<td>0.37</td>
<td>0.30</td>
<td>0.22</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.42</td>
<td>0.37</td>
<td>0.32</td>
<td>0.25</td>
<td>0.17</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.34</td>
<td>0.28</td>
<td>0.20</td>
<td>0.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.28</td>
<td>0.21</td>
<td>0.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.25</td>
<td>0.16</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.22</td>
<td>0.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.19</td>
<td>0.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.17</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. **Low listing cost (c=0.05)**
   - **Very small subpopulation (P=0.025)**

<table>
<thead>
<tr>
<th>$P_2/P_0$</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>0.64</td>
<td>0.61</td>
<td>0.58</td>
<td>0.54</td>
<td>0.49</td>
<td>0.44</td>
<td>0.37</td>
<td>0.28</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.51</td>
<td>0.46</td>
<td>0.40</td>
<td>0.33</td>
<td>0.23</td>
<td>0.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.44</td>
<td>0.37</td>
<td>0.28</td>
<td>0.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.39</td>
<td>0.30</td>
<td>0.17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.35</td>
<td>0.24</td>
<td>0.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.32</td>
<td>0.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.29</td>
<td>0.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.27</td>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. **Low listing cost (c=0.05)**
   - **Not-so-small subpopulation (P=0.075)**

<table>
<thead>
<tr>
<th>$P_2/P_0$</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>0.78</td>
<td>0.75</td>
<td>0.73</td>
<td>0.70</td>
<td>0.66</td>
<td>0.61</td>
<td>0.54</td>
<td>0.44</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.69</td>
<td>0.65</td>
<td>0.59</td>
<td>0.51</td>
<td>0.40</td>
<td>0.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.64</td>
<td>0.57</td>
<td>0.47</td>
<td>0.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.60</td>
<td>0.50</td>
<td>0.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.57</td>
<td>0.43</td>
<td>0.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.54</td>
<td>0.35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.52</td>
<td>0.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.49</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 6.14. Optimum allocation for estimating means and similar statistics over a rare population: ratio of sampling rates (Stratum 2 / Stratum 1)

#### (1) High listing cost (c=0.25)  
**Very small subpopulation (P=0.025)**

<table>
<thead>
<tr>
<th>(P_2/P_1)</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>1.40</td>
<td>1.44</td>
<td>1.50</td>
<td>1.57</td>
<td>1.66</td>
<td>1.78</td>
<td>1.97</td>
<td>2.31</td>
<td>3.11</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.71</td>
<td>1.83</td>
<td>1.99</td>
<td>2.25</td>
<td>2.74</td>
<td>4.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.97</td>
<td>2.20</td>
<td>2.60</td>
<td>3.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2.21</td>
<td>2.61</td>
<td>3.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2.43</td>
<td>3.10</td>
<td>5.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2.66</td>
<td>3.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2.88</td>
<td>4.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>3.11</td>
<td>6.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>3.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### (2) High listing cost (c=0.25)  
**Not-so-small subpopulation (P=0.075)**

<table>
<thead>
<tr>
<th>(P_2/P_1)</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>1.33</td>
<td>1.36</td>
<td>1.41</td>
<td>1.47</td>
<td>1.54</td>
<td>1.65</td>
<td>1.82</td>
<td>2.11</td>
<td>2.81</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.55</td>
<td>1.64</td>
<td>1.78</td>
<td>2.00</td>
<td>2.40</td>
<td>3.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.72</td>
<td>1.91</td>
<td>2.22</td>
<td>2.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.87</td>
<td>2.18</td>
<td>2.92</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2.01</td>
<td>2.51</td>
<td>4.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2.14</td>
<td>2.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2.27</td>
<td>3.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2.41</td>
<td>5.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### (3) Low listing cost (c=0.05)  
**Very small subpopulation (P=0.025)**

<table>
<thead>
<tr>
<th>(P_2/P_1)</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>1.25</td>
<td>1.28</td>
<td>1.32</td>
<td>1.36</td>
<td>1.42</td>
<td>1.51</td>
<td>1.65</td>
<td>1.89</td>
<td>2.48</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.40</td>
<td>1.48</td>
<td>1.58</td>
<td>1.75</td>
<td>2.07</td>
<td>3.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.51</td>
<td>1.65</td>
<td>1.89</td>
<td>2.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.61</td>
<td>1.84</td>
<td>2.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.69</td>
<td>2.06</td>
<td>3.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1.77</td>
<td>2.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1.85</td>
<td>2.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1.94</td>
<td>4.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### (4) Low listing cost (c=0.05)  
**Not-so-small subpopulation (P=0.075)**

<table>
<thead>
<tr>
<th>(P_2/P_1)</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>1.14</td>
<td>1.15</td>
<td>1.17</td>
<td>1.20</td>
<td>1.23</td>
<td>1.28</td>
<td>1.37</td>
<td>1.52</td>
<td>1.90</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.20</td>
<td>1.24</td>
<td>1.30</td>
<td>1.40</td>
<td>1.59</td>
<td>2.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.25</td>
<td>1.32</td>
<td>1.46</td>
<td>1.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.29</td>
<td>1.41</td>
<td>1.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.32</td>
<td>1.52</td>
<td>2.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1.36</td>
<td>1.68</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1.39</td>
<td>1.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1.43</td>
<td>2.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6.6 Oversampling strata with higher concentrations
Table 6.15. Optimum allocation for estimating means and similar statistics over a rare population: ratio of relative efficiency (cost*variance) of Optimal-to-Proportional allocation

<table>
<thead>
<tr>
<th>(1) High listing cost (c=0.25)</th>
<th>Very small subpopulation (P=0.025)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{P}{P^*} )</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>0.99</td>
</tr>
<tr>
<td>3</td>
<td>0.98</td>
</tr>
<tr>
<td>4</td>
<td>0.95</td>
</tr>
<tr>
<td>5</td>
<td>0.93</td>
</tr>
<tr>
<td>6</td>
<td>0.90</td>
</tr>
<tr>
<td>7</td>
<td>0.87</td>
</tr>
<tr>
<td>8</td>
<td>0.84</td>
</tr>
<tr>
<td>9</td>
<td>0.81</td>
</tr>
<tr>
<td>10</td>
<td>0.77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(2) High listing cost (c=0.25)</th>
<th>Not-so-small subpopulation (P=0.075)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{P}{P^*} )</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>0.99</td>
</tr>
<tr>
<td>3</td>
<td>0.98</td>
</tr>
<tr>
<td>4</td>
<td>0.97</td>
</tr>
<tr>
<td>5</td>
<td>0.95</td>
</tr>
<tr>
<td>6</td>
<td>0.93</td>
</tr>
<tr>
<td>7</td>
<td>0.91</td>
</tr>
<tr>
<td>8</td>
<td>0.89</td>
</tr>
<tr>
<td>9</td>
<td>0.86</td>
</tr>
<tr>
<td>10</td>
<td>0.84</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(3) Low listing cost (c=0.05)</th>
<th>Very small subpopulation (P=0.025)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{P}{P^*} )</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>0.99</td>
</tr>
<tr>
<td>4</td>
<td>0.98</td>
</tr>
<tr>
<td>5</td>
<td>0.97</td>
</tr>
<tr>
<td>6</td>
<td>0.96</td>
</tr>
<tr>
<td>7</td>
<td>0.94</td>
</tr>
<tr>
<td>8</td>
<td>0.93</td>
</tr>
<tr>
<td>9</td>
<td>0.91</td>
</tr>
<tr>
<td>10</td>
<td>0.90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(4) Low listing cost (c=0.05)</th>
<th>Not-so-small subpopulation (P=0.075)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{P}{P^*} )</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>1.00</td>
</tr>
<tr>
<td>4</td>
<td>0.99</td>
</tr>
<tr>
<td>5</td>
<td>0.99</td>
</tr>
<tr>
<td>6</td>
<td>0.99</td>
</tr>
<tr>
<td>7</td>
<td>0.98</td>
</tr>
<tr>
<td>8</td>
<td>0.98</td>
</tr>
<tr>
<td>9</td>
<td>0.97</td>
</tr>
<tr>
<td>10</td>
<td>0.96</td>
</tr>
</tbody>
</table>
6.6.4 Obtaining large enough sample sizes for in-depth analysis

The conventional criterion for optimising sample allocation is to minimise variance for a given budget, or to minimise cost for achieving a given level of precision.

However, when the added objective is to obtain sufficiently large sample sizes for certain subgroups of the rare population, the allocation criterion is shifted to the following: to maximise the sample size (or sample sizes for particular subgroups) for a given budget.

The allocation procedure would thus involve two steps. First, the sample is allocated optimally to meet the more precisely defined quantitative objectives, such as estimating mean values and similar statistics for the rare population, but keeping some resources in hand. Then, the remaining resources are used to allocate the sample in a targeted manner. Where other substantive constraints permit, the extra sample is concentrated in strata with lower per unit costs, so as to maximise the sample size obtained with given resources. This would generally mean extra sample for strata where members of the subclass (or specific groups of it) are more concentrated.

The extra sample of course also contributes towards increased precision of the overall estimates, but less than what would have been achieved with optimised allocation. The extra sample compensates – but only in part – for the effect of reduced sample size at the first (optimal allocation) step.

The above considerations are likely to be very important for child labour surveys where the survey objectives go well beyond producing precise estimates for a few selected statistics, towards exploring and describing diverse and complex aspects of the situation of working children.

6.6.5 Compromise allocation

The issue of sample allocation arises when a survey is being designed to produce estimates for a number of different domains, for subclasses that cut across the domains, as well as for the total population. It is common for domains of interest to vary considerably in size, some of them being very small (i.e. rare) domains. Compromise allocation refers to optimising the sample allocation to compromise between different objectives, such as estimating for the total rare population and also for its individual domains separately. Since there can be only a single allocation in the end, it is not possible to optimise for each domain separately.

This problem may seem to have some overlap with the issue discussed in the previous subsection. The two sometimes use similar procedures but in fact are quite distinct. In the discussion in Section 6.6.4, the added criterion was to maximise the available sample size so that diverse aspects of the complex situation could be studied in sufficient detail. By contrast, here the objective is still optimising the allocation for producing precise estimates for certain statistics – except that the optimisation is for some appropriately weighted combination of different domains or different objectives.

A. Commonly used schemes

An example is provided by Kish (1976, 1988). Under the commonly made assumptions that element variances and survey costs are the same across domains, the optimum
All domains receive the same (equal) allocation if the objective is to produce results for each domain with the same precision. The two allocations are conflicting, especially if the domains are of very different sizes. However, a compromise allocation that falls between the two optimum allocations may meet both objectives reasonably well. With $W_h$ as the relative sizes of the stratum populations ($\sum W_h = 1$) and $H$ the number of strata, a compromise solution is to determine the domain sample sizes in the form:

$$n_h \propto \left( k W_h^2 + (1 - k) H^{-2} \right)^{0.5},$$

where $k$ and $(1 - k)$ represent the relative importance of the total population estimate and the domain estimates, respectively. If $k = 1$, the allocation is a proportionate allocation, which is optimal for the total population estimate. If $k = 0$, we have equal allocation, which is optimum for the domain estimates.

Verma (1991) proposed the following modification to the above:

$$n_h = n_0 \left( k^2 + (1 - k^2) \cdot M_h^{-\alpha} \right)^{0.5},$$

where $(1 - k^2)$ and $k^2$ represent the relative importance of the total population estimate and the domain estimates, respectively. $M_h$ is a (relative) measure of the domain population size, normalised to average 1.0 over domains.

Constant $n_0$ is determined in order to make the sample sizes $n_h$ add up to the desired overall total size, $\sum n_h = n$. Constant $n_0$ is the size of the sample received by a domain of average size ($M_h = 1$). This procedure imposes a minimum sample size for the smallest of the domains (for $M_h$ tending to zero) as $n_{h(\text{min})} = kn_0$.

Parameter $\alpha$ is introduced to impose a constraint on the maximum sample size which, with given $k$ and $n_0$, is determined by the largest value of $M_h$ encountered in the data, i.e. the largest domain.

To give an example of the procedure using (6.9): it has been used by Eurostat to determine national sample sizes for the EU Statistics on Income and Living Conditions (EU-SILC) surveys, in which around 30 European countries have been participating on an annual basis. In the EU-wide context of EU-SILC, countries form the ‘domains’. While different countries may require – despite differences in their population sizes - similar sample sizes for the same level of precision, there are many well-known reasons why it is meaningful and useful to have larger samples in larger countries. The added reason for increasing the sample size with increasing population size (but of course much less than proportionately) is the requirement for reporting at the EU level. For such reporting, the ideal would be to sample at a uniform rate throughout, i.e. increase the national sample size in proportion to the population size. However, this will be unacceptable for the production of national level statistics (which require more equal sample sizes). There are constraints on the minimum sample size which can be allocated to any country irrespective of its population size, and the maximum sample size which any country, however large, will be willing or able to accept. There are also practical constraints on the total sample size, summed over Member States. For reasons which need not be discussed here, it was decided to limit the total to around 80,000 households in EU-15, or around 120,000 households in the expanded EU. For EU-15, the following values of the parameters were used in determining the national sample sizes.
6.6 Oversampling strata with higher concentrations

sizes: \( k^2 = 0.25, n_0 = 5,785, \alpha = 0.75 \), giving \( \sum \alpha n_h = n = 80,000 \), \( n_{\min} = kn_0 = 3,000 \). Subsequently, individual figures were adjusted or rounded as desired.

B. Choice of sample allocation parameters in Equation (6.9): numerical illustration

Table 6.16 provides a numerical illustration of the effect of the parameters \((k, \alpha)\) of Equation (6.9) on sample allocation among domains of the study population.

We consider 10 domains; the domain sizes have been scaled to average 1.0: \( \bar{M} = 1, \sum M_h = 10 \). The total sample size to be allocated among the 10 domains is taken as \( n = 1,000 \), so that the allocation for the average sized domain \((M_1 = 1)\) is \( n_0 = 100 \).

Illustrations are provided for two assumed distributions of (relative) domain sizes:

1. A moderate range of variation, with the largest domain being 10 times bigger than the smallest domain: \( M_{10}/M_1 = 10 \), shown in Table 6.16A.

2. An extreme range of variation, with the largest domain being over 80 times bigger than the smallest domain: \( M_{10}/M_1 = 82 \), shown in Table 6.16B.

Results for combinations of values \((k = 0.1-0.7)\) and \((\alpha = 0.5-2.0)\) are shown in both cases.

<table>
<thead>
<tr>
<th>Table 6.16. Numerical illustration concerning choice of parameters in sample allocation according to Equation (6.9)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Table 6.16A. Moderate range (1:10) of variation in domain sizes</strong></td>
</tr>
<tr>
<td>( h = )</td>
</tr>
<tr>
<td>( n_h = )</td>
</tr>
<tr>
<td>( k = )</td>
</tr>
<tr>
<td>( \alpha = )</td>
</tr>
<tr>
<td>( k = )</td>
</tr>
<tr>
<td>( \alpha = )</td>
</tr>
</tbody>
</table>

| | 0.5 | 0.25 | 0.5 | 77 | 86 | 92 | 96 | 100 | 104 | 107 | 110 | 112 | 115 | 1,000 | 1.5 |
| | 1.0 | 0.77 | 0.83 | 0.88 | 0.94 | 0.99 | 1.03 | 1.08 | 1.12 | 1.16 | 1.20 | 1,000 | 1.5 |
| | 1.5 | 1.49 | 1.49 | 1.49 | 1.49 | 1.49 | 1.49 | 1.49 | 1.49 | 1.49 | 1.49 | 1,000 | 2.0 |
| | 2.0 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 1,000 | 2.6 |

| | 0.3 | 0.09 | 0.5 | 71 | 82 | 90 | 96 | 101 | 105 | 109 | 112 | 115 | 118 | 1,000 | 1.7 |
| | 1.0 | 0.77 | 0.83 | 0.88 | 0.94 | 0.99 | 1.03 | 1.08 | 1.12 | 1.16 | 1.20 | 1,000 | 1.7 |
| | 1.5 | 1.49 | 1.49 | 1.49 | 1.49 | 1.49 | 1.49 | 1.49 | 1.49 | 1.49 | 1.49 | 1,000 | 3.8 |
| | 2.0 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 1,000 | 5.1 |

| | 0.1 | 0.01 | 0.5 | 68 | 81 | 89 | 96 | 101 | 106 | 110 | 113 | 117 | 120 | 1,000 | 1.8 |
| | 1.0 | 0.77 | 0.83 | 0.88 | 0.94 | 0.99 | 1.03 | 1.08 | 1.12 | 1.16 | 1.20 | 1,000 | 3.1 |
| | 1.5 | 1.49 | 1.49 | 1.49 | 1.49 | 1.49 | 1.49 | 1.49 | 1.49 | 1.49 | 1.49 | 1,000 | 5.3 |
| | 2.0 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 1,000 | 8.8 |

The table shows sample sizes allocated to the 10 domains, for the total sample size \( n = 1,000 \).
The effect of parameter $k$ is pronounced at the lower end – on the sample allocated to the smaller domains. The minimum possible allocation to any domain equals $k = 0.0$, corresponding to a domain of vanishingly small size ($M_h = 0$). In practice, even the smallest domain has non-zero size, so that the minimum sample size allocated is $n_1 > (k \ n_0)$ to the smallest domain, as determined from Equation (6.9).

The choice of parameter $k$ is determined by the decision on minimum sample size which must be received by any domain, even the smallest. This minimum requirement is determined by substantive and statistical considerations based on the survey objectives. Sample size is the primary determinant of the sampling precision of the resulting estimates.

Parameter $\alpha$ determines how rapidly the sample allocated to large domains is allowed to increase with increasing domain size. A large value of $\alpha$ results in the sample being shifted away from small and middle-sized domains to larger domains. The choice of the parameter should take into account two considerations:

1. What is the minimum sample size required for small domains for the results to have sufficient precision to be useful?
2. What is the desirable sample size required so that the results can be analysed with the desired level of detail, at least for middle-sized domains?

In practice, the effects of the two parameters ($k, \alpha$) interact, depending on the distribution of the domain sizes $M_h$ in the population, as can be seen from Tables 6.16A-B. Table 6.16C provides some summary statistics of the allocation according to Equation (6.9).
Table 6.16C. Some summary statistics of the allocation according to Equation (6.9)

<table>
<thead>
<tr>
<th>Range of the ratios of allocated sample sizes = ( n_{10}/n_1 )</th>
<th>10.0</th>
<th>82.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k = 0.7 )</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>( \alpha = 0.5 )</td>
<td>1.3</td>
<td>1.5</td>
</tr>
<tr>
<td>( 1.0 )</td>
<td>1.6</td>
<td>2.0</td>
</tr>
<tr>
<td>( 1.5 )</td>
<td>1.8</td>
<td>2.6</td>
</tr>
<tr>
<td>( 2.0 )</td>
<td>2.1</td>
<td>3.2</td>
</tr>
</tbody>
</table>

Effect of parameters \( (k, \alpha) \) on sample allocation:

<table>
<thead>
<tr>
<th>Range of domain sizes ( (M_{10}/M_1) )</th>
<th>10.0</th>
<th>82.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k = 0.7 )</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>( \alpha = 0.5 )</td>
<td>1.3</td>
<td>1.5</td>
</tr>
<tr>
<td>( 1.0 )</td>
<td>1.6</td>
<td>2.0</td>
</tr>
<tr>
<td>( 1.5 )</td>
<td>1.8</td>
<td>2.6</td>
</tr>
<tr>
<td>( 2.0 )</td>
<td>2.1</td>
<td>3.2</td>
</tr>
</tbody>
</table>

The top panel of Table 6.16C gives the ratio of the sample size allocated to the largest domain \( (h=10) \) to the sample allocated to the smallest domain \( (h=1) \). The ratio increases with declining \( k \) (declining minimum allocation), and also increases with increasing \( \alpha \) (increasing influence of domain size \( M_h \)). With very small \( k \) and large \( \alpha \), the allocation varies more nearly in proportion to domain size. These effects are more marked when the variation in domain sizes is more extreme.

The lower panel of Table 6.16C shows the actual sample sizes allocated to the smallest and the largest domains for the combinations of the largest and the smallest values of parameters \( k \) and \( \alpha \), i.e. \( k=(0.1, 0.7) \) and \( \alpha=(0.5, 2.0) \).

C. Some alternative allocation schemes

Kalton (2009) provides a review of some similar and some alternative sample allocation approaches, which are summarised briefly below. (See the above-mentioned source for reference to the literature.)

In a compromise solution similar to those described above, which has been termed ‘power allocation’, the domain sample sizes are determined from \( n_h \alpha W_h^q \), \( 0 \leq q \leq 1 \) (see Section 5.3.1B above). Parameter \( q = 0 \) corresponds to equal allocation and \( q = 1 \) to proportionate allocation. In an alternative approach, the allocation is determined by fixing a core sample that will satisfy one of the objectives and then supplementing that sample as needed to satisfy the other objective; the objective is to allocate sufficient sample to small domains to produce estimates at the required level of precision. (These may be variants of the idea discussed in Section 6.6.4.) In another approach, ‘inferential priorities’ such as \( P_d = N_d^\alpha \) are assigned to each domain \( d \) of size \( N_d \), with \( 0 \leq \alpha \leq 2 \), the lower limit corresponding to equal and the upper to proportionate allocation. A more general approach to sample allocation is
via mathematical programming. This approach can accommodate unequal variances across domains, intersecting domains, and multiple estimates for each domain.

When there are multiple domains of interest and multi-stage sampling is to be used, a variant of the usual measure of size for probability proportional to size (PPS) sampling can be useful for controlling the sample sizes in the sampled clusters. This is a composite measure of size that takes account of the differing sampling rates for different domains, and has been described in Section 5.3.5 in a particular context. For a wider use of the technique of rescaling unit size measures to facilitate sample selection, see Verma (2008, Section 4.4).

6.7 Counting, listing, screening, two-phase sampling

6.7.1 Distinct but closely-related operations

Counting, listing, screening and two-phase sampling refer to various operations which, though distinct, are closely related.

Counting is an operation aimed at estimating the size of a population or a subgroup in the population. In survey practice, counting methods have to be cheap, quick, non-intrusive, and hence necessarily approximate in estimating the population size. Counting normally involves identifying collectively elements which belong to the population of interest, but does not involve collecting information on their individual identities or characteristics.

Listing means identifying population elements individually, with sufficient information so that individual elements can be located, selected, contacted and enumerated as required in a survey following the listing operation. Often, information is also collected during listing on basic characteristics of units which may be useful for designing an efficient sample. Listing usually covers the total population at the location of listing.

The complexity and amount of information collected during listing can vary greatly. It may involve little more than preparing lists from visual inspection (such as of addresses or locations of occupied dwellings), in which case it may not be much more than careful counting.

Or it may involve much more complex information for the identification of particular subgroups in the population. In this case, it merges with screening.

Screening refers to an operation with the explicit objective of identifying members of a subpopulation. Like listing, but even more so, the amount and complexity of the screening information to be collected can vary greatly.

When the information required for screening becomes too complex or expensive to collect, the operation may be broken into two (or more) phases. This is two- (or multi-) phase sampling. The first phase may use simpler and less expensive data collection procedures in order to eliminate most of the units which are not members of the target population. The objective is to obtain a subset of units which (hopefully) contains the whole or at least a vast proportion of the target subpopulation members, and does not contain too many non-members. More complex and expensive methods are then used
in the second phase in order to refine the screening operation. One of the functions of two- (or multi-) phase sampling is that it can form a step in the screening operation. We return to this in Section 6.7.3 below.

### 6.7.2 Screening

#### A. Issues in quality of the screening operation

Screening is useful when the result contains few ‘false negatives’, meaning cases which actually belong to the subpopulation of interest, but are classified in error as non-members during screening. These false negatives will be missed from the survey coverage. The part which failed the screening test should be checked for the extent to which target population elements have been excluded mistakenly as non-members.

False negatives arise primarily from response errors in the screening interview. It is worth quoting Sudman (1972) at some length:

> “Since screening is so expensive there is a natural tendency to attempt to shorten and simplify the screening interview. This could be done by asking a series of screening questions, so constructed that a ‘no’ would end the interview. … The major drawback of this approach is that the respondent quickly recognizes that a very specific type of respondent is sought. It is simple for the respondent to act as a ‘gatekeeper’ and to prevent an interview from being conducted merely by saying ‘no’ to one of the questions. That is, the household is deliberately misclassified by the respondent on the screening interview to avoid the perceived bother of being interviewed. … The solution to this problem requires a longer screening interview [that] prevents the ‘gatekeeper’ from functioning … Sometimes, the interviewer is not told what the desired characteristics [for being selected for further interviewing] are, so that neither the respondent or the interviewer can act as ‘gatekeepers’ …”

Apart from the above ‘gatekeeper’ errors, screening can be subject to serious misclassification errors, errors which are not distributed randomly but tend to systematically miss members of the rare population:

> “… the magnitude of misclassification errors may easily be underestimated. Previous studies have shown misclassification errors of 5-10 per cent on many variables looked at singly. These errors multiply, however, when the rare population is defined by many characteristics. Thus for some complex populations as many as 1/3 to 1/2 of eligible respondents could be missed. These errors can be reduced by making the screening questions more detailed, but this adds to the length and cost of the screening interview. Alternatively, one could loosen the criteria for eligibility on the final interview, again adding to the cost of interviewing. Although misclassification cannot be avoided, it can be minimized by careful, thorough screening. …” (ibid).

There is also a problem of ‘false positives’, meaning cases which actually do not belong to the subpopulation of interest but have been classified in error as members during screening. “Loosening the criteria for eligibility” mentioned in the above quotation can increase the incidence of false positives. However, false positives are less of a problem.
They can usually be eliminated during the second, more intensive, phase – though at some extra cost.

**B. Controlling screening costs**

Sometimes it is possible to take steps to reduce screening costs. Possible measures include the following.

1. Separating out areas empty and nearly empty of members of the rare population of interest.

2. Separating out areas with greater concentrations of the rare population from areas of low concentration, and oversampling the former and under-sampling the latter.

3. Taking properly the cost of listing and screening into account in choosing the survey design, in particular in determining the sample size per cluster in a multi-stage design.

4. Using large clusters, so that travel and related costs can be reduced. See further comments below. Within households, collecting data on all eligible persons rather than selecting only one person can be economical, especially when the interest is in a subpopulation rather than the whole population.

5. Using proxy responses where possible with minimum compromise of the data quality. When data are to be collected on all persons in a household, one household member can often answer for all household members. Some surveys obtain information also from neighbours, but this is unlikely to be a practical or acceptable method in child labour surveys.

6. Using cheaper modes of data collection for listing. However, the scope for this is very limited in the context of child labour surveys where – apart from information collected solely through observation – generally face-to-face interviewing is essential for data collection. There is little scope for cheaper modes such as mail or telephone surveys.

7. Where possible, using the results of screening for more than one survey, so that the cost can be shared. Here an important consideration is whether the screening is used to classify stable units such as areas, or to identify more variable elements such as individual persons in the population. More stable, higher stage units are more suited to sharing across different surveys and survey rounds over time.

**C. Increasing the sample take per cluster**

It is a well-established advice that one should select a large sample per cluster when the interest is in a rare population, rather than in the total population.

In order to describe this idea in concrete terms, let us consider a two-stage design. As elsewhere, the term ‘cluster’ is used here to refer to area units to be selected at the first stage. The second stage may involve, within each selected cluster, the subsampling of elements of the total population (T), within which are identified (possibly through a screening operation) members of the rare population of interest (R). Often in practice, at the second stage we dispense with subsampling in order to obtain a sufficient number of cases if the subpopulation of interest is very rare. This is ‘compact cluster’ sampling.
The sample take per cluster (the average number of population elements selected per cluster) may be increased by (i) taking larger areas as the PSUs or clusters; and/or by (ii) increasing the subsampling rate within clusters (up to the point of compact cluster sampling).

The effect on efficiency of increasing the sample size per cluster may be seen in quantitative terms as follows. We are considering a situation when the rare population of interest is spread out across the total population, uniformly to a lesser or greater degree (in contrast to being concentrated in a few pockets). Let \( b_T \) be the optimal sample size per cluster for a survey of the total population, and \( b'_T \) the optimal sample size per cluster for a survey focused on the rare population (both these quantities being in terms of the number of elements of the total population). By ‘increasing the sample take per cluster’ we mean making \( b'_T > b_T \).

In a two-stage design, let \( C \) be the cost per sample cluster, and \( c \) the cost per interview once the interviewers are in the cluster. Per unit cost \( c \) includes any cost of listing and screening. First consider a survey based on the total population \((T)\). Optimum cluster size (meaning average sample size per cluster) is given by the following well-known expression (see Kish 1965, 8.3B, for instance):

\[
 b_T = \left( \frac{C}{c_T} \frac{1 - \rho_T}{\rho_T} \right)^{1/2},
\]

where \( \rho \) is the intra-cluster correlation and subscript \( T \) stands for ‘total population’.

Now consider a survey based on a rare population, say \( R \). The optimum cluster size becomes:

\[
 b_R = \left( \frac{C}{c_R} \frac{1 - \rho_R}{\rho_R} \right)^{1/2}, \quad b'_T = \frac{1}{P} b_R.
\]

Here \( c_R \) is the per unit cost for the rare class. It includes the cost of interview with a member of the rare class, as well as the cost of listing and screening to obtain the sample for the rare class. Obviously \( c_R > c_T \), the latter being per unit cost for a survey of the total population, which does not involve screening and is likely to need only a shorter, cheaper interview. Consequently, assuming \( \rho \) values for the rare class to be similar to those for the total population (see below), we have \( b_R < b_T \). However, in order to obtain \( b_R \) interviews with the rare class, it is necessary to have a cluster size \( b'_T = b_T / P \) in terms of the number of general population members per cluster, where \( P \) is the proportion of the rare class in the population. Assuming similar \( \rho \) values, and noting that \( c_R > c_T \), we have

\[
 \frac{b_T}{b'_T} = P \left( \frac{c_R}{c_T} \right)^{1/2} > P.
\]

It can be shown in most cases that

\[
b_T < b'_T < (b_T / P).
\]

Below is a brief comment on the assumption made above regarding the approximate equality of \( \rho_T \) and \( \rho_R \).

If a simple random sample (SRS) is taken from a cluster, then it has been established that the intra-class correlation is approximately the same for the total population of the
cluster and a subsample from the cluster (Hansen, Hurwitz and Madow, 1953, vol. I, 6.8; the proof is given in vol. II, 6.7). This essentially reflects the fact that average physical distance between elements in a cluster is not affected by simple random subsampling. A well-distributed rare class may be considered similar to a SRS from the total population. To the extent that subsampling departs from SRS (i.e. the rare class is unevenly distributed), we expect \( \rho_R \) to exceed \( \rho_T \) (Kish, 1965, 5.6B), thus adding to the effect of \( c_R \) exceeding \( c_T \) noted above. Nevertheless, we can still expect the effect of factor \((1/P)\) in Equation (6.11) to predominate, and hence \( b'_T > b_T \).

A simple numerical illustration would be useful. With \( c_T \) as the per unit cost in a survey of the total population\(^{26}\) and \( P \) as the proportion of the rare class in the population, let

\[
    c_R = c_T + \left( \frac{1-P}{P} \right) c_L \quad \text{or} \quad \frac{c_R}{c_T} = 1 + \left( \frac{1-P}{P} \right) \frac{c_L}{c_T}
\]

be the per unit cost of surveying the rare population. Here \( c_L \) is the per unit cost of listing and screening in the total population. This cost for proportion \((1-P)\) population units not in the rare population is shared among the remaining proportion \( P \) of the population units which are in the rare population. There would likely be some additional costs in \( c_R \) which, for simplicity, have been ignored in (6.14).

From (6.12) and (6.14) we have

\[
    \frac{b'_T}{b_T} = \frac{1}{P} \left( \frac{c_R}{c_T} \right)^{-0.5} = \frac{1}{P} \left( 1 + \frac{1-P}{P} \frac{c_L}{c_T} \right)^{-0.5}.
\]

Table 6.17 shows \((b'_T/b_T)\) as a function of \( P \) and \((c_L/c_T)\). This is the factor by which the optimal cluster size would be increased as we move from a survey of the total population to a survey of a rare population. For example, if it is a common practice to use a cluster size of \( b_T = 10 \), and \( P = 0.25 \), \((c_L/c_T) = 0.25 \), then \((b'_T/b_T) \approx 3 \) so that \( b'_T = 30 \). With \( b_T = 15 \), we would have \( b'_T = 45 \). Similarly for a larger class, \( P = 0.50 \), and relatively more costly listing and screening cost, \((c_L/c_T) = 0.50 \), we have \((b'_T/b_T) = 1.6 \), so that \( b'_T \approx 15 \) corresponding to \( b_T = 10 \), and \( b'_T \approx 25 \) corresponding to \( b_T = 15 \).

Table 6.17. Illustration of \((b'_T/b_T)\) as a function of \( P \) and \((c_L/c_T)\)

<table>
<thead>
<tr>
<th>( P )</th>
<th>0.10</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.40</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>((c_L/c_T))=</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td>7.25</td>
<td>4.23</td>
<td>3.51</td>
<td>3.00</td>
<td>2.33</td>
<td>1.91</td>
</tr>
<tr>
<td>0.20</td>
<td>5.98</td>
<td>3.73</td>
<td>3.16</td>
<td>2.75</td>
<td>2.19</td>
<td>1.83</td>
</tr>
<tr>
<td>0.25</td>
<td>5.55</td>
<td>3.54</td>
<td>3.02</td>
<td>2.65</td>
<td>2.13</td>
<td>1.79</td>
</tr>
<tr>
<td>0.30</td>
<td>5.20</td>
<td>3.37</td>
<td>2.90</td>
<td>2.56</td>
<td>2.08</td>
<td>1.75</td>
</tr>
<tr>
<td>0.40</td>
<td>4.66</td>
<td>3.10</td>
<td>2.70</td>
<td>2.40</td>
<td>1.98</td>
<td>1.69</td>
</tr>
<tr>
<td>0.50</td>
<td>4.26</td>
<td>2.89</td>
<td>2.53</td>
<td>2.26</td>
<td>1.89</td>
<td>1.53</td>
</tr>
</tbody>
</table>

When increase in sample size per cluster is achieved by taking physically larger areas as the PSUs (clusters), we can expect lower values of intra-cluster correlation in physically larger clusters (Hansen, Hurwitz and Madow, 1953, 6.27) – reflecting simply the

\(^{26}\) It includes the cost of listing and screening of the unit as well.
increased average physical distance between elements in the cluster. Increasing the physical size of clusters tends to inflate further the ratio \( \frac{b_T'}{B_T} \) of the subclass and total population cluster sample takes.

### 6.7.3 Two-phase sampling

As noted in Section 6.7.1, when the information required for screening is too complex or expensive, the operation may be broken into two (or more) phases. Forming the second step in the screening operation is one of the functions of two- (or multi-) phase sampling. Deming (1977) studied the ratio of the per-unit costs of the two phases of data collections, and concluded that it should be at least 6:1 for the procedure to be useful.

Multi-phase sampling is of course a general technique with wider application, going beyond this particular function in relation to screening. In more general terms, two-phase sampling refers to the selection of the final sample from a previously selected larger sample that provides information for improving the final selection. This also includes the possibility of using information obtained at the first phase as supplementary or auxiliary information for improving estimates for variables obtained only in the second phase.

The concept is explained as follows by Kendall and Buckland (1982):

“It is sometimes convenient and economical to collect certain items of information from the whole of the units of a sample and other items of usually more detailed information from a sub-sample of the units constituting the original sample. This may be termed two-phase sampling, e.g. if the collection of information concerning variate, \( y \), is relatively expensive, and there exists some other variate, \( x \), correlated with it, which is relatively cheap to investigate, it may be profitable to carry out sampling in two phases. At the first phase, \( x \) is investigated, and the information thus obtained is used either (a) to stratify the population at the second phase, when \( y \) is investigated, or (b) as supplementary information at the second phase, a ratio or regression estimate being used. Two-phase sampling is sometimes called ‘double sampling’. Further phases may be added if desired. It may be noted, however, that multi-phase sampling does not necessarily imply the use of any relationships between the variates \( x \) and \( y \). The expression is not to be confused with multi-stage sampling.”

There are thus three types of use of two-phase sampling:

1. to serve as the second, more precise, step in the screening operation
2. to provide information for improving the design and estimation of the second phase sample
3. to collect different but complementary information in the two phases - the first phase has a larger sample and the information collected is usually simpler, and the second phase usually collects more complex information.

The CLS-LCS distinction (Verma, 2008) provides a good example of two-phase sampling in child labour surveys involving all its three objectives. By CLS is meant a ‘child labour survey’ which is conducted over a relatively large sample of the general population of
children in order to estimate the proportion who are engaged in child labour and to study characteristics and variations in child labour rates. The LCS, ‘labouring children survey’, refers to a survey aimed at studying in-depth characteristics and conditions of working children. It is normally conducted over a relatively small sample of working children.

The LCS forms the second phase, and the CLS fulfils various functions as its first phase.

1. The CLS can be used for screening to identify samples of labouring children for the subsequent LCS. This does not necessarily mean that the particular working children identified in the first phase are individually passed on to the second phase. The identification may be in terms of units at a higher stage – e.g. of households containing working children, or of households containing any children, or even of areas containing concentrations of working children at the time of the first phase. The form of linkage between the two phases depends, among other things, on the relationship between the two phases in terms of the timing of the survey.

2. The CLS provides information for sample design and estimation for the second (LCS) phase.

Concerning design, the LCS may be based, for example, on a subsample of CLS sample areas. Information collected during the CLS can be used in selecting the LCS sample areas – information such as for improved stratification, on the number and proportion of working children in each CLS area, on areas which may be excluded from the LCS because they contain few working children, etc.

The distribution of working children by various characteristics in the larger CLS sample can be used to calibrate the sample weights of the smaller LCS sample. Other CLS data may also be useful for improving LCS estimates.

3. The two phases provide complementary statistics: the CLS primarily on the number, proportion and distribution of children who are engaged in labour, and the LCS primarily on characteristics and conditions of working children.

6.8 Cumulation over time

Cumulation of data from repetitions of the survey over time, when available, is an obvious method for increasing the sample size. Other procedures for enhancing selection probabilities of rare elements will be discussed comprehensively in the following chapters.

When survey data collection is repeated over time, we can take advantage of that feature by cumulating data for sampling and estimation for rare populations (Kish 1999). Cumulation is more efficient when the samples over time in the repeat survey are independent; cumulation is least efficient when the repeat survey is a panel with a largely unchanging sample over time.

When two or more data sources contain a set of variables each measured in a comparable way for the same type of units, then the information may be pooled either (a) by combining estimates from the different sources, or (b) by pooling data at the micro level. Technical details and relative efficiencies of the procedures depend on the
situation. The two approaches may give numerically identical results, or the one or the other may provide more accurate estimates; in certain situations, only one of the two approaches may be appropriate or feasible in any case.

Consider for instance the common case of pooling results over years. For linear statistics such as totals, pooling individual yearly estimates say $\hat{\theta}_i$ with some appropriate weights $p_i$ gives the same result as pooling data at the micro level with unit weights $w_{ij}$ rescaled as $w_{ij}' = w_{ij} \left( p_i/\sum w_{ij} \right)$. For ratios of the form $\hat{\phi}_i = \frac{\sum w_{ij} \cdot v_{ij}}{\sum w_{ij} \cdot u_{ij}}$, the two forms give very similar but not identical results, corresponding respectively to the ‘separate’ and ‘combined’ types of ratio estimate.

Let us consider some common issues involved in pooling of different sources pertaining to the same population or to largely overlapping populations. In particular, the interest is in pooling over survey waves in a survey in order to increase the precision of subpopulation estimates. Estimates from samples from the same population are most efficiently pooled with weights in inverse proportion to their variances (meaning, with similar designs, in direct proportion to their sample sizes). Alternatively, the samples may be pooled at the micro level, with unit weights inversely proportional to their probabilities of appearing in any of the samples. This latter procedure may be more efficient (e.g. O’Muircheataigh and Pedlow, 2002), but may be impractical as it requires information, for every unit in the pooled sample, on its probability of selection into each of the samples irrespective of whether or not the unit appears in the particular sample (Wells, 1998). Another serious difficulty in pooling samples at the micro level is that, in the presence of complex sampling designs, the structure of the resulting pooled sample can become too complex or even unknown to permit proper variance estimation. In any case, different waves of a survey normally do not correspond to exactly the same population. The problem is akin to that of combining samples selected from multiple frames, for which it has been noted that micro-level pooling is generally not the most efficient method (Lohr and Rao, 1996).

For the above reasons, pooling of wave-specific estimates rather than of micro data sets is generally the appropriate approach to aggregation over time from surveys.

With cumulation over time, the estimates produced are period, rather than point-in-time, estimates that can be difficult to interpret when the characteristics of analytic interest vary markedly over time (Citro and Kalton, 2007). Therefore, it may be argued that averaging and similar ‘manipulation’ is not acceptable, or at least that it introduces bias, since it alters the measures we obtain.

However, while it may be true in a literal sense that pooling alters the measures, this is not a sensible objection in many situations. We need a pragmatic and not an ideological approach to statistics. All statistical measures are constructed for the purpose of conveying some meaning, for providing some interpretation to real and complex situations. The particular forms of measures chosen are always determined by considerations of usefulness and practicality, are always compromises and in themselves not ‘sacred’. The objectives of pooling include searching for measures which convey essentially the same information as the ‘original’ un-pooled measures, but in a more robust manner, reducing random variability or noise. A related objective of pooling is trading dimensions – gaining in some more needed directions by losing something less needed for the particular purpose – such as permitting more detailed geographical
breakdown but with less temporal detail. A third objective is to summarise over different dimensions, providing more consolidated and fewer indicators. Such indicators are of course different from the more numerous ‘raw’ indicators, but are often more, or at least equally, meaningful and useful (Verma, Gagliardi and Ferretti, 2013).
Chapter 7

Multiplicity sampling

Multiplicity sampling offers design options which are not available in surveys based on traditional sampling. Sometimes these options can substantially improve the results of the survey in situations where traditional sampling is inefficient. The framework for multiplicity sampling was developed by Sirken and colleagues (see in particular, Sirken, 1970 and 2004). There have been diverse applications and developments of multiplicity sampling. In a sketch of its history, Sirken (2004) identifies the various phases in the development of different applications.

In this chapter we describe and illustrate fundamental features of the multiplicity sampling approach, and indicate situations where the approach may be useful in surveying rare populations of labouring children.

The basics of sampling with multiplicity were developed in Chapter 4. Sampling with multiplicity is the common basis of several different sampling techniques. The most direct application, in a somewhat specific form, is what is normally referred to as multiplicity sampling (or as ‘network’ sampling). This is described and illustrated in this chapter. Other applications of sampling with multiplicity include sampling from multiple frames and adaptive cluster sampling. Both these techniques are useful for surveying rare populations, and are discussed in Chapters 8 and 9 below. Other examples of sampling with multiplicity involve sampling in space and time, termed as time-location sampling, used for enumerating mobile populations (Chapter 10). The respondent-driven sampling estimator (Chapter 14) is also in the form of a multiplicity estimator.

7.1 Basics of multiplicity sampling

7.1.1 Relationship between sampling units and analysis units

The basis of the multiplicity sampling approach is the distinction between sampling units and analysis units. The process of sample selection is applied to the sampling units. Then, linking the sampling units to analysis units through multiplicity rules, a sample of the analysis units is obtained.

Example 1

Consider a household survey designed to estimate the total number of child domestic workers. A probability sample of households is drawn and information is obtained on children living in the household who are domestic workers (in one or more households, not necessarily including the household where they

---

27 This distinction was introduced in Sections 3.4.2 and 4.1.2B, and developed and illustrated in Sections 4.5 and 4.6.
7. Multiplicity sampling

live), as well as on children not residing in the selected household but who are employed as domestic workers by that household. In a conventional sample survey, estimation of the total number of child domestic workers would only use the information regarding members of the sample households, i.e. counting domestic workers only in the households where they live. In such sampling, it would be wrong to use the information on non-members of the sample household and extrapolate the results using the conventional weights. It would lead to over-counting the non-members, since they have the possibility of also being selected from other households.

In multiplicity sampling, estimation uses information both on members of the sample households, and on members of other households linked in some way to the sample households. In this example, households are the sampling units, and child domestic workers are the analysis units. There is a many-to-many relationship between units of these two types. A child domestic worker may be linked to more than one household – the household where the child lives, plus any other households where the child is employed as a domestic worker. A household may be linked to more than one child domestic worker living and/or working in the household. For instance, in a simple random sample of households the probability of a child appearing in the sample is proportional to the number of households with which the child is linked.

This example is developed and analysed in Section 7.6 below.

7.1.2 Reporting units and reported units

A very similar distinction is made between ‘reporting units’ (who provide information on other units, possibly including also themselves), and ‘reported units’ (units reported on, by other units in the population). Normally, the reporting units correspond to sampling units – they are identical or have one-to-one correspondence, such as one member of a selected household providing the information. Similarly, the reported units normally correspond to analysis units. But the concepts, though very similar, are not necessarily identical.

Let us consider multiplicity sampling in the context of a household survey. In a conventional sample survey, respondents are asked either about themselves and at most about other members of their household. However, in principle, there is no reason to limit the interview only to household members. Respondents could be asked about other persons, relatives, friends and acquaintances, neighbours, fellow members of social groups, co-workers, etc.

Multiplicity sampling seeks to make the survey design more efficient by utilising information that individuals not necessarily residing in the same household are able to report about one another by virtue of their relationship. That is, individuals linked to one another by specified relationships are eligible to serve as reporting units (respondents) for one another. The individual is linked to a group of related persons that are eligible to report on him/her, and the individual belongs to a group of related persons that he/she is eligible to report upon. The group of related persons an individual is eligible to report on has been termed as the individual’s ‘cluster’, and the group of related persons who
are eligible to report on the person concerned as the individual’s ‘network’. The size of the network is the person’s ‘multiplicity’. A person’s network and cluster may or may not contain the same individuals.

The concepts are illustrated in Figure 7.1. The direction of the arrow points from the reporting person to the person reported.

Thus C1 reports on N1-N3 (its cluster). This set of links is shown as dotted lines in the diagram.

N1 is reported on by C1-C3 (its network). These links have been shown as solid lines in the diagram. Often the pairs (N1, C1), (N2, C2) etc. refer to the same individual, but not necessarily; also the number of individuals in the two sets may be different.

Figure 7.1. ‘Networks’ and ‘clusters’ in multiplicity sampling

Surveys based on multiplicity sampling have three versatile design features:

- multiplicity rules that define relationships on the basis of which individuals become eligible to serve as informants about each other;
- estimators that take into account variations in network and cluster sizes; and
- multiplicity rule weights that adjust for differences in network sizes.

It may be pointed out that the sampling-analysis unit distinction is more fundamental and general compared to the reporting-reported unit distinction. The latter type of

---

28 It should be noted that in the literature, the terms ‘cluster’ and ‘network’ have, unfortunately, been used to mean different things in different contexts.
distinction depends not only on the multiplicity rules adopted, but is also on particular respondent rules.

7.1.3 Multiplicity rules and respondent rules

The respondent rules in a survey specify persons or other data sources which are eligible to provide information on the unit being surveyed. Surveys differ as to whether or not response by proxy is allowed. In a survey of individuals, for instance, information on a person may be (or may have to be, depending on the respondent rules) provided by the person concerned, or by some member of the person's household, or by persons outside the household who are connected to the person concerned in some specified manner. The set of units in the sample on which information is to be collected in the survey is not affected by the respondent rules adopted. A respondent rule only concerns who can give the requested information on a person in the sample; it does not affect the set of units, the sample, about whom the information is being collected.

By contrast, multiplicity rules specify the type of link which determines the set of units on which information is to be collected. This set is determined on the basis of its links with the sampling units selected into the sample. This is independent of the respondent rules which have been used.

With multiplicity sampling, one common arrangement is to obtain information from units originally selected into the sample on themselves and on the additional units they bring into the sample on the basis of links defined by the adopted multiplicity rules. Alternatively, in some surveys with multiplicity, the additional units brought into the sample are approached directly to provide information on themselves.

Example 2

The following is a brief description of the design of a study using multiplicity sampling to estimate the incidence of missing children. In addition to definitional problems that would be found in any study, there are several special difficulties in studying missing children (Sudman, 1986):

“a. The actual numbers of households in which a child is missing for any reason is small and becomes even smaller if one wishes to distinguish between alternative reasons for being missing and between population subgroups. Even very large samples may uncover too few cases to be sufficiently reliable. It should be noted, however, that rareness of the population is not unique to missing children but is found for many other populations that are important for policy evaluation.

“b. Answering questions about missing children may be threatening to some respondents, thus leading to substantial underreporting. This would be the case if the child either is a victim of parental kidnapping or is a runaway, the two major reasons for a child being missing. These problems may suggest that survey procedures should not be used, but the alternative methods are probably even less valid. [Surveys on crime, for instance, obtain] far higher levels of reported crime than are obtained from administrative records, and it is likely that careful surveys using multiplicity sampling ... procedures will
obtain better estimates of missing children than it is possible to obtain from administrative records or other sources.

“In the typical survey … respondents are asked either about only themselves or about all household members. For rare populations, the number located is small, often one or less per 100 contacts. Theoretically, there is no reason to limit the interview only to household members. Respondents could be asked about other persons, relatives, co-workers, neighbours, fellow members in organizations, friends, and acquaintances. To make the data useful, however, the respondent must be able to give reliable information about these additional persons and must also be able to report the size of the network so that it is possible to compute the probability of any individual being selected in the sample. If this can be done, it is possible to make unbiased estimates of the incidence of the rare population that are more reliable than simple household estimates. …

“Since households with missing children are rare (if one uses any policy-relevant definition of missing) and since the topic may be sensitive, the use of multiplicity sampling seems promising. There are, of course, several key questions: a. Will respondents be able to report accurately about missing children in other households? b. If yes, what types of network can be used to obtain accurate information about missing children? By network types we mean relatives, neighbours, co-workers, etc. Past research has indicated that as the network size increases and the frequency of contact decreases, reporting about other network members becomes less accurate. On the other hand, the larger the network, the greater the amount of information that is obtained. The optimum trade-off between quantity and quality of information must be determined by empirical research.”

Example 3

As noted, multiplicity sampling refers to the situation where the sampling units provide information not only about themselves but also about other units of the population. In order to describe the estimation procedure with multiplicity sampling, let us consider the following situation. As a simple example, suppose we wish to estimate the number of children working in a particular sector of activity. In a household-based survey, information is asked about the presence of working children in the household. But one could ask respondents not only about working children in their household but also about working children in households of, say, their brothers and sisters.

A child neither of whose parents has any brothers or sisters has only one chance of falling into the sample, which occurs if that child’s household is selected into the sample. A child whose father has two brothers and whose mother has one brother and two sisters, all living in different households, has six chances of falling into the sample, provided that these different households are in the sampling frame. If adult respondents report the information correctly, that child will be enumerated in the survey if his/her own household, or that of any of the five siblings of the parents are selected into the sample. To compute unbiased estimates, the data must be weighted. In an equal probability sample of households, if a child with no parental siblings gets a weight of
7. Multiplicity sampling

A child with five parental sibling households gets a weight of \((w/6)\) to compensate for a six-fold increase in his/her probability of appearing in the survey.

More specifically, in this example household head and spouse are each asked about the households of their siblings. They are requested to report the number of such households there are within the area of the study population, and how many working children they contain. Let household \(j\) be selected with probability \(f_j\) and contain \(n_j\) working children; let \(h_j\) be the number of ‘sibling households’ as defined above connected to this household and \(m_j\) the total number of working children those households contain. Then the conventional estimate of the number of working children obtained on the basis of data from household \(j\) is

\[
\hat{N}_j = \left( \frac{n_j}{f_j} \right) ,
\]

(7.1)

where factor \((1/f_j)\) is the inverse of the household selection probability and is taken as the weight of the sample household in the estimation equation. Summing the \(\hat{N}_j\) values over households in the sample gives an estimate of the total number of working children for the population in private households.

Different types of estimator are possible with multiplicity sampling. The estimators differ from each other with respect to the way the multiplicity information is used. The most commonly used is the multiplicity estimator, which counts every person in the target population as many times as he/she is reported by a different household in the sample, and weights each report by the inverse of the person's multiplicity.

Following Sirken (1970), the multiplicity estimate of the total number of working children based on the data from household \(j\) is

\[
\hat{N}_j = \left( \frac{1}{f_j} \right) \left( \frac{n_j + m_j}{1 + h_j} \right).
\]

(7.2)

Each of the \((n_j + m_j)\) children could have been reported from any of the \((1 + h_j)\) households, the selected household \(j\) and its \(h_j\) ‘sibling’ households. Factor \((1 + h_j)\) is the ‘multiplicity’ of reporting a working child linked to the sample household in question. This is the factor by which the probability of being reported is increased for each child in set \((n_j + m_j)\).

A practical remark on the above illustration: the illustration makes two important assumptions. Firstly, it is assumed that individuals are willing and able to report on all persons in the group as defined by the rules of the survey. Secondly, there is the assumption that the types of relationship chosen to define multiplicity are symmetrical: information about how many persons would report on A can be obtained by asking A about how many persons A would report on.

It is important to keep in view that the kind of social relationships which meet the above assumptions can be very culture-specific. The same rules cannot be applied to all populations or all situations.
7.2 Choice of multiplicity rules

It is important to emphasise that the amount and complexity of additional information to be collected for application of the multiplicity sampling method can vary greatly depending on the multiplicity rules chosen. This choice determines not only the complexity and cost of the operation, but also the quality of the resulting data.

7.2.1 A detailed illustration

In this subsection, we will give a detailed example of the application of multiplicity estimation in a survey, bringing out some practical issues concerning data quality, in particular the importance of the multiplicity rules adopted.

A. The study

The study (Czaja and Blair, 1990) examined a number of multiplicity counting rules, comparing them to a conventional counting rule in terms of victim reporting rates and error analysis.

A sample of known crime victims was selected from the police department records of a small area (in the USA). Three types of crime were covered: robbery, assault, and burglary. A general crime victimization interview was conducted with each victim. Data collection was conducted primarily by telephone, with face-to-face interviewing used for only a few respondents who were not reachable by telephone.

In the interview, respondents were asked both about crimes that had happened to them personally, and also about crimes that had happened to members of their pre-specified network. At the end of the interview, the name and telephone number of a randomly selected member of the victim’s network was elicited. These network members were phoned and the same interview was conducted with them. This design provided for simple comparisons of self and network reports. The victim and network member samples were combined with a general population (decoy) sample to mask the source of the original list from the interviewing staff and to provide anonymity to respondents. The sample was selected using a disproportionate stratified probability sampling procedure with systematic random sampling within strata. The stratification was by type of respondent (victim, network member, and decoy) and by type of victimization (robbery, burglary, and assault).

As noted, the primary difference between multiplicity and conventional surveys is the number of enumeration (sampling) units to which target respondents (analysis units) are linked. In conventional surveys, the multiplicity rule links each target respondent to one enumeration unit: e.g. his/her place of residence. In multiplicity sampling, respondents are linked to, i.e. can be reported (brought into the survey) and/or reported on by, additional enumeration units. This linkage is accomplished by using multiplicity counting rules. The most common types are rules which link individuals to the households of pre-specified relatives and close friends - rules 1-6 below provide examples. The resulting sample is a probability sample, although all elements do not have the same probability of selection. In data analysis, cases are appropriately weighted to adjust for the different probabilities of selection. The number of units eligible to report on the target respondent is elicited in the interview; the inverse of the
network size is typically the case weight. For multiplicity sampling to be successful, at a minimum two broad conditions must be satisfied: (i) the network respondents must be able and willing to report about the event or characteristic of interest (where relevant, they must also have a reasonable knowledge of the time period in which the event occurred); and (ii) they must know the size of the eligible reporting network.

B. Multiplicity counting rules

Which multiplicity rules to use in a network survey is a critical decision. Empirical knowledge is needed for deciding on how broad or narrow to make the reporting rules, depending on the specific situation and purpose. On one hand, the use of a broad multiplicity counting rule allows the target person or household to be reported on by multiple respondents which in theory minimizes the number of households to contact. However, as one expands the number of potential respondents, increases in reporting errors are also likely.

The study being discussed in this example assessed the six multiplicity counting rules listed below. Rule 1 is the conventional rule, where information on crime victims is sought only at their own household. Rules 2-6 add reporting at households of other persons step-by-step, from victim’s siblings (Rule 2) to all close relatives and friends (Rule 6).

- **Rule 1**: A conventional rule in which crime victims are linked only to their usual residence.
- **Rule 2**: A sibling rule in which the victims are linked to their usual residence and to the residences of their siblings.\(^{29}\)
- **Rule 3**: A parent and children rule in which the victims are linked to their usual residence and to the residences of their parents and children.
- **Rule 4**: A relatives rule in which the victims are linked to their usual residence and to the residences of their siblings, parents and children.
- **Rule 5**: A close friend rule in which the victims are linked to their usual residence and to the residences of their close friends.
- **Rule 6**: A combined rule in which the victims are linked to their usual residence and to the residences of their relatives and close friends.

In the study being discussed, the incidence of victimisation (\(v\)) for the population at risk (\(N\)) was estimated using the following multiplicity estimator:

\[
v = \frac{1}{N} \sum_{i=1}^{m} \left( \frac{1}{f_i} \sum_{j} \left( B_{(r)ij} / S_{(r)ij} \right) \right).
\]

In the above expression:

- \(m\) number of households in the sample;

\(^{29}\) By ‘linked to x’ is meant brought into the survey through or reported on by x, when x is selected into the sample.
7.2 Choice of multiplicity rules

\[ f_i \]  
selection probability of household \( i \);

\( (r) \)  
subscript \( (r=1-6) \) indicating which of the above rules is applied;

\[ B_{(r)j} \]  
=1 if crime victim \( j \) is reported in sample household \( i \), otherwise =0. Household \( i \) could be the victim’s own household, or any other household linked to the victim via the particular multiplicity rule \( (r) \). All reports from such households are counted, even if some of them refer to the same victim;

\[ S_{(r)j} \]  
total number of different households in which the victim \( j \) and close friends/relatives (as defined by the eligibility rule \( r \)) reside. \( S = 1 \) for Rule (1); \( S \geq 1 \) for Rules (2)-(6).

C. Lessons from the example

Some results are shown in Tables 7.1 and 7.2. The actual number of crimes are supposed to be known in this experiment from the police records. Against these numbers are compared the numbers reported within the victim’s own household (the conventional rule) and the additional number reported through households of close relatives and friends (reports from Rule 6, less reports from Rule 1). Only 66 per cent of the crimes known from police records were reported in the victim’s own household. By comparison, the reporting rate was 47 per cent from households of close relatives and friends, which is around 0.7 of the rate reported directly at victim’s own household. The reporting rates varied from a low of 26 per cent for siblings to a high of 59 per cent for the child-parent rule. There was considerable variability by type of crime, especially for the victim households (burglary and robbery were reported reasonably well at 84 per cent and 72 per cent, respectively, but assaults were grossly underreported at only 29 per cent).

There are some important lessons from this example.

(1) Clearly, reporting rates cannot be expected to be as good through other households in the victim’s network, compared to reporting rates at the victim’s own household.

(2) The reporting rates depend on the particular multiplicity links chosen. In the present example, the sibling rule clearly had the poorest reporting rates. However, such results cannot be automatically generalised to other surveys and conditions. This is an empirical question, and the answer has to be established in the particular circumstances of each survey.

(3) It is not always easy to determine what the ‘multiplicity’ of the event of interest is, though this is an essential item of information for the estimation procedure. This is how the authors of the study being discussed express this problem in relation to the ‘friend rule’ (Rule 5):

“The friend rule provides the broadest coverage but it also presents a problem in estimation. Often there is not reciprocity between friends. Person A may name person B as a close friend, but B, when asked the same question, may not name A. This creates a problem in estimation because individuals come into the sample with unknown probabilities. We tried to overcome this by asking for a set number of friends, three, but it did not work. Approximately 47 per cent of the network friends did not mention the crime victim as one of
their three closest friends. Our estimator does not take this into consideration. This issue must be addressed in future work before the friend rule can be an acceptable component in network designs. ... we believe that there is a need for more experimentation with network sampling”.

Table 7.1. Crime victimisation reporting rates according to multiplicity rule

<table>
<thead>
<tr>
<th>TOTAL</th>
<th>Actual number of victims</th>
<th>Number reported</th>
<th>Proportion reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 1 victim’s household</td>
<td>580</td>
<td>383</td>
<td>0.66</td>
</tr>
<tr>
<td>Households of close relatives and friends *</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rule 2 siblings</td>
<td>76</td>
<td>20</td>
<td>0.26</td>
</tr>
<tr>
<td>Rule 3 children &amp; parents</td>
<td>42</td>
<td>25</td>
<td>0.59</td>
</tr>
<tr>
<td>Rule 4 (=2+3) close relatives (siblings, children &amp; parents)</td>
<td>118</td>
<td>52</td>
<td>0.44</td>
</tr>
<tr>
<td>Rule 5 close friends (excluding close relatives)</td>
<td>67</td>
<td>34</td>
<td>0.51</td>
</tr>
<tr>
<td>Rule 6 (=4+5) all close relatives and friends</td>
<td>185</td>
<td>87</td>
<td>0.47</td>
</tr>
</tbody>
</table>

* Excluding reports already covered under Rule 1, i.e. reported at the victim’s own household

Source: Czaja and Blair, 1990.

(4) Under-reporting can be a serious problem in multiplicity sampling, a problem which is likely to increase with expansion of the multiplicity rules. Ideally, this problem should be taken into account in the estimation. This requires information on completeness of reporting under different rules, which of course is not easy to come by. In principle, correction should be applied on the following lines. Suppose that we used rules (1) and (6) in combination, and a particular reported victim could have been reported by own household, and by 3 other households under rule (6). The person’s theoretical multiplicity is (1+3) = 4; but using the figure on completeness of reporting in Table 7.1, the person’s effective multiplicity is (1*0.66 + 3*0.47) = 2.07.

(5) Of course, the merit of multiplicity sampling is to increase the number of reports, from the same number of sample households: in the data of Table 7.1, from 580 under rule (1), to (580+185) = 765 under multiplicity rule (1)+(6).
### 7.2 Choice of multiplicity rules

#### Table 7.2. Crime victimisation reporting rates by characteristics of the victim and the crime: reporting from victim’s own household versus reporting from households of close relatives and friends

<table>
<thead>
<tr>
<th>Victims by:</th>
<th>Reporting by victim’s household</th>
<th>Reporting by close relatives &amp; friends</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Rule 1)</td>
<td>(Rule 6)</td>
<td>(Rule 6 / Rule 1)</td>
</tr>
<tr>
<td></td>
<td>actual number</td>
<td>reported number</td>
<td>reporting rate</td>
</tr>
<tr>
<td>TOTAL</td>
<td>580</td>
<td>383</td>
<td>0.66</td>
</tr>
<tr>
<td>SEX</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>235</td>
<td>159</td>
<td>0.68</td>
</tr>
<tr>
<td>Female</td>
<td>346</td>
<td>224</td>
<td>0.65</td>
</tr>
<tr>
<td>RACE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>414</td>
<td>304</td>
<td>0.71</td>
</tr>
<tr>
<td>Nonwhite</td>
<td>166</td>
<td>79</td>
<td>0.46</td>
</tr>
<tr>
<td>AGE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 35</td>
<td>275</td>
<td>167</td>
<td>0.60</td>
</tr>
<tr>
<td>≥ 35</td>
<td>305</td>
<td>216</td>
<td>0.70</td>
</tr>
<tr>
<td>CRIME TYPE*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burglary</td>
<td>718</td>
<td>383</td>
<td>0.53</td>
</tr>
<tr>
<td>Robbery</td>
<td>243</td>
<td>204</td>
<td>0.84</td>
</tr>
<tr>
<td>Assault</td>
<td>379</td>
<td>110</td>
<td>0.29</td>
</tr>
</tbody>
</table>

*Multiple crimes to the same victim are apparently not recorded separately in the survey.

Source: Czaja and Blair, 1990.

#### 7.2.2 An important principle in choosing multiplicity rules

Since counting rules used in multiplicity sampling do not generally ensure that every analysis unit is uniquely linked to a sampling unit, some analysis units may not be linked to any sampling unit, while others are linked to several units. Perhaps the easiest and the most direct way of avoiding the situation that the rule fails to link some analysis units to any sampling unit is to always incorporate the traditional counting rule within the multiplicity counting rule. For instance, multiplicity sampling based on the rule which links individuals to the households of their siblings would miss individuals with no siblings. But the problem is avoided if individuals are linked not only to households of their siblings (if any) but always also to their own household (the latter being the traditional counting rule). With this rule, individuals with $m$ siblings will have a multiplicity of $(m + 1)$, and those with no siblings will be included with multiplicity of 1 (arising from their own household).

Hence a practical useful class of multiplicity rules is one which has the property of supplementing the conditions of a conventional rule with other conditions for linking elements to enumeration units. Such multiplicity rules have two desirable properties.

1. They guarantee that every element is linked to at least one enumeration unit, if that is guaranteed in the conventional rule they incorporate.
(2) Such rules permit the survey to produce several sets of estimates – one set based on the conventional rule and the other based on the different multiplicity rules which incorporate the conventional rule. In this manner, the estimate based on the conventional rule is preserved, and provides a point of comparison.

Sirken (1974), for instance, provides the following example.

“The rule that links persons to their own residence as well as the residences of their siblings and children would produce four sets of estimates based on the following rules:

(1) persons are linked to their own residence;
(2) persons are linked to their own residence and to the residence of their siblings;
(3) persons are linked to their own residence and to the residence of their children;
(4) persons are linked to their own residence and to the residence of their siblings and children.

The first is a conventional rule. The others are multiplicity rules.”

### 7.2.3 Examples of multiplicity rules

The following provides examples of multiplicity rules in three hypothetical surveys.

The first (Table 7.3) is a survey to enumerate deaths during a certain reference period. A sample of households is selected. In the conventional design, deaths are assumed to be reported at a sample household only if the deceased person had that household as his/her last place of residence. The death may have occurred at the household, or at a place other than a private household to which the person concerned moved before death. In the multiplicity design, a death is also eligible to be reported by a certain type of relative of the deceased person who lives at a sample household, irrespective of the last place of residence of the dead person.

<table>
<thead>
<tr>
<th>Multiplicity rule</th>
<th>Conventional design</th>
<th>Multiplicity design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Events covered at enumeration unit (sample household)</td>
<td>Deaths of persons who died, or had their last private residence, at the selected household</td>
<td>Deaths of persons (1) who died, or had their last private residence, at the selected household, or (2) who had a specified type of relationship with any current resident of the sample household</td>
</tr>
<tr>
<td>Multiplicity and Source of information on eligibility and multiplicity</td>
<td>1.0</td>
<td>1.0 for the dead person’s last private residence, plus 1.0 for each household containing a relative of the specified type.</td>
</tr>
<tr>
<td>Source of information on eligibility and multiplicity</td>
<td>Is this the dead person’s last place of residence? If no, does a relative of the specified type live in this household?</td>
<td></td>
</tr>
<tr>
<td>Number of other households</td>
<td>How many other households are there with a relative of the specified type as resident?</td>
<td></td>
</tr>
</tbody>
</table>

The second (Table 7.4) is a survey to enumerate patients receiving a certain type of treatment at a medical centre. The patient may be receiving treatment from several medical centres,
but one of those is the centre with the primary responsibility for the patient’s treatment. A sample of medical centres is selected. In the conventional design, count is taken of patients for which the selected centre has the primary responsibility. In the multiplicity design, all patients receiving treatment at the centre are counted, irrespective of whether or not the centre concerned has the primary responsibility for their treatment.

The third (Table 7.5) is a household-based survey of children working in an establishment in a certain sector “A”. A sample of households is selected. In the conventional design, children in the household who work in an establishment in sector A are enumerated. In the multiplicity design, each of these children is asked additional information about other children working in the same establishment as them. In particular, information is obtained from each working child in the household on the total number of children working in their establishment, and how many of those live in a household within the study area. Other child co-workers may be living outside the study area, or may not be living in a private household anywhere. (It is assumed in this simplified example that a child works in at the most a single establishment, and that most establishments of interest are small so that a child can supply the requested information on other children working in his/her establishment.)
7.2.4 Multiplicity counting rule strategy

The following very useful practical advice is surmised from the work of Sirken, Sudman and others.

As noted by Sirken (1974):

“The objective of the [multiplicity] counting rule strategy is to select the optimum rule for linking elements [analysis units] to enumeration [sampling] units and to assign the optimum set of weights to enumeration units that are linked to the same element. The rules and the weights are considered optimum in the sense that they minimise the combined effect of sampling and measurement errors on the survey estimates subject to the cost constraints of conducting the survey”.

Below are described some conditions which may make multiplicity rules preferable to conventional rules. Consideration must be given to the joint effect of sampling and measurement errors in selecting the counting rules for multiplicity sampling. Furthermore, choice of the multiplicity design and counting rules requires information comparing the cost coefficients as well as the error components associated with different designs.

As noted by Sirken (2004), the basic idea is to foster multiplicity sampling in household surveys by using multiplicity counting rules to form linkage configurations which simultaneously satisfy, to the extent possible, these three conditions:

Condition 1. The multiplicity of every individual is equal to or greater than one.
Condition 2. The distribution of the multiplicities has large mean and small variance.
Condition 3. Individuals are linked to households that are able and willing to report the multiplicities and the variables of interest about those individuals.

Condition 1 controls coverage bias. This condition is usually satisfied by using multiplicity rules that link individuals to their own households as well as other households.

Condition 2 maximizes the sampling efficiency of multiplicity sampling. It is often satisfied by using multiplicity rules involving a multiplicity of relationships.

Condition 3 minimizes the response error effects of multiplicity sampling. It may be satisfied, for example, by multiplicity rules which link individuals to households of relatives, friends, neighbours, and colleagues with whom they have close and well-defined relationships.

It needs to be emphasised that Condition 3 underlies the basic requirements in applying multiplicity sampling. The potential gains of multiplicity sampling over conventional sampling can be realised only if certain basic requirements are met. These include the following.

i. The respondent must be able to give reliable information about these additional persons which are brought into the study through links with that respondent.

ii. The respondent must also be willing and able to report the number of different respondents who could be providing information on the same persons.
iii. When the survey involves identifying (through links with the initial sample) and then directly approaching additional respondents, the procedures adopted must meet ethical and confidentiality standards.

The quality of the information obtained depends on (i) above; it determines the presence and magnitude of different types of non-sampling error in the results obtained. Information (ii) defines the ‘multiplicity’ of the individual; this information makes it possible to compute the probability of an individual being reported upon in the sample. If this can be done, it is possible to make unbiased estimates about the rare population, which may turn out to be more accurate than estimates obtained from a conventional survey costing the same amount.

Points (i) and (iii) concern alternative modes of application of the multiplicity sampling method – namely, gather information on additional units indirectly through a given sample, versus bringing into the survey additional units for direct interviewing. Requirement (ii) is of course common to both these modes.

A. Sampling error

Since the multiplicity rules specify the conditions for linking analysis units (population elements) to sampling (enumeration) units, the population distribution of the analysis units among the sampling units is a function of those rules. Changing the multiplicity rule modifies the population distribution and thereby alters the sampling distribution and the sampling variance of the survey estimates. It would be a mistake to conclude, however, that the estimates based on multiplicity rules are necessarily subject to smaller sampling variance than estimates based on conventional rules. They are not, except under special conditions. Nevertheless, when the population distribution based on the conventional rule is highly skewed, multiplicity rules offer a strategy for redistribution of the analysis units among the sampling units and thereby decreasing the variance of the population distribution. Multiplicity rules may also exist which have the effect of redistributing the weighted elements among the sampling units so that the resulting distribution is less skewed.

B. Coverage errors

Coverage errors occur in conventional sampling when the sampling frame does not include some of the units to be sampled (under-coverage), or when the frame includes duplicates or extraneous units. Additional coverage errors occur with multiplicity sampling when sampling units either fail to report analysis units to which they are linked, or erroneously report analysis units to which they are not linked by the multiplicity rule. In most surveys, under-reporting appears to be a far more serious problem than over-reporting.

In certain situations, use of multiplicity counting rules can in fact reduce the undercoverage error. For instance, if by their nature analysis units have multiple links with sampling units, then it is natural and often more efficient to adopt multiplicity counting rules, rather than try and convert the frame to have a one-to-one correspondence between units of the two types. For instance, if working children visit multiple locations, we may obtain a sample of children from a sample of locations using
multiplicity counting rules, rather than associate each child in the target population uniquely with only one location.

Similarly, if the attribute being measured in a survey concerns a population that does not reside in the area covered by the survey, it is difficult and inefficient to implement the conventional coverage rules, such as selecting individual persons only from their current place of residence. Selection using multiplicity rules on the basis of links between persons and households can broaden and improve coverage of the target population. For instance, the procedure may be used to improve the coverage of working children who have left their family home. The rule would be to allow children to be selected into the sample both from their current place of residence (as in a conventional household-based survey), and also from their family home even if they no longer live there.

Also, multiplicity rules may sometimes be subject to smaller coverage errors than conventional rules in surveys which collect sensitive information. This happens, for instance, when the attributes or the situation to be reported are such that the respondents are more reluctant to report those about themselves than they are to report those about others, such as about friends or relatives linked to them through the multiplicity rules adopted in the survey.

C. Response errors

Response errors represent incorrect or incomplete information provided by the ‘reporting units’ on the ‘reported units’ linked to them through the counting rules of the multiplicity design. Sometimes, one among the multiple reporting units linked to a reported unit is the “best” in the sense of providing the most complete and accurate information about the reported unit concerned. The “best” reporting unit may or may not be the unit linked to the reported unit by the conventional rule. Consider for example a conventional household-based child labour survey. The conventional rule would be to select children from the household where they live. Suppose this is extended to a multiplicity design in which the selection of children is allowed also from the location where they are found to be working. For some items of information (such as the child’s working conditions) the best source may be the child’s work place rather than his/her household as in the conventional design, while for some other items (such as the child’s family characteristics) the best source may be the household where they live.

D. Survey costs

A multiplicity rule is not necessarily more efficient than the conventional rule it incorporates even if, for a fixed sample size of sampling units, it is subject to lower total error. This is because the cost of conducting a survey for a fixed sample size in terms of sampling units is inevitably larger for a multiplicity design than a conventional one. There are two reasons for this.

(1) The cost per sampling unit is greater because the average number of analysis units linked to one sampling unit is greater for multiplicity than for a conventional design.

(2) The estimator based on the multiplicity rule and estimation of its variance require additional information to determine the weights assigned to the units which are not involved in the conventional estimator. Collecting this information in the survey adds to the survey cost and burden.
7.3 Situations where multiplicity sampling may be useful

Like most sample design options, multiplicity sampling is more advantageous in some circumstances than in others. It has potential advantage in conditions which tend to present problems with conventional sampling, such as the following.

(1) Analysis units of interest have multiple links with sampling units in the frame

Often such ‘multiplicity’ already exists as a result of the nature of the sampling frame; it may also be introduced by the survey design adopted.

An example of this problem existing as a result of the nature of the available sampling frame is when we want a sample of ‘visitors’ from a frame of ‘visits’ to a facility. More than one visit may represent the same individual. One may sample child labourers from places they visit, but any one child may visit several places, and hence have a multiple chance of being selected into the sample. A simple example was given earlier: child domestic workers being reported from their household of residence as well as from households where they work.

(2) Estimates are needed for rare populations and small domains

The problem in estimating for small population domains arises from insufficient sample sizes for small domains. Multiplicity sampling attempts to reduce this problem by increasing the effective sample size by increasing the units’ selection probabilities by permitting reporting on them from multiple sources. In this manner the available sample size of the rare population may be increased. This has been illustrated in the above examples.

Though multiplicity counting rules enumerate more persons than the conventional rules, sampling variances associated with multiplicity counting rules are not necessarily smaller because other factors are relevant, including the extent of clustering and variability in the multiplicities. In some applications of the multiplicity procedures, sampling variances are increased owing to the weights used to account for the differential selection probabilities. The additional information required to apply the design adds to the cost of multiplicity sampling. Many variations have been developed with the objective of limiting the additional information required.

However, in some uses of multiplicity sampling it has been found that there are only moderate increases in sampling variances while there are very large cost reductions. Thus, taking both cost and variance into account, multiplicity samples can be much more efficient for rare populations than are conventional samples.

The answer often depends on the particular multiplicity rule applied, which is also situation and topic specific. Here is a positive example. Nathan (1976) examined relative efficiency of the following three multiplicity rules in the reporting of births and marriages, the conventional rule and two multiplicity rules:

i. the conventional rule, covering births in the mother’s household (similarly, marriages of residents of the household);
ii. births linked to households of mother and of mother’s mother (similarly, marriages of residents of the household, and also of residents in their parents’ households);

iii. births linked to households of mother, of mother’s sisters and of mother’s mother (similarly, marriages of residents of the household, and also of residents in households of their siblings or parents).

The author concludes that

“... there are clear indications that in this situation, for the ranges of sample sizes generally considered, the full multiplicity rule offers definite superiority over the conventional rule, for the same size sample. Cost considerations may reduce this superiority somewhat, although experience shows that the addition of reporting relatives and [questions] on multiplicity increases costs only marginally....”.

(3) It is necessary to locate and estimate for inaccessible populations, especially populations hidden or well-dispersed in the general population

Just as in the case of estimating for rare populations and small domains, insufficient sample sizes is also a common problem in surveying inaccessible populations. Difficulties in locating and accessing members of the target population often present additional problems. With multiplicity sampling it is possible to obtain information for locating members of the target population so that they can be interviewed directly. Thus, these procedures can be used not just to measure the size and basic characteristics of difficult-to-access populations, but they can also serve as an effective method for locating them for obtaining more detailed information directly.

Note that potential advantage (2), namely low cost expansion of the sample size using multiplicity sampling with indirect reporting, and objective (3), namely using multiplicity sampling to locate and directly approach additional respondents, cannot both be realised in the same survey. One has to choose between the two, depending on particular requirements.

(4) Individuals are reluctant or unable to provide information about themselves

Multiplicity procedures can help when the survey questions are sensitive and respondents are reluctant to report information about themselves or their household members, especially when the interview is held at home. These respondents may be more willing to report about others whom they know outside their own households. The problem concerns non-response, specifically refusals, with direct questioning. In certain circumstances, it is easier or more successful to seek information from other persons rather than from the person concerned directly – an option available under the multiplicity sampling procedure. Individuals that are missed by traditional sampling because their own households are unwilling (sometimes even unable) to report on them are linked by multiplicity sampling to other households that are expected to be more willing/able to report on those individuals.

“Examples of topics where this might be the case are child beating and alcoholism. This may also be the case for missing children. In the case of sensitive questions, there may be ethical problems with asking respondents
to report about others who can be identified, but there are no problems if
the data are used simply for estimation purposes and not to locate the rare
population” (Sudman, 1986).

In the literature, an application commonly reported concerns illicit drug use. Questions
on illicit drug use by friends appear to be less threatening and hence less subject to
under-reporting than questions on self-use of drugs. Multiplicity sampling may handle
this problem by using counting rules that link drug users to the households of their
friends. Another common application is provided by surveys on mortality in situations
with incomplete death registraction. Households containing surviving relatives of a
deceased person appear to be more knowledgeable in reporting the death than persons
in the household where the death occurred. Furthermore, they are the only possible
source of information in cases where the deceased person did not live in a private
household at the time of death, or where his/her household no longer exists.

Examples where such procedures might be useful in our present context include cases
of mistreatment or abuse of children, children performing hazardous or illicit activities,
children being prevented from attending school, or run-aways and other children
reported as missing from home.

(5) The sampling frame is incomplete

This problem concerns non-coverage. Some forms of multiplicity linkages can bring
into the survey members of the rare population who are otherwise missing from the
sampling frame and would not be covered using conventional procedures.

Here is an example of how multiplicity sampling may, with appropriate counting rules,
help in reducing coverage problems. In conventional household surveys, individuals are
uniquely linked to their place of residence where they are eligible to be enumerated.
Therefore, those who do not reside at households covered by the sampling frame are
missed by the survey. Similarly, individuals who, for one reason or another, are not
reported on by their own household would also be missed. Examples are persons
without fixed addresses, residing in unregistered or unlisted households, or living in
institutions. If such individuals could be linked to households that are covered in the
frame and that contain persons willing and able to report on them, then the problem
of under-coverage would be reduced. Multiplicity sampling attempts to achieve this by
adopting counting rules that link missed individuals to households that are covered by
the frame.

With an appropriate form of questioning, a household-based child labour survey, for
instance, may be able to cover working children living outside the household sector but
who can be linked in some definable way to a private household – e.g. own-children
who have left home, even children of close relatives and friends who have left home
and have no current private address from which they could be sampled.
7. Multiplicity sampling

7.4 Limitations and problems

Improvement in efficiency with multiplicity sampling over conventional sampling cannot be taken for granted – this has to be evaluated empirically. It is necessary to take a balanced view of the potential uses as well as limitations of the multiplicity sampling procedure. Possible limitations and problems concern the following areas.

1. Lack of salience of relationships linking individuals in the population

Successful implementation of the multiplicity sampling procedure depends on the salience of the family and social relationships on which its multiplicity rules are based. Are those relationships mutually recognised? Are they strong enough to support sufficient interaction among members for them to have knowledge about each other on the questions being asked in the survey? What types of relationship are more suited than others for the particular objective being pursued? Culture-specific and situation-specific solutions need to be identified and empirically verified.

2. Lack of salience of events and of objects of the enquiry

Some events such as deaths, births, marriages and other demographic events tend to be salient in people’s minds, and are well-reported in multiplicity surveys as demonstrated by many studies. However, events of many other types are less salient, and are difficult to report accurately, especially in relation to persons outside one’s own household. This may well apply to the reporting of labouring activities of children by households of relatives and friends.

Considerations similar to events apply also to individuals – some individuals are more important (‘salient’), others remain unknown or are more easily ignored or forgotten. This applies especially to persons who are not themselves members of any reporting network but are the object of the enquiry in which other persons or households report on them. Children often fall in this category – they may not be asked to report anything on themselves or on others, only others may be asked to report on them.

3. Ethical, confidentiality and privacy concerns

As noted by Kalton and Anderson (1986), an ethical question with multiplicity sampling is

“... whether it is appropriate to collect the survey data from an informant who is not even a member of the linked person’s household. In some cases this question may not raise much concern. However, with private information... there may be serious difficulties, and especially so if the survey design calls for a subsequent interview with persons identified ...”.

For instance, a child or a child’s parents may be upset, may consider it inappropriate, for an interviewer to ask the child’s friends or distant relatives personal questions about the child, about the child’s parents, their circumstances, behaviour, attitudes, etc.

4. Response errors

As already noted, reporting biases are often larger for multiplicity counting rules than for ordinary unitary counting rules. Because of the difference in the relative magnitudes of
7.4 Limitations and problems

Sampling errors and reporting bias associated with the multiplicity and unitary counting rules, the relative efficiencies of composite multiplicity rules typically decrease with increasing sample size. As noted by Sudman (1986),

“... some users of survey data have been concerned about multiplicity methods because respondents do not give completely accurate information about persons in other households. It must be remembered, however, that reports about persons in the respondents’ households are not perfect either. The question is whether there is differential accuracy of reporting, and if there is, what the magnitude of the difference is. In several applications … reporting about others in a network has been shown to be only very slightly less accurate than reporting about household members”.

However, the assessment is not uniform in the literature. Kalton (2009), for instance, reviews a number of studies which illustrate some serious limitations of the procedure, and concludes as follows.

“[Multiplicity sampling ] has not been widely used in practice for surveys of rare population members. … There is the risk that the sampled informant may not accurately report the rare population status of other members of the linkage, either deliberately or through lack of knowledge. Non-response for the main survey data collection is another concern. In addition, ethical issues can arise when sampled persons are asked about the rare population membership of those in their linkage when that membership is a sensitive matter”.

(5) Increased complexity

An interview in a multiplicity-based survey is likely to be more complex than a conventional interview on the same topic. Multiplicities have to be established, and information about households and persons away from the location of the interview has to be asked. It is not a matter simply of the additional information to be collected; more importantly, the more complex and sensitive interview content and survey situation requires better trained and more skilled field staff.

(6) Lack of ‘portability’ of methodological details

The methodology of a multiplicity-based survey tends to be very specific to the particular situation and topic of the survey, compared with similar surveys using conventional procedures. This means that details of the application tend to be less portable across time, cultures, conditions and topics. There is a greater need for local validation of parameters of the design such as the multiplicity rules chosen. This requires additional resources and technical knowledge often not available to researchers of child labour in developing countries.

(7) Lack of symmetry in social relationships

Symmetry in social relationships is the basic assumption of the method. A network is composed of all analysis units having the same linkage configuration. Networks are non-overlapping and exhaustive partitions of the population. Inclusion probabilities are supposed to be identical for all analysis units in a network. Obviously, if unit A is
in the network of unit B, then B is assumed to be in the network of A, their roles in
the network being indistinguishable. But real social relationships usually lack such
symmetry. Better-off members are often better-known in extended family circles than
are members in poorer circumstances. Younger siblings are often more knowledgeable
about children of their older siblings, than the other way round. Person B being among
close friends of person A does not necessarily mean that A is counted among close
friends of B. And so on. Such asymmetries tend to be culture-specific. They affect the
completeness of reporting and the appropriate multiplicities to be applied.

7.5 Estimation with multiplicity sampling

Multiplicity sampling involves many-to-many linkage between sampling and analysis
units. As explained in Section 4.6, the commonly used multiplicity estimator is a
Hanson-Hurwitz type of estimator, and may be expressed as a summation over the
sampling units selected into the sample, or as a summation over the analysis units
associated with the selected sampling units.

A. The estimator in the form of summation over a sample \( S^4 \) of the sampling
units

In this form the estimator is as given in Equation (4.5):

\[
\hat{Y} = \sum_{i \in S^4} \left( \frac{y_i}{f_i} \right) \quad \text{with} \quad y_i = \sum_{j \in b_i} \left( \frac{y_j}{m_j} \right),
\]

(7.4)

where subscript \( i \) refers to a sampling unit (e.g. a household), and subscript \( j \) to an
analysis unit (e.g. a child worker) linked to the sampling unit, and:

- \( f_i \) is the selection probability of a sampling unit \( i \);
- \( b_i \) is the number of analysis units linked to it under the multiplicity rules used;
- \( m_j \) the multiplicity of analysis unit \( j \); and
- \( y_i = \sum (y_j / m_j) \), with the sum being over the \( b_i \) analysis units linked to sampling unit \( i \).

Incomplete reporting

Equation (7.4) is the standard form. It requires information on selection probabilities
\( f_i \) only for the sampling units which have been selected into the sample. Multiplicity
\( m_j \) of analysis unit \( j \) equals the number of sampling units associated with it – each
associated sampling unit contributing 1 to this multiplicity. This assumes that all
information is fully reported, irrespective of the type and distance of the link between
units of the two types. There is ample empirical evidence, however, that the reporting
tends to be incomplete, and that the degree of incompleteness depends on the type
and distance of the relationship between the analysis unit and the associated sampling
units. See for instance, the results reported in Tables 7.1-7.2 above from the study of
crime victimisation.
7.5 Estimation with multiplicity sampling

Now looking from the point of view of a particular analysis unit \( j \) (e.g. a child), suppose that it is possible, from past surveys and pilot studies etc. to estimate the degree of completeness in reporting, say \( p_i \leq 1 \), for each of the \( i \in m_j \) sampling units through which the analysis unit \( j \) has come into the sample. (By this we mean that \( p_i \leq 1 \) is the chance that an actually existing link between, say, household \( i \) and child \( j \) is reported/realised in the survey.)

We may improve Equation (7.4) by replacing theoretical multiplicity \( m_j \) in it by realised (effective) multiplicity \( r_j \):

\[
r_j = \sum_{i_{(r)}} p_{i_{(r)}} \leq m_j. \tag{7.5}
\]

Subscript \( i_{(r)} \) indicates that the particular sampling unit \( i \) is linked to analysis unit \( j \) through a relationship of type “\( r \)”. The \( p_{i_{(r)}} \) values are determined by the type and proximity of relationships involved in the multiplicity rules chosen. Let us assume that the degree of completeness in reporting \( p_{i_{(r)}} \leq 1, i \in m_j \) does not depend on the particular \( i \) and \( j \) (particular sampling and analysis units), but only on the type \( (r) \) of the relationship between them. On this basis, values of \( p_{i_{(r)}} \) may be determined through small-scale studies collecting information on the following lines. The study should cover analysis units of different types in a variety of situations.

For each unit \( j \) in a sample of analysis units, determine the set \( (i \in m_j) \) of sampling units which should be reporting on \( j \). Then by approaching each of the sampling units in the set, check whether or not they report on the analysis unit concerned. The observations arranged by type of relationship \( (r) \) provide estimates of the \( p_{i_{(r)}} \) values, which can be averaged to obtain \( p_{i_{(r)}} \) for the particular type of relationship \( (r) \). If necessary, the estimation may be performed separately for different categories of analysis units.

B. The multiplicity estimator in the form of summation over analysis units

As explained in Section 4.6, exactly the same estimator as (7.4) can be expressed in terms of summation over analysis units \( j \) in sample \( S^B \) as given in Equation (4.10):

\[
\hat{Y} = \sum_{j \in S^B} \left( \sum_{i \in S^B} \left( m_{ij} / f_i \right) \right) \frac{y_j}{m_j} = \sum_{j \in S^B} \left( \frac{m'_j}{m_j} \right) y_j, \text{ say,} \tag{7.6}
\]

where, as defined in Equations (4.11) and (4.12),

\[
m'_j = \sum_{i \in S^B} \left( m_{ij} / f_i \right) \text{ weighted sum of the links of analysis unit } j \text{ to sampling units in the sample, with each link weighted by } w'_i = (1/f_i), \text{ the inverse of the selection probability of the concerned sampling unit } i,
\]

\[
m_j \text{ total number of links of analysis unit } j \text{ to all the sampling units in the population.}
\]

C. Essential requirements for application of the procedure

Before the procedure described above can be applied, it is very important to resolve some fundamental issues:
(1) Will respondents be able to report accurately about working children in other households?

(2) If yes, what types of relationship between persons/households (different relatives, friends, neighbours, co-workers etc.) can be used to obtain accurate information about working children? These are referred as ‘networks’. There is evidence that as the network size increases, or as the frequency of contact between its members decreases for other reasons, reporting about other network members becomes less accurate. On the other hand, the larger the network, the greater the amount of information that is obtained. As has already been emphasised, appropriate trade-off between quantity and quality of information must be determined on empirical basis, in the circumstances of each survey.

7.6 Numerical illustration of the multiplicity estimation procedure

7.6.1 Illustrative data

In this illustration we discuss in detail the situation depicted in Example 1 of Section 7.1.1 above. We considered a household survey designed to estimate the total number of child domestic workers from a probability sample of households. Child domestic workers who were living and/or were working in a sample household were taken into the sample, rather than only those who were residents of the selected household. As noted, in a conventional sample survey, estimation of the total number of child domestic workers would use the information only regarding members of the sample households. In multiplicity sampling, estimation uses information on all child domestic workers found in sample households, irrespective of whether or not they were resident of the particular household where they were found.

In this example, households are the sampling units, and child domestic workers are the analysis units. There is a many-to-many relationship between units of these two types. For this sampling procedure to give unbiased estimates the relationship between the sample households and the target units should be clearly defined and the questionnaire should be designed to provide the necessary information for estimation with accuracy.

The following illustration is based on Mehran (2012). In the illustration, we consider a very small sample of child domestic workers. A sample of households (say, $S^4$) is selected, in which each household has at least one (i) child member who works as a domestic worker within or outside the household, and/or (ii) a child who is a non-member engaged to work as a domestic worker by the household. The two populations - namely (A) the household population and (B) the child domestic worker population – are linked by two types of relationship. These are household membership and employer-employee relationship. The first relationship is one-to-many: a household may have more than one child domestic worker as member, but each domestic worker is member of one and only one household. The second relationship is many-to-many: any child domestic worker may be working in more than one household and any household may be engaging more than one child domestic worker.
To fix ideas, consider a population $S^A$ of 6 households ($Hi$, $i=1-6$) and a population $S^B$ of 5 child domestic workers ($Dj$, $j=1-5$), related to each other as follows.

- Domestic worker $D1$ lives and works as a domestic worker in household $H1$, and does not work as a domestic worker in any other household.
- $D2$ lives and works as a domestic worker in $H2$, and does not work as a domestic worker in any other household.
- $D3$ also lives in $H2$, but works as a domestic worker only in $H4$.
- $D4$ lives and works as a domestic worker in $H4$ and also works as a domestic worker in $H3$.
- $D5$ lives in $H5$ and works as a domestic worker in $H4$ and $H6$, but in not his/her own household.

This information is summarised in Table 7.6A. The following symbols have been used:

- LW: the child lives and works in the household; the child may also be working as a domestic worker in one or more other households;
- W: the child only works (but does not live) in the household as a domestic worker; the child may also be working as such in one or more other households, possibly also including the household where he/she lives;
- L: the child only lives (but does not work) in the household as a domestic worker; the child is working as such in one or more other households.

Any of these relationships would result in the selection of the child into sample $S^B$ of children if the associated household is selected. The last line of the table is the count of non-blank entries in the column concerned: this is the multiplicity ($m_j$) of the child domestic worker.

### 7.6.2 Multiplicity estimate of the total number of child domestic workers

Consider an SRS of 2 households out of 6. The household selection probability $f_i$ in Equation (7.4) is a constant: $(1/f_i) = (1/f) = 3$. Also, in estimating the count of child domestic workers, $y_j \equiv 1$. Hence the contribution of sample household $i$ to estimate of the total population count is

$$\left( \frac{y_i}{f_i} \right) = \frac{1}{f} \sum_{j \in h} \left( \frac{y_j}{m_j} \right) = 3 \sum_{j \in h} \left( \frac{1}{m_j} \right),$$

which is shown in the penultimate column of Table 7.6A. The last column of the table shows $b_i$, the number of child domestic workers to which household $i$ is linked. Table 7.6B shows the results for all 15 possible samples of 2 households out of 6. A row ($i$) and a column ($k$) identify the two households in a sample. The estimate of total number of child domestic workers shown in the cell defined by the two households is the sum $(Hi + Hk)$ from the penultimate column of Table 7.6A. Thus for $i = 3$ and $k = 4$ (the sample consisting of households H3 and H4), the estimate is the sum of the 3rd and the 4th row of Table 7.6A, namely $(1.5+4.0) = 5.5$. 
Table 7.6. Multiplicity estimate of number of child domestic workers from a sample of households

A. Households containing child domestic workers; multiplicity estimate from each household

<table>
<thead>
<tr>
<th>Households containing child domestic workers</th>
<th>Child domestic worker (j)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household (i)</td>
<td>D1</td>
</tr>
<tr>
<td>----------------</td>
<td>----</td>
</tr>
<tr>
<td>H1</td>
<td>LW</td>
</tr>
<tr>
<td>H2</td>
<td>LW</td>
</tr>
<tr>
<td>H3</td>
<td></td>
</tr>
<tr>
<td>H4</td>
<td>W</td>
</tr>
<tr>
<td>H5</td>
<td></td>
</tr>
<tr>
<td>H6</td>
<td></td>
</tr>
<tr>
<td>( m_j )</td>
<td>1</td>
</tr>
</tbody>
</table>

B. Multiplicity estimate of number of child domestic workers in the population from sample of 2 households (\( H^i + H^k \))

<table>
<thead>
<tr>
<th>Households</th>
<th>H2</th>
<th>H3</th>
<th>H4</th>
<th>H5</th>
<th>H6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household (i)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H1</td>
<td>7.5</td>
<td>4.5</td>
<td>7.0</td>
<td>4.0</td>
<td>4.0</td>
</tr>
<tr>
<td>H2</td>
<td>6.0</td>
<td>8.5</td>
<td>5.5</td>
<td>5.5</td>
<td></td>
</tr>
<tr>
<td>H3</td>
<td>5.5</td>
<td>2.5</td>
<td>2.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H4</td>
<td>5.0</td>
<td>5.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.0</td>
</tr>
</tbody>
</table>

The actual number in the population of 6 households is 5 child domestic workers.

It can be observed from the table that the mean of the 15 possible estimates of the total number of child domestic workers is 5.0 (with a variance of 3.2), precisely the number of child domestic workers in the target population. This result verifies that the estimator is indeed unbiased. It is possible to verify that this also holds for estimates of the number of live-in and live-out domestic workers separately. Live-in domestic workers (domestic workers 1, 2, and 4) are working as domestic workers in the household in which they live, although they may also be working as domestic workers in other households. Live-out domestic workers (domestic workers 3 and 5) are working as domestic workers only in households other than their own.

### 7.6.3 Survey questions for identifying multiplicity

It is important to identify the type of questions required for obtaining the necessary information on multiplicity of the target (analysis) units encountered in the sampling units selected for the survey in a household survey. Table 7.7 identifies a possible sequence of question for the purpose. Child workers encountered in a sample household are classified into one of the three categories, distinguished by the symbols used in Table 7.6: as LW (1a in Table 7.7); as L (2a in the table); or as W (3a in the table). Each such encounter contributes 1 to the child’s multiplicity. To this is added the number of (other) households the child works in. In the cases LW and L, the above sum gives the final multiplicity of the child worker. In the case W, the above sum is augmented by 1 (in order to cover the child’s own household, whether or not the child works there). For example, if household H1 is in the sample, we encounter there only child domestic worker D1. The child does not work in any other household, and hence
the child worker’s multiplicity is \((1+0)=1\). With \(H2\) in the sample, the same as the above applies to child \(D2\) (multiplicity = 1). We also encounter \(D3\) there, who works in some other household as a domestic worker. This gives the child’s multiplicity as 2: the child can come into the sample from the household of residence (even if he/she does not work there as a domestic worker), or from the other household where he/she works as a domestic worker. (That ‘other household’ is in fact \(H4\), but this information is not required for the multiplicity estimator.\(^{30}\))

With household \(H4\) in the sample, we encounter three child domestic workers, \(D3\), \(D4\) and \(D5\). Child \(D4\) lives and works in \(H4\), and works in one other household, giving a total multiplicity of 2. Child \(D3\) works in \(H4\), and lives in another household (\(H2\), but we need not know the identity of the household), these two items giving a total multiplicity also of 2. Child \(D5\) is working in \(H4\) and in another household (\(H6\), but is living in a household different from these two (\(H5\)) - these three items giving the child a multiplicity of 3.

### Table 7.7. Example of items of information required to identify multiplicity of child domestic workers in a household survey

<table>
<thead>
<tr>
<th>Household</th>
<th>H1</th>
<th>H2</th>
<th>H3</th>
<th>H4</th>
<th>H5</th>
<th>H6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.a</td>
<td></td>
<td></td>
<td></td>
<td>D1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2.a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D3</td>
<td>D5</td>
</tr>
<tr>
<td>2.b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3.a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D4</td>
<td>D3</td>
</tr>
<tr>
<td>3.b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>3.c</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Notes:**
There is one column for each household in the population. The column identifies which particular child domestic workers (CDW) we would encounter as living and/or working in the household, if the household were selected into the sample. The column also gives responses to applicable questions for each CDW encountered. The last row gives the multiplicity of each CDW, as determined by the responses by or on behalf of the child (depending on the interview procedure), and the multiplicity rules described in the text.

In enumerating the number of households the child is working in (such as in questions 1b, 2b or 3c), it is important to be very clear as to what is or is not to be included. For instance, the purpose of question 3b is to make clear that in 3c, the child’s own household is excluded whether or not the child works in that household.

\(^{30}\) Identification of specific households involved would be needed if the household selection probabilities vary and Horvitz-Thompson type of estimator is being used. The multiplicity estimator here is of the Hanson-Hurwitz type, and requires less information in this respect.
7. Multiplicity sampling

7.6.4 Relative efficiency

For a given sample size of households, the multiplicity design collects information on a larger number of child domestic workers than would a conventional design which uses information only on members of the sample households. Therefore, we can expect the multiplicity design to have lower variance, but of course at increased cost because of the additional information to be collected – not only for a larger number of children, but also some additional information on each child in order to identify the child’s multiplicity. In the present illustration, the variance of the multiplicity and the conventional designs are, respectively, 3.2 and 6.8. Both sampling schemes are essentially unbiased, and the estimated total number of children averages 5.0 (which is the actual population value) over all possible samples in either case.

With equal probability of selection of sample households, the inclusion probability of the linked domestic workers can be calculated exactly with the available information, and the Horvitz-Thompson estimate can thus be derived. The probability of child domestic worker \( j \) falling in the sample is given by

\[
\pi_j = 1 - \binom{6 - m_j}{2} / \binom{6}{2},
\]

where \( m_j \) is the multiplicity of, i.e. the number of households linked to, child domestic worker \( j \). The variance of the estimates in this case equals 3.2, essentially the same as that quoted above for the multiplicity estimator. Under unequal probability of selection of the sample households, the calculation of the Horvitz-Thompson estimate is, however, more difficult or sometimes even impossible if information on the specific households linked to child domestic workers in the sample is not available. The multiplicity estimate, on the other hand, can be derived and used in a more straightforward manner as described above.

More efficient estimates, using unequal probability samples, are also possible. But they tend to require more information, and hence are often less practical. A couple of examples may be given.

The first example is for a sampling of households with probability proportional to the number of child domestic workers linked to it (\( b_i \) shown as the last column of Table 7.6A). Accordingly, the probabilities of selection of the sample households are \( \pi^A = (0.22, 0.44, 0.22, 0.67, 0.22, 0.22) \) for the given sample size \( 2 = \sum \pi^A \). Under balanced sampling (introduced in Chapter 12 below) with unequal probabilities of selection with fixed sample size (Tillé 2006), the variance with multiplicity sampling is found to be 1.5 (compared to 5.6 with conventional household sampling). This design, however, cannot be implemented in practice, as information on the number of links is generally not known in advance.
As another example, an alternative design can be considered assuming that the areas of concentration of units with high number of links are known. In the child domestic worker example, suppose it is known that households 2 and 4 have a relatively high number of links with domestic workers and households 1, 3, 5 and 6 have a relatively low number. Two strata can thus be formed accordingly and one household selected from each stratum by simple random sampling. The probabilities of selection of sample households would thus be $\pi^A = (0.25, 0.5, 0.25, 0.5, 0.25, 0.25)$. Variance with stratified multiplicity sampling is found to be 1.2 (compared to 5.0 with conventional household sampling noted above).

### 7.6.5 Alternative form of the multiplicity estimator

We return to the small population illustrated in Tables 7.6-7.7, this time expressing the multiplicity estimator in terms of its equivalent estimator based on the weight-share method. Table 7.8A, following Table 7.6A, shows $m_{ij} = 1$ if there is link between a sampling unit $i$ (household) and an analysis unit $j$ (child domestic workers) in the population. The sum of each column gives, in the last row, the total number ($m_j$) of links child worker $j$ has with households in the population.

In Table 7.8B, each row corresponds to a simple random sample of two out of six households, giving $w_i' = (1/\pi_i) = 3$ as the inverse of the uniform selection probability. There are 15 possible samples. The figures in cell $(k,j)$ give the total number of links a child domestic worker (column $j$) has with the two households in the sample (row $k$). From Equation (7.6), this quantity is in fact $(m_j/3)$ appearing in Equation (7.6). In each column (child $j$), the above value is divided by the child’s multiplicity $m_j$ from Table 7.8A to obtain the shared weight $w_j = (m_j'/m_j)$ corresponding to each child in a given sample. With $y_j \equiv 1$ as a count variable in Equation (7.6), the row total $\Sigma_k (m_j'/m_j)$ for each sample gives an estimate of the total number of child domestic workers in the population of 6 households. The estimates are shown in the last column of Table 7.8B. It can be seen that these estimates, expressed here in the weight-share form, are identical to the multiplicity estimates shown above in Table 7.6B.
## Table 7.8. Sample of households containing child domestic workers; weight-share estimate

### A. Links between analysis units (j) and sampling units (i) in the population

<table>
<thead>
<tr>
<th>Child domestic worker (j)</th>
<th>Household (i)</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
<th>D5</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H2</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H4</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>H6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>(m_j)</td>
<td></td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

### B. Links between analysis units (j) and sampling units (i) in the sample

<table>
<thead>
<tr>
<th>Weight-share</th>
</tr>
</thead>
<tbody>
<tr>
<td>(S^A) ( (m_j'/3) \mid S^A ) ( \Sigma_j (m_j'/m_j) )</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>H1+H2</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>D1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>7.5</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>H1+H3</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>D2</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>4.5</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>H1+H4</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>D3</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>7.0</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>H1+H5</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>D4</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>4.0</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>H1+H6</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>D5</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>4.0</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>H2+H3</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>D1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>6.0</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>H2+H4</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>D2</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>8.5</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>H2+H5</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>D3</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>5.5</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>H2+H6</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>D4</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>5.5</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>H3+H4</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>D5</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>5.5</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>H3+H5</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>D1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2.5</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>H3+H6</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>D2</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2.5</td>
</tr>
<tr>
<td>13</td>
</tr>
<tr>
<td>H4+H5</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>D3</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>5.0</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>H4+H6</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>D4</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>5.0</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>H5+H6</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>D5</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>2.0</td>
</tr>
</tbody>
</table>

**Notes:**

Each row in Panel B corresponds to a sample of 2 households. The value in a cell is the sum of corresponding cells for the two households in Panel A. Thus for example, cell(H2+H4,D3) of Panel B = cell(H2,D3)+cell(H4,D3) of Panel A = \((1+1) =2\). These values are multiplied by \((1/f) = 3\) to obtain \(m_j'\) for the child concerned, and then divided by \(m_j\) from the corresponding cell in the last row of Panel A. Summing the resulting figures across the row (which identifies the CDW encountered in the particular sample) gives the last column – the estimate from the particular sample of the total number of CDWs in the population.
Chapter 8

Multi-frame sampling

8.1 Multiple frames in the context of child labour surveys

8.1.1 Potential for use of multiple frames

In order to sample a rare population efficiently a sampling frame is required, ideally with the following features.

(1) The frame is complete – all elements (members) of the target population appear in it, whether explicitly (as in a list) or implicitly (such as within area units included in the frame), without necessarily having pre-existing lists of individual elements within areas. (See Section 4.1.2.)

(2) The frame, or an identified part of it, contains a high concentration of the rare class, and also accounts for a large part of the rare class.

Normally such an ideal frame does not exist. Nevertheless, it may be possible to take steps towards creating a frame with the desired characteristics. We have already discussed in Chapter 6 the potential for using existing economic and population censuses and surveys for the purpose. Another possibility is to seek a supplementary frame or frames which can be added to the existing frame to reduce its deficiencies. This is the multi-frame sampling approach. When no single sampling frame exists that can provide a sufficiently complete representation of the target population, a procedure to avoid coverage errors is the use of multiple frames. For example, we may seek one principal frame with near complete coverage of the target population and a supplementary frame providing coverage of the elements absent in the principal frame. More than one supplementary frame may also be envisaged. Generally, the multiple sampling frames overlap and steps need to be taken to deal with the overlap, for example by constructing a new single frame without duplicates or by accounting for the duplicates in the estimation process.

In the following sections we discuss the use of multiple sampling frames in the context of child labour surveys. When we have overlapping frames, some units of the population are present in more than one frame, and their chance of appearing in the sample is increased since they receive a probability of selection from each frame in which they appear. The main technical problem in multi-frame sampling is to determine and take into account in the estimation these changed unit selection probabilities. Again, as in the case of several other procedures discussed in this book, this problem can be addressed within the common framework of sampling with multiplicity, elaborated in Section 4.5.2.
In this chapter, we describe the main methods of removing the duplicates and constructing non-overlapping frames, as well as the main procedures for accounting for duplications and estimation from overlapping frames. The procedure is illustrated with numerical examples, involving multiple frames targeting working children, for instance in a situation with three overlapping frames - a frame of households, a frame of establishments and a list frame.

There are many examples of the use of multiple frames in surveys in diverse fields, including surveys in agriculture and industry, household surveys, and surveys of special populations. However, many of the published applications are from developed countries. Experience in developing countries, particularly in the field of child labour, is limited. Nevertheless, it is a technique which has the potential to be usefully employed in surveying labouring children and other elusive populations in developing countries. It should be exploited to the maximum extent possible for this purpose.

8.1.2 Characteristic features of the situation

The situation likely to be of interest in our present context has the following characteristics.

(1) The main frame for the survey is a frame of area units. The sampling procedure involves selecting a sample of area units, and then within each selected area unit, listing and selecting ultimate units, e.g. locations, establishments, households, or children.

(2) In principle, the area frame provides a reasonably complete coverage of the target population, though only implicitly of the final analysis units of interest. Using information from censuses and surveys and similar sources, steps may have been taken to target the frame to the rare population of interest, such as by identifying areas with higher concentrations of the rare population, separating out empty or near-empty areas, collecting and compiling auxiliary information for more efficient design, etc., as discussed in Chapter 6. Notwithstanding those efforts, the area frame usually remains a general one, lacking identification of areas with concentrations of the rare population.

(3) Lists of elementary units can be found, each with a high concentration of particular subgroups in the rare population but not necessarily accounting for a large proportion of those subgroups. We assume that the lists are targeted at particular subpopulations, but that generally each provides only a (very) partial coverage of the target population.

(4) The lists may overlap.

(5) The lists may contain information for the identification and linkage of units at the micro-level, for example name, sex, age and other basic characteristics in lists of persons. However, it is common that such information is not complete, unique or sufficiently precise for reliable identification and linking of the units. This is particularly likely to be the case in relation to geographically-based listings in the basic frame. It is usually difficult to link individuals identified from a listing operation in the area-based sample to existing administrative lists. The process in reverse - namely the process of linking individuals from administrative lists to listings prepared
in the area-based frame - is likely to be even more difficult, if not impossible. In any case, concerns of privacy and confidentiality often preclude such linkages.

(6) Another complexity to be taken into account is that the supplementary list frames may not be lists of persons but of units of other types - units at levels intermediate between persons and areas, such as economic establishments of different types where children work, bars, clubs, shelters and charity or official organisations providing support and services to working children. However, the presence of such intermediary bodies in an area can provide strong auxiliary indicators for the identification of areas with above-average concentrations of the rare population of interest.

8.2 Some examples

In this section, a couple of examples of the use of multiple frames are given. These are not from child labour surveys in developing countries, but are nevertheless useful for understanding what the use of multiple frames may involve. Following that, we will briefly explain a technical procedure of estimation in multi-frame sampling before turning to some practical aspects of implementation.

8.2.1 Supplementary area frame for a rare population

The following example is from the Canadian Survey of Household Spending, as reported by Lapierre, Nadeau, Tremblay and Gaudet (2004). The survey was designed to provide reliable provincial estimates of household spending. A multi-stage sample was selected from an area frame covering the entire population, and the data were collected by personal interview. During the 2003 survey, dual frame sampling was used to obtain improved estimates for a subpopulation of low income households representing approximately 2.5 per cent of all households in the Canadian province of Quebec.

Towards meeting the additional objectives of the 2003 survey, three options were considered.

(1) Use of the existing area frame to select the supplementary sample

The supplementary sample would use the same sampling frame as that used for the normal survey. The sample could consist of households which had already participated in the previous surveys, or of new households but selected using the existing design, or it could be based on a new design but using the same sampling frame. However, none of these options were considered efficient because the general frame used for the regular survey was not sufficiently targeted on the special population of interest, which was a rare population comprising only 2.5 per cent of the total population. As the authors note, “this approach was rejected because of the difficulty of assigning relevant and up-to-date information to geographical units in the area frame that are small enough to effectively target the population of interest”.
(2) Constructing a new list frame

The second option explored was to construct a frame of dwelling units from the 2001 Population Census, with census information on households occupying those dwellings at that time providing the auxiliary variables to target the sample to the rare population of interest. However, empirical investigation of the results showed that this option was not efficient, especially in comparison with the area-based approach described below.

(3) Targeting high concentration geographical areas

Households in the target population have certain common characteristics such as having low income and living in rented accommodation, and tend to have higher concentrations in particular geographical areas. Such geographical areas were identified from 2001 census data (the data covered household income, work, socio-demographic characteristics, and geographical location). The geographic units used were considered small enough in size to permit targeting for members of the rare population, but at the same time large enough in size “to be able to attribute some reliability to the estimates of prevalence” of the rare population. To the areas selected were attached dwelling lists extracted from Statistics Canada’s Address Register. The last-mentioned frame is kept up-to-date in areas with above-average population densities.

Three area frames with associated dwelling lists were evaluated. The first frame consisted of area units with the highest prevalence of the rare population according to 2001 census data. The second frame included the first frame, plus areas with the next highest prevalence. Similarly, the third frame included the second frame, plus areas with the next highest prevalence level. Table 8.1 shows the extent of geographical concentration of the rare population achieved.

The prevalence falls from 9.5 per cent to 4.4 per cent as more areas are added to the frame from frame (1) to frame (1+2+3). However, the share of the total rare population accounted for by the frame increases from around 40 per cent to around 90 per cent. The last column shows \( P_c/P \), the degree of concentration of the rare population in the frame areas compared with the average (\( P=2.5 \) per cent).

The final choice (frame 1+2+3) was made on the basis of empirical comparison of efficiencies. Even though the average prevalence rate in this frame is not so much higher than that in the total population (4.4 per cent against 2.5 per cent), the share of the rare population covered in this frame is indeed very high at 90 per cent.

The illustration shows that good auxiliary information can succeed in identifying strata and areas of high concentration and high coverage of a rare population. This is an example of combining multiple frames in order to create a more efficient frame which is better targeted at the survey objectives, and then using the targeted frame in combination with the general frame used for regular surveys.

This is multi-frame sampling. While the whole population receives a chance of selection from the general frame, a part of that population receives an additional chance of selection through its presence in the targeted frame. Both parts include members of the rare population (and also members of the rest of the population), but the efficiency is increased if the second part, which receives increased selection probability, contains a bulk of the rare population.
### Table 8.1. Illustration of separation of areas in the sampling frame according to concentration of the rare population

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,083</td>
<td>10</td>
<td>9.5</td>
<td>38</td>
<td>3.8</td>
<td></td>
</tr>
<tr>
<td>1+2</td>
<td>2,125</td>
<td>20</td>
<td>7.3</td>
<td>59</td>
<td>2.9</td>
<td></td>
</tr>
<tr>
<td>1+2+3</td>
<td>5,060</td>
<td>50</td>
<td>4.4</td>
<td>88</td>
<td>1.8</td>
<td></td>
</tr>
</tbody>
</table>

**Column headings:**

1. Geographical frames: “1” with the highest concentration, “2” with the next lower, etc.
2. Number of area units in the frame.
3. Percentage of the total population accounted for by the frame.
4. Average prevalence of the rare population in the frame areas (overall average in the population, \(P = 2.5\) per cent).
5. Share of the total rare population captured in the frame.
6. Relative concentration of the rare population in the frame, compared to overall average \(P (2.5\%)\), \(\frac{(4)}{P} = \frac{(5)}{(3)}\).

#### 8.2.2 Combining several list frames

The second example, from The National Incidence Study of Child Abuse and Neglect, is quoted below from Kalton (2009):

“The National Incidence Study of Child Abuse and Neglect ... used many frames to increase its overall coverage of abused and neglected children. Child Protective Services (CPS) agencies in the sampled PSUs [primary sampling units] were the basis of the main sampling frame, while police, hospitals, schools, shelters, day care centres and other agencies were the sources of other frames. The samples from CPS agencies were selected from list frames, but the samples from other agencies were drawn by sampling agencies, constructing rosters of relevant professional staff, and sampling staff who acted as informants about maltreated children. With these procedures, duplication across agencies cannot be ascertained, except in the case of CPS agencies and any of the other agencies. The design was therefore treated as a dual-frame design, with CPS as one frame and the combination of the other frames as the second frame (i.e. assuming no overlap between the other frames)".

An arrangement such as the above can be possible and useful in certain situations for surveying child labour. The main sample may be area-based in which households or establishments and child workers within those are listed, sampled and interviewed. The area-based frame may be able to provide a representative sample at the general level, but not an efficient sample for specific target groups of child workers. Furthermore, child workers in certain difficult situations or types of activity may remain unrepresented partly, or even completely, in the frame. To supplement the general frame, one may seek lists of working children, which can come from diverse sources. Sometimes many such lists can be found, but with the lists overlapping and each one being very incomplete. Rather than try and work with many different list frames, a more practical strategy

---

31 In other words, while duplications could be removed among CPS agencies frames, and presumably also among frames of other non-CPS agencies, duplications could not be removed between the resulting CPS and non-CPS consolidated frames.
can be to work on the lists carefully in order to link them at the micro-level, remove duplicates, and create a single unified list of working children (or of other suitable micro-level units such as households or small establishments). Thereafter, the general area-based frame and the consolidated list can serve as two supplementary frames.

**8.3 Dealing with multiple sampling frames**

With multiple frames, some members of the target population will be included in several frames, and can appear in the final sample from any one or more of those. Sampling from multiple frames requires appropriate estimation techniques to take this into account. The probability of selecting a unit from each frame must be determined and any duplication of units in the multiple frames accounted for.

There are basically two approaches to dealing with multi-frame sampling.

**(1) Removing duplications**

This involves removing duplicates (i.e. leaving any unit in no more than one of the frames) so that the frames no longer overlap. Every unit in the population will belong, at the most, to one frame (or to no frame in the case of units not covered). After removing duplicates, the frames become merely components of a single frame from which samples can be drawn independently as required.

As an example, in an integrated baseline survey aimed at hazardous child labour sectors in Bangladesh, two sampling frames were used – a list frame and an area frame. While surveying sample areas from the area frame, all the units of the list frame found in the area were excluded, irrespective of whether or not those units had been selected from the list frame. In this way, overlap of the area frame with the list frame was removed.

**(2) Accounting for duplications**

This involves retaining duplicates, but taking them into account in determining the resulting unit selection probabilities and the corresponding sample weights for estimation.

Each of these two approaches has some variants. These are described below in turn.

In clarifying the issues involved in the construction and use of multiple frames, we will use as an example the simple illustration in Figure 8.1.
In Figure 8.1, there are two frames, represented by the area of the two circles A and B. The frames overlap (area C), but do not cover the whole target population. The rectangle enclosing the two circles represents the target population. In the figure on the left, some units are covered in both frames (part C), others in only one of the frames (A or B), while some others are not covered by either of the frames (the area in the rectangle outside the circles A and B). The figure on the right shows a special – but commonly encountered – case. Here, frame A covers all that is covered of the population. Frame B is merely a subset of frame A.

8.4 Constructing non-overlapping frames

Consider the case when the two frames are list frames. Individual units appear in each frame with their identifiers (name, sex, age, place of residence, possibly name of a parent, etc.). Removing duplicates requires identification and linkage of units across frames. Such record linkage can be expensive, uncertain and error prone, depending on the quality and completeness of the identification information. Nevertheless, in certain situations, this may be the most efficient or convenient approach.

Several forms of the procedure are possible.

(1) Assigning all duplicates to only one of the frames

All duplicates between two frames may be assigned to one of those frames and altogether removed from the other. This option is shown in Figure 8.2(1). All units covered in both frames (C) have been assigned to frame B, and removed from frame A, leaving it as (A-C). Consequently we have two non-overlapping frames B and (A-C). Their total is (A+B) - C.

The figure on the right is a special case where B is a subset of A, so that the overlap is C=B. The whole of the overlap C is assigned to B, and removed from A – the two resulting frames do not now overlap.

In more general terms, in this scheme frames are assigned an order of priority. A unit present in a set of several frames is considered eligible in only one of the frames – the frame assigned the highest priority in the set.
Figure 8.2(1). Removing duplication: assignment of the overlap (C) to one of the frames (B)

Partial overlap

Complete overlap (B = C)

(2) Assigning each duplicate to a particular frame

Alternatively, the overlap may be considered case by case, and each case (unit) assigned to one and only one of the frames. In other words, the overlap between two frames is divided into two non-overlapping parts, and one of those parts is assigned to each of the two frames. This option is shown in Figure 8.2(2). The overlap C is divided into two non-overlapping parts C1 and C2. One part C1 is assigned to A only (it is removed from frame B); conversely, the other part C2 is assigned to B only (it is removed from frame A). This gives non-overlapping frames (A-C2) and (B-C1). Their total is, as before, \((A+B) - (C1+C2) = (A+B) - C\).

Figure 8.2(2). Removing duplication: partitioning of the overlap \(C = C1 + C2\), and assignment of a part to each frame
The work can be simplified if it is possible to use a unique identification rule that assigns each member of the target population to only one of the frames (Kish, 1965, Section 11.2). For a unit appearing in more than one frame, its identifiers in only one of the frames is marked as the unique one. When this can be done, it is not necessary to physically remove duplicates from the frames. Furthermore, it also may not be necessary to ‘mark’ for each unit which of its identifiers is unique. Often it is possible to formulate rules which can be used to determine whether or not an identifier is ‘unique’ for the unit concerned, if and when that becomes necessary (such as when the unit falls into the sample). A sample can be selected separately from each frame; any selected units which are without the unique identifier are simply dropped from the sample.

Even though the above procedure based on unique identifiers is much simpler than the alternative of removing duplicates from all frames, it requires clear identifiers and linking of units across list frames. That limits its application in typical child labour surveys.

The essential difference between procedures (1) and (2) is that in the former priority is defined at the level of the frame, while in the latter priority is defined at the level of individual units through individual identifiers. This more or less limits the application of procedure (2) to list frames. By contrast, procedure (1) applies equally to area frames, including area frames in combination with list frames, which makes it particularly useful in our context.

A common application was illustrated in the diagram on the right hand side of Figure 8.2(1). Frame A is a general area-based frame which covers the whole population but lacks specific focus on the rare population of interest. Frame B (which is a subset of A, and hence equals the overlap C by definition) is a list frame more clearly focussed on the target population but its coverage could be seriously incomplete.

A unit selected from frame A is removed from the sample if it is also present in the list frame B (irrespective of whether it is selected from frame B or not). Thus we have two non-overlapping frames: frame (A-B), i.e. A excluding B (or C), and frame B (=C). The linking requirement is that all units selected in A are searched for presence in frame B, and are deleted from the sample from A if they are found in frame B. Unit identification information must be available to permit such linkage. Note that no linkage with A is required for units in B.

(3) Separating out overlap as additional frames

Another way to obtain non-overlapping frames is to divide up the given frames such that every overlap is considered a separate frame in its own right. The overlap is removed from both A and B, resulting in three non-overlapping frames: (A-C), (B-C), and (C). This is illustrated in Figure 8.2(3).
Hartley (1974) developed an estimation procedure in multi-frame surveys. The procedure is to divide $Q$ overlapping frames into (up to) $D = (2^Q - 1)$ disjoint (non-overlapping) domains. Sample data from the resulting $D$ non-overlapping frames are used to produce domain totals, which are then aggregated to produce estimates for the whole population. The maximum possible number of disjoint frames can be seen from the following argument. Units can be classified according to the pattern of their presence or absence in each frame, giving $2^Q$ possible patterns; one of these patterns represents absence from all frames and any units with this pattern are not covered in the set of frames. Of course in a real situation, only a subset of the $2^Q$ possible patterns may be actually present.

As an illustration, in the case of $Q=2$ frames, say frames $A$ and $B$, the population covered can be divided into $(2^2 - 1 = 3)$ non-overlapping domains:

$$A \cap \overline{B}, \quad B \cap \overline{A}, \quad \text{and} \quad A \cap B$$

meaning, respectively, ‘only in $A$’, ‘only in $B$’, and ‘in both $A$ and $B$’. This is illustrated in Figure 8.2(3).

The sample from the last-mentioned domain ‘both $A$ and $B$’ can be separated into units selected into the sample from frame $A$ and units selected from frame $B$. The total estimate for this domain is a weighted sum of estimates from its two samples. Different estimation procedures differ according to the choice of weights for this purpose.

Different types of multi-frame estimator rely on the partition into non-overlapping domains as above. Practical application of these procedures requires correct identification of the particular domain each sample unit belongs to. This implies that every unit in the sample should provide information, in addition to the survey variables being collected, on its membership status for each frame. This is a lot of information to ask the respondent or compile from other sources, and raises concerns of data quality, response rates, sensitivity, confidentiality, privacy, as well as ethics. In any case, respondents are often unable to provide such information. We also need to know each sample unit’s selection probabilities for all the frames. Often selection probabilities are known only from the frame(s) from which the unit has been actually selected, but not
8.5 Estimation using overlapping frames

Figure 8.3 depicts the option of retaining overlap between frames, and accounting for it in the estimation procedure. In the figure on the left, the overlapping part between frames A and B remains a part of both the frames. In the figure on the right, frame B is a subset of frame A, and therefore equals the overlap C. It is treated as a separate frame, though it also continues to be counted as a part of frame A.

8.5.1 Commonly used estimation procedures

In Section 4.6.1, we described commonly used estimation procedures which are applicable when sampling with multiplicity is involved. Below they are applied to our simple illustration of Figure 8.3.

Let a unit be sampled from two, possibly overlapping, frames 1 and 2 with probabilities \( f_1 \) and \( f_2 \). (It is not necessary to retain subscript \( j \) to identify a particular unit as we are considering only one unit at a time.) We assume that the selections from the two frames are independent.

(1) Horvitz-Thompson (H-T) estimator

The procedure based on Horvitz-Thompson estimator calculates the overall selection probability of the unit taking into account its possible selection from any of the frames. The inverse of the selection probability so computed is used as the weight in estimation if the unit appears in the combined sample from all the frames. In the present example, the inclusion probability of a unit is

\[
\pi = \left[1 - (1 - f_1)(1 - f_2)\right] = \left(f_1 + f_2 - f_1f_2\right) = \frac{1}{w}. \tag{8.1}
\]
8. Multi-frame sampling

Term \((1 - f_1)\) is the probability of the unit not being selected from frame 1, and similarly \((1 - f_2)\) of not being selected from frame 2. Assuming independence, the product of these two terms, \((1 - f_1)(1 - f_2)\), is the probability of not being selected from either of the frames. The complement (\(\pi\)) of that is the probability of being selected from one or both of the frames. The inverse of this selection probability is the unit weight \(w\) if it appears in the final sample. The weight is of course zero if the unit is not selected from any of the frames.

(2) Hansen-Hurwitz (H-H) estimator

Another common procedure, based on the Hansen-Hurwitz estimator, calculates the expected number of times a unit is selected into the sample, taking into account its selection from any of the frames. The inverse of this expected number of selections is used as the weight applied to every appearance of the unit in the sample.

In the present example, the expected number of times a unit is selected, considering both the frames, equals

\[ p = (f_1 + f_2) = \frac{1}{w}. \]  

(8.2)

Here \(p\) is the draw-by-draw selection probability. The inverse of this is the unit’s weight \(w\) irrespective of which frame the unit has been selected from. If the unit has been selected twice, once from each frame, or twice from the same frame, it gets double that weight.

(3) Multiplicity estimator

A much simpler and practical solution is provided by the multiplicity estimator, described in detail in Section 4.6. The multiplicity of a unit is the number of frames in which the unit appears. In the present illustration, if a unit is present in both frames its multiplicity \(m=2\). If the unit is selected from frame 1 sampled at rate \(f_1\), the weight given to the unit is

\[ w_1 = \frac{1}{mf_1} = \frac{1}{(2 f_1)}. \]  

(8.3)

If the unit is selected from frame 2 sampled at rate \(f_2\), the weight given to the unit – irrespective of its selection from the other frame – is

\[ w_2 = \frac{1}{mf_2} = \frac{1}{(2 f_2)}. \]  

(8.4)

\[ ^{32} \] Duplication of a unit in the same frame also augments the unit’s multiplicity in the same way. For instance, a unit appearing twice in frame 1 and once in frame 2 has a multiplicity of \(2 + 1 = 3\).
It can be seen from Table 8.2 that all the above three are unbiased estimators. The requirement for this is that the product of the selection probability and the weight given, summed over all possible outcomes for any unit, equals 1. In our case there are four possible outcomes for any unit: unit not selected; selected only from frame 1; or selected only from frame 2; or selected from both frames. Nevertheless, the different estimators differ in their variance.

Despite this formal equality among the estimation procedures in terms of expected value, normally the multiplicity estimator is the most practical one in our context. It requires information on selection probability only for the frame(s) where a unit is actually selected. It does not need information on which particular frames the unit appears in; and certainly no micro-level matching across frames is required. Information on identification and matching of units across frames is very difficult – often impossible – to collect. It is easy to imagine that these difficulties are particularly severe in the context of child labour surveys in developing countries.

The procedure does however require information on the number of frames a sample unit is present in. This item of information is much simpler than that which is required with other procedures. Nevertheless, it is not always easy or possible to obtain even this. The determination of multiplicity requires information on the presence or absence of each sample child worker in the different sampling frames. Generally, such information has to be obtained by formulating appropriate questions for the survey interview.

### 8.5.2 Further comments on the multiplicity estimator for multi-frame sampling

The multiplicity estimator in more general terms may be stated as follows (see for instance Mecatti (2007), which presents it in the form of a ‘single frame multiplicity estimator’). The multiplicity of a unit is the number of frames \( m_j \) in which the unit \( j \) appears. If \( f_{jk} \) is the selection probability of unit \( j \) from frame \( k \), then in estimating from the sample from that frame, the unit is given a weight

\[
w_{jk} = \left( \frac{1}{m_j} \right) \left( \frac{1}{f_{jk}} \right)
\]  

(8.5)
The important thing is that the weights applied to sample data from a frame depend on the selection probabilities \( f_{jk} \) in that frame \( (k) \) only, and not those in other frames. The only information, additional to that required in a normal single frame situation, concerns the number of frames \( (m_j) \) the unit is present in.

The use of multiple frames can involve multiplicity in the selection of units at two levels: a unit may appear in more than one frame; and within any of those frames, the unit may appear more than once. This is not an unusual situation. For instance, we may use more than one frame for the selection of child labourers – within the study area, a listing of households, a listing of establishments employing children, a list of child workers compiled from administrative sources, and records of visits by working children to some facility providing care and assistance. There may also be multiple appearances of children in some of these frames, such as children working in more than one establishment, appearing in more than one administrative list, and very commonly, making multiple visits to the facility. We may write the multiplicity estimator (8.5) in a more general form as follows to cover such situations.

Let subscript \( l \) denote the \( l^{th} \) appearance in frame \( k \) of an analysis unit \( j \) and let

\[
\delta_{jkl} = 1 \text{ for } l > 0, \quad \delta_{jkl} = 0 \text{ for } l = 0
\]

be an indicator equalling 1 for each presence of a unit in any of the frames. The total number of times a unit is present in any of the frames is its multiplicity

\[
\sum_k \sum_l \delta_{jkl} = m_j.
\]

Note that every case \( \delta_{jkl} = 1 \) corresponds to a particular sampling unit \( (jkl) \) appearing (explicitly or implicitly) in one of the sampling frames. Let \( f_{jkl} \) be the probability of selection of sampling unit \( (jkl) \) and \( \delta_{jkl} = 1 \) an indicator of the selection of the unit into the sample. The multiplicity estimator (8.5) in a more general form is

\[
w_{jkl} = \left( \frac{\delta_{jkl}^{(r)}}{m_j f_{jkl}} \right).
\]

With these weights, contribution of analysis unit \( j \) to estimates of the population size and of population aggregate of some value \( y \) are given by (or are proportionate to), respectively,

\[
\hat{N}_j = w_j = \sum_k \sum_l w_{jkl}, \quad \hat{Y}_j = \sum_k \sum_l w_{jkl} y_j, \quad \hat{N}_j, \quad \text{total } \hat{N} = \sum_j \hat{N}_j,
\]

where \( y_j \) is some value measured for the unit, assumed to be the same for all appearances of the unit in the sample. Note that the summations above are over all selections \( (jkl) \) appearing in the sample. Hence, strictly speaking, it is not necessary to know whether different selections in fact refer to the same unit or frame \( (jk) \). This is an important point, since in some circumstances, such information cannot be obtained or cannot be obtained reliably. For the same reason, it is possible that quantities \( m_j \) and \( y_j \) even for the same unit \( j \) are not identical as enumerated in different appearances of the unit in the sample. Hence it is more practical and general to express Equations (8.8) and (8.9) as follows. The form allows the measured multiplicity \( m \) and value \( y \) of unit \( (j) \) to be different for different selections \( (jkl) \) of the unit:
8.6 Illustration 1. A three-frame example

In this and the next sections we consider some practical aspects of implementing the multiplicity estimation procedure in the context of a child labour survey, illustrating the procedure by two detailed numerical examples.

8.6.1 An illustrative scenario

In Section 8.1.2, we noted six characteristics of the situation we are likely to encounter concerning the sampling frame, which were briefly as follows.

(1) The main frame for the survey is a frame of area units.

(2) In principle, the area frame provides a reasonably complete coverage of the target population.

(3) Lists of elementary units can be found, each with a high concentration of particular subgroups.

(4) The lists may overlap.

(5) The lists may contain information for the identification and linkage of units at the micro-level for consolidation into fewer lists, or even into a single list.

(6) The supplementary list frames may not be lists of persons but of micro units of other types at the intermediate level.
Keeping those in view, let us consider some concrete options. Let us assume the following scenario for illustration and discussion. Suppose that after defining the study area and the target population, we construct a set of frames consisting of the following three.

(1) A general frame

A general frame, in principle covers the whole population. Let this be an area frame, with lists of households in the areas (or at least in the areas selected into the sample). A very desirable property of such a frame for the purpose of sampling a rare population is that it be targeted on the population of interest to the extent possible. ‘Targeting’ the frame means identifying, as much as possible, strata and areas of concentration of the target population, identifying and removing areas which contain few population elements of interest, and improving the frame through identification of boundaries, maps, size measures, etc., especially for areas with higher concentrations of the rare population. Several of these issues have been discussed in Chapter 6.

(2) An establishment frame

This refers to a frame covering establishments in the area where children are employed. Let us assume that such a frame provides a reasonably good – but by no means a complete – coverage of medium-to-large establishments employing children in the area. Coverage of small establishments is much less complete. (Issues in sampling establishments, in particular small and informal sector establishments, were discussed in detail in Chapter 5.) Establishments generally do not provide ready-made lists of working children from which samples of children may be directly selected. But we assume here that establishments are generally accessible for interview, and are willing to allow the identification, selection and interviewing of children working in them.

(3) A list frame

Finally, we have a list frame of children, or rather several different lists compiled by different institutions, organisations and facilities with which working children are in contact for receiving help and assistance. These may include hospitals and other health facilities, shelters, care centres, child protection services, employment-related services, even police and other law enforcing bodies. The lists of working children these organisations are able to provide are likely to be patchy, some heavily overlapping, others largely unique. It may be possible to contact personnel working in these agencies in a systematic manner in order to improve understanding of the child labour situation in the area and seek additional information for improving the quality and coverage of the lists. Often the most effective approach may be to try and compile various lists in a uniform format, identify and remove duplicates, and construct a single list frame to supplement the area-based household frame described under (1) above, and the establishment frame described under (2).

33 We use the term ‘general’ to indicate two common features of such frames: (i) the frame covers, explicitly or implicitly, a high proportion of the target population; but (ii) it is not targeted to capture that population efficiently.
Let us assume that it is possible to link various lists and compile from them (3) in the form of a single list without overlaps, but that it is not possible – nor is it considered desirable - to use individual identifiers to link children on the consolidated list (3) with children identified or interviewed in frames (1) or (2) during the survey.

8.6.2 Numerical illustration for the three-frame example

In this subsection, the three-frame example presented above is taken and numerical values are given to it as follows.

The numerical illustrations which follow have been developed by Mehran (2012).

(1) The population

The target population is 25 child workers, identified in Table 8.3 with numbers 1-25. Each column of the table corresponds to a child worker. Rows of the first panel provide information on sex and age of each child. There are 14 boys and 11 girls, and their average age is 12 years.

(2) Sampling frames

Three sampling frames are available covering all the child workers: a frame of households; an establishment frame; and a list frame of individual child workers.

We assume that every child in the target population appears in at least one of the household and the establishment frames, but that neither of these frames covers all the child workers. Some of the children may appear in both these frames, and/or in the list frame. Also, it is assumed that a child does not appear more than once in any given frame. The household frame contains 100 households (numbered 1-100), and 20 of the 25 child workers are members of the households covered by the household frame. The establishment frame contains 20 establishments (numbered 1-20), and 15 of the child workers are working in the establishments represented in the establishment frame. Finally, the list frame contains names of 10 of the child workers (numbered 1-10, in the same order as the 25 child workers in the population). Some of the children in the list frame are members of households in the household frame, some are working in the establishments of the establishment frame, and some are included in all the frames.

The data concerning the frames are presented in the second panel of Table 8.3. For example, the number 23 on the first row and first column of the panel means that child worker 1 is a member of household number 23 of the household frame. It can be observed from the same row that child number 9 is also a member of household 23. Household 18 and household 69 have both also two working children. The data on the other two frames may be similarly read in the next two rows.

(3) The sample

A sample is drawn from each of the three frames. A sample of 25 households is drawn with equal probability by systematic sampling from the sampling frame of households. Similarly, a systematic sample of 10 establishments is selected from the establishment
frame and a systematic sample of 5 children is selected from the list frame. The results are shown in the third panel of the Table 8.3.

Many of the sample households contain no child workers; 4 child workers are identified in the sample of households (child workers 1, 3, 9 and 18). Similarly, 6 child workers are identified in the sample establishments (child workers 10, 12, 16, 19, 22 and 25); and all five (child workers 1, 5, 14, 16 and 21) in the list frame of working children. Child worker 1 is both in the list sample and in the household sample; similarly, child worker 16 is both in the list sample and the establishment sample; the other children selected appear only in one sample. There are altogether 13 distinct child workers in the sample, from a total of 15 (=4+6+5) selections of child workers from the three frames.

The bottom panel of the table gives the sample weight for each of 15 selections of children according to Equation (8.8). For a child selected from more than one frame, the weight is different depending on the frame from which the child is selected. The last row of the table gives the weight a sample child receives according to Equation (8.9) - it is the sum of weights over the child's selections if more than one. For example, the weight of the sample child worker 1, member of household 23 and at position 1 of the list frame, is calculated as (2.0+1.0) = 3.0. Similarly for child number 16, which was selected from the establishment frame as well as the list frame.

The sum of the sample weights (over children or over selections in the sample) gives an estimate of the total number of child workers in the population (Equation 8.9):

\[ \hat{N} = \sum_j w_j = \sum_{jk} w_{jkl} = 23.7. \]

Similarly, separately for working boys (b) and girls (g) appearing in the sample,

\[ \hat{N}_{(b)} = \sum_{j(b)} w_j = \sum_{jk(b)} w_{jkl} = 13.0, \quad \hat{N}_{(g)} = \sum_{j(g)} w_j = \sum_{jk(g)} w_{jkl} = 10.7. \]

The average age of the child workers is estimated by the weighted sum of the age of the sample child workers:

\[ \hat{Y} = \frac{\sum_j w_j y_j}{\sum_j w_j} = 13.5. \]

Similar calculations have been carried out for all possible systematic samples of size 25 for the household sample, all possible systematic samples of size 10 for the establishment sample, and all possible systematic samples of size 5 for the list sample. The average value of the resulting estimates of the total number of child workers, with breakdown by male and female, and the average age of the child workers are shown in the left column of Table 8.4. The corresponding values based on single frames are shown in the next columns. The columns in the right panel give the standard deviations of the estimates.
Table 8.3. Three-frame example: the frames, samples and weights

<table>
<thead>
<tr>
<th>Frame</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>k=1 Household</td>
<td>23 18 59 66 69 82 94 23</td>
</tr>
<tr>
<td>k=2 Establishment</td>
<td>5 11 11 19 6 15 16 10</td>
</tr>
<tr>
<td>k=3 List</td>
<td>1 2 3 1 2 2 1 2 1 2 2 3 1 1 3 2 3 1 4 5 10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Links to frames $m_{ij}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household</td>
<td>2 3 1 1 2 2 1 1 2 2 1 2 2 1 1 2 2 1 3 2 3 1 4 5</td>
</tr>
<tr>
<td>Establishment</td>
<td>1 1 1 1 1 1 6</td>
</tr>
<tr>
<td>List</td>
<td>1 1 1 1 1 5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>child number</td>
<td>1 3 5 9 10 12 14 16 18 19 21 22 25 13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1/f= selected (=1) from frame:</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household</td>
<td>1 1 1 1 4</td>
</tr>
<tr>
<td>Establishment</td>
<td>1 1 1 1 1 1 6</td>
</tr>
<tr>
<td>List</td>
<td>1 1 1 1 1 5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Weight 1/(f$^{mj}$)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household</td>
<td>2.0 4.0 2.0 4.0 12.0</td>
</tr>
<tr>
<td>Establishment</td>
<td>1.0 1.0 1.0 1.0 2.0</td>
</tr>
<tr>
<td>List</td>
<td>1.0 0.67 1.0 0.67 2.0</td>
</tr>
<tr>
<td>Total</td>
<td>3.0 4.0 1.0 1.0 2.0 2.0 23.7</td>
</tr>
</tbody>
</table>

8.6 Illustration 1. A three-frame example
Compared with the actual values (25 child workers, 14 boys, 11 girls, and average age 11.6 years), the results shown above indicate that the multiplicity estimates based on the three frames in combination are unbiased. The estimates based on single frames (Household, Establishment and List) are also unbiased for the respective sampling populations represented in those frames, but biased for the whole target population.

Presence of ‘blanks’ in a frame tends to inflate standard deviations of the estimates. By blanks we mean listings in the frame which do not contain any working child such as a household without a working child or an establishment not employing children. The list frame has no blanks – it is simply a list of working children. Estimates from this frame therefore have the lowest standard deviations. Next is the establishment frame: the frame contains 15 child workers in 20 establishments, i.e. roughly 25 per cent blank entries. The highest standard deviations are for estimates from the household frame: the frame contains 20 child workers in 100 households, i.e. roughly 80 per cent blank entries. The standard deviations of the estimates for the three-frame samples are close to those for the single-frame sample of households, since the high proportion of blanks in that frame dominate the overall results.

| Table 8.4. Average of estimates and standard deviation: three-frame sampling of child workers |
|---------------------------------------------|------------------------------|---------------------------------------------|------------------------------|
|                                      | Average of estimates         | 3-frame sample                             |
|                                      |                              | Household | Establishment | List | Household | Establishment | List |
| Child workers                        |                              | 25        | 20            | 15   | 10        | 4.1           | 4.4  | 2.1  | 1.4 |
| - Boys                               |                              | 14        | 9             | 10   | 4         | 1.9           | 2.0  | 2.0  | 1.2 |
| - Girls                              |                              | 11        | 11            | 5    | 6         | 4.2           | 4.2  | 0.8  | 2.3 |
| Average age                          |                              | 11.6      | 11.3          | 11.5 | 11.6      | 1.0           | 1.1  | 1.1  | 0.7 |

8.7 Illustration 2. A refined frame with a part of the overlap removed

8.7.1 ‘Refining’ the frames by removing overlap

In certain situations, it may be possible and worthwhile to remove duplicates between frames on the basis of case-by-case matching. Using the same illustrative data as in the previous example, the following example describes the situation when overlap between the area-based household and the establishment frames is removed.

For the purpose of the illustration, let us assume the following about the situation.

i. As above, we assume that, between themselves, the household and the establishment frames cover the whole target population; that is, all working children in that population appear in at least one of the two frames, and some children appear in both the frames.

ii. The establishment frame is more focussed on the target population of interest, while the household frame is more general and contains many units (households) with no
working children. Also, it is more difficult to obtain precise information about the
nature and conditions of children’s work through a household-based survey, than it
is at the place of work. (These are quite realistic assumptions.) Consequently, it is
preferable to use the establishment rather than household frame for selecting the
sample, of course this is possible only when a choice exists between the two options.

It is therefore better to eliminate any duplicates between the two frames from the
household frame and retain them in the establishment frame. In applying this
requirement, we will need to eliminate duplicates only from the sample from the
household frame. More precisely, the procedure is as follows. If a working child is
selected from the household frame, it is checked whether the child exists in one of
the establishments covered in the establishment frame, irrespective of whether or
not the household has been selected from that frame as well. The selection from
the household frame is eliminated from the sample if the child is found to exist also
in the establishment frame. In addition, in such a situation the multiplicity of any
unit in another frame should not be augmented by the unit’s presence in the original
household frame. The above procedure is equivalent to removing duplicates from the
whole household frame before sample selection. This is what has been done in Table
8.5. Shaded cells in the table correspond to children who have been eliminated from
the household frame because they already appear in the establishment frame.

iii. We assume that we are able to identify whether a child selected in the sample from
the general area-based household frame (1) works in one of the establishments
covered in frame (2). We assume that this information is fairly precise so that linkage
at the level of the individual child in the cases involved is possible. This also requires
that in frame (2) a complete list of working children is available (or can be created
by listing) covering all the establishments in that frame, not just those in the sample
selected from that frame.

iv. On the basis of the above information, suppose that we are able to reconstruct
the household frame (1) so that it covers the whole population except for the
population covered in frame (2). Let us call this frame ‘(1-2)’. By definition, this
frame and frame (2) do not overlap. Actually, we do not require this operation to be
conducted on the whole of frame (1); it is sufficient for our purpose to do so only
for a sample from (1). The reduced frame (1-2) can be constructed by selecting a
sample of children from original frame (1) checking, as noted above, whether or
not each selected child can be linked to a child listed in frame (2) – the whole of
frame (2), not just a sample of establishments or children from it.

v. Let us assume that it is not possible, nor considered desirable, to use individual
identifiers to link children on the consolidated list (3) with children identified or
interviewed in frames (1) or (2) during the survey.
8.7.2 A numerical illustration of the ‘refined’ frame

The data are presented in Table 8.5. The population, and hence the first panel, is the same as Table 8.3. The ‘refined’ household frame has now 10 child workers, the other 10 child workers in the original household frame (Table 8.3) having being eliminated because they already appear in the establishment frame. There is no change in the other two frames: the establishment frame has 15 child workers, and the list frame 10 child workers (second panel of the table).

The third panel of Table 8.5 shows results for exactly the same sample as used in Table 8.3. This is a sample of 25 households drawn using with equal probability systematic sampling from the households frame, a sample of 10 establishments drawn also by systematic sampling from the establishment frame, and a sample of 5 child workers drawn systematically from the list frame. The sampling weights (bottom panel) are calculated using the multiplicity estimator as described in the previous section.

Two alternative procedures give the same results for the sample in Table 8.5:

1. Either working children are eliminated from the household frame if they also appear in the establishment frame; or

2. The above procedure is not performed on the whole frame but the following operations are performed on the samples selected.

The second alternative is likely to be more economical for sampling from large frames. It involves the following steps.

- A working child selected from the household frame is eliminated from the sample if the child also appears in the establishment frame (whether or not selected from the establishment frame).
- For children selected from the establishment frame, the multiplicity of the child is not increased for being present in the original household frame.
- For children selected from the list frame, the multiplicity of the child is augmented for being present in the original household frame only if the child is not present in the establishment frame.

If the sampling rates applied to the three frames are kept unchanged, the ‘refined’ frame results in the selection of a smaller number of children compared to the number selected from the original three frames. This is because children selected from the household frame are dropped if they also appear in the establishment frame. This generally increases variance of the resulting sample of children. However, if the sample size in terms of the number of children selected is kept the same, the ‘refined’ frame generally gives lower variance because more of the sample comes from the establishment frame rather than the household frame. An establishment frame is usually a more targeted and concentrated source for sampling working children than a general household-based frame.
### Table 8.5. Two-frame (“refined frame”) example: the frames, samples and weights

<table>
<thead>
<tr>
<th>Child Worker</th>
<th>Population</th>
<th>Frame</th>
<th>Sample</th>
<th>Weight</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Male</td>
<td>Household</td>
<td>23</td>
<td>2.0</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>Establishment</td>
<td>59</td>
<td>1.0</td>
<td>50</td>
</tr>
<tr>
<td>Age years</td>
<td>13</td>
<td>List</td>
<td>66</td>
<td>1.0</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td>82</td>
<td>1.0</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td></td>
<td>94</td>
<td>1.0</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td>59</td>
<td>1.0</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td>11</td>
<td>1.0</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td></td>
<td>19</td>
<td>1.0</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td></td>
<td>16</td>
<td>1.0</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td></td>
<td>15</td>
<td>1.0</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td></td>
<td>6</td>
<td>1.0</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td></td>
<td>8</td>
<td>1.0</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td></td>
<td>7</td>
<td>1.0</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td></td>
<td>20</td>
<td>1.0</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td></td>
<td>7</td>
<td>1.0</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td>12</td>
<td>1.0</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td></td>
<td>10</td>
<td>1.0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
<td>15</td>
<td>1.0</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td>9</td>
<td>1.0</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td>8</td>
<td>1.0</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td></td>
<td>17</td>
<td>1.0</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td></td>
<td>12</td>
<td>1.0</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td></td>
<td>16</td>
<td>1.0</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td>14</td>
<td>1.0</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td></td>
<td>11</td>
<td>1.0</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td></td>
<td>11</td>
<td>1.0</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td></td>
<td>14</td>
<td>1.0</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td></td>
<td>16</td>
<td>1.0</td>
<td>16</td>
</tr>
</tbody>
</table>

Note: The table shows the number of child workers selected from each frame, the number of selected child workers, and the weight for each selected child.
8.7.3 Estimating multiplicity (or overlap between frames)

Using the multiplicity estimator requires information on the multiplicity of each analysis unit (say, person), i.e. on the number of sources or frames from which the person could be selected. In surveys, normally this information has to be obtained from personal interview. A possible alternative is to match the frames at the micro-level, and from that deduce the information on the multiplicity for each person present in any of the frames.

There are circumstances when neither of these options is possible nor permitted.

Let us take as an example our 3-frame illustration, refined to a 2-frame situation by matching the establishment and area-based (household) frames at the level of the individual child worker. This frame refinement is possible based on the assumption that respondents are able and willing to provide the necessary information for identification and matching at the individual level between the two frames. This may not be an unreasonable assumption in many situations.

The picture can be quite different when it comes to matching between the area-establishment refined frame and the list frame. Matching between these frames may not be possible, permitted, and/or ethical. For instance, lists for the frame may have been constructed from sources which may be perceived as being threatening, embarrassing, or otherwise undesirable in some sense. Consequently it might be quite inappropriate to identify children in the general population as being on such lists. Also, often people may have no idea of whether or not they are on a list.

When individual multiplicities cannot be measured precisely and directly by asking questions to individuals, it is necessary to try and estimate their approximate values.

In the illustration of Section 8.7.2, the refined household frame (1) and the establishment frame (2) are two non-overlapping frames covering the whole population. Multiplicity ($m_i > 1$) arises because, in addition to appearing in one of these frames, some children are also present in the list frame (3).

Let subscript $i$ indicate an individual or a group of individuals or more generally some function of characteristics (of persons, circumstance, conditions and type of work, etc.), on the basis of which the target population has been divided according to the degree of overlap between the frames. Ideally, the degree of overlap should be uniform within categories identified by $i$. Let $a_i$ be the average proportion in subpopulation $i$ of working children in the refined household frame (1) who are also present in the list frame, and $e_i$ the same for working children in the establishment frame. By definition, the average multiplicities by category $i$ of children are as follows.

<table>
<thead>
<tr>
<th>Frame</th>
<th>Multiplicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1-2) Refined area frame</td>
<td>$1 + a_i$</td>
</tr>
<tr>
<td>(2) Establishment frame</td>
<td>$1 + e_i$</td>
</tr>
<tr>
<td>(3) List frame</td>
<td>$1 + 1 = 2$</td>
</tr>
</tbody>
</table>

The first two frames are non-overlapping and cover the whole population. Hence children in the list frame are always present in the list frame itself, and in one of the other two frames – their multiplicity is $(1+1=2)$. In the refined household frame and the establishment frame, average multiplicity exceeds 1 by an amount equal to the proportion of children in those frames who are also in the list frame.
Exact multiplicities of units cannot be known without matching children individually across the frames. The issue is to estimate \( a_i \) and \( e_i \) when this cannot be done by direct questioning in the survey.

Consider the following additions to assumptions (i)-(v) of Section 8.7.1 above.

(vi) Suppose that the following quantities can be computed or estimated for each subpopulation \( i \) of working children. We assume that these quantities can be estimated from samples drawn from the available frames.

<table>
<thead>
<tr>
<th>( N, N_i )</th>
<th>the total number of working children in the population in the combined frames (1)+(2), and its breakdown by subgroups ( i ), ( \sum N_i = N ).</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A, A_i )</td>
<td>the total number of working children in the population in the refined household frame (1), and its breakdown by subgroups ( i ), ( \sum A_i = A ).</td>
</tr>
<tr>
<td>( E, E_i )</td>
<td>the same for the population in the establishment frame (2), ( \sum E_i = E ).</td>
</tr>
<tr>
<td>( L, L_i )</td>
<td>the same for the population in the list frame (3), ( \sum L_i = L ).</td>
</tr>
<tr>
<td>( p_i )</td>
<td>estimated proportion in the establishment frame of the total population of working children, ( p_i = (E_i/N_i) ).</td>
</tr>
</tbody>
</table>

(vii) Finally, let us assume that the list frame is quite limited in coverage but provides detailed information on each child covered in the frame, on the basis of which we can estimate the proportion, say \( p_{Li} \), among those in the list frame who are also present in the establishment frame. For instance, we may know that the list frame is confined to (but covers well) particular areas or particular types of child labour activity. The proportion of children in the list frame with such well-covered characteristics may provide a reasonable estimate of the quantity \( p_{Li} \). Generally this proportion will differ from one subpopulation \( i \) to another. This information is likely to be much less precise than that on the overlap of the original area-based and establishment frame discussed in Section 8.7.1 above. Recall that linkage of the list frame with other frames at the level of individual children is being precluded here. The purpose of the information is to estimate, among children in a certain class \( i \), the proportion \( p_{Li} \) in the list frame who are likely to be working in an establishment of the type covered in frame (2).

Given \( p_{Li} \), the estimated number of children in the list frame who are also in the establishment frame is \( p_{Li} L_i \). By symmetry this is also the estimated number of children in the establishment frame who appear in the list frame, giving

\[
e_i E_i = p_{Li} L_i \text{ or } e_i = \left( \frac{L_i}{E_i} \right) p_{Li} = \left( \frac{L_i}{N_i} \right) \left( \frac{p_{Li}}{p_i} \right).
\]

The two factors on the right in Equation (8.12) are ratios of corresponding quantities between the population covered in the list frame and the total population of working children. Similarly,

\[
a_i = \left( \frac{L_i}{A_i} \right) (1 - p_{Li}) = \left( \frac{L_i}{N_i} \right) \left( \frac{1 - p_{Li}}{1 - p_i} \right).
\]

Hence the problem of estimating multiplicities \( (1 + a_i) \) and \( (1 + e_i) \) of units selected from the combined frame (1)+(3) and (2)+(3), respectively, is reduced to obtaining
some reasonable estimates of $p_{Li}$—the proportion of the list frame who are also present in the establishment frame.

When it is not possible to estimate $p_{Li}$ directly, we may have to approximate it as

$$p_{Li} = p_i \text{ giving } e_i = a_i = \frac{L_i}{N_i}. \quad (8.14)$$

Equation (8.14) assumes that the proportions present in the establishment frame among children from the list frame are the same as the proportions present in the establishment frame among all child workers in the population. This crude approximation does not have to be applied uniformly: it may be necessary for some subgroups $i$, but better estimates may be possible for some other subgroups.

Table 8.6 provides a simple numerical illustration of the above procedure. We assume that the population of child workers has been divided into 7 classes (row 1), which are expected to differ in the degree of overlap between frames, and to be internally homogeneous in that respect. Rows 2-5 show given data, determined from the three frames. Row 6 is supposed to be obtained from close examination of the information in the list frame, and any external information which may be helpful. The multiplicities in rows 7 and 8 are computed using Equations 8.12-13. Row 9 gives the value of the multiplicity according to Equation 8.14, applying to units in both the household and the establishment frames, when no information is available to estimate $p_{Li}$.

<table>
<thead>
<tr>
<th>(1)</th>
<th>$i$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>total/mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2)</td>
<td>$N_i$</td>
<td>59.2</td>
<td>94.9</td>
<td>109.8</td>
<td>276.1</td>
<td>176.0</td>
<td>66.6</td>
<td>217.4</td>
<td>1,000</td>
</tr>
<tr>
<td>(3)</td>
<td>$E_i$</td>
<td>46.5</td>
<td>63.9</td>
<td>61.6</td>
<td>118.7</td>
<td>55.9</td>
<td>16.8</td>
<td>36.6</td>
<td>400</td>
</tr>
<tr>
<td>(4)</td>
<td>$L_i$</td>
<td>5.1</td>
<td>36.0</td>
<td>17.6</td>
<td>32.2</td>
<td>44.1</td>
<td>30.8</td>
<td>34.1</td>
<td>200</td>
</tr>
<tr>
<td>(5)</td>
<td>$p_i = E_i/N_i$</td>
<td>0.79</td>
<td>0.67</td>
<td>0.56</td>
<td>0.43</td>
<td>0.32</td>
<td>0.25</td>
<td>0.17</td>
<td>0.40</td>
</tr>
<tr>
<td>(6)</td>
<td>$p_{Li}$</td>
<td>0.20</td>
<td>0.27</td>
<td>0.13</td>
<td>0.36</td>
<td>0.01</td>
<td>0.35</td>
<td>0.22</td>
<td>0.22</td>
</tr>
<tr>
<td>(7)</td>
<td>$1 + a_i$</td>
<td>1.02</td>
<td>1.15</td>
<td>1.04</td>
<td>1.10</td>
<td>1.01</td>
<td>1.65</td>
<td>1.21</td>
<td>1.17</td>
</tr>
<tr>
<td>(8)</td>
<td>$1 + e_i$</td>
<td>1.32</td>
<td>1.85</td>
<td>1.32</td>
<td>1.13</td>
<td>1.36</td>
<td>1.40</td>
<td>1.15</td>
<td>1.36</td>
</tr>
<tr>
<td>(9)</td>
<td>$1 + L_i/N_i$</td>
<td>1.09</td>
<td>1.38</td>
<td>1.16</td>
<td>1.12</td>
<td>1.25</td>
<td>1.46</td>
<td>1.16</td>
<td>1.23</td>
</tr>
</tbody>
</table>

Notes to rows of table:
(1) Relatively homogeneous classes of child workers.
(2) Number of working children in the population.
(3) Number of working children in the establishment frame.
(4) Number of working children in the list frame.
(5) Of children in the list frame, proportion who are also in the establishment frame.
(7) Average multiplicity of children in the ‘refined’ household frame.
(8) Average multiplicity of children in the establishment frame.
(9) Crude estimate of average multiplicity of all children when (6) cannot be estimated.

A final remark. In the standard multiplicity estimator, the multiplicity of a unit is fixed, determined by the number of sources from which the unit could be selected. It is independent of the particular source from which the unit may have been selected. For
8.8 Choosing among options

In the introductory section, we identified two basic approaches to dealing with the use of overlapping multiple frames: removing duplicates and constructing non-overlapping frames; and accounting for duplications in estimation from overlapping frames.

What are the pros and cons of using one method rather than the other?

The main requirement (and the main difficulty) in removing duplicates among sampling frames concerns the need for micro-level matching of listings representing the same unit in different frames. There may be cost, sensitivity, ethical and other practical difficulties in accomplishing such matching.

The main requirement (and the main difficulty) in adjusting the estimation procedure for overlaps between sampling frames concerns the need to obtain information on the respondent’s multiplicity. Respondents may not be able or willing to provide such information.

The relative pros and cons of these procedures tend to be very specific to the survey situation.

(1) There are situations when removing duplicates between frames is feasible and is indeed a reasonable option.

An example relevant for surveys of child labour is when we have two overlapping frames: an area-based frame in principle covering the whole target population, but in a general and unfocused way; and a targeted but selectively partial list of child workers. We may select a sample from each of the two frames, and remove the effect of frame overlap by eliminating from the area-based sample any children who are present in the list frame. Another example of possible application to child labour surveys is the use of multiple lists of child workers coming from different types of sources which in major part are non-overlapping, though in part overlaps do exist. If matching among lists is possible, it may be easier to eliminate duplicates than to question children about their multiplicities. Also, many children may not be aware of which lists they may have been included in.

(2) By contrast, there are situations in which it is preferable to estimate using samples coming from overlapping frames, rather than try to eliminate duplicates through micro-level matching before sample selection. This is feasible when the multiplicity of units in the sample can be assessed with affordable effort.

An example is the use of a household-based and an establishment-based frame covering the study population which overlap in the child workers covered. After putting together the samples selected from the two frames, it may not be difficult to determine from each child selected his/her presence in the frame other than...
the one from which he/she has been selected. For instance, one may determine for each child selected from the household-based frame whether he/she works in a type of establishment included in the establishment frame; and similarly, determine for each child selected from the establishment-based frame whether his/her household is in the study area.

(3) In practice, it is often more effective to use the two approaches in combination. We may begin by consolidating the available frames into fewer frames, each new frame putting together one or more existing frames after removing duplicates on the basis of micro-level matching. The consolidated frames may overlap among themselves. Estimation from the combined sample selected from them would require information on the units’ multiplicities in the consolidated frames.

Example 2 in Section 8.2.2 provides an illustration. Various list frames from CPS (Child Protective Services) agencies were consolidated into a single non-overlapping list. The same was done for various list frames from non-CPS agencies. There was overlap between the two consolidated lists, estimation from which could be made using information on whether a sample unit was present in only one or in both of the two frames.

(4) Consolidation of the frame in the above example involves micro-level linkage among frames in order to remove duplicates. Another, often very useful, option is to enrich an available frame by incorporating into it information from other frame(s) – and possibly also from other types of sources – but without involving micro-level linkage. For instance, we may have two area-based frames: one a population-based frame of areas and households; and the other an economic activity based frame of areas and establishments. Through linkage between the frames even if only at the area level, we may be able to incorporate information from the second frame to obtain an enriched frame of households, with improved data for sample allocation, stratification, multistage design, etc.

Example 1 in Section 8.2.1 provides an illustration. Area-level information on socio-economic characteristics from the preceding population census was used to identify areas of concentration of the target population, on the basis of which geographical units in the existing frame could be stratified. In fact the study also provides an example of option (2) above. Units in the population received a chance of selection from the original general frame, while a part with high concentrations of the target population received an additional chance from its presence in the targeted frame.

(5) Finally, it is also possible to consider “doing nothing”. There are situations when an imperfect solution has to be accepted. For example, the overlap between frames may not be so large as to make a significant difference, or the cost of dealing with it may be too high in relation to the expected benefit. Or the effect of the overlap between frames may be significant, but removing it may not be feasible or affordable. Sometimes approximate solutions may be found to reduce the extent of the problem without eliminating it. Approximate estimation of multiplicity (the extent of overlap between frames) in Section 8.7.3 provides an example.
Chapter 9
Adaptive cluster sampling

9.1 Introduction

Adaptive sampling is a technique designed to obtain more adequate and efficient samples for a population which is rare and very unevenly distributed, compared to what is possible using a conventional sampling design. The technique is most effective when the population of interest is concentrated in relatively few and large clusters but little information is available on the extent, location and patterns of its concentration. Examples of such populations include street children, children engaged in street trades, child beggars.

9.1.1 The approach

To begin with, let us summarise briefly what the basic technique involves. We start from an initial sample selected by using some conventional probability sampling design. For example a random sample of area units may be selected, followed by a survey of child labour in each selected area unit. Such a design in itself may not yield a reliable sample if the population of interest is not well-distributed but is confined to rare and large clusters. By ‘not reliable’ we mean that the sampling error is too large for the survey results to be considered useful or representative of the study population. The sample may miss most of the clusters, or by chance too many of them may fall into the sample. Information on the location and patterns of these concentrations is necessary for improving the design. When no such information is available prior to the selection of the initial sample, the only way to improve the sample is to modify it after initial selection, making use of the information collected in the course of implementation of the initial sample.

Adaptive sampling aims to do just that – but in such a way that the probability nature of the sample is retained and statistically valid estimates can be provided from it. This is in contrast to the conventional sampling design where the sample selection process is fixed and predetermined, and not affected by what is found in the course of implementing the sample.

Different approaches are possible under the general idea of adaptive sampling. Among these, the most basic and pertinent in our context is the so-called adaptive cluster sampling. The discussion in this chapter is confined to this particular form of adaptive sampling.

The adaptive cluster sampling technique involves selecting an additional sample at (in the neighbourhood of) points where concentrations of the population of interest are found during implementation of the initial sample. In this way the resulting sample captures more cases from the population of interest. This over-representation is taken into account at the estimation stage in order to produce valid (unbiased) estimates from the sample.
As remarked by Thompson (1992), the adaptive sampling approach is that if at any point (selected in the sample), “fishing is good”, then “keep fishing”.

In more concrete terms, the basic procedure may be described as follows.

(1) We begin with an initial sample of units, selected with some random, conventional procedure.

(2) In the sample (as conceptually in the whole population), units are divided into two types: (a) units with a ‘sufficient concentration’ of the rare population of interest; and (b) other units, lacking such concentration.

(3) Type 2(b) units are retained in the sample as originally selected under (1).

(4) However in the case of type 2(a) units, additional units are added as follows. For each unit a ‘neighbourhood’ is defined on the basis of some specified rules (such as areas sharing a boundary). All units in the neighbourhood of an originally selected type 2(a) unit are added into the sample if they are also of type 2(a). The process is repeated for each added unit – all type 2(a) units in its neighbourhood are added to the sample, and so on.

(5) The result is that each original sample unit with a certain specified ‘sufficient concentration’ of the population of interest can bring in a whole cluster of such units into the sample. Selection of any unit in such a cluster results in the inclusion of all the units in the cluster – thus providing an over-representation of the population of interest in the sample.

(6) With pre-specified rules defining the neighbourhood and degree of concentration of units, the probability nature of the adaptive sample is preserved. The resulting unit selection probabilities can be taken into account at the estimation stage to yield statistically valid (unbiased) estimates.

(7) In practical application, it is also necessary to have rules which would terminate the process of adaptive additions to the sample. The objective is to avoid uncontrolled expansion of the sample as a result of adaptive additions. Of course, special steps to terminate the adaptive procedure are not always needed; often the procedure terminates automatically when no more units eligible for addition to adaptive clusters are found. However, given limited information about the population being surveyed, the sampler cannot always predict and control the adaptive process, and rules are needed to terminate the process if necessary. Also, it can be the case in certain situations that efficiency is improved by terminating the process earlier than its full application. As an environmentalist may say, “it may not be good to keep fishing, even if fishing is good for the moment”.

9.1.2 Potential advantages

Potential advantages of adaptive sampling over conventional sampling include the following three:

(1) The possibility of obtaining more precise estimates of population values such as means or totals for a given cost or level of effort.
(2) More information on the extent and patterns of concentration of the population of interest, which can be useful in identifying areas for closer study and targeted intervention.

(3) A larger yield of cases of special interest, on the basis of which their characteristics can be studied with greater detail and precision, including also the possibility of separate more intensive follow-up studies.

In relation to the relative efficiency, aspect (1) above, it should be noted that an adaptive design is not necessarily or always more efficient (or more cost-effective) than a conventional design. In any case its efficiency depends on how rare the population of interest is, and the pattern of concentration of that population. Of course the efficiency depends also on the parameters of adaptive design, such as the criteria defining what constitutes the ‘neighbourhood’ of a unit, the threshold (in terms of the level of concentration of the rare population) which triggers adaptive additions to the original sample, and, if applicable, the criterion for terminating the adaptive additions. These issues are considered in more detail in the following sections.

9.1.3 Limitations

(1) Adaptive sampling tends to be technically more complex in design than traditional sampling. A great deal of practical experience and expertise already exist concerning the traditional sampling approach.

(2) Higher levels of technical skills and survey experience are also required for the implementation of an adaptive design. For instance, interviewers have to be trained to decide how to continue sampling neighbouring units and when to terminate this procedure. Often special forms must be designed for controlling this process.

(3) The adaptive design requires additional information for the choice of design parameters. Often it is difficult to obtain reliable information on relative variances and costs of the different procedures.

The last-mentioned difficulty concerns in particular the choice of sample size. The choice of sample size involves both of the initial sample size, and the expected size of the final sample.

(4) The initial sample size being too small makes it likely that the sample misses significant clusters of the subpopulation of interest. Adaptive sampling can be useful only to the extent the initial sample hits such clusters. But too large an initial sample size may result in the final sample size becoming unaffordable and unmanageable.

(5) The final sample size is dependent on, but is not fixed by, the initial sample size. The expected final sample size depends, for a given initial sample size, on the extent and the pattern of concentration of the subpopulation of interest. In so far as these parameters are not known, the expected sample size is also not predictable. In addition, the actual sample size encountered is subject to chance variations. The final sample size may turn out to be too small to provide sufficient numbers of cases from the target population, especially for detailed surveys. A more likely
problem, however, is the sample size turning out to be too large to be manageable, affordable or necessary.

Steps are needed to keep control over the sample size with adaptive sampling. These are discussed in Section 9.7.4 below.

9.2 Basics of the technique

In this section, basics of adaptive cluster sampling are described in technical terms. Some further, more complex aspects will be introduced in subsequent sections.

To make the discussion more concrete, let us consider a specific situation.

Let the study area be a geographically or administratively delimited region of a country.

Suppose that the objective is to carry out a survey of children engaged in a particular sector in the study area. These children form a relatively small proportion of the total child population and tend to be concentrated in pockets over the study area. However, little prior information is available on the size and location of these concentrations.

The survey objectives are to estimate the number or proportion of children working in the sector of interest, to identify where and how they tend to be concentrated, and to obtain a sufficiently large sample of such children to study their characteristics and conditions through a detailed follow-up survey. Suppose that, given this scenario, adaptive cluster sampling appears an attractive and feasible alternative to a conventional approach.

9.2.1 Constructing adaptive clusters or networks

Suppose that the study area is divided into small non-overlapping area units, and we take a simple random sample of these units as our initial sample. Units in the sample are divided into two groups: group C consisting of area units where the number or proportion of children labouring in the sector of interest equals or exceeds a certain specific threshold; and the remainder, group U, composing of area units which do not reach the specific threshold of concentration.

For each unit in this initial sample, its neighbourhood is defined as consisting of the set of area units which share a boundary with it. The neighbourhood relationship is symmetric: if unit $i$ is in the neighbourhood of unit $j$, then unit $j$ is in the neighbourhood of unit $i$.

We start constructing an adaptive cluster from each type C unit selected into the initial sample by adding to it all type C units in its neighbourhood, and keep expanding the cluster by adding type C units in the neighbourhood of units already in the cluster. The process stops when there are no type C units left in the neighbourhood of units in the cluster.

The term network has been used to refer to clusters of the type described above. In fact, the whole population can be seen as divided exhaustively into non-overlapping networks: the networks being clusters of type C units, with each type U unit seen as a network of size 1.
Thus the population can be seen as divided in terms of non-overlapping units of two different types: (1) individual area units; or (2) networks of area units as defined above.

A conventional design is based on (1) as the sampling units – in our example, a simple random sample (SRS) of such units.

An adaptive design also begins with an initial sample of individual area units (1) - also assumed to be an SRS in our example - but then each selected unit is replaced by the whole network to which it belongs.

### 9.2.2 Clarifying terminology

Below we explain the use of the following terms: (1) adaptive clusters versus networks; (2) units, clusters, and elements or ultimate units.

1. **Adaptive clusters versus networks**

In the literature, the two terms ‘adaptive cluster’, and ‘network’ in the context of adaptive cluster sampling, are used to describe somewhat different concepts. It is useful to clarify our usage of these terms here.

Let the concept of neighbourhood of a unit and the condition C for additional adaptive selections in the neighbourhood be defined as above. The concepts and procedures are explained well by Thompson (1992, Section 24.1):

“When a selected unit satisfies the condition, all units within its neighbourhood are added to the sample and observed. Some of these units may in turn satisfy the condition and some may not. For any of these units that does satisfy the condition, the units in its neighbourhood are also included in the sample, and so on.

“Consider the collection of all the units that are observed under the design as a result of initial selection of unit \( i \). Such a collection, which may consist of the union of several neighbourhoods, will be termed a [adaptive] cluster when it appears in the sample. Within such a cluster is a sub-collection of units, termed a network, with the property that selection of any unit within the network would lead to inclusion in the sample of every other unit in the network …

“Any unit not satisfying the condition but in the neighbourhood of one that does is termed an edge unit. While selection of any unit in the network will result in inclusion of all units in the network and of all associated edge units, selection of an edge unit will not result in the inclusion of any other units. It is convenient to consider any unit not satisfying the condition a network of size one, so that … the population may be uniquely partitioned into networks.”

Note that while the networks form exhaustive and non-overlapping partitions of the population, this is not so in the case of adaptive clusters as defined above. The same edge unit may be included in more than one cluster, if it lies at the edge of more than one network.
Edge units are relevant to the adaptive design because they have to be identified in order to define the boundaries of networks and adaptive clusters. This identification involves collection and analysis of data, with associated costs. However, it is complex to include the information on all edge units in the estimation procedure. This is for the following reasons. The adaptive sample consists of units of the following types.

i. Networks in the sample, all units in which satisfy condition C.

A unit not satisfying condition C for adaptive additions can come into the sample in two ways:

ii. It is selected directly as a part of the initial sample. In this case, the selection probability of the unit is known from the initial sample design.

iii. It comes into the sample because it forms the edge unit of one or more adaptive clusters. In general, not all those adaptive clusters of which the unit in question forms an edge unit will be in the sample, and hence their probabilities of selection will not be known. Hence at least some of the edge units have unknown selection probability and therefore cannot be included in the estimation procedure in a straightforward manner.

Thus in practical application, the estimation is based on (i) networks in the sample, all units of which satisfy condition C, plus (ii) only those (edge) units (not satisfying condition C by definition) which are selected into the initial sample directly. As noted, the latter may also be considered ‘networks’, but always of size one and not satisfying condition C.

In the following, we will always assume the estimation procedure to be as described above, i.e. including only those edge units which were selected directly into the initial sample.

In the literature, distinction has been made between ‘networks’ and ‘clusters. A network is taken to refer to a neighbourhood of type C units as described above. A cluster is taken to include in addition type U units which were not in the initial sample, but have come into the sample by virtue of being in the neighbourhood of a network in the sample.

However, with the estimation procedure disregarding edge units not directly selected into the sample, the terminological distinction between the two terms - ‘cluster’, and ‘network’ in the context of adaptive cluster sampling - is no longer important for practical application. Estimation from the sample is based on the set (i)+(ii) of non-overlapping units with known selection probabilities.

Hence this terminological distinction is not relevant in our present discussion, and we would use the more familiar term cluster to refer to a ‘network’ in the above-mentioned sense. While sometimes it may be convenient to use the terms ‘clusters’ or ‘networks’ interchangeably, our clear preference is to use the term clusters, or more properly adaptive clusters, because the term ‘networks’ is also used frequently in contexts other than adaptive cluster sampling, such as for multiplicity sampling, or for sampling links between ‘networks’ of units as distinct from sampling individual units.
9.2 Basics of the technique

(2) Clusters, units, and elements or ultimate units

In this discussion of adaptive sampling, we use the term ‘unit’ to mean an area segment or some entity similarly representing a grouping of individuals.

The term ‘element’, or equivalently ‘ultimate unit’, normally refers to individuals associated with units as defined above, such as households, persons or children living or working in an area.

‘Cluster’ refers to a grouping of units. In the present context, it is an area segment, or a group of area segments in a neighbourhood satisfying certain conditions relating to the target population of interest. In our simplified terminology, it also stands for the term ‘network’ used in the literature, as explained above.

9.2.3 Basic estimation procedure

The following procedure assumes the initial sample to be a simple random sample. The precision of an adaptive sample developed around this initial sample is compared to a conventional sample, the latter also assumed to be a simple random sample.

Let subscript $i$ denote a cluster and $(i,j)$ any area unit in that cluster. We use the double subscript $(i,i)$ to indicate the particular area unit in the initial sample through which cluster $i$ has come into the sample. It would be helpful to clarify this notation further. The population is divided exhaustively into non-overlapping clusters. A cluster is identified by index $i$; $x_i$ is the set of units belonging to cluster $i$. One of these units, the one through which this cluster came into the sample is identified as $(i,i)$. The remaining $(x_i - 1)$ units in the cluster are identified as $(i,j), j \neq i$; there are no such units in a cluster of size one. Note that $(j \in x_i)$ indicates all the units in cluster $i$, including the originally selected unit $j = i$.

In addition, we will assume here – as is mostly the case in practice – a one-to-one correspondence between initial sample area units and the associated clusters. It is of course possible - especially in the presence of high sampling rates - for more than one initial sample unit to belong to the same cluster, or even for the same area unit to be selected into the sample more than once if the sampling is with replacement. Brief comments of such complications are made later in the text.

With a conventional SRS of size $n$ of area units (as for the initial sample in an adaptive design), an estimate of a mean per area unit and its variance is:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i ;$$  \hspace{1cm} (9.1)

$$\text{var}(\bar{y}) = \frac{(1 - f)}{n} \left[ \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{(n-1)} \right].$$  \hspace{1cm} (9.2)

Variable $y$ refers to any of the characteristics being measured in the survey. Here $f$ is the sampling rate ($n/N$ with $N$ as the number of areas in the population), often negligibly small.
With adaptive sampling on the lines described here, it can be easily shown that the resulting sample is equivalent to a SRS of size \( n \), with value \( y_i \) for any individual unit \( i \) in the initial sample replaced by the average value over the \( x_i \) units in the adaptive cluster formed around that unit:

\[
 z_i = \frac{1}{x_i} \sum_{j \in x_i} y_j. \tag{9.3}
\]

This gives the adaptive sample estimate of the mean per unit and its variance as:

\[
 z = \frac{1}{n} \sum_{i=1}^{n} z_i, \tag{9.4}
\]

\[
 \text{var}(z) = \frac{(1 - g/n) \left[ \frac{\sum_{i=1}^{n} (z_i - z)^2}{(n-1)} \right]}{n}, \tag{9.5}
\]

where \( g \) is the sampling rate \( n/G \), with \( G \) as the number of clusters in the population (normally \( G < N \) hence \( g > f \)).

Note that \( \bar{y} \) in (9.1) and \( Z \) in (9.4) estimate the same parameter: mean \( y \) value (say \( \bar{Y} \)) per unit in the population. Statistically, the two have the same expected value.

Also note that while the number of units (individual area segments) in the initial sample is \( n \), the number of units with adaptive sampling is larger, \( X = \sum_{i=1}^{n} x_i \).

The more heavily clustered the population of interest, the larger is this number expected to be compared with \( n \). Adaptive cluster sampling has replaced in this example the original SRS of \( n \) units with the same number of clusters of these units, each cluster constructed around a unit in the initial sample.

### 9.3 Relative efficiency of adaptive versus conventional sampling

Now we turn to the more complex question concerning how efficient in terms of variance and cost is adaptive sampling compared to conventional sampling. The answer depends on characteristics of the population and details of the sampling design chosen.

Starting with an initial sample of size \( n \), adaptive sampling brings in additional units into the sample and involves additional costs, in terms of money and effort. This additional money and effort could have been used instead to increase the original sample size (from \( n \) to, say, \( m > n \)), and then to stick with conventional sampling without introducing adaptive sampling.

Hence in comparing the two sampling approaches, we have to compare adaptive design constructed from initial sample size \( n \), with a conventional design with a larger sample size, \( m > n \), determined such that the overall cost of the two approaches is the same.
First consider the comparison in terms of variance. Comments on the relationship between \( m \) and \( n \) in term of cost will follow.

### 9.3.1 Variance

We can write equations (9.1) and (9.2) in terms of the new sample size. For a simple random sample of size \( m \):

\[
\bar{y} = \frac{1}{m} \sum_{i=1}^{m} y_i; \\
\text{var}(\bar{y}) = \frac{(1-f)}{m} \left[ \sum_{i=1}^{m} \frac{(y_i - \overline{y})^2}{(m-1)} \right], \quad f = \frac{m}{N}.
\]

(9.6)

(9.7)

Subscript \( i \) in (9.6)-(9.7) refers to individual units in the simple random sample of area units.

For adaptive cluster design, we still have \( n \) clusters formed around an initial simple random sample of \( n \) units. Equations (9.3)-(9.5) give estimates of the mean and variance.

Variance (say \( \sigma^2 \)) among units in the population can be decomposed into two components: (1) variance among units within clusters, \( \sigma_w^2 \); and (2) variance between cluster means, \( \sigma_b^2 \). The identity

\[
\sum_{i,j} (y_{ij} - \overline{y})^2 = \sum_{i,j} (y_{ij} - z_i)^2 + \sum_{i,j} (z_i - \overline{y})^2,
\]

(9.8)

dividing by \( N \), gives:

\[
\sigma^2 = \sigma_w^2 + \sigma_b^2.
\]

(9.9)

The term in square brackets in (9.2) or (9.7) is an estimate of the total variance \( \sigma^2 \), while the term in square brackets in (9.5) estimates only the between-cluster component \( \sigma_b^2 = (\sigma^2 - \sigma_w^2) \).

From the above equations it follows that adaptive cluster sampling is more efficient than conventional sampling if:

\[
\frac{\sigma_b^2}{\sigma^2} = 1 - \frac{\sigma_w^2}{\sigma^2} < \left( \frac{1-f}{1-g} \right) \left( \frac{n}{m} \right).
\]

(9.10)

The first factor on the right is >1, and the second factor is <1. Recall that \( f \) is the sampling rate for the initial sample, and \( g \) the same for the bigger adaptive sample constructed from an, albeit somewhat smaller, initial sample.

In order to assess relative efficiency of conventional versus adaptive design in terms of variances, we need an estimate of \( \sigma_b^2 / \sigma^2 \), the proportion which between-cluster variation forms of the total variance. The more heterogeneous is the cluster, the smaller is this proportion.
Disregarding the difference in the finite population correction between the two schemes, variance of the adaptive sample is essentially \( \frac{\sigma^2_b}{n} \), and that of the conventional sample is \( \frac{\sigma^2}{m} \). Hence the adaptive design is more efficient if

\[
m \left( \frac{\sigma^2_b}{\sigma^2} \right) = m \left( 1 - \frac{\sigma^2_b}{\sigma^2} \right) < n,
\]

(9.11)

where \( n \) is the size of the initial sample in the adaptive design which we would get for the same cost as a conventional sample of size \( m > n \).

Hence adaptive sampling is more efficient with more heterogeneous clusters (\( \sigma^2_b \) large, i.e. \( \sigma^2_b \) small), i.e. with clusters made up of dissimilar units. In the extreme case, when a cluster is made up of identical units, by including more than one unit of the cluster into the sample – as done in adaptive sampling – we add absolutely no new information, and such addition in the adaptive sample is therefore wasteful.

### 9.3.2 Cost

The following simple model is helpful in providing an idea of the relative costs of the two sampling schemes. In a simple random sample of area units, a cost model can be of the form:

\[
C_{srs} = C_0 + m \left( (1 - p) C_1 + p C_2 \right),
\]

(9.12)

where the first term \( C_0 \) is the overall fixed cost, and the second term is the additional cost dependent on the sample size. Per unit cost may differ between type C units, which are the elements of interest and would normally require more intensive data collection, and type U units which do not contain elements of special interest in the survey. These per unit costs are, respectively, \( C_2 \) and \( C_1 \). Here \( p \) is the proportion of type C among the \( m \) sample units.

Now let us compare this with an adaptive design with a simple random sample of \( n \) units as the initial sample. Suppose that the adaptive scheme adds \( n_c \) units to the sample; by design, all these are type C units. Their per unit cost is \( C_2 \) as defined above. The last-mentioned per unit cost applies to \( (p n + n_c) \) type C units in the sample. Per unit cost \( C_1 \) as defined above applies to \( n (1 - p) \) type U units in the sample.

Next, introducing an adaptive design involves additional fixed costs, for example due to increased overall complexity; it also involves the additional per unit cost of identifying unit type, boundaries and neighbourhood, etc. Let these costs be, respectively, \( \Delta C_0 \) and \( \Delta C_1 \). Apart from the \( (n + n_c) \) units in the adaptive clusters, forming adaptive clusters also requires identifying what are termed ‘edge units’ (see Section 9.2 above), which are type U units demarcating cluster boundaries. Let there be \( n_u \) such units for the sample clusters. The additional per unit cost \( \Delta C_1 \) applies to the edge units as well, meaning that it applies to a total of \( (n + n_c + n_u) \) units.

Hence a cost equation for adaptive sampling, similar to (9.12) above for conventional sampling, may be expressed as:

\[
C_{ads} = \left( C_0 + \Delta C_0 \right) + n (1 - p) C_1 + (np + n_c) C_2 + (n + n_c + n_u) \Delta C_1.
\]

(9.13)
For a given situation (the criteria used to determine adaptive additions to the sample, and pattern of concentration, size of the population, etc.), expected values of numbers like $\Pi_c$ or $\Pi_u$ vary with the initial sample size $n$ – albeit in complex ways.

In order to assess the relative efficiency of an adaptive design it is necessary to have some – even if only approximate – idea of these relationships, and also of the various cost components. On this basis one may be able to form an idea of the size of cost $C_{AdS}$ for a choice of initial sample size $n$, using equation (9.13). Then from equation (9.12), we can estimate $m$, the size of the conventional sample which would be obtained with the same resources $C_{SPS} = C_{AdS}$. With given initial sample size $n$ for adaptive design and the conventional sample size $m$ for the same cost, we can evaluate the relative efficiencies of the two designs in terms of variance using equation (9.10) or (9.11).

### 9.3.3 Factors influencing relative efficiency

In the following summary we draw on Thompson and Seber (1996, Section 5.7), but with some reinterpretations and explanatory comments added.

The following characteristics tend to increase the efficiency of adaptive cluster sampling relative to conventional random sampling.

1. The population of interest forms a small proportion of the general population from which it has to be screened out.

2. The distribution of the population of interest (type C units) in the general population is patchy and uneven. The implication is that starting from a type C unit there is a higher chance of finding another type C unit in the neighbourhood, than would be the case in a more uniformly distributed target population. This reduces the relative number of type U units in the neighbourhood of clusters – such ‘edge’ units increase survey costs without making any contribution to the precision of the estimation.

3. The population is aggregated into large and heterogeneous clusters, whose size and location are not all known before sample selection. The efficiency of adaptive cluster sampling is increased to the extent individual clusters reflect the heterogeneity of the whole population. Such a pattern implies that the within-cluster variance is a high proportion of the total variance of the population, and consequently the between-cluster variance is low.

4. There is saving of fieldwork costs due to geographical clustering of units resulting from adaptive additions to the initial sample, compared to the more dispersed conventional sample (which is usually similar in design to the initial sample).

5. The cost of enumerating non-members of the population of interest – even though normally lower than the cost of interviewing members of that population – is relatively high.

6. Cheap methods (e.g. use of easy-to-measure auxiliary variables) are available for screening to separate out members from non-members, for defining adaptive clusters and identifying edge units demarcating cluster boundaries.

7. An adaptive design facilitates the use of more efficient estimation procedures.
9. Estimation

With adaptive sampling, a cluster of units is selected through the selection into the sample of one or more units belonging to the cluster. The following are the commonly used estimators, expressed here in terms of adaptive cluster sampling. The reader may refer to the discussion in Chapter 4, in particular Section 4.6. In describing its application to adaptive cluster sampling, it is convenient to use the notation introduced in Section 9.2.3. Subscript \( i \) denotes a cluster and \((ij)\) an area unit in it. With \( x_i \) as the number of units belonging to cluster \( i \), set \((j \in x_i)\) indicates all the units in cluster \( i \). Let \( f_{ij} \) be the selection probability of unit \( j \) in cluster \( i \).

9.4.1 Horvitz-Thompson (H-T) estimator

Suppose that a sample of \( n \) units out of \( N \) is selected, and adaptive clusters are formed around these selected units. Since some clusters may contain more than one selected unit, the number of distinct clusters in the sample is \( k \leq n \).

The inclusion probability \( (\pi_i) \) of a cluster is the probability that at least one of the units \((j)\) comprising it is selected. With \( f_{ij} \) as the selection probability of unit \( j \) in cluster \( i \), the inclusion probability of the cluster is given by Equation (4.1), namely:

\[
\pi_i = 1 - \prod_{j \in x_i} (1 - f_{ij}); \text{ with corresponding weight } w_i = \frac{1}{\pi_i}.
\]  

(9.14)

With \( y_{ij} \) as the value of a variable for unit \( j \) in cluster \( i \), and \( y_i = \sum_{j \in x_i} y_{ij} \) its cluster aggregate, Horvitz-Thompson (1952) estimator of its population total is:

\[
Y_{HT} = \sum_{i=1}^{k} \left( \frac{y_i}{\pi_i} \right).
\]  

(9.15)

Means per cluster and per unit are, respectively \( Y_{HT}/K \) and \( Y_{HT}/N \), where \( K \) is the number of clusters and \( N \) the number of units in the population. These numbers are normally not known. A more stable and practical version of the above is (Hájek, 1971):

Mean per cluster:

\[
\sum_{i=1}^{k} \left( \frac{y_i}{\pi_i} \right) / \sum_{i=1}^{k} \left( \frac{1}{\pi_i} \right),
\]  

(9.16)

Mean per unit:

\[
\sum_{i=1}^{k} \left( \frac{y_i}{\pi_i} \right) / \sum_{i=1}^{k} \left( \frac{x_i}{\pi_i} \right),
\]

the denominators being estimators, respectively, of the number of clusters \( K \) and the number of units \( N \) in the population.

In the context of adaptive sampling, the estimator tends to be complex because it involves inclusion probabilities of clusters which come into the sample indirectly through the selection of units comprising them. In any case, variance estimation tends to be complex, as it involves not only inclusion probabilities \( \pi_i \) for individual clusters, but also their joint inclusion probabilities.
9.4.2 Hansen-Hurwitz (H-H) estimator

As noted in Section 4.6.1B, this estimator is simpler, hence more practical and useful for our purpose, than estimator (9.14-9.16) above. Again, using the notation introduced in Section 9.2.3, let $f_{ij}$ be the selection probability of unit $j$ in cluster $i$, and $\delta_{ij}$ a $(0,1)$ indicator with $\delta_{ij} = 1$ if unit $(ij)$ has been selected into the initial sample and $\delta_{ij} = 0$ otherwise. Assuming independence between selection of units, the expected number of times a cluster appears in the sample is, as given in Equation 4.2:

$$p_i = \sum_{j \in Z_i} f_{ij}.$$  \hspace{1cm} (9.17)

By definition, the actual number of appearances of the cluster in the sample is

$$s_i = \sum_{j \in Z_i} \delta_{ij}.$$  \hspace{1cm} (9.18)

Hence the total weight (i.e. the combined weight of all its appearances, whether 0, 1 or more than 1) given to a cluster in the sample is

$$w_i^{(s)} = \left( \frac{\sum_{j \in Z_i} \delta_{ij}}{\sum_{j \in Z_i} f_{ij}} \right) = \frac{s_i}{p_i}.$$ \hspace{1cm} (9.19)

Subscript $(s)$ indicates that this is the total weight given to the $s_i$ appearances of cluster $i$ in the sample. The weight received by each of the $s_i$ appearances of the cluster is

$$w_{ij} = \left( \frac{1}{\sum_{j \in Z_i} f_{ij}} \right) = \frac{1}{p_i} w_{ij}^{(s)}; \quad w_{ij} = w_i.$$  \hspace{1cm} (9.20)

The weights (9.19, 9.20) are also the weights of each of the $x_i$ units constituting the cluster:

$$w_{ij}^{(s)} = w_{ij}^{(s)}; \quad w_{ij} = w_i.$$

Population total, population size, and means per cluster and per unit are estimated as:

- Population total: $\Sigma_{i=1}^n w_i y_i$;
- Number of clusters: $\Sigma_{i=1}^n w_i$;
- Number of units: $\Sigma_{i=1}^n w_i x_i$; \hspace{1cm} (9.21)
- Mean per cluster: $\Sigma_{i=1}^n w_i y_i / \Sigma_{i=1}^n w_i$;
- Mean per unit: $\Sigma_{i=1}^n w_i y_i / \Sigma_{i=1}^n w_i x_i$.

9.4.3 Multiplicity estimator

The multiplicity estimator has been described in some detail in Section 4.6. Here we note the main points and express the estimator specifically in terms of adaptive cluster sampling.
With multiplicity estimator, each appearance of the cluster \((i)\) in the sample is weighted inversely proportional to the probability of selection \((f_{ij})\) of its particular unit \((ij)\) through which it has been selected, taking into account the number of units \((x_i)\) from which that cluster could have been selected. The weight a particular appearance of the cluster receives is

\[
w_i = \frac{\delta_{ij}}{x_i f_{ij}},
\]

where, as defined earlier, \(\delta_{ij}\) is a \((0,1)\) indicator with \(\delta_{ij} = 1\) if unit \((ij)\) has been selected into the initial sample and \(\delta_{ij} = 0\) otherwise.

Note that, unlike estimator (9.20), the particular unit of the cluster through which the cluster has been selected into the sample does matter, and of course it also matters how many times the cluster appears in the sample. In the special case when the selection probabilities are uniform for all units \((ij)\) of the cluster \((i)\), \(f_{ij} = f_i\) say, estimators (9.20) and (9.22) are the same.

The total weight received by all appearances of the cluster in the sample is

\[
w_j^{(s)} = \frac{1}{x_i} \sum_{j \in s} \left( \frac{\delta_{ij}}{f_{ij}} \right). \tag{9.23}
\]

The advantage of the multiplicity estimator is that the selection probability (and the corresponding weight applied to sample data) is determined separately for each unit through which the cluster comes into the sample, and does not depend on the selection probabilities of other units in the cluster. The estimator requires information on selection probability only for the units which are actually selected into the initial sample. In fact it is not necessary to know whether the clusters selected through different units are the same or different clusters. For these reasons, this estimator is the most practical one in our context of child labour surveys in developing countries.

### 9.4.4 Weight-share estimator

The weight-share method of estimation was described in Section 4.6.3. As is the case generally, for adaptive cluster sampling as well the weight-share estimator is a form of the multiplicity estimator outlined above. This can be seen from the following.

In the context of adaptive cluster sampling, let \(A\) indicate the population of area units from which the initial sample is drawn, and \(B\) the target population of the same units grouped into adaptive clusters. The relationship between \(A\) and \(B\) is defined in terms of the relationship between the target units and the initial sampling units, and by the neighbourhood relationship defining the clusters among the initial sampling units.
Figure 9.1. Illustration of weight-share estimator for adaptive cluster sampling

In the diagram on the left, the links shown are between corresponding units in population $A$ (of sampling units), and population $B$ (of analysis units). Unit $j = 2$ is selected from population $A$. In conventional sampling, with one-to-one correspondence between units in populations $A$ and $B$, the above selection simply implies the selection of unit $j$ in $B$ (shown by the connecting solid line). With adaptive design, units 1-3 form a cluster, and the selection of any one of units brings into the sample all units in its cluster. Corresponding clusters in the two populations can be seen to have 3 links to each other, one link coming from each pair of corresponding units in the clusters. The diagram on the left shows these three links.

An entirely equivalent way of looking at the situation is to consider every unit in the cluster in population $B$ to be linked with every unit in the corresponding cluster in population $A$. This is depicted by showing all the links in the diagram on the right. All units ($j = 1, 2, 3$) in the cluster in population $B$ come into the sample through selection of unit $j = 2$ in $A$ (hence all these units are connected by solid lines). Each unit in $B$ is linked to 3 units in the cluster in $A$. There are $3 \times 3 = 9$ links between corresponding clusters in the two populations.

Figure 9.1 shows how this model fits the adaptive cluster sampling design. Population $A$ consists of units identified by subscript $(ij)$; these units are grouped into adaptive clusters (identified with subscript $i$) which implicitly exist, though need not be (and are normally not) known at the time of the sample selection. Each unit $(ij)$ in population $B$ can be seen as being linked to (selected through) its own representation in population $A$. This is shown on the left in Figure 9.1. For the illustration, a solid line indicates a link with a unit in $A$ which has been selected into the sample $S^A$; a dotted line indicates a unit in $A$ which has not been selected. The selection of any unit $(ij)$ in $B$, through its link with a corresponding unit in $A$, brings into the sample the whole cluster $i$ to which the unit $(ij)$ belongs.

According to the weight-share method, in estimating from sample $S^B$ of analysis units, the weights given to a cluster $i$ or any of its units $(ij)$ is given by the ratio of two factors:
9. Adaptive cluster sampling

### Factor

<table>
<thead>
<tr>
<th>Numerator: weighted sum of the links of cluster (i) in analysis population ((B)), to sampling units ((ij)) in the sample from population ((A)), with each link weighted by (\left(\frac{\delta_{ij}}{f_{ij}}\right)), the inverse of the selection probability of the concerned sampling unit ((ij)).</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ \sum_{j \in x_i} \left( \frac{\delta_{ij}}{f_{ij}} \right) ]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Denominator: total number of links of cluster (i) in population ((B)), to sampling units ((ij)) in the whole population ((A))</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ x_i ]</td>
</tr>
</tbody>
</table>

Using the notation introduced above (Section 9.2.3), it can be easily seen that the above two factors are given by the terms in the column “Expression (1)” according to the representation of links in the diagram on the left of Figure 9.1. Their ratio, which is the weight given to adaptive cluster \(i\) in the final sample according to the weight-share method, is numerically exactly as given in equation (9.23) for the multiplicity estimator:

\[
W_i^{(ws)} = \frac{1}{x_i} \sum_{j \in x_i} \left( \frac{\delta_{ij}}{f_{ij}} \right). \tag{9.24}
\]

Since a unit \((ij)\) in (\(B\)) is included in the sample \(S^B\) as a result of the selection of any of the \(x_i\) units in cluster \(i\) in \(S^A\), the links between the two populations can also be viewed as in the diagram on the right of Figure 9.1. (See Lavallée, 2007, Section 6.5.) For cluster \(i\) in \(S^B\), now there are a total of \(x_i^2\) links between units in the two populations: \(x_i\) links for each of the \(x_i\) units in \(i\). It can be seen that the two factors forming the numerator and the denominator of the weight of \(i\) are given by the column “Expression (2)”. Their ratio is exactly as before, as given in equation (9.23).

Though the original multiplicity estimator, (9.22), and the weight-share estimator, (9.24), are identical in content, in certain practical situations the form of the latter can be less convenient than that of the multiplicity estimator for the following reason.

If selection of different sampling units \(j\) lead to the selection of the same analysis unit \(i\), the form of (9.24) involves consolidation of the information over the multiple selections of \(i\). However, sometimes it is not possible or easy to identify that such multiple appearances in the sample refer in fact to the same analysis unit, rather than to different analysis units. An example is the selection of children through sampling at locations where they work, when a child may work (and hence be selected) at more than one location. It would be necessary to collect sufficient information of a personal nature in order to identify whether selections from different locations refer to the same child. The collection of such information may not be possible, easy, convenient, cheap, or even ethical.

By contrast, the multiplicity estimator (9.22) refers to each selection separately. It is not necessary to identify whether different selections refer to the same analysis unit or to different analysis units.
9.5 Numerical illustration

9.5.1 Illustrative population

A. The data

The small illustrative population in Tables 9.1-9.6 consists of 60 units (small area segments). The variable of interest ($y$) is the number of child street workers in each area.\footnote{The illustrations in this section are drawn in part from Mehran (2012).} For each unit information on variable $y$ is shown in Table 9.1. The table shows these values for each of the 60 units, and the average value for each of the 20 clusters (defined below). The total number of child street workers in these study areas is 654. In reality, such values are not known for the whole population and our objective is to illustrate how well samples with different designs can be expected to predict these values: the total number or mean per area of labouring children ($y$), the extent and pattern of their geographical concentration, etc. Adaptive clusters are identified and defined after the selection of the initial sample, and only around the units selected into that sample.

Suppose that we chose the criteria for adaptive sampling to be $C_y = 5$; that is, for each selected unit with $y \geq C_y = 5$, other units with $y \geq 5$ in the neighbourhood of the first unit selected are grouped together to form an adaptive cluster. (These are the ‘C’ type of units defined earlier). Each of the remaining units with $y < C_y = 5$ forms a separate cluster of size one. (These are the units of type ‘U’.)

The classification between U and C type of areas and clusters depends on the choice of criterion $C_y$. We have constructed a somewhat artificial example with $U=C=10$ (resulting in a total of $G=U+C=20$ adaptive clusters) in order to limit the size of the illustrative table. In real situations, one would normally choose criterion $C_y$ to have many more clusters of type U than of type C.

Suppose that the application of these parameters results in the grouping of 60 units in the population into $G=20$ clusters each represented by a row of Table 9.1. A row lists $y$ values (the number of child labourers) in areas forming the cluster.

In 10 cases, each cluster consists of only one type U area ($y < C_y = 5$). They are each considered a separate cluster by itself; no units are added to these to form adaptive clusters. (In the table, these units are labelled ‘S’ for small.)

The remaining 50 units meet the requirement of having $y \geq 5$ (these are ‘type C’ units), and potentially serve as starting points for forming adaptive clusters. Each unit in an adaptive cluster may share a boundary with (i.e. be in the neighbourhood of) one or more type C units from the same cluster, but not with any type C units from a different cluster. In this way, 10 adaptive clusters have been formed in the illustration as shown in Table 9.1. These include 6 medium sized (‘M’) clusters each with up to four units; and 4 large (‘L’) clusters each with five or more units.

The last two columns of Table 9.1 show:

$$x_i$$, the number of area units in cluster $i$, and
\[ y_i = \frac{1}{x} \sum_{j=x} y_{ij}, \] the average per unit of individual \( y_{ij} \) values for the cluster.

<table>
<thead>
<tr>
<th>( j= )</th>
<th>Type of adaptive cluster</th>
<th>Unit IDs in the cluster</th>
<th>Unit number in the cluster ((j): x_i)</th>
<th>( y_i ) mean per unit</th>
<th>( x_i ) no. of units in cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M</td>
<td>01-04</td>
<td>9 6 12 14</td>
<td>4</td>
<td>10.3</td>
</tr>
<tr>
<td>2</td>
<td>S</td>
<td>05</td>
<td>2</td>
<td>1</td>
<td>2.0</td>
</tr>
<tr>
<td>3</td>
<td>L</td>
<td>06-15</td>
<td>12 14 17 12 14 10 17 11 12</td>
<td>10</td>
<td>13.1</td>
</tr>
<tr>
<td>4</td>
<td>S</td>
<td>16</td>
<td>3</td>
<td>1</td>
<td>3.0</td>
</tr>
<tr>
<td>5</td>
<td>L</td>
<td>17-24</td>
<td>18 12 13 10 17 15 13</td>
<td>8</td>
<td>14.4</td>
</tr>
<tr>
<td>6</td>
<td>S</td>
<td>25</td>
<td>2</td>
<td>1</td>
<td>2.0</td>
</tr>
<tr>
<td>7</td>
<td>L</td>
<td>26-32</td>
<td>17 13 14 11 10 16</td>
<td>7</td>
<td>14.0</td>
</tr>
<tr>
<td>8</td>
<td>S</td>
<td>33</td>
<td>3</td>
<td>1</td>
<td>3.0</td>
</tr>
<tr>
<td>9</td>
<td>S</td>
<td>34</td>
<td>3</td>
<td>1</td>
<td>3.0</td>
</tr>
<tr>
<td>10</td>
<td>M</td>
<td>35 10</td>
<td>1</td>
<td>1</td>
<td>10.0</td>
</tr>
<tr>
<td>11</td>
<td>S</td>
<td>36</td>
<td>3</td>
<td>1</td>
<td>3.0</td>
</tr>
<tr>
<td>12</td>
<td>M</td>
<td>37-39</td>
<td>11 7 10</td>
<td>3</td>
<td>9.3</td>
</tr>
<tr>
<td>13</td>
<td>M</td>
<td>40-41</td>
<td>5 5</td>
<td>2</td>
<td>5.0</td>
</tr>
<tr>
<td>14</td>
<td>S</td>
<td>42</td>
<td>2</td>
<td>1</td>
<td>2.0</td>
</tr>
<tr>
<td>15</td>
<td>M</td>
<td>43-45</td>
<td>12 12 14</td>
<td>3</td>
<td>12.7</td>
</tr>
<tr>
<td>16</td>
<td>S</td>
<td>46</td>
<td>2</td>
<td>1</td>
<td>2.0</td>
</tr>
<tr>
<td>17</td>
<td>L</td>
<td>47-55</td>
<td>15 13 11 17 18 17 16 10 16</td>
<td>9</td>
<td>14.8</td>
</tr>
<tr>
<td>18</td>
<td>S</td>
<td>56</td>
<td>4</td>
<td>1</td>
<td>4.0</td>
</tr>
<tr>
<td>19</td>
<td>M</td>
<td>57-59</td>
<td>8 12 5</td>
<td>3</td>
<td>8.3</td>
</tr>
<tr>
<td>20</td>
<td>S</td>
<td>60</td>
<td>1</td>
<td>1</td>
<td>1.0</td>
</tr>
</tbody>
</table>

| Total | Adaptive clusters | 20 |
| Area units | 60 |
| Child street workers | 654 |

S units not satisfying the adaptive clustering criterion (type ‘U’ units - see text)
M medium sized adaptive clusters (formed around type ‘C’ units in the initial sample)
L large adaptive clusters

To summarise, our illustrative population consists of \( N = 60 \) area units, grouped into \( G = 20 \) clusters, in three domains, L, M and S. Of the clusters, 10 are type U units, i.e. with \( y < C_y = 5 \) labouring children. Therefore these form 10 clusters each with only a single unit; the average \( y \) value per unit in this group is 2.5. We call them ‘small clusters’. The remaining 50 are type C units, i.e. with \( y \geq C_y = 5 \) labouring children. These form 10 clusters. Of these, 6 contain one to four area units each (‘medium clusters’). The remaining 4, containing seven to ten units each, are ‘large clusters’.
B. Graphical illustration of geographical order

Table 9.1 reflects some kind of geographical order of the units and clusters. A possible configuration of the population is illustrated graphically in Figure 9.2. The target population of 654 child street workers is spread in 60 area segments of a city, shown in the figure by rectangles or near rectangles with solid borders. There are fifty area segments with 5 or more child street workers shown in grey; the remaining are the ten area segments with less than 5 child workers.

In the square for each area unit, area unit serial number appears in the small top left rectangle, and number of child street workers in the centre of the square.

The area segments with 5 or more child street workers form 10 contiguous areas, each called a cluster. Area segments with less than 5 child street workers each form a ‘cluster’ of its own, i.e. a cluster of size 1. In the illustration shown in Figure 9.2, there are altogether 20 clusters: nine of size more than 1 and one of size 1, all containing only type C units; and 10 of size 1, each containing only a type U unit.35

The area segments are also related to each other by sample design. According to the design of adaptive cluster sampling, if an area segment selected in the sample satisfies the given condition C (5 or more child street workers), the neighbouring area segments are also included in the sample: two area segments are ‘neighbours’ if they are geographically adjacent to each other and each has 5 or more child street workers. In the present context, geographically adjacent means sharing one or more common boundaries (north, south, east and west).

For example, in Figure 9.2, area segments 1 and 2 are ‘neighbours’ as they share a common boundary and each has 5 or more child street workers (9 child street workers in area 1 and 6 in area 2). Area segments 2 and 5 are not ‘neighbours’ although they share a common boundary. This is because one of them (area 5) has less than 5 child street workers.

35 As noted earlier, in the literature the area segments defined above are often called ‘networks’ as distinct from ‘clusters’. Type U ‘edge units’ surrounding each network composed of type C unit(s) are included along with the network to form a ‘cluster’. Data for edge units are normally not used in the estimation from the adaptive sample. Hence we have used the terms ‘cluster’ or ‘adaptive cluster’ in preference to ‘network’.
In the square for each area unit, the area unit serial number appears in the small top left rectangle, and number of child street workers in the centre of the larger square.

Two area segments may not be immediate neighbours of each other for them to form parts of the same cluster, i.e. they may be neighbours of neighbours. In Figure 9.2, area segments 1 and 4 are not immediate neighbours in the sense defined above, but share a common ‘neighbour’ (area 2 is neighbour of both area 1 and area 4, so is area 3). They are therefore part of the same cluster.

C. Rearrangement of data to facilitate exposition and analysis

In order to facilitate our exposition and analysis, the data have been rearranged by cluster type in Table 9.2. This information is summarised in the bottom panel of the table. Larger clusters are seen to contain not only more units (area segments) but also a greater concentration of variable \( y \) (numbers of labouring children) per unit. The table also shows total variance of \( y \), decomposed into within and between cluster components.
The number of areas in the adaptive sample is greatly increased, from 6.0 in the initial sample to an expected value of 34.2 – by a factor of nearly six in our (somewhat unrealistic) illustration. Note that this factor is twice as large as the average number (3.0) of units per cluster in the population. This is because the adaptive sample is designed to heavily over-represent large clusters.

An expected proportion \(28.9/34.2 = 85\) per cent of areas in the adaptive sample come from the 4 large (‘L’) clusters. These large clusters are expected to contain \(332/600 = 55\) per cent of all the 600 labouring children in the whole population.

Thus adaptive sampling has the potential to help identify and locate large concentrations of the population of interest by increasing the chance of their appearing in the sample.

The adaptive sample as a whole is expected to contain \((2.5 \times 1.0 + 33.2 \times 11.5)/600 = 64\) per cent of the 600 labouring children in the population.

Consequently, the adaptive sample also helps in obtaining a larger number of elements of interest (such as children working in a particular sector), thus providing a basis for more detailed surveys of those.

### Table 9.2. Data of Table 9.1 rearranged for exposition and analysis

<table>
<thead>
<tr>
<th>Type of Adaptive Cluster</th>
<th>Unit Number in the Cluster (j):</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unit Values ((y_{ij})):</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i= 1</td>
<td>S</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>8</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>12</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>11</td>
<td>7</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>9</td>
<td>6</td>
<td>12</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>17</td>
<td>13</td>
<td>17</td>
<td>14</td>
<td>11</td>
<td>10</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>18</td>
<td>12</td>
<td>13</td>
<td>10</td>
<td>17</td>
<td>17</td>
<td>15</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>18</td>
<td>17</td>
<td>16</td>
<td>10</td>
<td>16</td>
<td>11</td>
<td>17</td>
<td>15</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>12</td>
<td>14</td>
<td>17</td>
<td>12</td>
<td>12</td>
<td>14</td>
<td>10</td>
<td>17</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Summary of the population

<table>
<thead>
<tr>
<th>Type</th>
<th>no. of units</th>
<th>no. of clusters</th>
<th>units per cluster</th>
<th>mean (y) per unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>10</td>
<td>10</td>
<td>1.0</td>
<td>2.5</td>
</tr>
<tr>
<td>M</td>
<td>16</td>
<td>6</td>
<td>2.7</td>
<td>9.5</td>
</tr>
<tr>
<td>L</td>
<td>34</td>
<td>4</td>
<td>8.5</td>
<td>14.0</td>
</tr>
<tr>
<td>Total</td>
<td>60</td>
<td>20</td>
<td>3.0</td>
<td>10.9</td>
</tr>
</tbody>
</table>

#### Variance of unit \(y\) values

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>total</td>
<td>24.3</td>
</tr>
<tr>
<td>within clusters</td>
<td>4.9</td>
</tr>
<tr>
<td>between clusters</td>
<td>19.5</td>
</tr>
</tbody>
</table>
9.5.2 Simple random sample of area units

Suppose we select an initial simple random sample of \( n = 6 \) areas. Every area in the population has the same \( \frac{n}{N} = \frac{6}{60} = \frac{1}{10} \) chance of selection. In this initial sample of area units, the number of child street workers in each sample area is counted. For each selected unit that contains five or more child street workers, all neighbouring units with 5 or more child workers are included in the sample as additional units, and the process goes on until no new units can be added to the sample.

First we will consider the results of one arbitrary sample with a simple random sample (SRS) design for the initial sample. Then average or expected characteristics of samples under the SRS design will be considered. We will compare these with the set of all possible systematic random samples. An example of stratified design will be discussed in the subsequent subsection.

A. One simple random sample

Let the first column of Table 9.3 be the unit sequence numbers drawn in an initial sample of 6 units. These units are selected at random with equal probability \( f_i = \frac{6}{60} = 0.1 \). An adaptive cluster sample is formed around the units selected in the initial sample as described above, and the number of child street workers in each area in the two samples is counted.

The number of child street workers in the initial sample is \( 73 = 12 + 12 + 15 + 3 + 15 + 16 \). Therefore, based on the initial sample, the estimate of the total number of child street workers in the city is \( 730 = 10 \times 73 \), where 10 is the weight calculated as the inverse of the sampling probability \( f_i = 0.1 \). (This may be compared to the actual population value of 654 child street workers.)

<table>
<thead>
<tr>
<th>Initial sample</th>
<th>Adaptive cluster sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) area unit selected</td>
<td>(1) areas in cluster</td>
</tr>
<tr>
<td>(2) selection probability</td>
<td>(2) number of areas</td>
</tr>
<tr>
<td>(3) sample weight</td>
<td>(3) sample weight</td>
</tr>
<tr>
<td>(4) no. of child workers in area</td>
<td>(4) no. of child workers in cluster</td>
</tr>
<tr>
<td>(5) estimate of total number</td>
<td>(5) estimate of total number</td>
</tr>
<tr>
<td>( f_i )</td>
<td>( \chi_i )</td>
</tr>
<tr>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>18</td>
<td>0.1</td>
</tr>
<tr>
<td>23</td>
<td>0.1</td>
</tr>
<tr>
<td>33</td>
<td>0.1</td>
</tr>
<tr>
<td>47</td>
<td>0.1</td>
</tr>
<tr>
<td>53</td>
<td>0.1</td>
</tr>
<tr>
<td>Total</td>
<td>73</td>
</tr>
</tbody>
</table>

*Second selection of same cluster

The right hand panel of the table shows the adaptive cluster sample resulting from the above initial sample, drawn from the population displayed in Figure 9.1. The first column shows the adaptive sample derived from the initial sample. It consists of 22
area segments (with a total of 292 child street workers), grouped into 4 clusters (area sequence numbers 1-4, 17-24, 33 and 47-55). In line with the design of adaptive cluster sampling, the edge units are ignored. Two of these clusters are selected twice: cluster “17-24” selected through unit numbers 18 and 23 in the initial sample, and cluster “47-55” selected through units 47 and 53. Column (2) shows the cluster size \( x_i \) (number of area units in it). Column (3) shows the multiplicity weight for each selection of a cluster according to equation (9.22), with \( \delta_{ij} = 1 \) for each unit selected by definition, and \( f_{ij} = 0.1 \), i.e. \( w_i = 10/x_i \).

In the present context, the weights are equal for all child street workers in the same adaptive cluster. Thus, the 41 child street workers in the cluster consisting of area segments (1-4) have all the same weight (=2.5). For the 133 child workers in the second cluster, the weight is 1.25 for each of the two selections of the cluster, giving a total weight of 2.5. The 3 child workers in the single unit cluster (33) have all weight 10. For the 133 child workers in the cluster consisting of area segments (47-55), the weight is 1.11 for each of the two selections of the cluster, giving a total weight of 2.22.

The estimated total number of child street workers is obtained by the multiplication of the weights, column (3), with the number of child workers in the adaptive cluster, column (4), and added over all selections of the clusters. The result obtained here is:

\[
\text{Estimated total number of child street workers} = 715.6.
\]

Table 9.4 shows the same results using the weight-share formulation. The left and the right panels of the table correspond, respectively, to “Expression (1)” and “Expression (2)” of Section 9.4.4. We compare each of these results with those from standard application of the multiplicity estimator in the right hand panel of Table 9.3. Concerning the “Expression (1)” panel, the only difference from Table 9.3 is that multiple selections of an adaptive cluster are now consolidated, each consolidated figure appearing in a single row. The number of links of an adaptive cluster in population \( B \) to units in population \( A \) is simply the number of units in the cluster \( (x_i) \), shown in column (1). The weighted number of links to selected units in \( A \) is the number of units selected from that cluster (1 or 2 in the case of the four clusters in our example), multiplied by \( (1/f) = 10 \); the result \( (s_i) \) is shown in column (2). The ratio \( (2)/(1) \) gives the required weight. Multiplying this by the number of child street workers in the cluster gives an estimate of their total number from the cluster.

In the alternative formulation in the “Expression (2)” panel of Table 9.4, the number of links is seen to equal \( (x_i^2) \), rather than \( (x_i) \) as in the previous case. Hence columns (1) and (2) from the left panel are both multiplied by the number of units in the cluster \( (x_i) \), so that their ratio and the rest of the estimation procedure are not affected in any way.
Table 9.4. Estimate for the adaptive cluster sample using the weight-share formulation
Sample of 4 distinct adaptive clusters (from 6 selections in the initial sample)

<table>
<thead>
<tr>
<th>areas in cluster</th>
<th>“Expression (1)”</th>
<th>“Expression (2)”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td></td>
<td>number of links in population</td>
<td>weighted number of links in sample</td>
</tr>
<tr>
<td></td>
<td>$x_i$</td>
<td>$s_i$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$s_i / x_i$</td>
</tr>
<tr>
<td>1-4</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>103</td>
</tr>
<tr>
<td>17-24</td>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>115</td>
</tr>
<tr>
<td></td>
<td></td>
<td>288</td>
</tr>
<tr>
<td>33</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>47-55</td>
<td>9</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>133</td>
</tr>
<tr>
<td></td>
<td></td>
<td>296</td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>292</td>
</tr>
<tr>
<td></td>
<td></td>
<td>716</td>
</tr>
</tbody>
</table>

“Expression (1)” and “Expression (2)”; see text.

B. Expected (average) value of all possible initial simple random samples of size 6

Table 9.5 shows average results for all possible samples, when the initial sample consists of a simple random sample of $n=6$ units. This corresponds to a uniform sampling rate of $(6/60)=0.1$. The table shows the expected number of units in different parts of the sample (clusters of size categories S, M and L). The population values are the same as in Table 9.2.

Table 9.5. Expected value of illustrative conventional and adaptive samples
Initial sample: SRS of size $n=6$ units

<table>
<thead>
<tr>
<th>Population (60 units, 20 clusters)</th>
<th>Conventional sample (6 units)</th>
<th>Adaptive sample (6 clusters)</th>
</tr>
</thead>
<tbody>
<tr>
<td>no. of units</td>
<td>no. of clusters</td>
<td>units per cluster</td>
</tr>
<tr>
<td>S</td>
<td>10</td>
<td>1.0</td>
</tr>
<tr>
<td>M</td>
<td>16</td>
<td>6</td>
</tr>
<tr>
<td>L</td>
<td>34</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>60</td>
<td>20</td>
</tr>
</tbody>
</table>
9.5 Numerical illustration

The conventional initial sample is expected to consist of 6 units, distributed among strata S, M and L in proportion to the number of units in the stratum in the population. The numbers shown in the table are the expected values of the number of sample units in the strata, i.e. the numbers averaged over all possible samples. Hence these can be fractional. These will of course be whole numbers in any particular sample.

Formation of adaptive clusters around each unit in the initial sample in strata M and L brings additional units into the final sample. There are a total of 5 units expected to be selected from strata M and L, each yielding one adaptive cluster (though it is possible for different selected units to result in the same adaptive cluster). No units are added to the one unit selected from stratum S – that unit forms a cluster by itself. In the end there are 6 clusters in the sample, containing a total of up to 34.2 units. (The last-mentioned number would be smaller when multiple selections of the same clusters occur.) Units in the final sample represent 10 per cent of the units in stratum S, over 25 per cent of those in stratum M, and as many as 85 per cent of the units in stratum L. Such over-representation of large concentrations of the variable of interest is precisely the objective of adaptive sampling.

The small illustration given here is unrealistic in that, in practice, one would expect many more type U units (here grouped into stratum S) in a population than type C units (here grouped into strata M and L). Recall that adaptive clusters are formed only around the latter type of units satisfying the chosen criteria for the purpose.

Table 9.6 shows a more realistic case with 100 type ‘U’ units in stratum S (in place of 10 in the previous example), with other characteristics of the population unchanged. In order to capture the same proportions of units in the different strata as in the previous example, the initial sample size has to be increased, in our example, from \( n = 6 \) to \( n = 15 \).
### Table 9.6. Illustration with many small (type U) units

<table>
<thead>
<tr>
<th>group</th>
<th>no. of units</th>
<th>no. of clusters</th>
<th>units per cluster</th>
<th>mean y per unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>100</td>
<td>100</td>
<td>1.0</td>
<td>2.5</td>
</tr>
<tr>
<td>M</td>
<td>16</td>
<td>6</td>
<td>2.7</td>
<td>9.5</td>
</tr>
<tr>
<td>L</td>
<td>34</td>
<td>4</td>
<td>8.5</td>
<td>14.0</td>
</tr>
<tr>
<td>Total</td>
<td>150</td>
<td>110</td>
<td>1.4</td>
<td>5.9</td>
</tr>
</tbody>
</table>

**Initial sample size = 6 units**

Sample size as in Table 9.5.

<table>
<thead>
<tr>
<th>Conventional sample</th>
<th>Adaptive sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>(6 units)</td>
<td>(6 clusters)</td>
</tr>
<tr>
<td>number selected</td>
<td>number selected</td>
</tr>
<tr>
<td>units</td>
<td>clusters</td>
</tr>
<tr>
<td>S</td>
<td>4.0</td>
</tr>
<tr>
<td>M</td>
<td>0.6</td>
</tr>
<tr>
<td>L</td>
<td>1.4</td>
</tr>
<tr>
<td>Total</td>
<td>6.0</td>
</tr>
</tbody>
</table>

**Initial sample size = 15 units**

Sample size increased to keep numbers selected from strata M & L unchanged from Table 9.5.

<table>
<thead>
<tr>
<th>Conventional sample</th>
<th>Adaptive sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>(15 units)</td>
<td>(15 clusters)</td>
</tr>
<tr>
<td>number selected</td>
<td>number selected</td>
</tr>
<tr>
<td>units</td>
<td>clusters</td>
</tr>
<tr>
<td>S</td>
<td>10.0</td>
</tr>
<tr>
<td>M</td>
<td>1.6</td>
</tr>
<tr>
<td>L</td>
<td>3.4</td>
</tr>
<tr>
<td>Total</td>
<td>15.0</td>
</tr>
</tbody>
</table>
C. Sampling distribution with initial systematic samples of size 6

With the given ordering of 60 units in the population, there are 10 possible systematic samples of size 6. (These are obtained by starting with each of the ten numbers 1-10, and then taking every 10th unit in the list.) Table 9.7 gives the adaptive sample estimates of the total number of child street workers based on 10 initial area units drawn by systematic sampling. Mean of the estimates is 654, validating the fact that the estimation procedure is unbiased. Standard deviation of the estimates is 110.5, about 7 per cent lower than standard deviation of the estimates based only on the initial samples.

<table>
<thead>
<tr>
<th>Systematic sample</th>
<th>Initial sample units</th>
<th>Estimate based on adaptive sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1  11  21  31  41  51</td>
<td>715.0</td>
</tr>
<tr>
<td>2</td>
<td>2  12  22  32  42  52</td>
<td>685.0</td>
</tr>
<tr>
<td>3</td>
<td>3  13  23  33  43  53</td>
<td>681.7</td>
</tr>
<tr>
<td>4</td>
<td>4  14  24  34  44  54</td>
<td>681.7</td>
</tr>
<tr>
<td>5</td>
<td>5  15  25  35  45  55</td>
<td>545.4</td>
</tr>
<tr>
<td>6</td>
<td>6  16  26  36  46  56</td>
<td>391.0</td>
</tr>
<tr>
<td>7</td>
<td>7  17  27  37  47  57</td>
<td>739.2</td>
</tr>
<tr>
<td>8</td>
<td>8  18  28  38  48  58</td>
<td>739.2</td>
</tr>
<tr>
<td>9</td>
<td>9  19  29  39  49  59</td>
<td>739.2</td>
</tr>
<tr>
<td>10</td>
<td>10 20  30  40  50  60</td>
<td>622.5</td>
</tr>
</tbody>
</table>

Mean 654.0
Standard deviation 110.5

9.5.3 Adaptive cluster sampling with stratification

A. Example of a stratified sample

Suppose that some information on variations in the concentration of child street workers is available making it possible to divide the study area into strata - say into two strata, one with low concentration of child street workers and one with relatively high concentration. The two strata are shown with dashed lines in Figure 9.3. Starting at the top left corner of the city and proceeding to the bottom right, the dashed lines divide the area into four segments. The second and the fourth segments together form the high concentration stratum, and the other two segments form the low concentration stratum.
Altogether, the low concentration stratum is composed of 24 area units and the high concentration stratum of 36 area units. An initial sample of 6 area units is selected by systematic sampling, 2 from the low concentration stratum and 4 from the high concentration stratum. The results are shown in Table 9.6. The procedure in the table is identical to that of Table 9.3 already discussed. The only difference is that the sampling rates and hence the weights in the initial sample differ by stratum.
### Table 9.8. Illustration of stratified adaptive cluster sample

<table>
<thead>
<tr>
<th></th>
<th>Initial sample</th>
<th>Adaptive cluster sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) area unit selected</td>
<td>(1) areas in cluster</td>
</tr>
<tr>
<td></td>
<td>(2) selection probability</td>
<td>(2) number of areas</td>
</tr>
<tr>
<td></td>
<td>(3) sample weight</td>
<td>(3) sample weight</td>
</tr>
<tr>
<td></td>
<td>(4) no. of child workers in area</td>
<td>(4) no. of child workers in cluster</td>
</tr>
<tr>
<td></td>
<td>(5) estimate of total number</td>
<td>(5) estimate of total number</td>
</tr>
<tr>
<td></td>
<td>( f_i )</td>
<td>( \chi_i )</td>
</tr>
<tr>
<td><strong>Stratum 1</strong></td>
<td></td>
<td>= ( (3)*(4) )</td>
</tr>
<tr>
<td>1</td>
<td>0.083</td>
<td>12</td>
</tr>
<tr>
<td>40</td>
<td>0.083</td>
<td>12</td>
</tr>
<tr>
<td><strong>Stratum 2</strong></td>
<td></td>
<td>= ( (3)*(4) )</td>
</tr>
<tr>
<td>6</td>
<td>0.111</td>
<td>9</td>
</tr>
<tr>
<td>15</td>
<td>0.111</td>
<td>9</td>
</tr>
<tr>
<td>24</td>
<td>0.111</td>
<td>9</td>
</tr>
<tr>
<td>47</td>
<td>0.111</td>
<td>9</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\*Second selection of the same cluster

B. Sampling distribution with stratified systematic samples of size 6

All possible adaptive samples with an initial stratified sample of size 6 drawn by systematic sampling were obtained. It generated 108 adaptive samples.\(^{36}\) The corresponding adaptive estimates of the total number of child street workers were calculated. Their frequency distribution is shown in Figure 9.4 below. As expected, the mean of the distribution is 654.0, indicating that the estimates are unbiased. The standard deviation is 65.8, almost half the standard deviation reported earlier (110.5) for systematic sampling of the initial sample without stratification. It can be observed in the figure that the frequency distribution is not symmetric. The median of the distribution is 674.3 and the mode 689.3.

![Figure 9.4. Frequency distribution of estimates of total number of child street workers with stratified adaptive cluster sampling](chart)

\(^{36}\) The systematic sampling interval in the low concentration stratum is 24/2 = 12, giving 12 possible starts (from numbers 1-12) for systematic sampling. For the high concentration stratum, the sampling interval and the number of possible systematic samples equals 36/4 = 9. From the two strata together, we have 12x9 = 108 distinct samples.
C. Remark on stratified adaptive cluster sampling

The above illustration is simplified in an important way: it assumes that adaptive clusters formed with a stratified initial sample do not cut across the initial strata.

However, clusters according to an adaptive sampling scheme may cut across the original strata boundaries by incorporating neighbouring areas from strata other than from where they started in the initial sample. The sample selection and the estimation procedures have to take this complexity into account. See further comments in Section 9.6.2 below.

9.6 Further technical aspects

In describing the basics of the adaptive cluster sampling method in the preceding sections, we assumed a simple design, namely that the initial sample consists of a simple random sample of units such as area segments, and within each selected unit, all elements of interest (such as children working in a particular sector) are included in the survey. The adaptive design was compared to a conventional design of the same type, i.e. a simple random sample (SRS) of compact clusters.

In practice, the design is (and generally has to be) more complex. In this section we briefly describe some common variations from single stage simple random sampling.

9.6.1 Unequal unit selection probabilities

Often it is efficient to select units with unequal probabilities, most commonly with probability proportional to some measure of unit size (PPS). For example, area units may be selected with probability proportional to the population size, or to the number of children in the area - or better still, the number of children working in the sector of interest, where such information is available in the sampling frame even if it is only approximate.

Adaptive sampling differs from conventional sampling in that for any unit, its selection probability in the conventional design is replaced by the sum of selection probabilities of all units in its adaptive cluster, as defined by the adaptive sampling criterion chosen. (The adaptive cluster always includes the original unit itself.)

Starting with a simple random sample as the initial sample, the cluster selection probability (and the selection probability of each unit in the cluster) in the adaptive sample is proportional to \( x_i \), the number of units in the cluster. By the same token, if the unit selection probabilities for the initial sample are proportional to a variable \( a_{ij} \), the cluster selection probability is \( a_i = \sum_{j \in x_i} a_{ij} \) summed over units \( j \) in the cluster \( i \). \(^{37}\)

In estimating from the sample, unit values have to be weighted inversely proportional

\(^{37}\) Note that, more precisely speaking, the term ‘probability of selection’ used above refers to the number of times a unit is expected to be selected, which can exceed 1.0, for instance for large clusters. This is handled in the Hansen-Hurwitz or the multiplicity estimator simply by including a cluster (and units in it) in the estimate as many times as the cluster appears in the sample as a result of selection of units into the initial sample. Index \( i \) in the above equations refers to distinct selections, rather than to distinct clusters, in the sample. This has been explained in Section 9.4. This point is important because, in the presence of large concentrations of the target population, multiple selection of adaptive clusters can occur frequently.
to their selection probabilities. Defining $w_i = x_i/a_i$ for a cluster, the estimated mean in equation (9.4) is now:

$$z = \frac{\sum_{i=1}^{n} w_i z_i}{\sum_{i=1}^{n} w_i}.$$ (9.25)

The expression for unit variance in equation (9.5) $\sum_{i=1}^{n} (z_i - z)^2 / (n - 1)$ is replaced by:

$$\frac{n}{n - 1} \left[ \sum_{i=1}^{n} w_i (z_i - z)^2 / \sum_{i=1}^{n} w_i \right].$$ (9.26)

Note that SRS for the initial sample is a special case of the above, with $a_i = x_i$, so that $w_i = 1$, and $\sum_{i=1}^{n} w_i = n$.

### 9.6.2 Stratification

In practice, sample design almost always involves some form of stratification. The initial sample is selected independently from each stratum. However, clusters according to an adaptive sampling scheme may cut across the original strata boundaries. The sample selection and the estimation procedures have to take this into account.

There are two ways of dealing with this situation.

1. One is to modify the adaptive rule such that the resulting clusters are not allowed to cross strata boundaries. That is, any cluster is formed by adding units only from within the stratum of its initial unit. Additions to a cluster are terminated where it reaches a stratum boundary, and no more units are added to it from the neighbouring stratum. This means that the adaptive sample, just as the traditional sample, remains independent between strata. Hence stratification involves no new problems at the estimation stage. However, some efficiency is lost by not allowing adaptive clusters to cut across stratum boundaries.

2. The other, possibly more efficient option is to ignore strata boundaries in the formation of adaptive clusters. Some resulting clusters may cut across strata boundaries and this has to be taken into account at the estimation stage. The estimation formulae are more complex with this design, though they can be worked out in most situations. See for example Thompson and Seber (1996, Section 4.9), and Thompson (1992, Chapter 26). The derivations are rather cumbersome and it does not appear useful to elaborate them here. The principles are simple:

   a) A unit in the adaptive sample contributes to the estimate only for the stratum to which it belongs. This means that a cluster cutting across strata contributes to more than one stratum.

   b) As always, a unit contribution is weighted in inverse proportion to its selection probability. The unit’s probability of selection into the sample equals the total probability of selection of the adaptive cluster to which it belongs. The latter in turn is the sum of original selection probabilities of all the units contained in the cluster, irrespective of the strata from which those units come.
9.6.3 Multistage sampling

Multiple stages of sampling may be introduced in one or both of the following ways:

i. subsampling within units, and/or

ii. selection of units in two or more stages.

Before discussing this issue, it is useful to reiterate the use of terms ‘cluster’, ‘unit’ and ‘element’ in the following.

As noted in Section 9.2.2(2), in the context of adaptive sampling, we use the term ‘unit’ to mean an area segment or some entity similarly representing a grouping of individuals. The term ‘element’, or equivalently ‘ultimate unit’, normally refers to individuals associated with units as defined above, such as persons living or working in an area. The term ‘cluster’, or equivalently ‘network’, refers typically to a group of area segments in a neighbourhood satisfying certain conditions relating to the target population of interest.

Subsampling within units refers to taking a sample of elements (persons) within units (area segments). The selection of units in two or more stages refers to the situation when some larger area units are selected first, within which area segments are selected as second or lower stage units. Nevertheless, adapted clusters are formed around area segments, and not the higher stage units.

A. Subsampling within units

This means taking a sample of, for example, households, persons, children or working children within sample areas (units), rather than including all such elements in the area in the survey.

Scheme (1)

Scheme (1) represents the ‘ideal’ scheme applied in conventional two-stage sampling. The scheme has been elaborated in Section 3.3.1 and further in Section 5.5.1.

In conventional sampling, subsampling is introduced for the purpose of controlling the final sample size and the selection probabilities of the ultimate units, such as children in a household-based child labour survey. A common procedure is to select areas with probability proportional to some measure of size (PPS sampling) and then subsample ultimate units within each selected area with probability inversely proportional to that measure of size (inverse-PPS sampling). This scheme can yield uniform selection probabilities for the ultimate units, and at the same time an approximately constant sample size from each area in the sample. In more precise terms, the sampling stages are as follows.

Units are selected with probability proportional to some measure of size \( a_i \), say with probability \( p_{1i} = (f/b)a_i \), and then sub-sampled with inverse-PPS say at \( p_{2i} = (b/a_i) \). This gives the overall selection probability for the ultimate units to be a constant \( f \). Normally, size measure \( a_i \) reflects the number of elements (persons, children, working children) in unit \( i \). When \( a_i \) equals the actual number of elements in the unit (say \( a'_i \)), then at the second stage the expected number of selected elements per unit is also a constant \( b \). In general it is variable \( b_i = b(a'_i/a_i) \)
9.6 Further technical aspects

Scheme (2)

This is scheme (1) modified for adaptive cluster sampling.

With an adaptive design we have units of two types in the sample: unit types U and C. Type U units, as defined earlier, do not meet the criterion for the selection of additional units in the neighbourhood. Hence these appear in the sample in the same way as units in a conventional design. For these units, conventional methods of subsampling of the type described above in (1) may be appropriate.

The sampling scheme, along with the notation used, is shown in Table 9.9. Note that \((x, a, b, f, \ldots)\) are actually arbitrary constants, but it is helpful to give them a meaning as done here so as to make the relationship between the different sampling schemes clearer.

Type C units defined earlier are those which meet the condition for the addition of neighbouring units to form adaptive clusters. Here \(a_i\) represents not the size measure of the individual unit initially selected, but the sum of size measures of all the \(x_i\) units in its cluster. Apart from this change, scheme (2) has the same structure as scheme (1): units are selected with probability proportional to measure of size \(a_i\), \((p_{1i} = (f/b) a_i)\), and then sub-sampled with inverse-PPS \((p_{2i} = (xb/a_i))\), resulting in constant \((fx)\) overall selection probability for the ultimate units.

However, this scheme gives a constant sample size per cluster, and hence is not a suitable scheme under adaptive cluster sampling. This is because the whole point of adaptive sampling is to retain the oversampling of units which appear in large clusters in the population. This can be achieved by subsampling within units as shown in schemes (3) and (4).

<table>
<thead>
<tr>
<th>Scheme</th>
<th>(p_{1i})</th>
<th>(p_{2i})</th>
<th>(p_i = p_{1i} \cdot p_{2i})</th>
<th>(c_i = p_{2i} \cdot a_i)</th>
<th>(c_i / x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1), (2)</td>
<td>(\left(\frac{f}{b}\right) a_i)</td>
<td>(\left(\frac{b}{a_i}\right) x)</td>
<td>(fx)</td>
<td>(bx)</td>
<td>(b)</td>
</tr>
<tr>
<td>(3)</td>
<td>(\left(\frac{f}{b}\right) a_i)</td>
<td>(\left(\frac{b}{a}\right) x)</td>
<td>(fx\left(\frac{a_i}{a}\right))</td>
<td>(bx\left(\frac{a_i}{a}\right))</td>
<td>(b\left(\frac{a_i}{a}\right))</td>
</tr>
<tr>
<td>(4)</td>
<td>(\left(\frac{f}{b}\right) a_i)</td>
<td>(\left(\frac{b}{a_i}\right) x)</td>
<td>(fx\left(\frac{x_i}{x}\right))</td>
<td>(bx\left(\frac{x_i}{x}\right))</td>
<td>(b\left(\frac{x_i}{x}\right))</td>
</tr>
</tbody>
</table>

Variables:
- \(x_i\) Number of units (area segments) in cluster \(i\)
- \(a_i\) Number of elements (persons, children, working children) in cluster \(i\)
- \(c_i\) Sample size (number of elements selected) in cluster \(i\)

Constant parameters:
- \(x\) Average number of units (area segments) per cluster
- \(a\) Average number of elements (persons, children, working children) per cluster
- \(b\) Average sample size (number of elements selected) per unit (area segments)
- \(fx\) Overall selection probability of an element (person, child, working child, etc.)

In conventional sampling, all quantities above refer to units (areas), rather than to clusters; \(x \equiv 1\) by definition.
9. Adaptive cluster sampling

Scheme (3)

The scheme for selecting area segments in the initial sample is the same as in (1) or (2). Then adaptive clusters are formed around the selected units. Elements (households, persons, children, etc.) are selected within clusters at a constant rate, rather than with inverse PPS as in the previous schemes. The overall selection rate becomes proportional to the measure of size $a_i$, the number of elements (persons, children, working children) in cluster $i$. The same applies to the sample size (in terms of the number of elements selected) per cluster. The last column in Table 9.9 gives the equivalent figure per unit, which is of the same order of magnitude in the different schemes.

Scheme (4)

An alternative scheme would be to have the subsampling rate as $f_{2i} = b_i(x_i/a_i)$. The overall selection rate becomes proportional to $x_i$, the number of units (areas) in the adaptive cluster. The same applies to the sample size (in terms of the number of elements selected) per cluster.

The choice between schemes (3) and (4) may be guided by the desirability of avoiding extreme variations in overall selection probabilities or sample-takes per unit.

B. Selection of units in two or more stages

Suppose that, as the initial sample for an adaptive design, we first select a sample of primary sampling units (PSUs) such as large areas, and within each of those selected areas, select one or more area segments as the final stage area units. Then the adaptive sampling procedure is applied at the level of these final stage area units. In other words, adaptive sampling is still based on characteristics of the last stage area units, rather than of the PSUs. Some of the clusters formed may cut across boundaries of the PSUs, possibly going into PSUs not selected into the original sample.

The procedure for estimating mean per unit is straightforward in principle, as follows.

For each unit in the initial sample, the probability of selection is known: it is the product of the probability of selection of the PSU in which it lies, and the probability of selection of the unit within the PSU. The probability of selection of the adaptive cluster formed around a selected unit, and hence the final probability of selection of every unit in that cluster, is the sum of probabilities of selection of all the units in that adaptive cluster.\(^{38}\)

For estimating a mean, a weighted average of unit values in all the adaptive clusters in the sample is taken, with weights taken as inverse of the unit final selection probabilities.

We can sum these quantities over units ($j$) in each cluster, and then over clusters ($i$), noting that $w_j = w_i$ uniformly for all units $j$ in cluster $i$:

$$\frac{\sum_j w_j y_j}{\sum_j w_j} = \frac{\sum_i w_i \sum_j y_j}{\sum_i w_i x_i} = \sum_i w_i y_i \quad .$$

\(^{38}\) The term ‘probability of selection’ used above actually refers to the number of times a unit is expected to be selected. See note in Section 9.6.1.
Here $X_i$ is the number of units in the cluster. $y_i = \sum_j y_{ij}$ is the total value of the variable for cluster $i$, summed over units $j$ in the cluster.

### 9.6.4 Multivariate and auxiliary criteria

Adaptive sampling is considered where large concentrations of the variable of interest are found in the initial sample. It therefore appears natural to define the criteria for introducing adaptive additions to the sample points in terms of some critical value of the main variable of interest $y_i$, say the number of working children in area $i$:

$$\text{Adaptive additions if } y_i \geq C_y.$$  \hspace{1cm} (9.28)

Variations are of course possible on the above simple specification. The condition for adaptive additions can be based on multiple variables and/or auxiliary variables. This in no way affects the method or the estimation procedure. It is the cost and variance, and hence efficiency of the procedure, which depend on the criterion or criteria chosen.

The criteria may involve more than one variable if the primary objective of the survey concerns several variables. When multiple variables are involved, the conditions for introducing adaptive sampling may be in terms of any one of the variables reaching a critical value, or all of them reaching their critical values, or this may apply to certain subset(s) of variables.

Note that adaptive addition stops only when all the units in the neighbourhood of a unit selected or added into the adaptive cluster fail to meet the critical limit for continued addition.

Criteria in terms of auxiliary variables are needed when prior information on the actual variables of interest in the survey is not available, and/or when auxiliary variables can be more easily and cheaply collected during the survey operation. Of course, auxiliary variables are useful for the purpose only in so far as they are correlated with the actual variables of interest in the survey.

A particularly attractive option is to use a simple auxiliary variable in order to start and terminate adaptive additions to an initial sample point. This is because information on the variable used has to be collected and analysed during the survey operation itself, which is easier to do if the variable is simple and readily collected.

The edge units, determining boundaries of adaptive clusters, are not normally included in estimation from the survey. Yet information has to be collected on them in order to determine whether the adaptive additions can be terminated. It is wasteful if such information is too complex or costly to collect and analyse.

Some examples of simple questions or procedures which may be useful for the purpose are: asking some ‘key’ respondent whether the problem being surveyed exists in the area (e.g. concerning street children, or street child workers); quick observation and assessment, without involving interviewing; noting the presence of the sort of establishments or other locations normally employing children in the concerned sectors.
9.6.5 Order statistics

In place of using some pre-specified criterion, such as a value in unit $i$ reaching or exceeding a certain level, $y_i \geq C_y$, for determining whether adaptive additions should be made, one may simply take a certain number (say $n_c$ out of $n$) of the largest $y$ values in the initial sample for the purpose. That is, the $n_c$ largest $y$ values determine the points in the initial sample where adaptive additions will take place. If $y_{(i)}$ represents ordered values with $y_{(i)} > y_{(i+1)}$, then $n_c$ adaptive clusters are formed around the sample units $y_{(1)}$ to $y_{(n_c)}$ - those are treated as type C units – and no adaptive clusters are added to the remaining $(n - n_c)$ initial sample units $y_{(n_c+1)}$ to $y_{(n)}$.

Adaptive clusters are formed in the usual way. All units in the neighbourhood of a selected type C unit are examined; any unit with $y \geq y_{(n_c)}$ are added to the cluster. The above procedure is repeated for every added unit until no units satisfying the above criterion are found in the neighbourhood of any units in the cluster.

There can be two types of reason for replacing the normal criterion like $y_i \geq C_y$ with that choosing a specified number of largest values:

i. in order to keep a control on final adaptive sample size when sufficient information for the choice of appropriate threshold $C_y$ for the purpose is not available; and

ii. in order to draw a profile of extreme values and study more closely the largest values – in addition to the normal objective of estimating population means and totals.

In the usual adaptive procedure based on a condition such as $y_i \geq C_y$, the structure of the population in terms of adaptive clusters is implicitly fixed prior to sample selection, even though the structure may not be known to us explicitly for the entire population. The sampling process only determines which particular clusters happen to be selected.

However, the situation is more complex when adaptive clusters are formed around a certain specified number of units with the highest values found in the initial sample. Here the structure in terms of adaptive clusters is not pre-fixed independently of the sampling process, but is determined by the outcome of the initial sample.

Nevertheless, despite this complexity, it can be shown that procedures based on order statistics can provide an unbiased estimator of the population mean. However, estimation of variance can be much more complex than, say, standard procedures which apply for ‘normal’ adaptive sampling. Simply applying the standard procedure may, nevertheless, provide a useful indication on the magnitude of sampling error (Thompson and Seber, 1996, Chapter 6).

9.6.6 Imperfect detectability

Imperfect detectability means that not all elements of interest to the survey within the selected units (area segments) may be successfully identified. For example, we may select a sample of area segments, but may not succeed in enumerating all the child workers in the selected areas.
Imperfect detectability is an important source of non-sampling error, especially in surveys of an elusive population such as children living and working outside their home.

Imperfect detectability introduces a downwards bias in the estimation of the population total. It also increases variance of the estimates. The variance is further increased if the probability of detection itself has to be estimated from existing surveys or other sources.

The above considerations apply to all types of design whether conventional or adaptive. However, there are additional and potentially more serious effects of imperfect detectability with adaptive sampling. Imperfect detectability may result in the decision not to introduce adaptive additions at some points in the population or the initial sample, when in fact the decision would have been the opposite if detectability were perfect. Thus imperfect detectability affects not only coverage and effective selection probabilities, but also the sampling distribution (Thompson and Seber, 1996, Chapter 9).

The increase in variance as a result of incomplete detection tends to be similar in conventional and adaptive designs. Hence it tends to attenuate differences in efficiency between the two designs.

9.7 Implementation issues

This section considers issues in the implementation of an adaptive cluster design.

9.7.1 When should adaptive sampling be introduced?

This, of course, is the first and most fundamental issue. For a given amount of resources, would sampling efficiency and data quality be enhanced by introducing an adaptive design?

If there are any known large concentrations, they should be automatically included in the survey rather than being subject to chance selection into the sample. The issue then needs to be considered only for the remainder of the population.

The initial answer to the question on whether to introduce adaptive sampling depends on our assessment of how rare and clustered is the target population, and our knowledge about the location of its concentrations. Adaptive sampling is most effective when there is reason to believe that quite large concentrations of the population of interest do exist, but the concentrations are irregular and their extent and patterns are unknown. Adaptive sampling can turn out to be more efficient than ordinary stratified disproportionate sampling even when concentrations of the population of interest are known.

The justification for introducing adaptive sampling is strengthened when only a limited number of sample points can be afforded.

Three aspects in particular should influence the decision to introduce adaptive sampling.
9. Adaptive cluster sampling

(1) Relative efficiency in terms of variances and costs of adaptive versus conventional sampling, as concerns the estimation of means, totals and similar overall parameters of the population as a whole.

(2) The objective of identifying locations and patterns of concentration and obtaining a more detailed profile of extreme values; also the need to ‘capture’ a sufficient number of cases from the target population for in-depth studies.

(3) Technical skills and survey arrangements required for the implementation of an adaptive design.

Concerning (1), usually, it is difficult to obtain reliable information on relative variances and costs of the different procedures. Nevertheless, often differences are large and can provide a reasonably clear basis for choice.

Similarly, (3) is a matter of judgement about the available skills and experience of the survey staff.

Condition (2), when present, is likely to be more clear and decisive. It refers to the frequent situation when conventional sampling procedures cannot guarantee that sufficient numbers of cases of the rare population would be captured in the survey.

9.7.2 Criteria, definitions, rules for adaptive sampling

The next choice concerns the criteria to trigger additional adaptive sampling at points in the initial sample. The criteria must be objective, i.e. must not be dependent on which particular points happen to be selected into the initial sample. The criteria should be clear and not too complex, so as to ensure uniform application by different persons. The criteria can be in terms of several variables, all, some or none of which may be the same as the target variables to be measured in the survey. Information for applying the criteria has to be collected not only for units to be included in the (expanded) final sample, but also for the edge or boundary units delimiting sample clusters – units which are normally excluded from the final sample at the estimation stage.

This makes it attractive to use cheaper and easy-to-collect simple, ideally dichotomous, auxiliary variables for determining the conditions for adaptive additions to the sample.

Typically the criterion for this purpose is in the form:

\[ y_i \geq C_y, \]

where \( y_i \) is the target variable of interest. Adaptive sampling is introduced if \( y_i \) for a sample point reaches or surpasses the critical value \( C_y \).

The choice of threshold \( C_y \) can be critical in determining the final sample size. If \( C_y \) is too low, it could trigger too much adaptive sampling making the final sample size too large. Too large a value would restrict the introduction of adaptive sampling, and the final sample may turn out to be too small.

When information for the choice of \( C_y \) is lacking, an alternative is to specify the limit in terms of a fixed number of clusters to be included for the introduction of adaptive sampling. Uncertainty in the final sample size is reduced, as the number of clusters is
now fixed, but variability remains depending on the size of clusters encountered during the application of adaptive sampling.

In any case, it is prudent to place some (arbitrary) limit on how large any cluster is allowed to grow. The process of adaptive additions at the point concerned is terminated when the limit is reached.

The definition of what constitutes a *neighbouring unit* also needs to be clear, and should not be too complex to permit uniform application. The definition should be a symmetrical definition: if unit B is defined to be in the neighbourhood of A, then A is in the neighbourhood of B. This condition ensures that structure of clustering in the population is predetermined (once the choice of \( C_y \) has been made); what units any cluster would comprise is pre-determined, independently of the sampling process and irrespective of the particular initial sample unit from which the cluster has been constructed.

For area units, the usual definition of neighbourhood is in terms of shared physical boundaries. More complex criteria may also be used for convenience and greater effectiveness. Restrictions may be placed on the procedure for constructing clusters, for example by not permitting extension of clusters over domain or strata boundaries, or by selecting systematic samples covering the population and permitting linkage between units only within the same systematic sample (thus avoiding very large or compact clusters).

Alternative criteria may be used in place of geographical proximity for defining the neighbourhood, such as family or friendship relationship between individuals.

### 9.7.3 Size of the initial sample

The choice of sample size has to be concerned with both of the *initial sample* size (which determines the number of sample locations to be covered), and the expected size of the *final sample* (which determines the overall workload).

The final sample size is not fixed even with fixed initial sample size. The expected value of the final sample size, of course, depends on the initial size, but not linearly. With large sampling rates, some large clusters can be selected more than once (i.e. more than one initial sample unit resulting in the selection of the same cluster). Repeated selection does not add to the effective sample size, nor to the precision of the results.

On the other hand, a small sample size makes it likely that the sample misses most of the significant clusters of the subpopulation of interest. Adaptive sampling can be useful only to the extent the initial sample hits such clusters.

### 9.7.4 Controlling sample size

We refer here to the size of the final sample.

The expected sample size depends, for a given initial sample size, on the extent and pattern of concentration of the subpopulation of interest. In so far as these parameters are not known, the expected sample size is also not predictable. In addition, the actual sample size encountered is subject to chance variation around the expected value.
The sample size may turn out to be too small to provide sufficient numbers of cases from the target population, especially for detailed surveys. A more likely problem, however, is the sample size turning out to be too large to be manageable, affordable or necessary.

Steps are needed to keep control over the sample size with adaptive sampling. The following are the main practical steps one may take for the purpose (Thompson and Seber, 1996, Section 5.8).

Firstly, one may do something before the survey to avoid or minimise the problem:

1. The design may be based on information on parameters collected from pilot studies or from similar studies in the past. Most relevant is the information on size and distribution of the target population, in particular the extent and patterns of its concentration into clusters.

2. The study area may be split into parts, and the survey designed and applied in steps from one part to the next. Design for later stages in the operation may be adjusted on the basis of information collected during application at earlier stages.

3. The population may be partitioned into domains or strata, and the adaptive clusters restricted, by definition, not to cut across domain/strata boundaries. This in no way affects probability nature of the sample, though it may be less efficient than a design allowing for unrestricted growth of adaptive clusters across strata boundaries.

The idea can be extended to define more detailed ‘partition neighbourhoods’ which can be used to limit the potential size of adaptive clusters. For instance, we may divide the study region into ‘blocks’ of some required average size, and not permit the growth of any adaptive cluster to cross the block boundaries. That is, when any unit already in the sample meets the conditions for adaptive additions, units to be added can come only from within the block (the ‘partition neighbourhood’) of that unit. Thus no clusters can exceed the size of a block, and many clusters may be smaller than that.

4. When little prior information is available for the choice of design parameters (such as concerning the criterion for initiating adaptive sampling), a different type of criterion may be used so as to keep a control over the final sample size. For instance, as noted earlier, we may pre-specify the number of the largest values of the criterion encountered, rather than a fixed threshold of it, for the introduction of adaptive sampling at any point in the initial sample. This would fix the number of adaptive clusters in the sample, and control (but not completely fix) the number of units in the final sample.

Adjustments may also be made in the course of survey implementation. For example:

5. The procedure may be adjusted while the survey is in progress, and subsequently post-stratified estimates may be used to take this into account.

6. The initial sample may be selected in a sequential manner (rather than with a pre-specified sample size). The sampling is terminated when the desired final sample size is achieved. With this plan, each time a unit in the initial sample is
enumerated, the associated adaptive additions of surrounding units are carried out. If the cumulative total sample size resulting from these additions is still below the specified limit, the next initial unit is added and associated adaptive additions are carried out. When the cumulative total sample size has reached or surpassed the specified limit, no more initial units are selected.

With such a procedure it is important to ensure that despite the termination, the initial sample remains a probability sample, representative of the entire population. Nevertheless, as noted by Thompson and Seber (1996), “the sequential selection of initial units introduces a bias into the estimates of population total”. Also, “although the overall sampling effort is quite well controlled by this method, the need to sequentially select initial units at random precludes using the most efficient travel path between selected units in the study region”.

(7) In some situations, some departure from strict probability sampling may be unavoidable. Often it is possible to use procedures to reduce the impact of departures from probability sampling on the results. Model assisted approaches – for example maximum likelihood estimates with a good model for the population – may reduce the impact of the departures from probability sampling.

(8) Arbitrary limits may be put to terminate the growth of clusters. The process of adaptive sampling involves a sequence of steps: (i) selection of initial sample units; (ii) addition of units in the neighbourhood of these units; and (iii) inclusion of units in the neighbourhood of the added units, and so on. A limit may be put in the maximum number of steps or ‘waves’ allowed.

The last-mentioned option – putting a limit on the number of ‘waves’ of adaptive additions allowed – can be an important practical tool for keeping a control over the final sample size. In the following section we discuss and provide numerical illustrations of it.

9.8 Variations on the standard adaptive cluster sampling design

9.8.1 Constructing adaptive clusters in waves

As has been noted above, one of the important practical requirements in implementing an adaptive sampling scheme is to have clear criteria for controlling the adaptive development of the sample. In particular, any uncontrolled expansion of the sample size has to be avoided.

One way is to divide the adaptive addition of units to the sample into ‘waves’ (steps), and stop the procedure when some appropriate number of waves has been achieved. This procedure involves at least two additional steps.

(1) Choice of criteria for the division of the adaptive procedure into waves. This is in addition to the basic criterion chosen for the formation of adaptive clusters as discussed above.
Determination of the number of waves for which to carry out the adaptive additions to the sample.

The idea is as follows. The basic criterion chosen for the formation of adaptive clusters determines the potential set of additions to be made to the initial sample points to form adaptive clusters around them. These potential additions are divided into waves on the basis of the additional criteria (1) above. The number of waves involved in completing the construction of a potential adaptive cluster may differ from one cluster to another. For instance, some clusters may require only one wave of adaptive additions to be completed; others may require two, three or four waves, etc., as the case may be. The process may be stopped after a certain number of waves as chosen in (2). This number may be smaller than the number of waves required to complete the construction of some adaptive clusters. They would be included in the final sample as incomplete adaptive clusters. The characteristics and sampling distribution of the resulting sample depends on the choice of criteria (1) and (2).

Formally the above procedure may appear like some form of link-trace sampling, in particular a type of snowball sampling design with its choice of “names” and “waves”. There can be some similarity in the mathematical description of the procedures, in so far as both involve sampling with multiplicity. However, in the conditions of application, objectives and practical details, the two procedures are quite different. (Snowball and related link-tracing sampling procedures will be discussed in Chapters 13 and 14.)

9.8.2 Estimation with overlapping adaptive clusters

The estimation procedure described below (Verma and Betti, 2003) was outlined by the present author in the context of the development of an improved sampling strategy for the conduct of Surveys of Family Health Facilities serving the general population such that the facilities selected are operationally and substantively linked to sample areas of national Demographic and Health Surveys (DHS) of women and children in the population. That is, the facility surveys were concerned with institutions providing family health and related facilities to the population of households, women and children investigated in detail in the demographic survey. Subsequently, versions of the scheme have been elaborated in a sampling manual (MEASURE Evaluation, 2001), and applied in a number of countries, including Guatemala (Wilkinson and Verma, 1998), Tanzania (Tanzania, 2000), and also Rwanda, Egypt and some other countries (unpublished reports).

The sampling procedure may be formalised as follows. We assume that sampling units \(i\) in the initial sample are area-based units selected in one or more stages with probability proportional to some size measure \(A_i\).

The objective is to associate with each selected unit a larger area (a ‘cluster’ of the type of units selected in the initial sample). Specifically, we associate with each unit \(i\) in the population a unique set (cluster) \(I\) of units in the ‘neighbourhood’ of \(i\) (to be defined). The set \(I(i)\) associated with a particular unit \(i\) is referred to below as its ‘associated cluster’. Unlike the normal application of adaptive cluster sampling, the associated clusters are generally overlapping (i.e. they share some units in common), and are not necessarily all distinct (i.e. several clusters may each be composed of
9.8 Variations on the standard adaptive cluster sampling design

exactly the same set of initial units). Nevertheless, there is one-to-one correspondence between units and their associated clusters.

The sample selection method is as follows. The sampling plan consists of:

- step (SA), the selection of a sample of units; and
- step (SB) taking into the sample the clusters associated with each selected unit.

The first step provides Sample A: the initial sample of units. The second step provides Sample B: the sample of associated clusters.

In order to obtain a probability sample of associated clusters, the following conditions must be observed.

**Condition (i)**

The rules and procedures determining the set (cluster) \( I \) to be associated with a unit \( i \) must be objective, in that the outcome is independent of whether or not the particular unit has been selected into the sample. Given that Sample A would normally involve selecting units with probability proportional to some measure of population size \( A_i \), the probability of a cluster appearing in Sample B is given as follows. The probability of selection of cluster \( I \) is the same as that of the unit \( i \) with which it is associated (through which it is brought into the sample), namely \( A_i \). The probability of a unit appearing in Sample B is the sum of probabilities of the clusters which contain that particular unit:

\[
B_i = \sum \left( A_j \left| i \in J \right. \right),
\]

where \( J \) refers to the set of units forming the associated cluster \( j \). The condition for inclusion in the above is that unit \( i \) lies within cluster \( J \).

To be practically useful, we define a cluster associated with a unit so as to meet the following condition:

**Condition (ii)**

\[
i \in J \quad \text{if} \quad j \in I,
\]

that is, if any unit \( i \) is in the cluster of unit \( j \), then unit \( j \) must also be in the cluster of unit \( i \). With this condition, (9.29) becomes:

\[
B_i = \sum \left( A_j \left| j \in I \right. \right),
\]

that is, the probability of selection of any unit \( i \) in Sample B is simply the sum \( (B_i) \) of probabilities \( (A_i) \) in sample A of all the units comprising its cluster.

Condition (ii) is quite general. It is satisfied, for instance, by the following diverse arrangements.
(1) The ‘associated clusters’ may simply refer to some higher level, non-overlapping clusters each composed of a distinct set of units (for instance, districts or localities, with census EAs as area units).

(2) Adjacent units in a geographically arranged list frame may be appropriately grouped to form non-overlapping clusters.

(3) The cluster associated with a unit may consist of the unit itself, plus a certain specified number of units before it and the same number after it in a geographically arranged list frame. The resulting clusters will be overlapping.

(4) The cluster associated with a unit may consist of the unit itself, plus all units which have a common boundary with it.

(5) The above may be extended to more than one ‘tier’: the above set, plus units having a common boundary with any of the units in that set gives a two-tier system. Adding units with a common boundary with any units in the two-tier set gives three-tiers, and so on.

(6) Clearly, Condition (ii) is also satisfied if the associated cluster is an area in the form of a circle of some fixed physical size around the ‘centre’ of the original unit.

(7) We should also mention the ‘null’ option, which may be a reasonable choice in certain practical situations: simply using the original Sample A units as the ‘clusters’ for Sample B.

Each arrangement has its range of useful applications, and also its advantages and difficulties. In practice the choice among them will depend on the survey objectives and circumstances.

Options (1) and (2) define clusters as non-overlapping units, and hence are easiest to comprehend and implement (Verma, 1998). Clusters can be defined from a list frame without recourse to detailed maps. However, a major disadvantage is that the resulting clusters are not centred around the original units, thus weakening the link between the two samples A and B. In option (1) administrative or other grouping of units in the frame already provide the associated clusters of interest. Option (2) offers less geographic control but provides more control in the choice of the size of resulting clusters. Clusters can be made more uniform in size, if that is a desirable feature.

Option (3) is similar; the resulting clusters are more centred around the units with which they are associated, but there is less control over their size.

Options (4)-(6) are more complex. List frames need to be supplemented by maps to identify the associated clusters. Their main attraction is that the clusters are properly centred around the associated units. Options (4) and (5) are rigid in that the size of the resulting clusters cannot be controlled. Option (6) is clearly more complex in implementation – it is necessary to define in operational terms the ‘centre’ of each unit, rules concerning whether to include fully or in part units lying across boundaries of the circle defining its cluster, the boundaries of the circle in the field, etc. However, it has the advantage of flexibility, in that the size of the circle used for defining the cluster of interest can be chosen more freely to meet the sample size and other requirements...
for the survey; the size can also be varied from one sampling domain to another, or gradually across the population.

Indeed, different arrangements may be used in different parts (strata) of the population. For instance, the ‘null’ option (7) may suffice in strata where the Sample A units are already large enough for the purpose of capturing the rare population of interest; option (4) may be used in strata with intermediate sized units; and the multiple tier option (5) when the units are small in size.

Importance of Condition (ii)

It should be noted that, while Condition (ii) is quite general, it is not trivial nor automatically satisfied in all designs. For instance, if clusters of constant population size rather than of constant physical size (as in option (6) above) are created, the condition will not be satisfied in the presence of variations in population density. Similarly in option (3), the condition is violated if the number of units taken into the area before and after the reference unit is not the same. Given the selection method, Equation (9.29) is always valid, but may be intractable in practice. Its transformation into the more practical and implementable Equation (9.30) requires that the procedure satisfies Condition (ii).

Multiple selections

With this scheme, a unit may be selected into the sample more than once. Equations (9.29) and (9.30) in fact do not give selection probabilities but the expected number of selections of a unit, which may exceed 1.0. Taking a distinct ‘selection’ as the unit of analysis, multiple selections of the same unit are handled by the Hansen-Hurwitz estimator or the multiplicity estimator in a straightforward manner, as outlined below.

Estimation

We can take individual units in the final sample B as the analysis units. Equation (9.30) gives expected number of selections of unit \( i \) into Sample B. As explained in Section 9.4, the inverse of the expected number of selections, \( w_i = (1/B_i) \), is the weight applied to every appearance of the unit in the sample:

\[
B_i = \Sigma \left( A_j \right)_{j \in l}, \ w_i = (1/B_i).
\]

Alternatively, we can consider the clusters in sample B as the analysis units. Let \( i \) be a unit selected into the sample. The cluster associated with unit \( i \) is the set \( I \) of units centred on it, \( I \) being also the number of ‘links’ the cluster has with the sampling population. The multiplicity estimator replacing (9.30), is

\[
B_j = I A_i \left( j \in l \right), \ w_j = (1/B_j). \tag{9.31}
\]

This weight is uniform for all units \( j \) which lie in the selected cluster \( I \) (i.e. in the cluster of a selected unit \( i \)). If the cluster has been selected into the sample more than once, this weight is applied to its every appearance.
The practical advantage of (9.31) over (9.30) is that for the former, we need to know the sampling probabilities only for the units selected into the initial sample \( A \). By contrast, (9.30) requires selection probabilities of all units in the clusters of the selected units; such information is not always available in practice.

In the following illustrations, we use the multiplicity estimator (9.31). The illustrations are based in part on Mehran (2012).

### 9.8.3 Illustration 1. Constructing adaptive clusters in waves

Let us consider the same illustrative population as in Section 9.5.1 above. The small population consists of 60 units (area segments). The variable of interest \( (y) \) is the number of child street workers in each area. The basic criterion for forming adaptive clusters is as stated earlier: for each unit with \( (y \geq C_y = 5) \), the potential adaptive cluster includes that unit plus other units with \( (y \geq 5) \) in the neighbourhood of that unit. The result was to give 10 such potential clusters. The remaining units had \( (y < 5) \), and each formed a single-unit cluster by itself.

‘Neighbourhood’ for the above purpose was defined as areas sharing a common physical boundary. The assumed physical arrangement of the 60 area units was shown in Figure 9.2, each area being a rectangle. Clusters formed by units with \( (y \geq 5) \) in the same neighbourhood are non-overlapping, among themselves as well as with single-unit clusters with \( (y < 5) \).

For the purpose of dividing the adaptive procedure into ‘waves’, we define ‘immediate neighbourhood’ of a unit to include up to four areas sharing a boundary with it in the four directions (north, east, south and west), corresponding to four sides of the rectangular area. Hence two area units are ‘immediate neighbours’ if they are geographically adjacent to each other (share a boundary in the north, east, south or west) and each has 5 or more child street workers.

The first wave adds only units in the immediate neighbourhood of the initial sample points with \( (y \geq 5) \), to the extent such units are available (the maximum number added is 4 by definition). The next wave consists of adding units in the immediate neighbourhood of the resulting clusters. All units considered must meet the criterion \( (y \geq 5) \) and should not have been already included in the adaptive sample.

It is easy to see that clusters formed on the basis of the ‘immediate neighbourhood’ criterion meet conditions (i) and (ii) of Section 9.8.2. If unit \( A \) is in the immediate neighbourhood of unit \( B \), then \( B \) is in the immediate neighbourhood of \( A \). However, the resulting clusters are not identical (they overlap) since the immediate neighbourhood of the two units do not include the same set of units, as do the normal adaptive clusters. The resulting clusters are made non-overlapping only by excluding units already included in a previous cluster.

The initial sample in this example is drawn using systematic sampling with interval=10 and random start=1. The resulting six sample area units are: area units 1, 11, 21, 31, 41 and 51. In the first stage of adaptive addition, up to 4 immediate neighbours are added to each of these units having 5 or more child street workers.
Thus, to area unit 1 in the initial sample are added its two immediate neighbours, namely area units (2, 3); to area unit 11 are added its three immediate neighbours, namely area units (8, 10, 14); to area unit 21 are also added its three immediate neighbours, area units (19, 22, 23); similarly, two immediate neighbours are add to area 31, namely units (28, 32); and finally just one unit each (unit number 40 and number 52, respectively) to area units 41 and 51. The resulting 12 additional units are listed in the first-wave column of Table 9.10.

The process continues and in the second wave of adaptive addition, 10 other area units are added to the sample. For example, unit 4 is added to unit 2 from the previous wave, unit 7 is added to unit 8, units (9 and 13) are added to unit 10, and so on.

<table>
<thead>
<tr>
<th>Initial sample</th>
<th>First wave</th>
<th>Second wave</th>
<th>Third wave</th>
<th>Fourth wave</th>
<th>Fifth wave</th>
<th>Sixth wave</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>9</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>19</td>
<td>17</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>28</td>
<td>26</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>52</td>
<td>53</td>
<td>54</td>
<td>49</td>
<td>47</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>55</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the third wave, 7 new units are added; in each of the fourth and fifth waves, 2 other units are added; and in the final sixth wave, just one new unit is added. Thus, the initial 6-unit sample has generated altogether 40 units in the final adaptive sample as listed in the table. The columns cumulated up to a point show the adaptive sample obtained at intermediate wave corresponding to that point. For instance, we have 6 units in the initial sample; the total adaptive sample after wave 1 is 6+12=18 units appearing in the first two columns of the table; the adaptive sample after wave 2 is 18+10=28 units appearing in the first three columns of the table, and so on.

It is interesting to compare how the precision of the resulting sample varies according to how many adaptive waves have been added. Using the multiplicity estimator (or its equivalent, the weight-share method), the estimates of the total number of child street workers in the target population have been calculated at different waves of adaptive sampling, and reported in the first row of Table 9.11.\(^{39}\) Given in the other rows are the corresponding estimates for different random starts of the initial systematic sampling.

\(^{39}\) Mathematically, the estimates may be calculated based on the link matrix developed for indirect sampling (Deville and Lavallée, 2006) reflecting the neighbourhood concept of this example. Matrix multiplication of the basic link matrix gives the relevant link matrix for each wave of the sample.
The mean and standard deviation of the estimates for each wave of sampling are reported in the last rows of the table. The figures in the last column refer to the mean and standard deviation of the completed adaptive cluster samples. It can be noted that the means of the estimates are all equal to 654, the total number of child street workers in the population of this example. The estimates are therefore all unbiased.

The standard deviation of the initial sample is 117. It is lower than the standard deviations of the estimates at the first and second waves. It then decreases to reach a minimum at the fourth wave of sampling (108). It is instructive to note that the minimum does not occur at the end of the process when we get the full adaptive cluster sample (111).

This means that, while adaptive cluster sampling may be a very efficient approach in appropriate circumstances, sometimes applying it partially (fewer than the full number of possible waves) may turn out to be more efficient. Indeed, as noted at the outset of this chapter, there are situations when it is not useful or appropriate to introduce adaptive sampling at all.

<table>
<thead>
<tr>
<th>Random start</th>
<th>Initial sample</th>
<th>First wave</th>
<th>Second wave</th>
<th>Third wave</th>
<th>Fourth wave</th>
<th>Fifth wave</th>
<th>Sixth wave</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>830</td>
<td>713</td>
<td>697</td>
<td>694</td>
<td>677</td>
<td>703</td>
<td>715</td>
</tr>
<tr>
<td>2</td>
<td>690</td>
<td>702</td>
<td>701</td>
<td>685</td>
<td>687</td>
<td>689</td>
<td>685</td>
</tr>
<tr>
<td>3</td>
<td>720</td>
<td>723</td>
<td>723</td>
<td>703</td>
<td>686</td>
<td>686</td>
<td>682</td>
</tr>
<tr>
<td>4</td>
<td>630</td>
<td>709</td>
<td>697</td>
<td>721</td>
<td>703</td>
<td>686</td>
<td>682</td>
</tr>
<tr>
<td>5</td>
<td>560</td>
<td>459</td>
<td>498</td>
<td>542</td>
<td>566</td>
<td>550</td>
<td>545</td>
</tr>
<tr>
<td>6</td>
<td>410</td>
<td>361</td>
<td>356</td>
<td>375</td>
<td>391</td>
<td>391</td>
<td>391</td>
</tr>
<tr>
<td>7</td>
<td>780</td>
<td>664</td>
<td>676</td>
<td>727</td>
<td>730</td>
<td>744</td>
<td>739</td>
</tr>
<tr>
<td>8</td>
<td>670</td>
<td>768</td>
<td>701</td>
<td>670</td>
<td>709</td>
<td>721</td>
<td>739</td>
</tr>
<tr>
<td>9</td>
<td>640</td>
<td>814</td>
<td>825</td>
<td>774</td>
<td>760</td>
<td>744</td>
<td>739</td>
</tr>
<tr>
<td>10</td>
<td>610</td>
<td>628</td>
<td>666</td>
<td>629</td>
<td>613</td>
<td>627</td>
<td>623</td>
</tr>
<tr>
<td>Mean</td>
<td>654</td>
<td>654</td>
<td>654</td>
<td>654</td>
<td>654</td>
<td>654</td>
<td>654</td>
</tr>
<tr>
<td>Std. dev.</td>
<td>117</td>
<td>140</td>
<td>132</td>
<td>117</td>
<td>108</td>
<td>109</td>
<td>111</td>
</tr>
</tbody>
</table>

**9.8.4 Illustration 2. Adaptive clusters formed with immediate neighbours**

In the following, we provide further numerical illustrations of adaptive cluster sampling using data for the same population but drawing samples differently to illustrate the procedure in detail. As to the common population, let us refer back to the small data set presented in Table 3.3. As noted there, we use this data set to illustrate and compare some sampling procedures. The numerical illustration is implemented in an Excel folder prepared by Mehran (2012) for selecting samples and producing estimates for a set of different sample designs. To remind, the illustrative population consists of 80 households spread over 20 geographical areas. Some households have no children and some have more than one. There are altogether 215 children, 20 of whom are working – 5 are unpaid and 15 are paid workers, with total earnings of 45 currency units. The list of households and their main characteristics was given in Table 3.3, with
summary statistics in Tables 3.4 and 3.5. In Section 3.4.4C (Tables 3.6-3.7) results were presented for a simple random sample of size $n = 13$ households, selected from the whole population of 80 households, and in Section 6.4 with the same number selected from a stratified sample confined only to areas which contain one or more working children, on the assumption that empty areas can be identified before sample selection.

The following illustrations show results for three different procedures or samples.

- Adaptive clusters formed in the ‘standard’ way, with geographically neighbouring households meeting the condition for adaptive addition.
- Potentially overlapping adaptive clusters formed with immediate neighbouring households.
- Estimation when adaptive clusters in the sample actually overlap.

A. Illustration 2(A). ‘Standard’ adaptive clusters with geographically neighbouring households

The design is as follows. A simple random sample of households is initially drawn from the sampling population. Households are said to be neighbours if their household numbers are consecutive. Thus for example, household 42 is a neighbour of household 41 as well as a neighbour of household 43. If a sample household is found to have one or more working children, the neighbouring households are also drawn in the sample, and the process continues until there remains no neighbouring household with working children.

Table 9.12 shows the illustrative population of 80 households with data on the number of children, number of working children, and the total wages received by the working children in the household. These data are the same as given earlier in Table 3.3. Table 9.12 also shows the adaptive clusters which would be formed according to the rule specified above, and two initial samples of size $n = 8$. The samples were selected by simple random sampling, with a uniform sampling rate $(8/80)=0.1$, meaning a uniform design weight $(1/0.1)=10$ for each selected sampling unit. If a selected unit has any working children, households with working children in its neighbourhood are added to form an adaptive cluster and are included in the final sample. We find that both the random samples include the large cluster of 7 households with a total of 9 working children – this cluster in fact has a high probability $(=7/10)$ of being included in any sample. The second sample happens to also include a second adaptive cluster of 3 households with a total of 3 working children.
### Table 9.12. The illustrative data set, showing adaptive clusters and two samples (n/N=8/80)

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>Sample 1</th>
<th>Sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>41</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>42</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>43</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>X</td>
<td>44</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
<td>45</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td>46</td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>47</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>X</td>
<td>48</td>
<td>12</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>X</td>
<td>49</td>
<td>13</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td>50</td>
<td>13</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
<td>51</td>
<td>13</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
<td>52</td>
<td>13</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td></td>
<td>53</td>
<td>14</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>54</td>
<td>14</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td></td>
<td>55</td>
<td>14</td>
</tr>
<tr>
<td>16</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>56</td>
<td>14</td>
</tr>
<tr>
<td>17</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>57</td>
<td>15</td>
</tr>
<tr>
<td>18</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>X</td>
<td>58</td>
<td>15</td>
</tr>
<tr>
<td>19</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>59</td>
<td>15</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>60</td>
<td>15</td>
</tr>
<tr>
<td>21</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>10</td>
<td>3</td>
<td>61</td>
<td>16</td>
</tr>
<tr>
<td>22</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>62</td>
<td>16</td>
</tr>
<tr>
<td>23</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>X</td>
<td>63</td>
<td>16</td>
</tr>
<tr>
<td>24</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>64</td>
<td>16</td>
</tr>
<tr>
<td>25</td>
<td>7</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
<td>65</td>
<td>17</td>
</tr>
<tr>
<td>26</td>
<td>7</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
<td>66</td>
<td>17</td>
</tr>
<tr>
<td>27</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
<td>67</td>
<td>17</td>
</tr>
<tr>
<td>28</td>
<td>7</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>X</td>
<td>68</td>
<td>17</td>
</tr>
<tr>
<td>29</td>
<td>8</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>X</td>
<td>69</td>
<td>18</td>
</tr>
<tr>
<td>30</td>
<td>8</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
<td>70</td>
<td>18</td>
</tr>
<tr>
<td>31</td>
<td>8</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
<td>71</td>
<td>18</td>
</tr>
<tr>
<td>32</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>72</td>
<td>18</td>
</tr>
<tr>
<td>33</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>X</td>
<td>73</td>
<td>19</td>
</tr>
<tr>
<td>34</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
<td>74</td>
<td>19</td>
</tr>
<tr>
<td>35</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
<td>75</td>
<td>19</td>
</tr>
<tr>
<td>36</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>76</td>
<td>19</td>
</tr>
<tr>
<td>37</td>
<td>10</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td></td>
<td>77</td>
<td>20</td>
</tr>
<tr>
<td>38</td>
<td>10</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td></td>
<td>78</td>
<td>20</td>
</tr>
<tr>
<td>39</td>
<td>10</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
<td>79</td>
<td>20</td>
</tr>
<tr>
<td>40</td>
<td>10</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td></td>
<td>80</td>
<td>20</td>
</tr>
</tbody>
</table>

**Column headings**

1. Household identification number
2. Area identification number
3. Number of children in household
4. No. of working children in household
5. Total wages of working children

For each sample, there are two columns. An ‘X’ in the first column indicates a unit (household) selected into the initial sample. The second column identifies the adaptive cluster formed around that selection, showing the number of child workers in each household in the cluster. (This is the same information as column (4).)
Tables 9.13 and 9.14 show, respectively, the two adaptive cluster samples in detail, and the estimates obtained from each. Each adaptive cluster is selected through a single household (sampling unit) in the initial sample. The inverse of the selection probability of a household in the initial sample is its weight, called “Weight A” in column (7). The number of households in a household’s cluster determines the multiplicity of that household as recorded in column (9) in Tables 9.13 and 9.14 for the two samples. Dividing “Weight A” in column (7) by the multiplicity in column (8) gives the weight for each household in the final sample, called “Weight B” in column (9). This weight is the same for all households in the same adaptive cluster. For further details, see notes to Table 9.13.

<table>
<thead>
<tr>
<th>Table 9.13. Estimates from adaptive cluster Sample 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SAMPLE 1</strong></td>
</tr>
<tr>
<td><strong>unit</strong></td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>(1)</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>17</td>
</tr>
<tr>
<td>18</td>
</tr>
<tr>
<td>19</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>21</td>
</tr>
<tr>
<td>22</td>
</tr>
<tr>
<td>23</td>
</tr>
<tr>
<td>28</td>
</tr>
<tr>
<td>29</td>
</tr>
<tr>
<td>49</td>
</tr>
<tr>
<td>53</td>
</tr>
<tr>
<td>58</td>
</tr>
<tr>
<td><strong>Total estimate</strong></td>
</tr>
<tr>
<td><strong>True (population) value</strong></td>
</tr>
</tbody>
</table>

Column headings:
(1) household in adaptive clusters selected in Table 9.12 (Sample 1)
(2) identification number of the adaptive cluster
(3) number of children in the household
(4) number of child workers in the household
(5) total wages received by child workers in the household
(6) ‘X’ indicated household selected in the initial sample
(7) sample weight – inverse of selection probability – of the initial sample household
(8) multiplicity – number of households in the adaptive cluster - of the initial sample household
(9) weight of a household in the final sample, = (7)/(8)
(3a), (4a), (5a) estimate of population values of the variables in columns (3)-(5)
### Table 9.14. Estimates from adaptive cluster Sample 2

<table>
<thead>
<tr>
<th>SAMPLE 2</th>
<th>Estimates from sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (3a) (4a) (5a)</td>
</tr>
<tr>
<td>8</td>
<td>2 3 0 0 X 0 10 1 10.0 30.0 0.0 0.0</td>
</tr>
<tr>
<td>9</td>
<td>3 3 0 0 X 0 10 1 10.0 30.0 0.0 0.0</td>
</tr>
<tr>
<td>16</td>
<td>5 4 1 3 1 1.1 4.4 1.1 3.3</td>
</tr>
<tr>
<td>17</td>
<td>5 4 1 2 X 1 10 9 1.1 4.4 1.1 2.2</td>
</tr>
<tr>
<td>18</td>
<td>5 2 1 0 1 1.1 2.2 1.1 0.0</td>
</tr>
<tr>
<td>19</td>
<td>5 4 1 4 1 1.1 4.4 1.1 4.4</td>
</tr>
<tr>
<td>20</td>
<td>5 4 1 2 1 1.1 4.4 1.1 2.2</td>
</tr>
<tr>
<td>21</td>
<td>6 5 3 10 3 1.1 5.6 3.3 11.1</td>
</tr>
<tr>
<td>22</td>
<td>6 3 1 5 1 1.1 3.3 1.1 5.6</td>
</tr>
<tr>
<td>33</td>
<td>9 3 0 0 X 0 10 1 10.0 30.0 0.0 0.0</td>
</tr>
<tr>
<td>41</td>
<td>11 4 1 4 X 1 10 3 3.3 16.7 3.3 6.7</td>
</tr>
<tr>
<td>42</td>
<td>11 5 1 2 1 3.3 3.3 3.3 6.7</td>
</tr>
<tr>
<td>43</td>
<td>11 1 1 0 1 3.3 3.3 3.3 0.0</td>
</tr>
<tr>
<td>62</td>
<td>16 0 0 0 X 0 10 1 10.0 0.0 0.0 0.0</td>
</tr>
<tr>
<td>65</td>
<td>17 3 0 0 X 0 10 1 10.0 30.0 0.0 0.0</td>
</tr>
<tr>
<td>66</td>
<td>17 0 0 0 X 0 10 1 10.0 0.0 0.0 0.0</td>
</tr>
<tr>
<td><strong>Total estimate</strong></td>
<td><strong>77.8 182.2 20.0 48.9</strong></td>
</tr>
<tr>
<td><strong>True (population) value</strong></td>
<td><strong>80.0 215.0 20.0 45.0</strong></td>
</tr>
</tbody>
</table>

Same as Table 9.13, except that it is for Sample 2.

### B. Illustration 2(B). Potentially overlapping adaptive clusters formed with immediate neighbours

Table 9.15 shows the results for an adaptive sample of $n=8$ as the initial sample size, selected with simple random sampling from a population of $N=80$ units. The final sample includes units of the following types.

**Type C.** The potential adaptive cluster is formed from each selected unit which has one or more child workers. One immediate neighbour on either side of this household (as determined by the sequence of household identification numbers) is included in the sample provided that it meets the condition of having at least one child worker. Adaptive additions are made from a type C unit only if the unit comes from the initial sample, but not from any type C unit which was not in the initial sample but has only been included as an adaptive addition to a type C unit in the initial sample.

**Type E.** If a neighbouring unit in the sense defined above does not meet the condition of having at least one child worker, then it is not included in the sample used for estimation. These are the ‘edge units’.

**Type U.** This refers to the initial sample units which did not meet the condition of having at least one child worker. Unlike type E (edge) units, these are included in the sample used for estimation. Unlike type C units in the initial sample, no adaptive additions are made from these units.
The procedure described in this illustration permits adaptive clusters to be overlapping, though this did not occur in the particular sample shown in Table 9.15. Such overlapping does not happen in ‘standard’ adaptive cluster sampling.

The survey is used to estimate a number of variables such as the number of households having children, number of children, number of working children, etc. The contribution of type E (edge) units is not included in the estimate even if they have a non-zero value for the variable being estimated (e.g. the number of children in the household). The contribution of type U units, if any, is always included.

In Table 9.15, column (2b) is the weight of each unit selected into the initial sample, 

\[ w = \frac{N}{n} = \frac{80}{8} = 10. \]

Columns (5) and (6) show values for units in the adaptive clusters formed and the type of unit in each case, following the procedure described above. We return to the original sample units in columns (7) and (8). Column (7) gives the number of units in the cluster formed around each such unit (its multiplicity) – excluding any type E units – and column (8) is the unit weight (2b) divided by this multiplicity, as required by the multiplicity estimator. This weight is then applied to every unit in the adaptive cluster, and the unit’s contribution to the estimate for any variable computed using this weight. Note that type C and U units can make a contribution, but not type E units. Bottom rows of the table compare adaptive sample estimates with the population values.

Population estimates in the second panel of Table 9.15 are obtained from the corresponding sample values multiplied by the final weights of the units concerned.

**Table 9.15. Estimates from an adaptive sample based on ‘immediate neighbourhoods’**

<table>
<thead>
<tr>
<th>Sample values</th>
<th>(1)</th>
<th>(2)</th>
<th>(2a)</th>
<th>(2b)</th>
<th>(3a)</th>
<th>(3)</th>
<th>(4a)</th>
<th>(4)</th>
<th>(5a)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
<th>(9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.10</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>U</td>
<td>1</td>
<td>10.0</td>
<td>10.0</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>0.10</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>U</td>
<td>1</td>
<td>10.0</td>
<td>10.0</td>
</tr>
<tr>
<td>26</td>
<td>7</td>
<td>0.10</td>
<td>10</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>U</td>
<td>1</td>
<td>10.0</td>
<td>10.0</td>
</tr>
<tr>
<td>38</td>
<td>10</td>
<td>0.10</td>
<td>10</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>U</td>
<td>1</td>
<td>10.0</td>
<td>10.0</td>
</tr>
<tr>
<td>41</td>
<td>11</td>
<td></td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>11</td>
<td>0.10</td>
<td>10</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>C</td>
<td>3</td>
<td>3.3</td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>11</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>12</td>
<td></td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>12</td>
<td>0.10</td>
<td>10</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>C</td>
<td>3</td>
<td>3.3</td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>12</td>
<td></td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>14</td>
<td></td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>14</td>
<td>0.10</td>
<td>10</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>C</td>
<td>1</td>
<td>10.0</td>
<td>10.0</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>14</td>
<td></td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>20</td>
<td>0.10</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>U</td>
<td>1</td>
<td>10.0</td>
<td>10.0</td>
<td></td>
</tr>
</tbody>
</table>
### Table 9.15 (cont.)

<table>
<thead>
<tr>
<th>Population estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>26</td>
</tr>
<tr>
<td>38</td>
</tr>
<tr>
<td>41</td>
</tr>
<tr>
<td>42</td>
</tr>
<tr>
<td>43</td>
</tr>
<tr>
<td>46</td>
</tr>
<tr>
<td>47</td>
</tr>
<tr>
<td>48</td>
</tr>
<tr>
<td>53</td>
</tr>
<tr>
<td>54</td>
</tr>
<tr>
<td>55</td>
</tr>
<tr>
<td>79</td>
</tr>
<tr>
<td>Total estimate</td>
</tr>
<tr>
<td>True (population) value</td>
</tr>
</tbody>
</table>

**Column headings**

1. Household identification number
2. Area identification number
2a. Selection probability
2b. Sample weight
3a. Whether has children
3. Number of children in household
4a. Whether has working children
4. No. of working children in household
5a. Has a paid child
5. Total wages of working children
6. Type of unit
   - C: initial sample unit, meeting condition for adaptive addition
   - U: initial sample unit, not meeting condition for adaptive addition
   - E: edge unit (not in the initial sample)
7. Adaptive cluster size (multiplicity)
8. Weight of selected unit
9. Weights of units in the adaptive cluster
### Illustration 3. Estimation when adaptive clusters in the sample overlap

In Illustration 2(B) the adaptive clusters formed with immediate neighbours were potentially overlapping, but this did not occur in the sample actually selected for that illustration.

In Table 9.16, we show a part of a hypothetical sample selected from the same population as in the previous illustration using the same procedure (adaptive clusters formed from only immediate neighbours). The column headings (1)-(9) have the same meaning as in Table 9.15.

<table>
<thead>
<tr>
<th>Sample values</th>
<th>Adaptive sample with overlapping clusters (initial sample includes units 40, 41, 42, 43)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) (2) (2a)</td>
<td>(2b) (3a) (3) (4a) (4) (5a) (5) (6) (7) (8) (9)</td>
</tr>
<tr>
<td>40 10</td>
<td>0.10 10 1 4 0 0 0 0 U 1 10.0 10.0</td>
</tr>
<tr>
<td>40 10</td>
<td>1 4 0 0 0 0 0 E</td>
</tr>
<tr>
<td>41 11</td>
<td>0.10 10 1 4 1 1 1 4 C 2 5.0 5.0</td>
</tr>
<tr>
<td>42 11</td>
<td>1 5 1 1 1 2 C</td>
</tr>
<tr>
<td>41 11</td>
<td>1 4 1 1 1 4 C</td>
</tr>
<tr>
<td>42 11</td>
<td>0.10 10 1 5 1 1 1 2 C 3 3.3 3.3</td>
</tr>
<tr>
<td>43 11</td>
<td>1 1 1 1 0 0 C</td>
</tr>
<tr>
<td>42 11</td>
<td>1 5 1 1 1 2 C</td>
</tr>
<tr>
<td>43 11</td>
<td>0.10 10 1 1 1 1 0 0 C 2 5.0 5.0</td>
</tr>
<tr>
<td>44 11</td>
<td>1 3 0 0 0 0 E</td>
</tr>
</tbody>
</table>

The column headings are the same as those in Table 9.15.

In this illustration we assume that household numbers (40, 41, 42 and 43) have been selected in the initial simple random sample of size 8 out of 80 households. Household number 40 is of type U (contains no child workers), and therefore forms an adaptive cluster by itself. The other households are of type C (i.e. contain child workers), and in each case any immediate neighbours if also of type C are added to form an adaptive cluster. If an immediate neighbour to a selected household is of type E (a unit neither directly selected in the initial sample nor contains a child worker), it is an edge unit and is ignored in the estimation from the sample. The resulting adaptive clusters overlap: household 40 appears in one cluster, but 41 and 43 appear in two clusters, and household 42 appears in three clusters. While the clusters overlap, the procedure for constructing them meets Conditions (i) and (ii) of Section 9.8.2, and therefore the following multiplicity-based estimation procedure can be used. It will be appreciated that despite the complexity of the design, the estimation procedure is fairly straightforward.

Each non-zero figure in column (9) represents one selection of a household in the final sample – we have 8 ‘household selections’ from an initial sample of 4 households. The
households are our sampling units, while the ‘household selections’ form the analysis units from which population estimates are formed using Equation (3.12) of Chapter 3, with unit weights as given in column (9). These weights have been computed using the multiplicity estimator (4.3) of Section 4.6.

To remind, the estimation procedure is as follows. Each appearance \((k)\) of unit \((j)\) in the sample (which forms one analysis unit) is weighted inversely proportional to its probability of selection \((f_{jk})\) in the particular source and sampling unit from which it has come into the sample, taking into account the number of possible sources \((m_j)\) from which that unit could have been selected. The multiplicity probabilities and weights are:

\[
\text{probability } p_{jk} = \frac{m_j}{f_{jk}};
\]

\[
\text{weight } w_{jk} = \frac{1}{p_{jk}} = \frac{1}{m_j} \left( \frac{1}{f_{jk}} \right), \text{ if the unit is selected, and } w_{jk} = 0 \text{ if not selected.}
\]

Subscript \(j\) refers to a selection (analysis unit) and \(k\) to the (selected) sampling unit with which the selection is linked. Each selection is, by definition, linked uniquely to only one sampling unit. Selections in an adaptive cluster are linked to the same sampling unit, the selection probability \(p_{jk}\) and weight \(w_{jk} = (1/p_{jk})\) of that unit are shown in columns (2a) and (2b), respectively. Multiplicity \(m_j\) of a selection in column (7) is the number of selections in its cluster (i.e. type C or type U units, excluding any type E units); it has the same value for all selections in the cluster. The weight in column (8) is the ratio of weights in column (2a) and multiplicity in column (7). This weight applies to every selection in a cluster.

### 9.9 Some guidelines and recommendations

#### 9.9.1 Concerning the approach

While the introduction of adaptive sampling brings in some additional considerations and complexities to survey design and estimation, and also in survey planning and implementation, the technique potentially also has important advantages in terms of efficiency and the additional information provided.

The technique has been widely used in environmental surveys, such as for estimating abundance of animals and plants. Its use has been much more limited in social surveys.

*In our view, there are many situations – including those concerning child labour – where the population of interest is quite rare and unevenly clustered, and the introduction of adaptive sampling will have important advantages.*

The advantages go beyond possible improvements in efficiency for estimating population means and totals. The technique yields more information on the pattern of distribution and characteristics of extreme values, and also provides larger samples of ‘interesting cases’ for more in-depth studies.
9.9.2 Concerning design

Beyond the conventional sample with which the process begins, adaptive sampling involves the choice of parameters defining criteria and procedures for the construction of adaptive clusters.

Choice of the parameters depends on knowledge about the size and distribution of the population, and about how the choice affects size and efficiency of the final sample. Normally the information available for these choices is limited. Caution is recommended in the choice: it is more important to avoid extreme situations (such as the final sample size being allowed to become too large or uncontrolled), than to maximise efficiency. Gradually the design efficiency can be improved with the cumulation of improved information for the purpose. It is likely to be a step-by-step process.

9.9.3 Concerning implementation

Adaptive sampling involves taking technical and operational decisions as well as analysis of the data collected on a continuing basis, during the course of data collection itself. Thus it requires higher levels of technical and practical skills in the field than required in a comparable survey with conventional sampling design.

In practice, this often means that too much decentralisation of the survey operations must be avoided. This is particularly important when the survey organisation lacks experience in the application of the adaptive sampling technique.

9.9.4 Concerning analysis

Adaptive sampling has the potential to yield two quite different types of information:

1. Estimates pertaining to the whole population such as population totals and means. This also includes estimates for different partitions of the population such as geographical domains, demographic class and other subpopulations.

2. Profiles pertaining to groups of individual units defined in terms of ‘exceptional’ values of the study variable(s). Examples are characteristics of areas with high (abnormally above average) concentration of labouring children, or the extent and pattern of concentration of such areas in the population. More extensive listing of individual units in the target population from which samples for more detailed studies can be obtained is also an important output. Information on profiles of population groups represents an important comparative advantage of adaptive sampling. Maximum use should be made of such information during data analysis when adaptive sampling has been introduced.

9.9.5 Concerning evaluation

Evaluation of adaptive sampling in comparison with conventional sampling should cover all the three main aspects of survey practice: cost, data accuracy, and value (usefulness) of the data generated.
(1) Cost should cover monetary cost, as well as cost in non-monetary terms such as ‘cost’ of the level of skill required, of the risks involved in the introduction of newer and more complex methods, and of the uncertainty introduced in the size of the operation.

(2) Data accuracy should cover estimates of variance (including the breakdown of total variance into between- and within-cluster variance), as well as bias – in particular bias which arises from errors or shortcomings of implementation.

(3) Analysis of the value of the data should cover information obtained on the profile of exceptional values and the pattern of their concentration. It should also include an assessment of the quality and usefulness of lists of members of the population of interest obtained as a by-product of the expansion of the initial sample with adaptive sampling. Such lists are potentially useful for drawing samples for more detailed studies.
IV. Mobile populations
10.1 Issues and framework

A. Mobility and mobile populations

This chapter addresses special problems and issues which arise in sampling mobile populations. It is useful to begin by clarifying the concepts of mobility and mobile population from the perspective of sampling procedures.

We may broadly distinguish between two sampling situations. (i) Frequently we are dealing with populations the elements of which are associated with fixed points, such that it is possible and preferable to use those fixed points as sampling units in order to obtain samples of the population elements. This is the ‘normal’ sampling situation. (ii) In contrast to this are situations when it is necessary or preferable to sample and enumerate units (say individual persons) through their mobility (movement). From the perspective of the sampling procedure, we refer to such populations as mobile populations.

The concept of ‘mobile population’ is more general than simply not having a fixed place of residence or work. There are at least three types of situation concerning mobile populations which require alternative sampling strategies of the type discussed in this chapter.

(1) The first is the situation when the very definition of the target population is based on some aspect of its movement. This applies, for instance, to surveys of passengers, migrants crossing border points, tourists, shoppers, and visitors to facilities of many other types (clinics, job centres, fairs and festivals, museums, libraries, and so on).

(2) A population is considered mobile because individuals in it are not associated with fixed points (locations, addresses, listings, etc.) through which a probability sample of them could be obtained. This may apply, for instance, to working children who neither have a home or other fixed living space, nor a fixed or regular place of work – locations which are known and could be used to obtain a sample of the children. Many children engaged in hazardous forms of labour are homeless. Also, many labouring children do not have a fixed place of work. Patterns of movement are very diverse depending on the situation and nature of work. Special but also very specific procedures are required to catch reasonably representative samples of them. Under-coverage and double-counting can both be problems, one not precluding the other. Another dimension of complexity is that many children move in and out of child labour depending on the season or other factors.

(3) Then there are situations when it is preferable to sample individuals through their movement, even when it is possible to sample them through their association with
fixed points. By ‘preferable’ we mean being more efficient, less costly, higher in quality, or more practical and manageable. It is true that, for instance, children living in households through which they can be contacted need not be considered ‘mobile’ for the purpose of sampling, even if they work on the street with no fixed place of work; similarly, children working in establishments at fixed, known or otherwise contactable locations may not be considered ‘mobile’ as concerns sampling procedures, even if they have no fixed place of residence. However, children moving to and from and between work places are ‘mobile’, even if they have fixed places of residence but which are too dispersed to be adequately captured through a normal household survey. The same can be the case with migrants who are too dispersed to be contacted through their households. The real issue is the efficiency (and other desirable features) of the sample design. Is it more efficient to sample the children from their fixed points of residence or fixed points of work, or is it more efficient to sample them through their movements? We say that we are sampling a mobile population if we have decided to sample the population through their movements, using special procedures as necessary. If we sample the same population through their fixed points of residence or through their fixed points of work, we would not need to deal with this as a mobile population requiring special sampling procedures.

B. Sampling framework

Special procedures are required to establish links between mobile units and fixed reference points through which a probability sample of the units could be obtained. For instance, mobile individuals may be contacted at places they visit, or ‘intercepted’ at points through which they pass, or they could be followed by interviewers who are themselves mobile. Sampling in space as well as in time is required to survey mobile populations.

It is useful to have a framework to organise the variety of circumstances, problems and solutions encountered in sampling mobile populations. Four important concepts in the framework are: the location, observation point, time segment, and ‘space-time primary sampling unit’ - the PSU.

A ‘location’ may refer to an area, a market, an air or land or sea passenger terminal, or various types of institution such as those providing commercial, entertainment, health, educational or employment-related services or advice. For child labour surveys, locations may refer to places where children’s work is concentrated, such as garbage dumps or bus and train terminals, and other places which are frequented by working children.

The concept of ‘observation point’ and ‘time segment’ are particularly relevant for surveys where the objective is to collect information on individuals visiting or passing through chosen locations. The concepts are also useful in the wider context of sampling mobile populations.

Typically a location will have several ‘observation points’ at which the flows under study can be most effectively observed. When there are many small locations, a location may itself constitute an observation point. The locations and observation points where the survey is conducted provide the necessary geographical units at which mobile
individuals in the target population can be sampled, screened as necessary, and observed or interviewed.

The flows to be studied are likely to vary over time, both in the numbers and in characteristics of the persons passing through the observation points. It is usually necessary to divide the period covered by the survey into time segments so as to properly cover variations in the flow. The choice of what sort of time segments to use depends on the nature of the survey and characteristics of the flows being studied. The segments may refer to periods during the day, such as short 1-2 hour periods, or longer periods such as the whole morning, the whole afternoon, or the whole day; or they may refer to particular times when large numbers of persons are entering or leaving the location; or to particular days of the week, possibly distinguishing week days from weekend days; or they may cover repetitions extending over several weeks.

A ‘time-location’ sampling framework is suitable for surveys of flows. In this framework, the cross of observation points by time segments form the primary sampling units (PSUs). A key aspect determining the sample design and implementation is the size and nature of the flow. The size of a PSU means the number of individuals flowing through the observation point and the duration of the time segment defining the PSU.

The population of interest can be mobile and difficult to locate and identify to varying degrees. Furthermore, it is often elusive in other dimensions as well: it may lack an adequate sampling frame; it may be rare, widely scattered and/or concentrated in a few unknown locations; it may also be reclusive – hidden, inaccessible, unwilling to participate in surveys.

The difficulties of enumerating such populations concern the identification of (i) who are the eligible respondents for the survey, and (ii) where and (iii) when to find them; also (iv) what information to ask them and (v) how to ask for the information. Another set of issues concerns (vi) how to use sample data to produce valid estimates for the population, and (vii) how to assess variances and biases to which those estimates are subject.

These problems are of course common to all sample surveys, but they tend to be more acute in the type of situation and survey involving mobile populations. In the following we will use several examples to define and describe sampling procedures for mobile populations. The examples differ in the nature and complexity of the problems involved. Situations can be compared and contrasted in terms of several aspects relating to the identification of ‘who’, ‘where’, ‘when’, ‘what’ and ‘how’ aspects of the methodology:

- definition of locations and whether and how to sample them;
- identification of observation points within each location in the survey;
- choice of time segments during which the information would be collected;
- identification and selection of PSUs (observation points by time segments);
- counting, screening and selecting individuals for the survey within each sample PSU; and
- special issues in estimating from the sample survey.
10.2 Location sampling

10.2.1 Introduction

Many mobile populations have a tendency to congregate in particular locations (such as places, territorial sites, centres for commercial, cultural or other activity, institutions providing facilities or services). Such locations of congregation may be used to obtain samples of the population when better alternative frames are not available. The sampling scheme used for such a purpose has been variously named, such as ‘centre sampling’ (Blangiardo, 1996), or ‘intercept point sampling’ (McKenzie and Mistiaen, 2007). The sampling scheme normally does not involve explicit time reference, but refers only to locations where individuals of the target population gather on a regular basis. Blangiardo used the term ‘centres’ to refer to either a partial list or a territorial place where units of interest congregate. The scheme has been used, for example, to study migrants in Italy. Many migrant groups use such centres in order to satisfy diverse needs concerning, for instance, religion, health, social interaction, recreation, shopping, and eating. McKenzie and Mistiaen also used a similar scheme to survey migrant households in Brazil, and compared the results with a conventional census-based household survey and snowball sampling. We will discuss this example in Section 10.4.3.

We will refer to the basic sampling scheme more simply as location sampling, as distinct from time-location sampling discussed in the next section.

Typical features of the location sampling scheme may be summarised as follows. In normal location sampling the study area consists of a small number of locations. It is assumed that all the locations have been identified individually. Often all of them are included in the survey without sampling. Individuals in the population are members of one or more of these locations. The method also assumes that at a certain appropriate time or times for each location, all its members can be found or located through the location concerned. Beyond that, there is no time dimension to the sampling. Normally individuals are sampled at each location. A screening operation may also be involved in order to identify location members from among all persons who may be present at the location. Sometimes variable sampling rates may be applied to different categories of persons, but normally individuals are sampled at a uniform rate. However, often the size of the location (total number of members) is unknown, and so also is the actual sampling rate applied; only the size of the sample selected is known. Hence the variation of the sampling rates across the locations is not known and cannot be controlled. The most important objective is to estimate the sampling rate at least in relative terms across the locations so that the data can be properly weighted for pooling over different locations. It is also useful to be able to estimate the sampling rates in absolute terms, and hence the location sizes so that totals (and not only proportions, means, ratios etc.) can be estimated. The main complexity arises from overlap or multiplicity: individuals are members of more than one location.

Various sampling designs are possible. The most important one is taking all locations into the sample, and taking a sample of persons within each. Other possibilities are: taking a sample of locations, and all persons within each; or sampling both of locations, and then of individuals within selected locations (Pratesi and Rocco 2005).
10.2 Location sampling

10.2.2 Strong assumptions

The location sampling scheme is based on certain quite strong assumptions about the pattern of movement of the mobile population being studied. These assumptions concern both the nature of locations of congregation and the sampling scheme applied to them. These are discussed in this subsection. The main technical scheme discussed subsequently concerns the common single-stage design where all locations in the study population are included and at each a sample of individuals regularly visiting the location is taken.

It is important to keep the strong assumptions behind the location sampling scheme in mind when considering application of the procedure to real life situations.

A. Locations

(1) All individual units of interest are ‘associated’ with a (very) limited number of locations (‘centres’). At each location, all units can be found on one occasion, or with a limited number of revisits. Hence, apart from choosing the right occasion(s), no sampling of time segments is involved.

The above is not meant to imply that individuals in the target population are always present at the location. It means that the relationship of the individual to the facility at the location is such that by visiting the facility at any time (when it is open), the individual can be found or an ‘appointment’ can be made to contact the person concerned. Thus the chance of contacting an individual is not determined by how frequently or at what times the person concerned may be visiting the facility.

(2) Each unit (say a person) has a fixed pattern of association with locations. The unit may be associated with any set of one or more specific locations, including all the locations. Units not associated with any location are not covered by the procedure; they remain outside the study population by definition.

(3) Each unit (j) in the survey is able to report on its pattern of association, i.e. for each location i, on whether it is associated with it \( r_{ij} = 1 \) or not \( r_{ij} = 0 \).

(4) For each individual unit \( j \) interviewed at location \( i \), a set of survey variables, generically denoted by \( y_{ij} \), is measured.

B. Screening and sampling

(5) The procedure has been developed for the situation when it is not possible or desirable to record and use individuals’ identities. The procedure does not depend on linkage of individual records across locations. Such anonymity may be necessary for ethical or practical reasons. When all units at a location can be found on one occasion, it is not necessary to have individual identifiers. However, individual identifiers (and screening as noted below) may be necessary when revisits are required.

(6) Screening may not be necessary if the location at the time of the survey contains only units associated with the location. If more than one visit has to be made by the interviewers to the location in order to contact respondents, then screening
10. Sampling mobile populations

may be required to identify who has or has not been already contacted. Generally screening would also be required to eliminate any non-eligible units if present at the location at the time of the survey.

(7) At each location, the total of associated eligible units may or may not be known, but it is assumed that a sample of units with known relative probabilities can be selected at each location: \( f_{ij}^* = k_i f_{ij} \), where \( f_{ij} \) are known relative probabilities of selection of units \( i \) at location \( j \). Parameters \( k_i \) vary by location and are unknown when the size of the population at the location is not known. With equal probability sampling within each location (which is a common design in practice), we can take \( f_{ij} = \text{constant} = 1 \) without loss of generality, so that the actual selection probabilities become \( f_{ij}^* = k_i \).

10.2.3 Single-stage sampling of persons at locations

Alternative designs for sampling locations and persons are possible, for example:

(i) all locations in the target population are included in the survey (i.e. no sampling of locations is involved), and at each location individual units are selected independently between locations;

(ii) selection of a sample of locations, but with all units at each selected location taken into the survey (single-stage compact cluster sampling); or

(iii) a two-stage design, involving selection of a sample of locations, and then samples of individual units at each selected location.

We begin in this section with the ‘basic’ and most commonly used design (i); variations such as (ii) and (iii) on the basic design have limited application in practice, and we will comment on those briefly in Section 10.2.5.

A. Data matrix

With the assumptions of the method noted above, the survey data may be represented by a matrix in the form we will present below. This form of presentation is convenient for explaining the method and the estimation procedure, and also the notation used.

As above, we identify a location with subscript \( i \) (1 to \( I \)) and an individual associated with it with subscript \( j \) (1 to \( N \)). The \( N \) units in the population can be grouped according to the ‘pattern’ of their association with the set of \( I \) locations. The pattern of a group of units specifies, for each of the locations, whether or not the units concerned are associated with (are members of) the location. For example, a pattern identified as “10110....” specifies that units with that pattern are associated with locations 1, 3 and 4, but are not associated with locations 2, 5, ... Subscript \( k \) will be used, where necessary, to indicate a pattern. There are \( K = (2^I - 1) \) possible patterns – with two (0,1) possible values at each of the \( I \) locations, with the case with all zeros (persons associated with no location) being out of the population covered with this methodology. Many of the patterns, especially with large \( I \), may have only a few units, or even none at all. Nevertheless, we must note that \( K \) increases rapidly as the number of locations \( I \) increases. For example: for \( I = 6 \), \( K = 63 \); for \( I = 13 \), \( K = 8,191 \). Hence the sampling scheme is suitable when there is only a small number of locations.
Large values of $I$ would also result in a large range of the weights to be given to units in the estimation formulae, which would tend to inflate sampling error.

There is also a practical reason for keeping the number of locations included in the study population small. The number of categories in the pattern of association with locations, which each respondent is required to remember and report, must not be allowed to become large. Note that this requirement applies to the number of locations in the study population; having a smaller number in the sample does not change the situation.

For each unit in the population, the following information is assumed to be 'obtainable' from the respondent or other sources.

- $r_{ij}$: a $(0,1)$ index indicating whether $(=1)$ or not $(=0)$ unit $j$ is associated with location $i$.
- $k$: The pattern to which the unit belongs. It summarises the vector of values $r_{ij}$: each $k$ represents a particular combination of $r_{ij}$ values, $i = 1$ to $I$. For instance, in the example given above, pattern “01101...” corresponds to some particular value of $k$ identifying that pattern. Similarly, pattern “01011...”, if present, would be identified by some other value of $k$.
- $f_{ij}$: The selection probability of unit $j$ at location $i$. As noted earlier, this may only be a relative value, within some unknown constant $k_i$ specific to the location: $f_{ij}' = k_i f_{ij}$.
- $y_j$: Refers to the value of any variable for unit $j$. It is assumed that for a given variable, the value is a characteristic of the individual unit concerned, and does not depend on the location where the variable is measured. If a unit is in the sample at more than one location, at any subsequent appearance after enumeration at the first location, the $y_j$ value for the unit is not measured again but assumed to be the same as that at the first measurement.
- $Y_{j(k)}$: In $Y_{j(k)}$, the subscript indicates in addition that unit $i$ belongs to a particular pattern $k$; numerically this is the same quantity as $y_j$, except for the identification of the pattern to which unit $j$ belongs. For some purposes, it is useful to aggregate values over units belonging to the same pattern $k$. Subscript $j(k)$ tells us that unit $j$ belongs to pattern $k$; another unit $j'$ belonging to the same pattern would be identifies as $j'(k)$ when it is necessary.

In the above, we have characterised the variables as obtainable rather than as ‘available’ a priori. This is to emphasise that though it is possible in principle to obtain the values of these variables for all units in the population, it is usually not necessary to do so. Normally a sample of units will be selected and the above variables measured only on units in the sample – even though the following presentation is in terms of the whole population.
Table 10.1. Data matrix for a location survey

<table>
<thead>
<tr>
<th>Individual unit</th>
<th>Location i</th>
<th>Pattern Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1 (r_ij values)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>N</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

Cells of the table show, for each unit \( j \) in the population of \( N \) units, whether \((=1)\) or not \((=0)\) the unit is associated with a location. This defines a pattern \( k \) for the unit. Subscript \( j(k) \) indicates that unit \( j \) has the pattern \( k \). A whole set of units may have the same pattern. \( y_j(k) \) is the value of some variable of interest for unit \( j \) with pattern \( k \).

The potential data matrix for the population is of the form of Table 10.1.

At each location \( i \) (column of the matrix in Table 10.1), a sample is taken. As noted above, the common situation is that the relative value of the sampling rate applied \((f_{ij})\) to unit \( j \) at location \( i \) is known but, in the absence of information on size of the population, the actual values of the sampling rates, \( f_{ij} = k_i f_{ij} \), are not known. The scale of \( f_{ij} \) is arbitrary and may be chosen for it to average 1.0 at each location. Often \( f_{ij} \) may vary only by location, being a constant for all units at a location. In this special case, we may take \( f_{ij} = 1 \) at each location, giving \( f_{ij} = k_i \).

Next, we can consolidate the sample data by pattern \((k)\). Each row of the consolidated table is obtained by summing over all individuals with the same pattern of membership of locations, as shown in Table 10.2. The consolidation may be done over the whole population, or for units in the sample. It is the latter type of consolidation which is shown in Table 10.2. In reality, the population is unknown, and it is only the distribution in the sample as shown in Table 10.2 which is available.

Quantity \( n_{ki} \) in the table is the number of sample units with pattern \( k \) at location \( i \). Quantity \( r_k \) is the number of ‘1s’ in pattern \( k \), that is, it is the number of locations of which individuals with pattern \( k \) are members, being the sum of \( r_{ij} \) values in the full data matrix for any one of the individuals with the pattern concerned. In the last row for sums, \( n_i = \sum k n_{ki} \) is the total sample size for location \( i \); \( N_i \) is the corresponding population size, which is generally unknown as noted above, though may be available in exceptional cases.
### Table 10.2. Data consolidation by the pattern of location membership

<table>
<thead>
<tr>
<th>Pattern ( k )</th>
<th>Location ( i )</th>
<th>Number of memberships</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>4</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( k )</td>
<td>.</td>
<td>( n_i )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( K )</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

Total (in sample) \( \ldots \) \( n_i \) \( \ldots \) \( r_k \)

Number in population \( N_i \)

Cells of the table show, for each pattern and each location, the number of units with a given pattern \( k \) which are associated with each location \( i \). \( n_{ki} = \sum_{j(k)} r_{ij(k)} \); the total number of units associated with location \( i \) (with whatever pattern) in the sample; \( N_i \) is the corresponding population value.

Number of memberships: with \( r_{ki} \) indicating whether \( (=1) \) or not \( (=0) \) pattern \( k \) includes, by definition, association with a particular location \( i \), we have \( r_k = \sum_i r_{ki} \); the total number of locations associated with membership pattern \( k \).

#### B. Estimation procedure for individual locations

At a given location \( i \), the selection probability of a unit in the sample is \( f_{ij} = k_i \, f_{ij} \). The inverse of these unit selection probabilities are used as weights for estimation for the particular location. The sum of weights of the \( n_i \) units in the sample estimates the population size at the location:

\[
\hat{N}_i = \sum_{j=1}^{n_i} \left( 1/f_{ij} \right) = \frac{1}{k_i} \sum_{j=1}^{n_i} (1/f_{ij}).
\]

(10.1)

Since the \( f_{ij} \) are relative quantities, they can be scaled arbitrarily within each location. Henceforth we will take the scaling to be such that

\[
\sum_{j=1}^{n_i} (1/f_{ij}) = n_i.
\]

(10.2)

It implies that values \( f_{ij} \) are scaled such that average \((1/f_{ij}) = 1\), separately for each location. If, for instance, we had an equal probability sample from a location, \( f_{ij} = \) constant = 1, this scaling gives \( k_i = (n_i/N_i) \).

We use the following notation:
\( N_{io} \) number of members of location \( i \) in the population who also belong to location \( o \) in the population. The first index \((i)\) is the reference location, the second \((o)\) stands for any other location. By symmetry, \( N_{io} = N_{oi} \) and the table is symmetrical around the diagonal running from top right to bottom left. This relationship applies only to the population; it does not necessarily apply to samples drawn from that population as explained below.

\( n_{io} \) number of members in the sample of location \( i \) which are members of the population at each of the other locations \( o \).

\( n_{oi} \) number of members in the population of location \( i \) which are found in the sample at each of the other locations \( o \).

Note that unlike the case \( N_{io} = N_{oi} \) in the population, generally \( n_{io} \neq n_{oi} \) in the sample. These numbers will be equal only if the same sampling rate has been applied to all the locations. In practice this cannot be done in so far as the location population sizes are unknown. It is only the sample sizes obtained which are known.

From Equation (10.1) applied only to units associated with location \( o \), the population size \( N_{io} \) is estimated as:

\[
\hat{N}_{io} = \sum_{j=1}^{n} \left( \frac{1}{f_{ij}} \right) = \frac{1}{k_i} \sum_{j=1}^{n} \left( \frac{1}{f_{ij}} \right) = \frac{n_{io}}{k_i}.
\]  
(10.3)

The last term in the above is based on the added assumption that the scaling for the whole sample at a location \( \sum_{j=1}^{n} \left( \frac{1}{f_{ij}} \right) = n_i \) holds approximately for part \( n_{io} \) of the sample:

\[
\sum_{j=1}^{n} \left( \frac{1}{f_{ij}} \right) = n_o.
\]  
(10.4)

Similarly \( \hat{N}_{oi} = (n_{oi} / k_o) \). Given that \( N_{io} = N_{oi} \), we get

\[
\frac{k_o}{k_i} = \frac{n_{oi}}{n_{io}}.
\]  
(10.5)

This is a simple and intuitively appealing result: for example, doubling the sampling rate at a location doubles the expected number of members of another location seen in the sample at this location, but the number of members of this location seen at the other location is not affected.

Averaging over all locations \( o = 1 \) to \( I \) gives

\[
k_i = \bar{k} / av_i (n_{oi} / n_{io}), \text{ where}
\]  
(10.6)

\[
av_i (n_{oi} / n_{io}) = \frac{1}{I} \sum_{o=1}^{I} (n_{oi} / n_{io}).
\]

(Note. In taking the average, cases with \( n_{oi} \) or \( n_{io} = 0 \) should be disregarded, also in counting \( I \).)

With these relationships, the set of \( I \) unknowns \( (k_i, i = 1 \) to \( I \) \) is reduced to a single unknown \( \bar{k} \). There is a major difference between the two situations. With unknown \( k_i \) values, the variation of the sampling rates across the locations is not known and cannot be controlled. In the absence of information on relative values of the sampling rate across the locations, data cannot be properly weighted for pooling over different
locations. This can be done once the $k_i$ values are connected through a constant parameter $k$, albeit still unknown. The next step is to estimate $k$, and hence the sampling rates and the location sizes in absolute terms, so that totals (and not only ratios) can be estimated from the sample.

The size of unknown $k$ can be determined even if we know just a single $k_i$ value. If more than one $k_i$ values are known, they can be averaged to obtain a more stable estimate for $k$.

For any location, $k_i$ value can be estimated if an estimate of $N_i$ is available, for example from
- lists maintained by the location;
- lists compiled from other administrative sources;
- rough estimates by persons knowledgeable about the location;
- or from quick counts at the location.

C. Combining samples over locations

Now consider the combined sample from all the locations. Different forms of estimation are possible, as described in Section 4.6 above.

(1) Horvitz-Thompson (H-T) type of estimator

This procedure calculates the overall selection probability of the unit taking into account its possible selection from any of the locations. Following Equation (4.1), unit selection probability is computed in the form:

$$\pi_j = 1 - \prod_{i \in I_j} \left(1 - k_i \cdot f_{ij}^* \right), \text{ weight } w_j = 1/\pi_j,$$

where $I_j$ is the set of locations among $I$ to which unit $j$ belongs.\(^{40}\)

The inverse of the selection probability so computed is used as the weight in estimation if the unit appears in the combined sample from the locations. The estimation procedure, especially variance estimation, is more complex than the following alternatives.

(2) Hansen-Hurwitz (H-H) estimator

The Hansen-Hurwitz type of estimator calculates the expected number of times a unit is selected into the sample, taking into account its selection from any of the locations. Following Equation (4.2), this number in the combined sample is given by the sum of its probabilities over the locations with which it is associated:

$$p_j = \sum_{i \in I} f_{ij}^* \cdot r_j = \sum_{i \in I} k_i \cdot f_{ij}^* \cdot r_j,$$

or simply summing over the $I_j$ locations of which unit $j$ is a member:

\(^{40}\) Alternatively, we can define $f_{ij} = 0$ if unit $j$ is not associated with location $i$, so that $I_j$ in (10.7) can be replaced by $I$, the total number of locations in the population, without affecting $\pi_j$ in the equation.
\[ p_j = \sum_{i \in l_j} f_{ij}' = \sum_{i \in l_j} k_i f_{ij}, \text{ weight } w_j = 1/p_j . \]  

The inverse of these quantities are used as weights for the combined sample from all locations, applied to each selection of the unit into that sample.

(3) Multiplicity estimator

Here the probability is specific to the location \((j)\) from which the particular enumeration \((j)\) of the unit has been obtained. Following Equation (4.3) we have

\[ p_{ij} = l_j f_{ij}' = k_i l_j f_{ij}, \text{ weight } w_{ij} = 1/p_{ij} . \]  

A unit is counted wherever it is selected – repeatedly if has been selected at more than one location. This form of the estimator is perhaps the most practical one in our context. It requires information on selection probability only for the location(s) where a unit is actually selected. It does not need information on which other locations the unit may have been selected from, nor even on if there are repeated selections of the unit from the same location. Certainly no micro-level matching across locations or within locations is required. All such information is very difficult – often impossible – to collect, especially (but not only) in the context of child labour surveys in developing countries. The procedure does however require information on the total number of locations in the population to which a sample unit belongs.

10.2.4 Numerical illustrations

Tables 10.3-10.5 provide numerical illustration of the procedures described in the preceding sections.

Table 10.3A shows the population used for this illustration. In a real situation this population is, of course, unknown, and the objective of the procedures discussed is to estimate the parameters of this population.
### Table 10.3. Illustration of location sampling: population and sampling rate unknown

#### A. Population (unknown)

<table>
<thead>
<tr>
<th>Members of location i</th>
<th>N_i</th>
<th>o=1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>total</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>i =1</td>
<td>260</td>
<td>260</td>
<td>80</td>
<td>60</td>
<td>40</td>
<td>30</td>
<td>0</td>
<td>30</td>
<td>10</td>
<td>50</td>
<td>20</td>
<td>580</td>
<td>2.2</td>
</tr>
<tr>
<td>2</td>
<td>290</td>
<td>80</td>
<td>290</td>
<td>10</td>
<td>40</td>
<td>60</td>
<td>10</td>
<td>60</td>
<td>50</td>
<td>90</td>
<td>700</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>140</td>
<td>60</td>
<td>10</td>
<td>140</td>
<td>20</td>
<td>80</td>
<td>70</td>
<td>10</td>
<td>40</td>
<td>80</td>
<td>580</td>
<td>4.1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>220</td>
<td>40</td>
<td>40</td>
<td>220</td>
<td>10</td>
<td>20</td>
<td>50</td>
<td>90</td>
<td>80</td>
<td>90</td>
<td>660</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>160</td>
<td>30</td>
<td>60</td>
<td>80</td>
<td>10</td>
<td>160</td>
<td>80</td>
<td>50</td>
<td>90</td>
<td>60</td>
<td>910</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>90</td>
<td>0</td>
<td>10</td>
<td>70</td>
<td>20</td>
<td>80</td>
<td>90</td>
<td>30</td>
<td>70</td>
<td>90</td>
<td>530</td>
<td>5.9</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>290</td>
<td>30</td>
<td>10</td>
<td>10</td>
<td>50</td>
<td>50</td>
<td>70</td>
<td>310</td>
<td>40</td>
<td>60</td>
<td>700</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>160</td>
<td>10</td>
<td>60</td>
<td>70</td>
<td>90</td>
<td>90</td>
<td>30</td>
<td>40</td>
<td>160</td>
<td>80</td>
<td>0</td>
<td>630</td>
<td>3.9</td>
</tr>
<tr>
<td>9</td>
<td>190</td>
<td>50</td>
<td>50</td>
<td>40</td>
<td>80</td>
<td>60</td>
<td>70</td>
<td>60</td>
<td>80</td>
<td>190</td>
<td>0</td>
<td>680</td>
<td>3.6</td>
</tr>
<tr>
<td>10</td>
<td>200</td>
<td>20</td>
<td>90</td>
<td>80</td>
<td>90</td>
<td>90</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>200</td>
<td>730</td>
<td>3.7</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2,000</td>
<td>580</td>
<td>700</td>
<td>580</td>
<td>660</td>
<td>710</td>
<td>530</td>
<td>700</td>
<td>630</td>
<td>680</td>
<td>730</td>
<td>6,500</td>
<td>3.3</td>
</tr>
</tbody>
</table>

#### B. Sample

<table>
<thead>
<tr>
<th>Members of location i</th>
<th>n_i</th>
<th>o=1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>total</th>
<th>(1/f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>i =1</td>
<td>65</td>
<td>65</td>
<td>20</td>
<td>15</td>
<td>10</td>
<td>8</td>
<td>0</td>
<td>8</td>
<td>3</td>
<td>13</td>
<td>5</td>
<td>145</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>97</td>
<td>27</td>
<td>97</td>
<td>3</td>
<td>13</td>
<td>20</td>
<td>3</td>
<td>3</td>
<td>20</td>
<td>17</td>
<td>30</td>
<td>233</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>12</td>
<td>2</td>
<td>28</td>
<td>4</td>
<td>16</td>
<td>14</td>
<td>2</td>
<td>14</td>
<td>8</td>
<td>16</td>
<td>116</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>31</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>31</td>
<td>1</td>
<td>3</td>
<td>13</td>
<td>11</td>
<td>13</td>
<td>94</td>
<td>116</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>18</td>
<td>3</td>
<td>7</td>
<td>9</td>
<td>1</td>
<td>18</td>
<td>9</td>
<td>6</td>
<td>10</td>
<td>7</td>
<td>10</td>
<td>79</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>45</td>
<td>0</td>
<td>5</td>
<td>35</td>
<td>10</td>
<td>40</td>
<td>45</td>
<td>35</td>
<td>15</td>
<td>35</td>
<td>45</td>
<td>265</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>97</td>
<td>3</td>
<td>3</td>
<td>17</td>
<td>17</td>
<td>23</td>
<td>103</td>
<td>13</td>
<td>20</td>
<td>23</td>
<td>233</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>160</td>
<td>60</td>
<td>70</td>
<td>90</td>
<td>90</td>
<td>30</td>
<td>40</td>
<td>160</td>
<td>80</td>
<td>0</td>
<td>630</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>32</td>
<td>8</td>
<td>8</td>
<td>13</td>
<td>10</td>
<td>12</td>
<td>10</td>
<td>13</td>
<td>32</td>
<td>0</td>
<td>113</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>25</td>
<td>3</td>
<td>11</td>
<td>10</td>
<td>11</td>
<td>11</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>91</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>597</td>
<td>144</td>
<td>219</td>
<td>183</td>
<td>201</td>
<td>231</td>
<td>150</td>
<td>223</td>
<td>261</td>
<td>222</td>
<td>167</td>
<td>2,000</td>
<td></td>
</tr>
</tbody>
</table>

\( f = \) sampling rate (unknown when the population size in not known)

Each row in the figure represents a location \((i)\). The second column shows the number of persons belonging to the location concerned, i.e. who are its members \(N_i\), and among those the number who also belong to each of the other locations \(o\). Thus there are \(N_1 = 260\) members of location \(i = 1\); of these \(80\) \((N_{12})\) also belong to location \(o = 2\), \(60\) \((N_{13})\) also to location \(o = 3\), etc. The two last mentioned groups may comprise of the same individuals or of different individuals. The 260 members of location \(i = 1\) have a total of 580 ‘memberships’, as shown in the column headed ‘total’, because of the individuals belonging to multiple locations.
Row 2 shows that $N_2 = 290$ persons are members of location $i = 2$, of which 80 are also members of location 1 ($N_{21}$). By symmetry, $N_{21} = N_{12}$.

Simply adding numbers of members found at the various locations, we get a total of 2,000. The number of persons in the population is much smaller because of multiple membership. The last column shows the average number of membership of individuals at each location. The average exceeds 3, indicating that the population size may be around $2000/3 \approx 600$ persons.

None of these figures are, of course, known if our observations are confined to a sample, and also when we are not linking individuals across locations.

For the illustration, we have drawn a sample shown in panel B from this population. The last column shows the inverse of the sampling rates which have been applied to each row of panel A. There is an equal probability sample of members at each location but the sampling fraction varies from one location to another.

As in a real situation, from now on we will assume that the size of the population and hence also the sampling rates applied are not known, except for the fact of equal probability sampling within each location. What is given are the sample data shown in panel B. The objective is to use this information to estimate the population size and the sampling rates applied at each location. After this exercise, we can compare the results obtained with the real situation given in panel A.

The figures shown in panel B are as follows. In row $i = 1$ for instance, there are 65 persons in the sample for location $i = 1$, of which 20 belong also to (the population of) location 2, and 15 belong to (the population of) location 3, and so on. This is $n_{i0}$ as defined above. The column shows how much of the population of a given location is present in the sample at each of the other locations (rows). For instance, 20 individuals in the population of location 2 are present in the sample of 65 persons at location 1. Of the population at location 2, 97 persons are present in the sample at location 2, which is obviously the whole of the sample there. Similarly, only 2 of that population are present in the sample of 28 persons at location 3.
### Table 10.4. Illustration of location sampling: cross-membership between locations

#### (A) Number of members of location \( i \) (in the population) who are in sample of location \( o \) \((n_{oi})\)

<table>
<thead>
<tr>
<th>Location ( i )</th>
<th>( o = 1 )</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>In sample of location ( o )</td>
<td>( i = 1 )</td>
<td>20</td>
<td>15</td>
<td>10</td>
<td>8</td>
<td>0</td>
<td>8</td>
<td>3</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
<td>97</td>
<td>3</td>
<td>13</td>
<td>20</td>
<td>3</td>
<td>3</td>
<td>20</td>
<td>17</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>29</td>
<td>4</td>
<td>16</td>
<td>14</td>
<td>2</td>
<td>14</td>
<td>8</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>31</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>13</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>7</td>
<td>9</td>
<td>18</td>
<td>9</td>
<td>6</td>
<td>10</td>
<td>7</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>5</td>
<td>35</td>
<td>10</td>
<td>40</td>
<td>45</td>
<td>35</td>
<td>15</td>
<td>35</td>
<td>45</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>3</td>
<td>28</td>
<td>3</td>
<td>60</td>
<td>8</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>60</td>
<td>70</td>
<td>90</td>
<td>30</td>
<td>40</td>
<td>160</td>
<td>80</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>13</td>
<td>10</td>
<td>12</td>
<td>10</td>
<td>13</td>
<td>32</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>11</td>
<td>10</td>
<td>11</td>
<td>11</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td></td>
</tr>
</tbody>
</table>

#### (B) Number of persons in sample of \( i \) who are members of location \( o \) in the population \((n_{oi})\)

<table>
<thead>
<tr>
<th>Location ( i )</th>
<th>( o = 1 )</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>In population of location ( o )</td>
<td>( i = 1 )</td>
<td>65</td>
<td>27</td>
<td>12</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>10</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>97</td>
<td>2</td>
<td>6</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>60</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>3</td>
<td>28</td>
<td>3</td>
<td>28</td>
<td>9</td>
<td>35</td>
<td>3</td>
<td>70</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>13</td>
<td>4</td>
<td>31</td>
<td>1</td>
<td>10</td>
<td>17</td>
<td>90</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>20</td>
<td>16</td>
<td>1</td>
<td>18</td>
<td>40</td>
<td>17</td>
<td>90</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>3</td>
<td>14</td>
<td>3</td>
<td>9</td>
<td>45</td>
<td>23</td>
<td>30</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>3</td>
<td>7</td>
<td>6</td>
<td>35</td>
<td>103</td>
<td>40</td>
<td>10</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>20</td>
<td>14</td>
<td>13</td>
<td>10</td>
<td>15</td>
<td>32</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>13</td>
<td>17</td>
<td>8</td>
<td>11</td>
<td>7</td>
<td>35</td>
<td>20</td>
<td>80</td>
<td>32</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>30</td>
<td>16</td>
<td>13</td>
<td>10</td>
<td>45</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
</tbody>
</table>

In Table 10.4, panels A and B show the basic figures we need in order to proceed with the computations. Panel 10.4A has exactly the same data as panel 10.3B; we have merely changed its description by reversing the role of indices \( i \) and \( o \). This is to make the form of table 10.4 panels A and B the same – the reference index \( i \) as columns, and the other index \( o \), referring to all other locations, as rows.

Panel 10.4A gives \( n_{oi} \) values. Its matrix is transformed in panel B to give \( n_{oi} \).

Table 10.5A shows cell-by-cell ratio of the two panels of Table 10.4, namely the ratio \((n_{oi}/n_{io})\). A number of statistics computed from this are shown in panel 10.5B. Row (1) of panel B gives the average of values in each column \( i \). This is the quantity \( a_{oi} (n_{oi}/n_{io}) \) of Equation 10.6. Its inverse is proportional to the sampling rate applied in the location corresponding to column \( i \). The constant of proportionality \( k \) is not known at this stage. The relative sampling rates given by Equation (10.6) are shown for all locations \( i \) in row (2). Note that some of the \( n_{io} \) values are zero, which happens when locations \( i \) and \( o \) do not have common members, and can also happen due to variability with small samples. The zero cells are excluded in taking the mean.

As noted, the predicted sampling rates in row (2) of Table 10.5B are relative. More information is required to convert these figures to rates in absolute terms.
Just to illustrate how the procedure has worked, we have shown in row (3) of Table 10.5B the actual sampling rates we had applied in order to construct this illustration. These were given in the last column of Table 10.3B. In any case, by scaling the predicted relative rates to have the same average as that of the actual rates in row (3) – which of course is not known to us in a real situation – we can see that, apart from an unknown constant of proportionality, the estimates are mostly very close to the actual values. Notable differences exist only in the case of one location (Location 10), in which the applied sampling rate was low; also this location had no overlapping membership with some of the other locations.

The ratio of average values of rows (2) and (3) is 0.32. This indicates the value of parameter $\bar{k}$.

Finally, row (5) of Table 10.5B shows the constant of probability estimated if the population size of at least one of the locations were known. In reality such information may be available from existing sources or it may be possible to collect it at some of the locations.

### Table 10.5. Illustration of location sampling: estimation of the unknown population

**A. Cell-by-cell ratio of the two panels ($n_{oi}/n_{io}$) in Table 10.4**

<table>
<thead>
<tr>
<th>Location i</th>
<th>i=1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>o=1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.00</td>
<td>0.75</td>
<td>1.25</td>
<td>1.75</td>
<td>2.25</td>
<td>.</td>
<td>.</td>
<td>0.75</td>
<td>0.25</td>
<td>1.50</td>
</tr>
<tr>
<td>2</td>
<td>1.33</td>
<td>1.00</td>
<td>1.67</td>
<td>2.33</td>
<td>3.00</td>
<td>0.67</td>
<td>1.00</td>
<td>0.33</td>
<td>2.00</td>
<td>2.67</td>
</tr>
<tr>
<td>3</td>
<td>0.80</td>
<td>0.60</td>
<td>1.00</td>
<td>1.40</td>
<td>1.80</td>
<td>0.40</td>
<td>0.60</td>
<td>0.20</td>
<td>1.20</td>
<td>1.60</td>
</tr>
<tr>
<td>4</td>
<td>0.57</td>
<td>0.43</td>
<td>0.71</td>
<td>1.00</td>
<td>1.29</td>
<td>0.29</td>
<td>0.43</td>
<td>0.14</td>
<td>0.86</td>
<td>1.14</td>
</tr>
<tr>
<td>5</td>
<td>0.44</td>
<td>0.33</td>
<td>0.56</td>
<td>0.78</td>
<td>1.00</td>
<td>0.22</td>
<td>0.33</td>
<td>0.11</td>
<td>0.67</td>
<td>0.89</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.50</td>
<td>2.50</td>
<td>3.50</td>
<td>4.50</td>
</tr>
<tr>
<td>7</td>
<td>1.33</td>
<td>1.00</td>
<td>1.67</td>
<td>2.33</td>
<td>3.00</td>
<td>0.67</td>
<td>1.00</td>
<td>0.33</td>
<td>2.00</td>
<td>2.67</td>
</tr>
<tr>
<td>8</td>
<td>4.00</td>
<td>3.00</td>
<td>5.00</td>
<td>7.00</td>
<td>9.00</td>
<td>2.00</td>
<td>3.00</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.67</td>
<td>0.50</td>
<td>0.83</td>
<td>1.17</td>
<td>1.50</td>
<td>0.33</td>
<td>0.50</td>
<td>0.17</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.50</td>
<td>0.38</td>
<td>0.63</td>
<td>0.88</td>
<td>1.13</td>
<td>0.25</td>
<td>0.38</td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
</tbody>
</table>

**B. Analysis**

<table>
<thead>
<tr>
<th>i=1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Mean (excluding zeros)</td>
<td>1.18</td>
<td>0.95</td>
<td>1.58</td>
<td>2.21</td>
<td>2.85</td>
<td>0.65</td>
<td>0.95</td>
<td>0.34</td>
<td>2.02</td>
<td>2.00</td>
</tr>
<tr>
<td>(2) inverse of above mean (relative sampling rates applied)</td>
<td>0.85</td>
<td>1.05</td>
<td>0.63</td>
<td>0.45</td>
<td>0.35</td>
<td>1.55</td>
<td>1.05</td>
<td>2.96</td>
<td>0.49</td>
<td>0.50</td>
</tr>
<tr>
<td>(3) (unknown) actual sampling rate applied</td>
<td>0.25</td>
<td>0.33</td>
<td>0.20</td>
<td>0.14</td>
<td>0.11</td>
<td>0.50</td>
<td>0.33</td>
<td>1.00</td>
<td>0.17</td>
<td>0.13</td>
</tr>
<tr>
<td>(4) if (2) were to be rescaled to have same average as (3), it gives</td>
<td>0.27</td>
<td>0.34</td>
<td>0.20</td>
<td>0.14</td>
<td>0.11</td>
<td>0.49</td>
<td>0.34</td>
<td>0.95</td>
<td>0.16</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Hence the procedure estimates well the relative sampling rates applied in different locations within an (unknown) overall constant estimated from the ratio of averaged values as $(0.32/0.99) = 0.32$.

**Ratio (3) / (2)**

| (3) / (2) | 0.30| 0.32| 0.32| 0.32| 0.32| 0.32| 0.32| 0.34| 0.34| 0.25 | 0.31 |

If the population size, hence the actual sampling rate applied were known for any location(s), the above ratios can be computed. This provides an estimate of the unknown constant. Values in (2) multiplied by this constant give sampling rates in absolute terms.
10.2.5 Sampling of locations

A. Single-stage sampling of locations

If there are many small locations in the study population, one may consider taking a sample of locations, and within each location selected, taking all individuals regularly visiting the location into the sample.

This can be a practical scheme if there are too many locations to be covered on a census basis. However, the practical utility of the scheme is limited because of the nature of the information required from respondents on the patterns of their association with the locations. Each person in the survey has to be able to report on his/her pattern of association, i.e. for each location in the study population, on whether the person is associated with it or not. At a minimum, it is necessary to know the number of locations in the population (not only in the sample) that a person is associated with. Often, respondents may not be aware of what the target population of locations is supposed to cover.

Given these requirements, the location sampling method works well when each individual in the target population is associated with a limited number of locations (for example ideally no more than 10), and at any location, all persons can be found on one occasion, or with a limited number of revisits.

In any case, technical details of the procedure are quite similar to those discussed in the previous section for the design with full coverage of locations and sampling within locations.

For instance, using the Hansen-Hurwitz type of estimator, the probability of a unit being in the combined sample may be taken as the sum of its probabilities over the locations with which it is associated. With notation already introduced:

\[ p_j = \sum_i \left( b_i r_{ij} \right), \]

or simply summing over the \( J_i \) locations of which unit \( i \) is a member:

\[ p_j = \sum_{i \in J_i} b_i. \]

Here \( b_i \) is the selection probability of location \( i \), and \( r_{ij} \) is a \((0,1)\) index indicating whether \((=1)\) or not \((=0)\) unit \( j \) is associated with location \( i \). The inverse of these unit selection probabilities are used as weights for the combined sample from all locations.

In fact the above expressions are much simpler than those discussed earlier with subsampling within locations. The location selection probabilities \( b_i \) are known in absolute terms, and there is no unknown scaling parameter \( k_i \) to deal with.

B. Two-stage sampling

If there are many big locations, it may become necessary to introduce sampling at both stages: first taking a sample of locations, and then within each selected location, taking a sample of individuals.
Technical details of the procedure are similar to but more complex than those discussed above. As before, using the Hansen-Hurwitz type of estimator for instance, the probability of a unit being in the combined sample may be taken as the sum of its probabilities over the locations with which it is associated. With notation already introduced:

\[ p_j = \sum_i b_i f'_i r_i = \sum_i b_i k_i f_i r_i, \]

or simply summing over the \( f_i \) locations of which unit \( i \) is a member:

\[ p_j = \sum_{i \in f_i} b_i f'_i = \sum_{i \in f_i} b_i k_i f_i. \]

The inverse of these unit selection probabilities are used as weights for the combined sample from all locations.

Here \( b_i \) is the selection probability of location \( i \), and \( r_{ij} \) is a \((0,1)\) index indicating whether \((=1)\) or not \((=0)\) unit \( j \) is associated with location \( i \). Within each sample location \( i \), the selection probability of the unit is \( f'_i = k_i f_i \). Generally, the relative selection probabilities \( f_{ij} \) are available, but not \( f'_{ij} \), the probabilities in absolute terms, because of lack of information on population sizes of the locations. Scaling factor \( k_i \) has to be estimated as described in detail previously.

### 10.3 Time-location sampling

#### 10.3.1 Basic characteristics

Among the diverse surveys of mobile populations, the examples described in this section represent, in certain respects, the opposite end of the spectrum from the location sampling discussed in the previous section. In normal location sampling the population consists of a small number of locations. The method also assumes that at certain appropriate time or times for each location, all its members can be found or located through the location concerned. Beyond that, there is no time dimension to the sampling.

Surveys of flows have a different structure. In the following, we begin by describing some basic features of flow surveys. Then some simple examples will be given to illustrate these features. These surveys monitor the flow of individuals through fixed points in space during specified time segments. The flow may comprise of individuals passing through chosen vantage points, arriving or leaving some place or institution, visiting a facility, etc.

#### A. Primary sampling units in surveys of flows

In the introductory section, the concepts of location, observation point and time segment were introduced. These concepts are particularly relevant for surveys where the objective is to collect information on individuals visiting or passing through chosen locations. They are also useful in the wider context of sampling mobile populations.
10.3 Time-location sampling

The locations and observation points where the survey is conducted provide the necessary geographical units at which mobile individuals in the target population can be enumerated. Typically a location will have several observation points at which the flows under study can be most effectively observed. When there are many small locations, a location may itself constitute an observation point.

It is common for flow surveys to involve many locations. However, in some examples of surveys of flows, the survey is confined to a single or a few locations. When that is the case, no explicit sampling of locations may be involved. In fact, sometimes the survey is designed only for one or a few specific locations. In other situations, there may be ‘implicit’ sampling involved, in that the particular locations taken into the survey are a small subset of possibly many similar (or assumed to be similar) locations, but the choice is made a priori, by judgement and not through random sampling. This assumes that the different possible locations are replaceable as concerns the survey objectives and the issues being studied.

In the time-location sampling framework for surveys of flow, the cross of observation points by time segments form the primary sampling units (PSUs). For example, suppose we have 4 locations all included in the survey, and these locations have a total of 15 observation points. The survey covers 5 days of a week, Monday-Friday, during working hours taken as 10:00-18:00. There are considerable variations in the size and nature of the flow of persons over days of the week and over times of the day, and therefore it is decided to conduct the survey on all the 5 days, and cover all the 8 working hours during the day divided into 4 segments of 2 hours each. This gives a total of $5 \times 4 = 20$ time segments. Observation points by time segments therefore form $15 \times 20 = 300$ PSUs. Let us assume that the available resources dictate that only 25 PSUs can be selected for enumeration. We would like a balanced sample where all – or at least the maximum number possible of – locations, observation points, days and time segments within a day are represented. But of course all possible combinations of the above (which, as noted, number 300) cannot be represented in a small sample of only 25 PSUs. Generally in flow surveys the sample includes all the observation points at the selected locations, and all the time segments covered by the survey period, but not all the combination of these two dimensions – including all the combinations would mean no sampling of PSUs within locations.

The controlled selection procedure described in Chapter 12 can be used to achieve such a balanced sample. When the number of PSUs to be selected is small, but the desired stratification and other control categories are more numerous, controlled selection and balanced sampling procedures can improve representativeness of the sample in terms of those controls.

The size of a PSU means the number of individuals flowing through the observation point and the time segment defining the PSU. Often the PSUs vary in size because of variations in the flow over space and time. Some observation points are busier than other points, and the flow is heavier at some times than others (during the day, between days etc.) The combined effect of these variations determines the variation in PSU sizes.

The usual sampling issues and solutions with variable unit size apply here. The first option is to take steps to reduce the variation in PSU sizes. As in conventional area
sampling, one may ‘split’ larger units, for example by making the time segments shorter during busier periods. Another option is to increase the selection probability of larger PSUs and correspondingly reduce the sampling rate at the later stage of sampling individuals within PSUs. This can be just an approximate or crude application of the standard probability proportional to size (PPS) sampling at the first stage and inverse-PPS sampling at the second stage. A more refined PPS sampling procedure can be applied, depending on what information on the unit size measures is available.

In any case, the effect of PSU size on the workload during fieldwork can be reduced by reducing the sampling rate for selecting persons within PSUs at the next stage, irrespective of the probabilities applied at the first stage of selecting PSUs. This reduces the final selection probability of individuals, which has to be compensated by appropriately increasing the unit weights at the estimation stage.

In the end, variations in workload have to be accommodated by adjusting the tempo of interviewers’ work – subjecting them to greater work pressure at busy times, and tolerating relaxed (or no) work at lax times when the flow of individuals through the observation points is low.

B. Selection of individuals within PSUs

Sampling of individuals at the observation points has to be done in the field since, obviously, no list frames of flowing populations can be prepared beforehand. In order to obtain an equal probability sample of individuals at an observation point, the easiest way is to take a systematic sample, say 1 out of every $I$ persons who pass through the point. The interval $I$ may vary from one observation point to another depending on the design. This scheme gives a uniform workload over time at an observation point if the flow is reasonably constant. The interviewers’ workload varies according to the rate of flow, i.e. the number of persons passing through the observation point during 1 unit of time. There can be problems if the flow varies greatly over time. The interviewers may be unable to handle the workload when the flow is high, or at least may be unable to sustain the effort except for a short time; conversely, they may be under-employed when the flow is low.

In order to reduce the problem, one may follow the conventional PPS design as already noted: select observation points expected to be busy with higher probability at the first stage and reduce the sampling rate at the second stage. This can work well only if (i) information on variation in flow rates is available prior to sample selection; (ii) no unexpected significant departures in the flow rates are found at the time of survey enumeration; and (iii) there are no large variations in the flow rate within time segments at a PSU.

The alternative to systematic sampling with a fixed interval is to have a ‘constant take’ design – the sample size enumerated at a point during a time segment is kept constant. This can be achieved by, say, interviewing 1 person every $T$ units of time. The rate of work and hence the size of the sample enumerated during a time segment remains constant; also within the time segment, variations in the flow do not affect the rate of work or the sample size obtained. However, the individuals’ selection probabilities vary in inverse proportion to the rate of flow. In order to know the sampling rate applied, it is necessary to know the rate of flow through the observation point: only if the flow rate,
say $F$, is known can we compute the sampling rate as $1/(TF)$, where $T$ is the number of time units separating successive interviews and $F$ is the amount of flow per unit of time. When the flow rate varies within the time segment, it is necessary to know these variations so that the variable sampling rates they imply can be taken into account.

A commonly used scheme is to take into the survey the next person passing the point as soon as the interview with the previous selected person has been completed. The objective of this scheme is to maximise the use of interviewers’ time. The scheme, however, does not produce an equal probability sample of persons, such as produced by the 1 in $I$ systematic sampling. In addition, the sampling rate also becomes dependent on the rate of work of individual interviewers. In fact, variations in the sampling rate cannot be monitored unless detailed records using very fine time intervals are kept of variations in the flow and the timing of individual interviews.

### C. Survey units: ‘visits’ vs. ‘visitors’

The characteristic feature of surveys of flows is that the study population is defined only in terms of visits to specified locations during specified time periods – and not in relation to particular areas or categories of individuals to be covered, as is the case in usual surveys. Those who never visit the locations (more strictly – those who do not visit during the period defining the population) are out of scope: such ‘reclusive’ persons are excluded from the frame.

The important point is that generally the units of observation are ‘visits’ to the sample locations, rather than the population of ‘visitors’. This is a very important distinction to be kept in view in the interpretation of the survey results. This means that if a person makes, say, $n$ visits to any location, then the person’s probability of being selected for the survey is increased $n$-folds. If the selection probabilities applied at different locations or observation points are different, then the resulting probability of selection of the individual is basically the sum of probabilities at the locations which the person visits during the period covered by the survey. Note that the person concerned need not be selected into the sample during different visits or at different locations for the probability to be so affected: the multiple visits and locations need only be within the population covered. For example, if a survey is conducted at a location twice a day (say during the morning and the afternoon) on two days, a child who comes to the location during all the 4 time segments has 4 times the chance of being selected for the survey, compared to a child who visits the site during only one of these time segments. This applies equally when the population has multiple sites and a random sample from them is selected for the survey. See Section 10.5 below for a discussion of how selection probabilities are affected by the pattern and frequency of visits.

In distinguishing between ‘visits’ and ‘visitor’, we may note that defining what constitutes ‘a visit’ also requires attention in some situations. What if the same individuals repeatedly exit and re-enter the same observation point? Usually it makes sense to consider any number of entries at a point during one time segment of the survey as a single visit.

The primary interest in child labour surveys is to obtain a probability sample of working children, irrespective of how those children may differ in their intensity of engagement
in the work. In short, our primary interest is to have a sample of ‘visitors’ rather than of ‘visits’.

Nevertheless, despite differences in selection probabilities to which visitors may be subject in a survey where the sampling unit is visits, a sample of visits can also be of interest in itself. Firstly, from the point of view of analysing the situation of institutions or facilities providing services to individuals, data on the number and characteristics of visits taking place are needed for policy, management or performance evaluation of the facility. Secondly, the frequency of visits a person makes can, in a certain sense, be taken as a measure of the ‘importance’ of the person for the topic being studied. For instance, a child intensively engaged in work may visit a work location every day, all day. Another child, more marginally engaged in that labour, may come to the work location say one afternoon a week. True, in the overall context of child labour, both children count equally; but it is also not unreasonable that for certain policy or research concerns, the child engaged in intensive labour counts more.

### 10.3.2 Applications in diverse fields

Kalton (1991) provides a useful review and discussion of several types of application of flow surveys from developed countries. Though the concerns and conditions of surveys concerning child labour, especially child labour in developing countries, are rather different, there are some common elements between different types of survey in the two situations. Both comparisons and contrasts between the two will be instructive. Also, the surveys briefly reviewed below help to clarify the basic concepts of time-location sampling of flows. The above-mentioned paper may be consulted for references to the literature giving more details on the examples below.

#### A. Survey of library use at a university (1984)

This example is useful in bringing out some most basic elements of the methodology. A small number (18) of libraries at University of Michigan, USA, formed the survey locations. Exits (one per library) provided the observation points. No sampling was necessary of the locations (libraries) or the observation points (exits). There was obviously no problem in defining or locating these units. Library exits by 2-hour long time segments formed the PSUs. With volume of flow of library users as the measure of size, a PPS sample of these PSUs was selected.

Persons exiting were counted, and a systematic equal probability sample of them interviewed. Eligibility for the survey was conditional on use of the library facilities. Screening for eligibility involved a simple question on whether the person had actually used the library facilities during that visit.

#### B. Survey of museum visits (1988)

A single institution, a museum in Washington D. C.; formed the location, and its 2 or 3 open exits formed the observation points. Time segments were two half days each day over the survey period. One time segment was selected per day, alternating between the morning and the afternoon during successive days.
Persons exiting the museum were selected for the interview. The commonly used scheme, of taking into the survey the next person exiting after the completion of the previous interview, was employed. The objective of this scheme is to maximise the use of interviewers' time. The scheme does not produce an equal probability sample of persons, but makes the sampling rate inversely proportional to the rate of flow and the interviewer's pace of work. Persons exiting were counted over 10 minute intervals to determine the rate of flow. Comparing this rate with the number of interviews completed gave an estimate of the sampling rate.

The screening operation was simple: the interview excluded museum staff and also children.

C. Exit polls in the USA

There are many surveys involving exit polls. In this example, the locations were voting precincts. A PPS sample of these was selected, with the (known) number of registered voters in the area forming the size measure. The observation points were the exits from the polling station (often 1-3 per location). At some locations, practical problems arose when the regulations prohibited interviewers from approaching within a certain distance the polling station exits.

The flow of exiting visitors, while not uniform, mostly did not have large peaks. Consequently, the normal equal probability systematic sampling could be used to select a sample of visitors. There is no ‘visit-visitor’ distinction here: it can be assumed that any person made at the most only one visit to the location.

D. Surveys of international passengers

There are different types of surveys of passengers. For example, there are in-flight surveys. In such surveys, airlines form the ‘locations’ in our terminology. Generally all the operators at an airport are included in the study. Scheduled flights crossed by weeks as the time segments may form the PSUs. Passengers on a selected flight are usually given self-completion questionnaires. There is no sampling necessary within PSUs (scheduled flights). However, while the procedure is simple in principle, the major practical problem with surveys using this type of methodology is the extremely low response rates achieved.

Alternatively, a survey of passengers may be carried out at the terminal, where a sample of arriving passengers are included in the study. One or more airports (or sea ports, or bus or train terminals, etc.) can be the locations. Scheduled flights may form the observation points. Part of the day, or a similar time segment in terms of flight schedules, may form the PSU. An equal probability sample of PSUs, followed by a systematic sample of passengers within PSUs, may be the simplest option if the flow is manageable. Otherwise, some alternative two-stage sampling scheme may be more suitable.

E. Survey of shopping centres in the UK

Four shopping centres where hypermarkets had been opened formed the locations. Trade outlets in a shopping centre formed the observation units. According to the study report, the creation of these observation units required more work than was the case
in the previous examples above. The work involved listing all retail outlets and their hours of opening. The total opening hours were divided into shorter time segments. Retail outlets by time segments formed the PSUs, of which a sample was selected. The next step involved counting the numbers of departing shoppers from the outlets in the sample during the sampled hours. The objective of this information was to determine the number, location and timing of the interviewers to be deployed. Shoppers were selected on the basis of taking the next one after the preceding interview was completed. The counts mentioned above could be used to determine the applied sampling rates.

F. Road traffic surveys

There are traffic surveys of many different types, but here we refer to surveys relating to traffic passing through specified set of locations. Observation points may be defined by going through a number of steps or stages of sampling – such as neighbourhoods, road intersections, sets of traffic lanes, individual lanes, etc. The lowest-stage units in such a hierarchy, crossed by suitably defined time segments, can form the PSUs. Finally, passing vehicles can be counted or sampled. A critical issue in this methodology is whether the survey methods require vehicles to stop. This requirement determines where and what type of survey is possible.

10.4 Sampling mobile populations in more difficult situations: examples

10.4.1 Introduction

In this section we will discuss in some detail a few examples of surveys of mobile populations in more difficult situations than the examples given in the preceding section. The populations of interest can be mobile and difficult to locate and identify to varying degrees. The difficulties of enumerating such populations concern the identification of who are the eligible respondents for the survey, and where and when to find them; also what information to ask them and how to ask it. Another set of issues concerns how to use sample data to produce valid estimates for the population, and to assess variances and biases to which those estimates are subject.

We will organise the following discussion around three aspects in sampling mobile populations identified earlier, namely:

- the identification and selection of PSUs (observation points by time segments);
- counting, screening and selecting individuals for the survey; and
- special issues in estimating from the sample survey.

The examples of surveys discussed in this section do not concern child labour but are from other fields. There is very limited experience and information on such surveys in the field of child labour. Nevertheless, common issues and problems are encountered often in surveys on different topics. It is therefore useful to study surveys in different areas, with the aim of identifying methods and procedures which may be adaptable to
surveys specifically concerned with child labour. In any case, these applications are of methodological interest in their own right.

This discussion will also provide the necessary background for the case study presented in Annex C, where we will consider issues in the development of a full-scale child labour survey drawing on the experience from a baseline survey of child labour with limited coverage.

Two surveys concerning the following areas are discussed below:
- homosexual young persons, hitherto and more appropriately termed as ‘young men who have sex with men’ (YMSM); and
- migrants residing in households scattered in the host community, who cannot be easily contacted at their normal place of residence or at another fixed place.

Both the surveys involve enumeration of a mobile population.

### 10.4.2 A survey concerning sexual behaviour of young persons (YSMS)

This is an illustration from Muhib et al. (2001). It is based on a survey under the Community Intervention Trial for Youth (CITY) project in the US, a six-year project begun in October 1996. The study was designed to evaluate the effects of a multi-component, community level intervention aimed at promoting safer sex among young men aged 15-25 years who reported having sex with other men. The intervention consisted of social marketing material, small group workshops, social events with embedded HIV prevention messages, development or enhancement of a peer outreach programme, and technical assistance designed to increase the capacity of local organizations to deliver HIV prevention and other services to the YMSM (young men who have sex with men). This intervention was evaluated through a series of annual, cross-sectional, confidential surveys of YMSM, who lived in the intervention or comparison communities. The surveys were conducted in monthly waves. The waves were essentially independent and here we can consider just a single wave.

**A. Locations**

Time-location sampling was implemented as part of the evaluation design in thirteen sites representing diverse race and ethnic populations in urban and suburban areas. While this method may not obtain information representing all members of a targeted population, it can produce a reasonably representative sample.

**B. Sampling frame of observation points and time segments**

The survey had a two-stage sampling design: appropriately defined observation points crossed by time segments formed the PSUs (the authors refer to them as ‘facility-day-time units’). The observation points (facilities of various types) were where members of the target population gathered. The facilities were public sex locations, sex businesses, non-gay businesses that cater to the general public (such as train stations), both informal and formal organizations, special events, and bars and clubs. Time segments referred to specific days and time periods when the target population congregated at each point. These points and their associated days and times were divided into standardized time segments (such as four-hour or other intervals depending on the point). First the PSUs
were randomly selected from a sampling frame, then members of the target population were systematically approached in each facility selected.

Identification of the universe of the ‘observation point by time segment’ PSUs is the cornerstone of this method. A reasonably complete and representative list of PSUs in the sampling frame is a key step to ensure that the method is carried out efficiently. Poor enumeration can misdirect resources to PSUs that may not yield adequate data, and productive PSUs may be missed if the earlier formative work is not thorough.

Also, the sampling design is time intensive. In this survey, construction of the initial sampling frame required two to three folds more time compared to the time for the main survey interviewing. The approach relies on identifying the universe of PSUs that a population attends, otherwise there could be a bias of over- or under-representation if the PSUs of some social networks are not identified.

Multiple sources provided data that were triangulated to form a holistic picture of PSUs. Points were initially identified by researchers who interviewed service providers and men from the community, read gay-oriented magazines and other publications, and kept in contact with the target population through word of mouth. They identified days and times when men might attend the points, and used the results of these investigations to decide which facilities to investigate further and when to investigate. The objective was to ensure that some facilities with low attendance were also included.

The formative work generated large numbers of prospective PSUs. Before entering a PSU into the sampling frame, it was evaluated for feasibility of conducting the survey in it, using criteria such as interviewer safety, whether physical layout of the facility allowed for conducting interviews confidentially, and the owner’s permission to collect data. The PSUs were further evaluated using what the study authors describe as Type I and Type II enumeration.

(1) Type I enumeration

The purpose of this was to assess whether the PSUs identified through formative data collection actually were attended by individuals from the target population. Individuals who appeared to be members of the target population were counted as they entered a facility during a designated day and time period. The number of individuals counted was used to estimate how many prospective respondents might enter the facility during a PSU of the standard time segment to be used in the actual survey. The number counted divided by the duration of counting gave an estimate of the rate of entry. The rate multiplied by the standard time unit gave the ‘standardized enumeration’ for the facility. The standardized enumerations were used to determine which PSUs had too few potential respondents to merit collecting data. This operation is not necessary when casual observation clearly indicates that a facility is attended by large numbers of the target population.

(2) Type II enumerations

The purpose of was to estimate the number of eligible individuals who enter the facility during a designated day and time period (the ‘effective yield’). During Type II enumerations, individuals who appeared to be eligible were counted (as in Type I enumeration) and brief interviews were completed to determine their eligibility for
10.4 Sampling mobile populations in more difficult situations: examples

the survey. To avoid selection bias that can result when interviewers choose whom to approach and whom to ignore, interviewers were required to approach potential respondents systematically as they crossed a predefined line or entered a predefined area. In large facilities, some potential respondents inevitably are missed when all field staff are busy interviewing.

Effective yield is estimated by multiplying the standardised enumeration by the screening fraction and the eligibility fraction. The screening fraction is the proportion of people who complete the screener among those who are approached. The eligibility fraction is the proportion of screened people who are eligible. Eligibility required that the respondent: (i) was 15 to 25 years old male; (ii) had had sex with a man in the past 12 months; (iii) was a resident of the catchment area; (iv) in some sites, self-identified as being part of a specific ethnic group or race; and of course (v) persons already screened were excluded. The effective yield estimates the number of interviews a PSU has the potential to generate. The effective yield was used to designate PSUs as large or small, and to stratify the PSU sampling frame by size in order to guarantee that men attending small PSUs were adequately represented in the sample. After facilities, days, and time periods had been identified, a set of unique PSU units was compiled.

In addition to the regular PSUs identified as above, a number of special PSUs were also created. These included special settings or events known to attract members of the target population that did not recur frequently enough or occurred during unpredictable times or at unpredictable locations. Special PSUs of three types were created: (i) a "wildcard", a one-time event that may attract a large number of eligible respondents, such as a ‘gay-pride’ event, an annual picnic, or a health fair; (ii) an event taking place during a specific day and time but at varied and possibly unknown location, such as a monthly house party sponsored by an individual or a club; and (iii) an event where neither the location nor the time was known at the time of sample selection.

The frame contained approximately 300 PSUs (from 140 different facilities) which were attended by the target population.

C. Selection of PSUs

After a sampling frame was developed, it was stratified into two size groups. PSUs were selected at random, with equal probability within the strata of large and small facilities, with the exception of ‘wildcards’ (see above), which were all taken into the sample. Following this, members of the population were systematically identified and surveyed within each selected PSU.

D. Screening and selection of members of the target population

Within each PSU, men for interviews were recruited using the screening procedures described for Type II enumeration. Prospective participants were approached, screened and, if eligible, asked to participate in the study. Screener information such as age, race, and residence was used to interview only persons in the target population. The final sample included 911 men aged under 21.

It should be noted that the procedure used in the construction of PSU also had a screening function. It identified places and times of concentration of members of the eligible population in the very formation of the PSU frame. This is an important strength
of the procedure. This procedure does not depend on the researchers gaining entry to specific social networks, unlike some other methods such as respondent-driven sampling (which of course may have their own rationale and advantages – see Chapter 14).

E. Limitations

A significant limitation of the procedure described above is that it assumes all members of the target population attend facilities that survey staff can access.

Some members of the population may not attend the facilities or may visit them very rarely. Because one is sampling from the universe of PSUs and not population members, individuals who do not attend or rarely attend any facility have a near zero probability of being sampled. ‘Any’ here means any facility included in the frame; it is not confined only to facilities in the sample, and certainly not to only the particular facility where the person is being interviewed.

Those who attend a facility less frequently have a lower probability of being selected compared with population members who attend more facilities and more often. Similarly, if the sample selection is done from the population present at the facility at a certain time (during a certain time interval), rather than at the entry point, those who spend less time on the average at a facility would have a lower probability of being contacted compared to those who tend to stay longer. In order to compensate for these differential selection probabilities, it is necessary to collect information on the frequency of attending, and if possible, also on the time normally spent at a facility. Of course those who never attend any facility remain excluded. For technical discussion of the weighting procedure, see Section 10.5.

Also, individuals entering heavily trafficked facilities may have a lower probability of being enrolled than those entering low-density facilities, unless the number of interviewers assigned to each facility compensates for population size at the facility.

Bias also can be introduced if researchers do not adhere to design and implementation procedures. As the study’s authors note, “the method will produce a systematic sample only if there is no deviation from key procedures, such as uniformity in constructing sampling calendars and interviewing respondents for each place, time, and sampling wave”.

10.4.3 A survey of migrants in Brazil

A. The survey

In this example, from McKenzie and Mistiaen (2007), we will see the potential use of the method of surveying members of the target population where they tend to congregate, but we will also see that the procedure is subject to selection bias in favour of individuals in whom ‘typical’ characteristics of the population of interest tend to be more pronounced.

The authors report the results of an experiment designed to compare the performance of three alternative survey methods in collecting data from Japanese-Brazilian families (Nikkei community groups), many of whom send migrants to Japan. The three surveys conducted were as follows.
(1) Households selected randomly from a door-to-door listing in selected census blocks from the Brazilian Census

This involved a stratified sample using the census to randomly sample census tracts, in which each household is then listed, and screened to determine whether or not it has a migrant belonging to the survey population, with the full length questionnaire then being applied in a second phase only to the households having a member of the target population. In the household-based survey, the main technical issues concern the fact that migrants form a rare subpopulation dispersed in the general population. The sampling issues involved have been discussed in Chapter 6.

(2) A snowball survey using Nikkei community groups to select the seeds

The common approach to snowball sampling is to choose the sample of individuals referred by friends or acquaintances already in the sample. Alternatively, or in addition, the survey may focus exclusively on areas of high concentration of the target population. In the present study, households with eligible migrants were asked to provide referrals to other households with migrant members. This part of the study in Brazil will be discussed in Chapter 13.

(3) A location survey

The survey was conducted at Nikkei community gatherings, ethnic grocery stores, sports clubs, and other locations where family members of migrants were likely to congregate. Individuals were sampled during set time periods at a pre-specified set of such locations.

An objective of the study was to investigate how closely well-designed snowball and location surveys can approach the more expensive conventional household survey method in terms of giving information on characteristics of the migrants, the level of remittances received, and the incidence of return migration. The real strength and utility of special methods for mobile populations such as the location survey discussed here is realised in situations where alternative methods such as a conventional household survey are not a feasible option. However, as this example shows, we may consider such special methods even when the target population is settled in households, which can be sampled in principle using the conventional household survey approach. This can be the case when, for instance, the conventional household survey would be prohibitively expensive for the rare and dispersed target population. “Although 11 out of the 15 locations in the survey were in two Nikkei neighbourhoods, only 18 per cent of the sample lived in these neighbourhoods .... in fact, individuals reported living in over 150 distinct neighbourhoods .... in fact, individuals reported living in over 150 distinct neighbourhoods. However, it must be admitted that such scatter also adversely affects the population coverage which can be achieved with location sampling.

B. The method

Our focus here is on the location sampling method. In sampling immigrants or ethnic minorities, this survey method makes use of the fact that immigrants often cluster at certain locations. Simple examples of this type of approach often carry out sampling at only one type of location. However, by sampling at only one type of location, the survey is likely to miss many migrants. Better coverage of the population of interest can be achieved by surveying at multiple locations. An issue which arises here is that
individuals frequenting more locations get a higher chance of being selected, which needs to be compensated for at the analysis stage. These issues have been discussed in Section 10.2; we will consider the estimation procedures further in Section 10.5.

In the present example, the survey was designed to carry out interviews at a range of locations frequented by the Nikkei population. It was conducted in Sao Paulo and another city. A listing was made of popular places where migrants tended to meet, such as places of worship, health care facilities, telephone call centres, shelters, and public squares. At each location migrants surveyed were asked how often they visited any of the other congregation points, allowing selection probabilities to be calculated for each individual surveyed.

C. Representativeness of the sample

In certain respects, the location sampling procedure is appealing compared to the other two methods explored in this study. On the one hand, it is likely to yield a sample which is more representative of the population of interest than can be found through the first few referral chains of a snowball sample. At the same time it requires less time and resources than a census-based screening and listing exercise followed by a household survey. However, a disadvantage of interviewing in public locations is that individuals will generally have less time to answer the survey than during a home visit which the other procedures employ. As a result, location surveys of this type have to use a much shorter questionnaire, thereby collecting less extensive data on the population of interest. (The mean interview duration for the location survey being discussed was only 9 minutes.)

D. Constructing the location frame

Consultations with Nikkei community organizations, local researchers, and officers of a bank which provided remittance services to this community, were used to select a broad range of locations. In Sao Paulo, the researchers chose nine “fixed-point locations” and six “events”. The nine fixed locations were: a sports club, a metro station, two Sunday open markets, a hospital focused on the community, two grocery stores specializing in Japanese foods, a Japanese cultural society offering language classes and evening events, and a bank in the neighbourhood. The six events were: an afternoon Japanese film event, a large cultural festival, two music and dance festivals, a Japanese food festival, a Japanese art exposition, and a Christmas concert.

The surveyors made an effort to gain the trust and support of the local community. This was done through contact with local associations, banks and researchers, facilitated by the use of Nikkei interviewers where possible. It should be stressed that a location survey requires (depends on) developing such contacts.

E. Counting, screening and sampling for the interview

Interviewers were assigned to visit each location during pre-specified blocks of time. Two field-workers were assigned to each location. One fieldworker carried out the interviews, while the other carried out a count of people of Nikkei appearance and apparently aged 18 or more.
A note was made of the number of individuals who were asked to answer the questionnaire because they appeared Nikkei, but who turned out not to belong to that community. The proportion of incorrectly identified Nikkei was used to adjust the count taken by the fieldworker to obtain an estimate of the number of Nikkei passing the intercept location.

Refusal rates were recorded, along with the sex and approximate age of the person refusing.

In the case of surveys carried out at events, a possible concern is that the same person might circle past the location multiple times, thereby invalidating the count. Therefore the fieldworker counted the total number of individuals passing during a short periods (10-minute duration) so as to minimise the chance of re-entries, and among them counted the number of persons thought to be Nikkei adults. Estimates of the total number attending the event obtained from the event organisers were then multiplied by the sample proportion observed to be adult Nikkei to get an estimate of the number of adult Nikkei attending the event.

Table 10.6. Some operational characteristics of the location survey in Sao Paulo

<table>
<thead>
<tr>
<th>Fixed point locations</th>
<th>(1) number of Nikkei at location</th>
<th>(2) number of interviews (attempted)</th>
<th>(3)=(2)/(1) sampling rate (%)</th>
<th>(4) number of interviews (completed)</th>
<th>(5) refusal rate (%)</th>
<th>(6) time spent in location (hours)</th>
<th>(7) time per interview (mins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>368</td>
<td>57</td>
<td>15.5</td>
<td>34</td>
<td>40</td>
<td>8.5</td>
<td>15</td>
</tr>
<tr>
<td>(2)</td>
<td>1,436</td>
<td>86</td>
<td>6.0</td>
<td>49</td>
<td>43</td>
<td>14</td>
<td>17</td>
</tr>
<tr>
<td>(3)</td>
<td>1,282</td>
<td>37</td>
<td>2.9</td>
<td>34</td>
<td>8</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>(4)</td>
<td>1,635</td>
<td>37</td>
<td>2.3</td>
<td>34</td>
<td>8</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>(5)</td>
<td>374</td>
<td>54</td>
<td>14.4</td>
<td>42</td>
<td>22</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>(6)</td>
<td>2,583</td>
<td>130</td>
<td>5.0</td>
<td>54</td>
<td>59</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>(7)</td>
<td>1,922</td>
<td>105</td>
<td>5.5</td>
<td>36</td>
<td>66</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>(8)</td>
<td>311</td>
<td>59</td>
<td>19.0</td>
<td>34</td>
<td>42</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>(9)</td>
<td>186</td>
<td>58</td>
<td>31.2</td>
<td>34</td>
<td>41</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>Events</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td>97</td>
<td>53</td>
<td>54.6</td>
<td>34</td>
<td>36</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>(2)</td>
<td>824</td>
<td>52</td>
<td>6.3</td>
<td>30</td>
<td>42</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>(3)</td>
<td>424</td>
<td>30</td>
<td>7.1</td>
<td>29</td>
<td>3</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>(4)</td>
<td>67</td>
<td>6</td>
<td>9.0</td>
<td>4</td>
<td>33</td>
<td>9</td>
<td>135</td>
</tr>
<tr>
<td>(5)</td>
<td>704</td>
<td>37</td>
<td>5.3</td>
<td>34</td>
<td>8</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>(6)</td>
<td>1,731</td>
<td>40</td>
<td>2.3</td>
<td>34</td>
<td>15</td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td>Total</td>
<td>13,944</td>
<td>841</td>
<td>6.0</td>
<td>516</td>
<td>39</td>
<td>128.5</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 10.6 provides some information on operational characteristics of the survey, such as the sample size, refusal rate and mean interview duration. The total number of Nikkei visiting the 15 locations during the sampling period was almost 14,000. In all, 516 interviews were collected, along with 325 refusals. The average refusal rate is thus as much as 40 per cent, with location-specific refusal rates ranging from only 3
per cent at the food festival (“event” (3)), to over 60 per cent at the two grocery stores (“fixed-locations” (6) and (7) in the table).

F. Frequency of visits and selection probabilities

At each location, individuals were asked whether or not they had visited any of the other fixed point locations during the past two weeks, and whether they had attended or were planning to attend any of the six events. Over 80 per cent of individuals had visited at least one other location, the average being 3.2 locations. As many as 12 per cent of individuals had visited 6 or more of the locations, with one individual going to 13 out of the 15 locations.

In order to examine the characteristics of individuals who visit more locations amongst those sampled, regression was performed on the number of locations visited as a function of several characteristics. It was found that females and older individuals visited more locations, as did return migrants. Individuals who refused to give their income range were found to visit fewer locations. These results showed that individuals who were more strongly linked to the Nikkei community were more likely to be included in the survey. Such bias needs to be reduced in the sample selection and/or the estimation procedure through weighting.

The probability that an individual appears in the sample is proportional to the number of locations which the individual visits from which he/she could be selected, and the sampling rates at the locations. Precise expressions covering various situations are given in Section 10.5 below.

The results clearly show that weighting reduces the proportion of individuals with strong linkages to Japan. Weighting is found to reduce the estimate of the proportion of Nikkei who are first generation migrants, and the proportion who are living in households where someone reads Japanese newspapers, listens to Japanese radio, watches Japanese television, or checks Japanese websites.

Differences in response rates are also important; indeed, they may be incorporated into the weights. Refusal rates for males and females were very similar (37 vs. 40 per cent), but refusal rates differed by age, from 25 per cent for persons aged above 50, to as high as 45 per cent for persons aged under 50.

Compared to the household-based (presumably the most representative) survey, the location survey shows a strong bias, though less extreme than the snowball sample. The location sample bias is reduced after weighting based on the frequency of visits to the locations observed. A few results are summarised in Table 10.7.
10.5 Probability of selection and sample weights of a mobile individual

10.5.1 The issue

When the sample is taken at an ‘observation point’ at a given ‘time segment’, the probability of an individual being included in the sample is affected by the person’s frequency and pattern of visits to the observation sites. When we need a sample not of visits but of individuals – as is the normal objective in a survey of child labour – it is necessary to take these variations in individual frequency and pattern of visits into account.

The situation can be complex. An individual may:

- visit different locations;
- with different frequencies (including never visiting);
- for different durations of stay;
- with a fixed, or a random, or more likely, a mixed pattern; and
- with patterns differing according to the type of location.

Sampling arrangements may also differ both in the periods covered and the sampling rates applied:

- all locations may be included, but each for a sample of the time segments covered in the survey;
- the time segments may not cover the whole period the location is available for visits, and in that case, the time periods may be a random sample, or a purposively selected one (for example during rush hours when most of the visits take place);
- different locations may be sampled at different times;
- only a sample of locations may be enumerated, locations selected using different
schemes, such as equal probability sampling, or probability proportional to size sampling;
- within a location, different observation points may be sampled at different rates;
- at a PSU (an observation point at specified time segment) different individuals may be sampled at different rates;
- at a PSU, the sampling rates may vary over time (e.g. according to the volume of the flow of visitors); and
- there may be more than one sampling occasion, for instance a survey repeated twice in a week, during one week day and one week-end day.

Complex situations require a lot of information to be collected in order to fully take into account the variations in selection probabilities. Collecting all the required information is often not possible; sometimes it is not even desirable to do so in an interview situation, so as to avoid excessive respondent burden and adverse effect on data quality. In view of this, it is necessary to have an alternative, more practical strategy. Such a strategy would have the following three elements.

(1) Firstly and most importantly, the problem cannot be ignored. At least the main source of variation in selection probabilities should be taken into account, hopefully minimising and simplifying the additional information which has to be collected for the purpose.

(2) When the objective of taking all the sources of variation in selection probabilities fully into account cannot be met, making some simplifying assumptions about the nature of the variation may make the problem more manageable – for example, assuming some actually non-random pattern to be a random one, or replacing some variable values by their average.

(3) Situations with different levels of complexity may be seen as forming a hierarchy with different levels of information needs. When the situation at a given level of complexity cannot be handled, one may ignore some sources of variation in probabilities and adopt a solution available for a some lower level of complexity.

**10.5.2 Some solutions**

In the following, some quantitative expressions are developed for variations in individual selection probabilities in a number of commonly encountered situations.

**A. The simplest cases**

(1) Suppose that we have only one location, and individuals are sampled at a constant rate at that location during one time segment selected at random from the time interval covered. Let us also assume that individuals visit that location during the time segment selected. However, individuals may also make varying numbers of visits to the location during other time segments within the interval covered.

Let us express the above in specific terms. Suppose the survey is being carried out at a worksite where children come to work and normally stay for the whole day. The site is open 5 days during the week, but children visit it for different number
of days – from 0 (never) to 5 (every day it is open). One day during the week is selected at random for the survey.

The probability of an individual being present at the location at the time of the survey is obviously \( p_i = (i/5) \). The relative weight given to such an individual compared to others in the sample is \( w_i = (1/p_i) = (5/i), i > 0 \). A person who never visits the location \((i = 0)\) can never be selected \((p_i = 0)\), and hence is not in the population covered. A person who visits the location every day during the week \((i = 5)\) is always available to be selected \((p_i = w_i = 1)\).

(2) Now suppose that the survey is carried out on two different randomly selected days, and that if an individual is selected on both days, the second selection is discarded. The other assumptions are the same as in (1) above.

Obviously, a person visiting the location 4 or 5 times during the week will always be available for selection during at least one of the survey days, hence \((p_4 = p_5 = 1)\).

For \( i < 4 \), the probability that a person is absent on both the sampling occasions is:

\[
\left(1 - \frac{i}{5}\right)\left(1 - \frac{i}{4}\right) = \frac{(5-i)(4-i)}{20}.
\]

The first factor is for the survey on the first occasion. The second factor is for the probability of not being present at the second occasion, which is on one of the remaining 4 days of the week. The probability of being present on at least one of the two occasions is the complement of the above:

\[
P_i = 1 - \frac{(5-i)(4-i)}{20}.
\]

(3) Similarly, if the survey is conducted on three different occasions during the week, then for \( i \geq 3 \), the person is present at least during one occasion. For \( i < 3 \), the probability of not being present at all the three occasions is:

\[
\left(1 - \frac{i}{5}\right)\left(1 - \frac{i}{4}\right)\left(1 - \frac{i}{3}\right) = \frac{(5-i)(4-i)(3-i)}{60},
\]

and of being present on at least one of the survey days is the complement of the above.

The results of the above three examples are shown in the table below.

<table>
<thead>
<tr>
<th>Table 10.8. Frequency of visits and selection probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of survey occasions</td>
</tr>
<tr>
<td>----------------------------</td>
</tr>
<tr>
<td>Probability of being present at the location during the survey ((p_i))</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>and the corresponding relative weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>
(4) Now consider a survey conducted on two occasions, but both on the same day of the week over two weeks. The situation is more complex than the above, and statistically less desirable. If the day of the week when a particular person visits the location is random, the situation is the same as repeating (1) above twice. The day of the visit being at random is the same thing as the survey occasion being at random. The probability that a person is absent on both survey occasions is \((1 - i_1/5)(1 - i_2/5)\)\(^{41}\).

But if a person visits the location always on a fixed day of the week, repeating the survey on the same day of the week does not help. The situation for persons in this category is the same as (1) above.

Usually, there is a mixture – for example many persons frequently coming on the same days of the week but not always. Such individual variations are not known to the survey and hence correct probabilities of selection cannot be determined.

Hence if a survey has to be repeated over two weeks, it should be on different days of the week.

B. Survey repeated during the week and the week-end

(5) Suppose the survey is carried out on one randomly chosen day during the 5 week days, and on one randomly chosen day during the 2 week-end days. Individual \(i\) is present at the location at random during \(i_1 = 0 - 5\) days during the week and \(i_2 = 0 - 2\) days during the week-end. Other assumptions are the same as those in the previous examples.

The probability of not being present during the week day survey is, as before,

\[(1 - i_1/5), \; i_1 = 0 - 5,\]

and of not being present at the location during the week-end survey is

\[(1 - i_2/2), \; i_2 = 0 - 2.\]

The probability of missing both days is the product of the above, and of being present at least on one occasion is the complement of that:

\[P_i = 1 - \left(1 - \frac{i_1}{5}\right)\left(1 - \frac{i_2}{2}\right).\]

The results are shown in Table 10.9.

<table>
<thead>
<tr>
<th>(i_1)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i_2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

\(^{41}\) The situation would be the same as case (2) above if the individual always changed the day of visit from one week to the next.
The corresponding weights are the inverse of the above.

(6) Now suppose that the sampling rates for selecting individuals at a location during the week day and the week-end are different, say $f_1$ and $f_2$ respectively. The week-end may be crowded for example, so that a lower sampling rate is taken, $f_2 < f_1$. The probability of a person being selected for the sample of the week day survey is $f_1 (i_1 / 5)$ and for week-end survey is $f_2 (i_2 / 2)$. With the same logic as (1) above, the probability of a person being included in the sample of at least one of the survey occasions is

$$P_i = 1 - \left(1 - f_1 \frac{i_1}{5}\right) \left(1 - f_2 \frac{i_2}{2}\right).$$

In estimation from the survey, the weight of the individual will be the inverse of the above.

C. Location or ‘centre’ sampling

(7) In Section 10.2 we discussed the situation when each person $(j)$ in the population visits a subset of locations in the survey, but can always be found at the location(s) which he/she does visit. Let $f_i$ be the sampling rate applied (to all individuals visiting) at location $i$. For a person who visits (is a member of) a subset $I_j$ of locations, the probability of being selected into the survey from one of these locations is

$$p_j = \sum_{i \in I_j} f_i,$$

i.e. the sum of selection rates over the locations that the person visits.

Of course, we may also have variation in individual selection probabilities within a location, giving the more general form of the above:

$$p_j = \sum_{i \in I_j} f_{ij}.$$

Actually $p_{ij}$ is not a ‘probability’ (since it can exceed 1), but the ‘expected number of times’ the individual is selected into the survey. The inverse of these probabilities are used as weights in the Hansen-Hurwitz type of estimator. Please refer to Section 10.2.3C where three different estimators have been described, namely:

- Horvitz-Thompson estimator, Equation (10.7);
- Hansen-Hurwitz estimator, Equation (10.8);
- Multiplicity estimator, Equation (10.9).

As noted, the multiplicity estimator is particularly convenient. Statistically this form is equally valid (is similarly unbiased) as the other two. A practical problem with the other two forms is that in order to compute the selection probability of an individual and hence its inverse, the weight to be applied in estimation, it is necessary to have information on the sampling rates $f_{ij}$ at all the locations $(i)$ the individual $(j)$ visits. This information may not be easy or even possible to obtain, especially if the sampling rate at a location is variable. An alternative expression for the selection probability is provided by the multiplicity estimator (10.9), which
in the present notation is:

\[ p_{ij} = l_j f_{ij}, \quad (10.10) \]

where \( i \) indicates the location where the data being used for producing the estimate are coming from. Information on sampling rates at other locations is not required.

**D. Individual variations in frequency of visits**

(8) Case (7) above assumed the situation that a person in the population visits a subset of locations in the survey, but can always be found at the location(s) which he/she does visit. This is a special case of the more general situation when individuals visit locations of interest in the survey at different frequencies. Account has to be taken of the probability \( g_j \) that person \( j \) is present at some location to be available for the survey. For example, if a person visits a location in the survey twice during the reference 5 week days, \( g_j = 2/5 \) as explained in Subsection A above.

The multiplicity estimator (10.10) is simply modified to:

\[ p_{ij} = l_j g_j f_{ij}. \quad (10.11) \]

To remind, the above quantities are as follows.

\( I_j \) Number of locations in the population individual \( j \) visits at some time.

\( g_j \) Proportion of the time segments he/she visits any of these locations. It is assumed in (10.11) that this factor depends only on the individual, and does not vary over locations for a given individual. This may be a convenient and practical application, or the factor may be seen as a value averaged over locations visited by the individual.

\( f_{ij} \) Probability of selection of individual \( j \) at location \( i \); \( i \) being the particular location where the data being used for the estimation are coming from.

In the case of multiple selections of the same individual, the process is repeated in the same way each time the individual is selected into the sample.

**E. Sampling of locations**

(9) Suppose that we have \( N \) locations and an equal probability sample of \( n \) locations is selected for the survey. Within the sample of locations, the situation is assumed to be the same as that in case A(1) discussed earlier. The results are the same but with one important difference. Quantity “\( i \)” refers not to the number of visits a person makes to the location at which he/she is selected for the survey, nor to visits to any of the \( n \) locations in our sample, but visits to any of the \( N \) locations in our population.

Only when \( N \) is trivially small can one actually specify to the respondent the full list of locations about which information is being sought, such as through a question like ‘how often do you visit any of the locations [list]?’. Normally, what is included in \( N \) has to be communicated to the respondent in vaguer terms, such as through a question like ‘how often do you visit a location [like the present one] in this
locality?’. If the survey covers a large area, we may have to ask even vaguer questions such as ‘how often do you visit any place like this one?’.

Another point to note is that in the survey questions it should be clear to the respondent that the number of visits over the reference period being asked is inclusive of visits to the particular place where the interview is being held. Sometimes it may be better to ask two separate questions, one for the place where the respondent has been selected and is being interviewed, and another to cover other similar places visited by the respondent, for example: ‘how often do you visit this place?’, followed by ‘how often do you visit other places like this one?’.
Chapter 11
Capture-recapture sampling

11.1 Introduction

11.1.1 Applications of the procedure in different fields

The capture-recapture sampling technique involves taking two (or more) independent samples from the same population and using the overlap found between the samples to estimate, under certain assumptions, the selection probabilities applied to obtain those samples and hence the total population size.

Extensive use of the capture-recapture method is found in ecology and biometrics, particularly in estimation of the size and density of animal populations. (A classic on this topic is the book by Seber, 1982). In enumerating wild life, obviously no lists are available for sample selection, and unit selection probabilities are not known even if a random sample has been selected.

The capture-recapture sampling technique has been increasingly adapted for health-related studies of human populations, especially concerning medicine and public health (Bohning 2008; Tilling 2001), for example in order to obtain more accurate estimates of the prevalence of disease and disability (various types of cancer, diabetes, stroke and HIV). Landmark papers on these applications include Hook and Regal (1995), and IWGDMF (1995a and 1995b). Similar applications include studies of injuries and traffic accidents (for instance, Morrison and Stone, 2000).

These applications involve estimating the total number of cases of interest using multiple sources of information, none of which is complete by itself. Typically the sources of information are registers and administrative records of various types, rather than independent surveys specially conducted for the purpose.

Of more direct interest in the study of child labour is the extension of the approach for studying the condition of human populations, especially of populations difficult to enumerate using normal sample survey procedures. The population of interest may be elusive because it is rare, mobile, hidden, or reluctant to participate in the survey, or because a frame for obtaining adequate samples of the population is lacking.

Below are some examples of the use of the capture-recapture methods for the study of an elusive human population:

- numbers of the homeless (Fisher, Turner, Pugh and Taylor, 1994; Williams, 2010);
- numbers of street children in a city (Gurgel et al., 2004);
- size of criminal population from police records (van der Heiden, Cruyff and Houwelingen, 2003);
size of the lesbian population (Aaron, Chang, Markovic and LaPorte, 2003);
family violence (Oosterlee, Vink and Smit, 2009);
prevalence of drug misuse at local level (Hay and McKeeganey 1996; Hay et al., 1997); and
census undercount (Hogan, 1993; Fienberg, 1992).
An early and well-known example in the population field is from Sekar and Deming (1949), who estimated the extent of under-coverage in birth and death registration at the local level.
Problems encountered in investigating children engaged in specific sectors or hazardous forms of child labour can be similar to those encountered in studying elusive populations of the type mentioned above, where capture-recapture methods have been used. Therefore we can expect these methods to be useful for studying child labour as well.
Capture-recapture applications in the field of ecology tend to be based on multiple small surveys, those in epistemology are often based on sets of administrative sources, while applications in the social field may be based on either type of source or – more likely in developing countries in particular – a combination of sample surveys and administrative sources.
The following are some examples of studies using the capture-recapture method on the basis of multiple types of source. A study using:
- 8 different administrative sources in a study of family violence (Oosterlee, Vink and Smit, 2009).
- 4 mailing lists in study of a lesbian population (Aaron, Chang, Markovic and LaPorte, 2003).
- Centres of 4 types providing services to homeless and mentally sick (Fisher, Turner, Pugh and Taylor, 1994).
- 4 samples of observations at bars to study sex workers (Watts, Zwi and Foster, 1995).
- 3 different sources, including 2 registers (for births and deaths) plus 1 survey to evaluate the completeness of registration (Sekar and Deming, 1949).
- 3 sources, including 2 surveys plus 1 administrative list for a street children survey in Brazil (Gurgel et al., 2004).
- 2 independent samples to investigate profile of street children in Cairo, Egypt (CAPMAS 2009).
- 2 samples of observation of female sex workers at bars in Bulawayo, Zimbabwe (Weir et al., 2003).
- 2 sources based on lists from agencies, repeated on 3 occasions, to study homelessness in two towns in the UK (Williams, 2010).
- 1 source only, estimating the size of a criminal population, based on truncated Poisson distribution (van der Heiden, Cruyff and Houwelingen, 2003).
11.1.2 Brief outline of the approach

A. The fundamental equation

It is best to begin with the simplest situation. Suppose we have a population of size $N$. Its size is not known and the objective is to produce a reasonable estimate of it. Suppose that we select two simple random samples of sizes $n_1$ and $n_2$ from the population. The two samples are drawn independently and the population remains unchanged between the two draws. Let $m_2$ be the number of units in the second sample which were also present in the first sample. These overlapping units can be identified by, for example, ‘marking’ all the units selected in the first sample, and then identifying the marked units in the second sample. (In surveys of individual persons, usually the marks are provided by the individuals’ basic socio-demographic characteristics, such as sex, age, name, location of residence, sometimes an official identification number.)

The basis of the method is the following.

The proportion of first sample cases found in the second sample ($m_2/n_2$) provides an estimate of the same proportion in the whole population ($n_1/N$). That is:

$$\frac{m_2}{n_2} = \frac{n_1}{N},$$

where $\hat{N}$ is an estimate of the required parameter $N$. This gives

$$\hat{N} = \frac{n_1}{m_2},$$

(11.1)

Note that by definition

$$\hat{N} = (n_1 - m_2) + (n_2 - m_2) + m_2 + n_0,$$

(11.2)

where terms on the right are, respectively,

- units present only in the first sample, but not in the second ($n_1 - m_2$)
- units present only in the second sample, but not in the first ($n_2 - m_2$)
- units present in both $m_2$
- units not present in any of the samples (say $n_0$).

It can be seen that by substituting for $\hat{N}$ from (11.1), (11.2) provides an estimate of the numbers missed in both samples as:

$$n_0 = \hat{N} - (n_1 + n_2 - m_2) = (n_1 - m_2)(n_2 - m_2)/m_2.$$  

(11.3)

For example, if size of the first sample is $n_1 = 100$ and of the second sample is $n_2 = 80$, and the overlap between the two is $m_2 = 32$ units, then estimated population size is:

$$\hat{N} = 100(80/32) = 250.$$  

The number not appearing in either sample is:

$$n_0 = (100 - 32) \times (80 - 32)/32 = 102.$$
The number appearing in at least one of the samples is, by definition,
\[(n_1 + n_2 - m_2) \text{ or } (\hat{N} - n_0), \text{ i.e. } (100+80-32) \text{ or } (250-102) = 148.\]

The above model is based on certain strong assumptions, which are discussed in later sections. In this way the capture-recapture approach differs from standard sampling in so far as the latter is free from assumptions concerning distribution of the underlying population. Some assumptions behind the capture-recapture method are more critical than others, in the sense that departures from them affect the validity of the method more seriously. If the assumptions - at least the most critical ones – are reasonably well satisfied in a real situation, then the method would yield useful estimates. Otherwise the results can be biased and misleading.

Fortunately, analytic methods have been developed to accommodate some of the common departures from the assumptions of the basic model. These methods have helped to greatly increase the range of situations to which the capture-recapture methodology can be fruitfully applied.

B. Modified version

Often a slightly modified version of (11.1) is used, which has less (statistical) basis, especially in dealing with situations where overlap may be zero, \(m_2 = 0\). It is
\[
\hat{N} = \frac{(n_1 + 1)(n_2 + 1)}{(m_2 + 1)} - 1.
\] (11.4)

Its variance estimation (assuming simple random sampling) has been given by Seber (1982) as
\[
\var(\hat{N}) = \frac{(n_1 + 1)(n_2 + 1)(n_1 - m_2)(n_2 - m_2)}{(m_2 + 1)^3(m_2 + 2)}.
\]

The important point is that the coefficient of variation of \(N\), \(c = se(N)/N\), is approximately
\[
c = \frac{1}{\sqrt{m_2}}.
\] (11.5)

This means that the precision of estimating \(N\) is almost solely dependent on the size of the overlap \(m_2\) between the two sources (Seber 1982, Section 3.1.2).

C. More general approaches

The basic model of equation (11.1) applies to the situation in which we have two random samples drawn from a closed population. Most survey-based studies of child labour would conform to this structure. More elaborate methods have been developed to deal with more than two samples drawn from the same population. Use of more than two sources is more likely to be the case when administrative source (registers, lists, etc.) are being used in place of, or to supplement, data from sample surveys.

A closed population means a population not subject to changes due to migration, births and deaths. An extension of the capture-recapture methodology is to allow for an open population subject to changes due to migration, births and deaths. However, in this chapter throughout, we will take the population to be fixed. This assumption is
more reflective of the situations generally encountered in child labour studies which are conducted mostly over a relatively short survey period. It is also simpler and more practical.

11.1.3 Sort of situations in which capture-recapture sampling might be useful

In equation (11.1), the estimated population size ($\hat{N}$) is inversely proportional to the extent of overlap ($m_2$) between two independent samples drawn from the same population. The overlap is a random variable, and can be subject to large fluctuations in relative terms if its magnitude is small. For example, $n_1 = 50, n_2 = 40, m_2 = 5$ would give $\hat{N} = 400$ as the estimated size of the population. It is highly likely that due to random variation $m_2$ turns to be a little larger, say $m_2 = 6$ giving a lower estimate $\hat{N} = 330$ approximately; or $m_2$ may turn out to be a little smaller, say $m_2 = 4$ giving a higher estimate of $\hat{N} = 500$.

From equation (11.1), the expected value of the overlap $m_2$ is:

$$m_2 = \left(\frac{n_1}{N}\right)\left(\frac{n_2}{N}\right)N.$$  \hspace{1cm} (11.6)

The first two factors on the right are the sampling fractions in, respectively, the first and the second samples. In order to avoid $m_2$ becoming small, these need to be reasonably large, unless $N$ is very large.

In fact small values of the sampling rates not only increase variance of the estimate $\hat{N}$, but also introduce a bias of overestimation.

“If the average probabilities of capture on each sampling occasion are less than 0.1 then $\hat{N}$ is generally significantly biased and the calculated confidence interval tends to be impractically wide. In this case the interval still provides relevant information, though about the failure of the experiment rather than about the population size. It is generally recommended that the average probability be greater than 0.2 ... for a reasonably short confidence interval, unless $N$ is very large. If the average probability of capture is too small, say less than 0.1, it may be a waste of time carrying out the experiment.” (Seber 1982, Section 12.8.6).

As noted earlier, the multiple sources of information used in applying the capture-recapture method may be sample surveys or administrative lists, or a combination of the two. In developing countries, usually administrative sources are lacking and we have to rely on sample surveys specially conducted for the purpose. In practice, such surveys cannot be large, say no more than a few hundred (or at the most a couple of thousand) cases. This means that, in order to keep the sampling rates large enough, the study population should also be small. For example with $n_1$ (or $n_2$) = 200, achieving $(n_1/N) \geq 20\%$ means that $N \leq 1,000$ - the target population should not exceed around 1,000 units. This implies that the capture-recapture method based on multiple sample surveys is best suited when the target population – such as of labouring children – is small, and preferably concentrated in well-defined local areas.
Somewhat different considerations may apply when the capture-recapture application is based on administrative lists or registers. A list can be seriously incomplete but may still cover a large proportion of the population, i.e. have large \( \frac{n_1}{N}, \frac{n_2}{N} \). However, problems can arise from the manner in which the lists are related.

Often different lists do not cover the same population. In the extreme case different lists may cover different, non-overlapping, parts of the target population. This would mean no overlap, i.e. \( m_2 = 0 \), so that capture-recapture based equation like (11.1) cannot be used to estimate \( N \). For example consider two lists covering child labour in two different sectors both of which are a part of the target population. There would be no overlap between the lists by definition, and we cannot use the capture-recapture estimate (11.1) based on these lists. In any case, if the populations covered overlap only partially, observed \( m_2 \) would tend to be too small, resulting in over-estimation of population size \( \bar{N} > N \).

The opposite problem also exists. Often lists are ‘dependent’ (for example, if they have been compiled using information from some common sources). This would tend to inflate overlap \( m_2 \) between them, resulting in under-estimation of population size \( \bar{N} < N \).

In the following sections we will elaborate on these and related issues.

### 11.1.4 An illustration: a profile of Cairo street children

The following is a brief description of the methodology of a recent study of street children in Cairo, Egypt, using the capture-recapture methodology. The study was based on two independent sample surveys covering the same population and conducted close to each other in time. In the following, we quote or paraphrase methodological description from the study report (CAPMAS 2009) in order to bring out some of the practical concerns in conducting a children’s survey using capture-recapture methodology.

Special effort was made in this survey to maximise the number of street children enumerated in each of the two surveys. The expectation is that such effort increases the achieved sample size (hence improved precision of the results) without essentially compromising the independence and SRS nature of the samples. The special effort may have also reduced bias by reducing the extent of undercoverage of the more-difficult-to-find street children.

**Methodology**

“Street children can be classified as a group of children in especially difficult circumstances. The circumstances of the work and the risks involved in many cases make these activities worst forms of child labour. There are broadly two groups of street children. Some street children are ‘on the street’, which means that they still see their families regularly and may even return every night to sleep in their family homes. Children ‘of the streets’, on the other hand, have no home but the streets. Even if they occasionally spend time in institutions for children or youths, they consider the streets to be their home. In this report, ‘street children’ is a term used to describe both children who work in the streets and markets of cities selling or begging and live with
their families, and homeless street children who work, live and sleep in the streets, often lacking any contact with their families. At highest risk is the latter group. ....

“Statistically speaking, street children are a ‘rare’ and ‘difficult to access’ population. ‘Rare’ because they represent a very small proportion of the overall child population, and ‘difficult to access’ because they tend to be elusive and to hide when they are not working. Several sampling methods have been devised by statisticians to study this type of population. A specific procedure called capture-recapture that makes it possible to generate data on the profiles of such population is used in this study. ....

“Sampling through capture and recapture ... requires producing two separate lists (one for the capture and one for the recapture), each representing a sample of the target population. The number of individuals on each of the lists is then computed, as well as the number of individuals who are on both lists. Analyses on the target population are then conducted on the basis of these three figures. ....

“In the case of this study in Cairo, the first ‘capture’ [Capture I] took place on 6th May 2009 in all the areas known for the presence of street children. Both the selection of the focus areas and the census of the street children were carried out with the help of key-informants. For the second capture (Capture II which took place on 11th May 2009), the sampling method was the same. Children included in both Capture I and Capture II were considered ‘recaptured’. Both in Capture I and Capture II several questions were posed to child beggars to gather data on their characteristics and their life conditions. This information was helpful in gaining a better understanding of the phenomenon of street children ....

“The validity of the capture-recapture estimates is based on four basic assumptions .... : (i) the population studied must be closed, meaning that it is not affected by births, deaths or migrations during the sampling process; (ii) each individual’s probability to be captured is different from zero; (iii) individuals already surveyed must be clearly identified; (iv) having been captured does not have any effect on the probability of being recaptured.

“In the case of street children in the Cairo area, the validity of the assumption concerning the closeness of the population can be argued. Street children cannot be considered as a close population as they are eminently mobile. The Capture I and Capture II phases were carried out in a short time, so that the number of children who could have left or arrived (or died) is negligible.

“In addition, researchers must make sure that for each street child the possibility of being captured is different from zero. During Capture I and Capture II all places where street children can be found were visited. However, other children who are not on the streets during the day or are more difficult to locate could be missed.

“Moreover, one of the main assumptions posits that researchers have to be able to clearly identify the children who have already been interviewed. This
The pilot survey

“On two successive days interviews were done with street children in Giza Governorate through using the capture and recapture method, the first interviews conducted on April 23rd, while the 2nd interviews conducted on April 27th, 2009….. Such above-mentioned areas were selected given the fact that they have the largest gatherings of street children… Work strategy adopted was to cover and survey all areas selected with an aim to train researchers on quick mobility and the ability to find all gathering points of street children to guarantee best results; same strategy was adopted in the main survey conducted in Cairo Governorate later. The total number of children interviewed through the two days was 332 [159 on day 1 and 173 on day 2 of the pilot] … Field work teams were selected from among the supervisors and field researchers of the well-experienced researchers of the Central Agency for Public Mobilization and Statistics (CAPMAS); four teams were chosen, each team consisting of six researchers and one field supervisor; with a total number of 4 supervisors and 24 field researchers. Each team was joined with 2 coordinators of NGOs that work with street children …. with an aim to integrate the community associations …. which have the selected areas under their jurisdictions …. This had a very good impact on facilitating the process … therefore the same was adopted in the main survey.

The main survey

“…. The field work conducted by 35 working teams distributed on all street children gathering points all around Cairo Governorate which were identified according to the experiences of NGOs dealing with street children and information sent by the Ministry of Interior Affairs. Each team was responsible to survey from 3 to 4 assembly points in the same geographical zone. Each team consisted of six members divided as follows; one coordinator from NGOs working with street children, a technical supervisor from CAPMAS and four field researchers from CAPMAS. [There were 114 Street children assembly points in the area covered by the survey population.] ..... “The movements of data-collecting researchers were according to a map that identified all the gathering points of street children; however all researchers were instructed to cover and to interview all street children whether found in groups or individually. Given that street children are a highly mobile population, all field researchers were told to scan all points identified as well as all neighbouring points with an aim to cover all street children. …

“Some external incidents on Day 1 of Cairo Street Children Survey affected the progress of work: the presence of many police raids aimed to catch street children which reduced the number of street children on these points; the funeral proceedings of the spouse of the Prime Minister that was held in a Mosque in Nasr City and tight security measures accompanied that
11.2 Assumptions in the capture-recapture method

11.2.1 Basic Assumptions

The capture-recapture procedure is based on certain assumptions about the population and the manner in which the samples have been drawn from it. The validity of the estimates drawn from the procedure depends on the validity of the underlying statistical assumptions. These assumptions are not always valid and, consequently, the results of capture-recapture studies should be interpreted with due caution. Wherever possible, the results should be cross-validated against other sources providing similar information.

Nevertheless, the usefulness of the capture-recapture technique arises from two factors. The first is the fact that often the procedure is found to be ‘robust’ against (insensitive to) departures from the assumed statistical model. Secondly, statistical procedures have been developed to accommodate (control the effect of) certain departures from the original simple model.

Capture-recapture studies are normally designed with the aim of meeting the following basic assumptions.

1. The target population is not subject to births, deaths, migration or other changes. This is referred to as the target population being ‘closed’.

2. Each source used in the capture-recapture estimation is a representative sample of the entire target population.

3. The different sources are linked at the micro-level, clearly and without ambiguity or errors. This means that for every unit, unambiguous information is available on its appearance or non-appearance in each source.

The above three can be considered as the fundamental assumptions for any capture-recapture application. A simple and commonly used model such as equation (11.1) makes some additional assumptions.

4. The different sources are ‘independent’. This means that the chance of a unit being present in one source is not dependent on whether or not the unit is present in the other source(s).
(5) All units have the same chance of being selected into a sample (or being present in another source). This chance may differ from one sample/source to another, but is uniform within each.

(6) More specifically, for the simple model, each source is an independent simple random sample.

Assumption (4)-(6) are unlikely to hold exactly in practice. Methods have been developed to take into account the effect on survey estimates of some departures from these assumptions.

At the same time, it is desirable to design the capture-recapture study so as to minimize any such departure. This helps to reduce complexities in the estimation procedure, and in the analysis and interpretation of the results.

As in practice there is almost always some departure from these assumptions, the issue is whether any departures significantly affect the conclusions reached from the data.

The assumptions are discussed in turn in the following.

### 11.2.2 Closed population

**A. Population changes**

The assumption is that, over the period to which the data sources refer, the population is not subject to changes due to births, deaths, migration, and other factors affecting units’ presence or characteristics.

Actually, the model can accommodate certain completely random changes in the whole population, changes not related to the inclusion of units into the samples. In addition, some systematic changes are automatically taken into account. For example, if additions to the population take place between two samples, then the overlap between the samples cannot contain any of the new units. Equation (11.1) estimates the population at the second sample, following the additions to the population. Similarly, if deletions (deaths, out-migration) from the population take place at random (i.e. are unrelated to the selection of units into the first sample), then equation (11.1) provides a valid estimate of the population size at the time of the first sample, before the deletions from the population.

**B. Differences in coverage of the population**

*When the capture-recapture application is based on sample surveys, the practical way to minimize changes in the population is to minimize the time interval between the surveys.*

Another factor in relation to the timing of the surveys is often also important. In many situations, even when the target population is fixed, the availability of individuals in it for inclusion in a sample has a cyclic pattern, and furthermore these patterns may differ for different parts of the population. For example, some children engaged in a certain sector may work in the morning, and others in the afternoon. Similar differences may be present in relation to working during the week versus working over weekends, working only during the weekends, etc. In the extreme case, different parts of the
11.2 Assumptions in the capture-recapture method

population may have completely non-overlapping timings of work – e.g. some children working only in the morning, others only in the afternoon. To be comparable, and for a relationship like (11.1) to apply, the two surveys must be conducted at the same points in the cycle of availability of individuals. Furthermore, in order to cover the whole population, each survey should cover representative samples from all parts of the cycle.

Similar problems of overlap between sources can arise for capture-recapture applications based on administrative sources. For the approach to be valid each list needs to cover the whole population, albeit incompletely. (The requirement is that it should be possible to regard the list as a random or stratified random sample from the whole population.) However, lists are often restricted to a specific subpopulation, for example:

- lists covering children in a particular sector, when other sectors of work are also a part of the study population;
- lists covering only homeless children from among all children working in a sector of interest;
- lists covering only those working children who are also attending school; or
- lists covering restricted geographical areas such as only the city centres.

Only lists meant to cover the same population can be regarded as parallel data sources for applying the capture-recapture method.

This may involve partitioning the population according to which data sources are available, and applying the capture-recapture method to each part separately based on the subset of data sources common to that part.

Alternatively, or in addition, lists covering different parts of the population may be combined together to form a single data source covering the whole population; in this case application of the capture-recapture method would require an additional data source with common coverage.

Sometimes different scenarios may be found in different parts of the target population: coverage of the lists overlapping in one part and disjoint in another. These should be separated out by stratification and treated differently. In more mixed situations which cannot be separated out, one may have to decide as to which of the two modes predominates: (a) ‘largely’ common coverage, or (b) ‘largely’ separate coverage.

Assuming (a), largely common coverage, when actually the coverage situation is mixed tends to overestimate the size of the population. If (b), largely separate coverage, is assumed when actually the situation is mixed, lists must be matched at the micro-level to eliminate duplicates.

C. Quantification of the effects of population differences

The simple illustrations below explain the above issues in quantitative terms. The same example has been used to illustrate both the aspects discussed above, namely changes in the population over time, and differences among sources in population coverage.
(1) Changing population

Consider population at time 1 to consist of two parts: a part of size \( N \), which remains unchanged at time 2; and a part of size \( A = cN \) (say) which is completely replaced in the new population at time 2 as \( A' = c'N \).

At time 1, a simple random sample (SRS) at rate \((n/N)\) is taken, thus giving samples of size \( n \) and \((c\cdot n)\) from the two parts \( N \) and \( A \) respectively, or the total sample as \( n_1 = n(1+c) \).

Similarly at time 2 a SRS at rate \((n'/N)\) gives \( n_2 = n'(1+c') \).

Let \( m_2 \) be the observed overlap between the two samples. If we ignored (or were unaware of) the turnover in the population, an estimate using equation (11.1) for the (assumed constant) population size would be:

\[
\hat{N} = \frac{n_1 n_2}{m_2} = \left\{ \frac{n \cdot n'}{m_2} \right\} (1+c) (1+c').
\]

The sample overlap can occur only in the overlapping part (intersection) \( N \) of the two populations. Therefore the first factor in (11.1) is the correct estimate of this overlapping population:

\[
\hat{N}_{I(1,2)} = \left( \frac{n \cdot n'}{m_2} \right).\]

(Subscript \( I \) (1,2) is used to indicate the intersection between the populations at times 1 and 2.)

If values of the parameters \( c = A/N \), \( c' = A'/N \) were known, estimates of sizes of the two populations would be:
11.2 Assumptions in the capture-recapture method

\[ \hat{N}_{(1)} = (N + A) = \hat{N}_{(1,2)} (1 + c), \]

\[ \hat{N}_{(2)} = (N + A') = \hat{N}_{(1,2)} (1 + c'), \]

giving the estimated size of their average as

\[ \hat{N}_{av} = \hat{N}_{(1,2)} \left[ 1 + \frac{c + c'}{2} \right]. \] (11.9)

Noting that (11.7) can be written as

\[ \hat{N} = \hat{N}_{(1,2)} (1 + c)(1 + c'), \] (11.10)

which over-estimates the average population size (11.9) by the factor

\[ \frac{\hat{N}}{\hat{N}_{av}} = \frac{(1 + c)(1 + c')}{1 + (c + c')/2} = 1 + \frac{(c + c')/2 + c c'}{1 + (c + c')/2} = 1 + c_a \left( \frac{1 + c_a^2/c_g}{1 + c_a} \right), \]

where \((c_a, c_g)\) are respectively the arithmetic and geometric means of \((c, c')\).

Its meaning becomes clearer if we simplify it by assume \(c = c'\), which gives the overestimation factor as

\[ \frac{\hat{N}}{\hat{N}_{av}} = (1 + c). \] (11.11)

Ignoring a population turnover of proportion \(c\), the capture-recapture formula (11.1) overestimates the average population size by factor \((1 + c)\). For example with 10 per cent turnover in the population between the two surveys, (11.1) would overestimate the average population size by around 10 per cent.

(2) Different coverage

The situation is very similar when we consider two lists, each with a partial coverage of the population. Suppose the population consists of \((N + A + A')\), and the scope of coverage of the first list is \((N + A)\) and that of the second list is \((N + A')\). Using the quantities defined above, the correct estimate of the population size (‘union’ of the two lists) would be

\[ \hat{N}_{U(1,2)} = (\hat{N}_{(1,2)} + A + A') = \hat{N}_{(1,2)} (1 + c + c'). \]

Estimating \(\hat{N}\) using equation (11.1) overestimates this size by the factor

\[ \frac{\hat{N}}{\hat{N}_{U(1,2)}} = 1 + \frac{(c + c') + c c'}{1 + (c + c')} = 1 + \frac{c c'}{1 + (c + c')} . \]

Again assuming \(c' = c\) in order to clarify the implication, the overestimation factor becomes

\[ \frac{\hat{N}}{\hat{N}_{U(1,2)}} = 1 + \frac{c^2}{1 + 2c} < 1 + c^2. \] (11.12)
The overestimation resulting from ignoring the mismatch in coverage of the two lists is small for moderate values of $c$, and is negligible for small values of $c$. For example, for $c=20$ per cent, the overestimation is by around 4 per cent; for $c=10$ per cent, it is only by 1 per cent.

It is for this reason that often in practice the method can give reasonable results even with lists differing to some extent in their coverage of the population. Of course the overestimation increases with increasing difference in the coverage of the lists.

### 11.2.3 Representative samples

That a sample is ‘representative’ is not a precise concept but is a useful one. It implies a certain level of confidence in the reliability of the estimates produced from the sample. This does not preclude the possibility that weighting of the sample data may be required for the purpose.

A representative sample means that it is obtained by applying a random procedure, so that it is a probability sample from which valid estimates can be obtained. Every unit in the population should have a non-zero probability of being selected, and the selection probabilities should be known at least in relative terms. ‘In relative terms’ means that values of the unit selection probabilities are known within some arbitrary (and possibly unknown) constant. The reason for using the capture-recapture approach is that the population size and hence the unit selection probabilities are not known in absolute terms. But if the relative selection probabilities are also not known, the capture-recapture method can be applied only by making some assumptions about them. Any parts of the population which are not represented in any of the sampling frames cannot be included in the estimation for the capture-recapture method.

Similar observations apply to the randomness of the selection procedure. If the sample selection procedure departs from strict randomness, application of the capture-recapture method requires ignoring any such departures.

Beyond the requirement of the sample being representative, the capture-recapture procedure does not impose any restriction on the structure of any of the samples. The samples can, for example, be stratified, multistage, with variable selection probabilities, and differ one from the other in structure. More strict restrictions on the sample structure are implied by assumptions (4)-(6) of Section 11.2.1 which are required by particular models such as equation (11.1).

Under some conditions the requirement of representative random sampling can be relaxed. For instance if the second sample is a strictly simple random sample (SRS), the first sample need not be random or representative for valid application of equation (11.1). But in general it is desirable to make all samples random and representative.

When the capture-recapture application is based on sample surveys, it is possible to aim at obtaining representative samples. However, this option is not available when existing administrative sources are used: how well a particular list represents the entire population is already determined by what is available. To proceed further, it is necessary to assume that each data source is representative of the target population, with appropriate weighting if necessary.
11.2.4 Linking of sources at the micro-level

A critical requirement of the method is that for each unit in any of the samples (or in other data sources), complete and correct information is available on whether or not this particular unit is present in each of the other samples/sources. That is, units must have unique identifiers to correctly link them across the data sources.

As can be seen from equation (11.1), failure to link data on the same unit across sources, i.e. under-estimating the size of overlap between samples \( m_2 \), results in over-estimation of the size of the population \( N \). Similarly, false links (overestimating \( m_2 \)) underestimates the size of the population.

In child labour surveys, children may be identified though their own name, names of parents (and especially of the mother), possibly place of birth, and of course sex and age or age group. These are very personal and confidential variables, and proper restrictions must be placed on their use in the survey operation (including, data collection, processing and reporting). Children’s privacy and rights must be respected even if that makes linking across surveys difficult.

11.2.5 Assumption concerning sample design and selection probabilities

Assumptions (4)-(6) are related, and all concern sampling design and selection probabilities, in particular the relationship between these across the samples.

Violations of these assumptions are common and cannot always be avoided. It may be possible to take steps to conform to the assumptions more closely, to minimise departures from them. Much effort has been made to develop analytical methods to reduce and control the effect of departures from those assumptions on the survey results.

A. Lack of independence between sources

Equation (11.13) – which is equation (11.1) written slightly differently - is based on the assumption that the two samples are independent: the probability of appearing in both of them, \( m_2 / N \), is simply the product of the probability of appearing in the first and the second samples:

\[
\frac{m_2}{N} = \left( \frac{n_1}{N} \right) \left( \frac{n_2}{N} \right).
\]

Factors resulting in lack of independence between the samples include the following.

1. Correlated unit selection probabilities

Different samples may be subject to similar variation in unit selection probabilities. It is very common that units having a higher (lower) chance of being present in the first sample also tend to have a higher (lower) chance of being present in the second sample.

Consider for example children visiting a particular site for work with different frequencies, some going every day, some every second day, etc. If we take samples on two arbitrary
days, those visiting the site more frequently would have a higher selection probability in both the samples.

A positive correlation in the selection probabilities leads to an under-estimation of the population size, and a negative correlation to an overestimation.

In applications based on sample surveys we are very likely to have positive correlations, because of common factors affecting all the samples, especially when the samples are drawn within a short period of time. Negative correlation between selection probabilities in different samples are much less likely, unless they have been introduced deliberately by design.

With administrative lists, the pattern can be more complex. We may have positive correlations or negative correlations between lists, or even a combination – positive in one part, and negative in the other part. For example, some groups of individuals may be under represented in all the lists, giving a positive correlation in selection probabilities, on the other hand, different lists may be focussed on different groups, thus have a negative correlation in unit inclusion probabilities.

(2) Correlated selections

This refers to the direct effect of the unit’s selection into one sample on its probability of selection into another sample. Most commonly this results from participation in one survey affecting the individual’s willingness to participate in the later survey (or surveys). Individuals may become less willing to respond in a second survey if the experience of the first participation was negative (e.g. negative reaction of the employer or peers to the participation, tension or embarrassment or fear at being questioned, loss of working time and earnings). This would result in reduced overlap \( m_2 \) and hence in Equation (11.1) giving an overestimation of \( N \). In any case equation (11.1) does not apply because of lack of independence of the samples.

It is also possible that individuals become more willing to be re-interviewed following positive experience of the first interview (e.g. due to gifts or other incentives received, satisfaction or comfort from the chance to talk about themselves and their situation). This would have the opposite effect – of increasing \( m_2 \) and hence under-estimating \( N \) from equation (11.1).

In the case of administrative lists, sometimes there is a direct dependence between lists – for a part of the population, inclusion in one list is determined by its inclusion in another list. This would increase \( m_2 \) and underestimate \( N \). The reverse also happens – inclusion in one list precluding a unit’s inclusion in the other list. For instance, lists may be for alternative entitlements, with any working child eligible for only one of them.

(3) Overlapping samples

A common example is using a multistage design (e.g. a sample of areas or work locations as the PSUs, followed by the selection of working children within each), in which the various samples are drawn from a common sample of PSUs. The motivation for a such design is to save costs and facilitate the work by using same PSUs (and possibly also other higher stage units) for different samples.
The overlap between samples ($m_2$) increases because of being confined to the same PSUs. Consequently simple application of equation (11.1) underestimates $N$. In fact, if the PSUs in the two samples completely overlap, equation (11.1) would provide an estimate of the population size of the selected PSUs only, and not of the whole target population. Some additional information and procedures would be required in order to produce correct estimates.

B. Non-uniform selection probabilities

Differences in unit selection probabilities is perhaps the most common and often unavoidable departure from the simple model of equation (11.1).

There can be many reasons why different units receive different probabilities of selection into the sample. Furthermore, these variations may be related across samples, which introduces dependency and additional complexity, as already noted.

Sometimes variation in selection probabilities are introduced deliberately by design (e.g. in order to obtain more detailed information on some parts of the population, or in response to large differences in per unit survey costs). More commonly, variations are not subject to easy control, being the result of inherent differences among population groups in their accessibility to the survey. For example, sampling frames for establishments of different types where children work may be incomplete to different (but unknown) degrees, varying the effective selection probabilities of children according to where they work. How ‘detectable’ to the survey children are may be highly dependent on characteristics such as sex, age, sector of activity and location of work. Mobility is another major cause of variation. More mobile groups, and groups visiting sample locations more frequently, may be caught more easily for the survey; or depending on the design and survey procedures, may be more difficult to identify and catch.

Issues relating to sampling of mobile populations have been discussed in Chapter 10.

Similar problems can exist with lists. Individuals’ likelihood of being included often depends on their characteristics.

C. Simple random sampling

That all samples used for the capture-recapture method are simple random is a strong requirement. It subsumes conditions (4) and (5) listed in Section 11.2.1 above, namely the requirement of the samples being equal probability and independent. It goes beyond them in requiring that inclusion into a sample/list is independent among individuals. (This is the technical meaning of the samples being ‘simple random’.)

Some analytical procedures depend on the assumption of simple random sampling, but often in practical application, this assumption has not been considered a critical one. In most practical survey situations, the sample designs used have been more complex than simple random sampling, involving features such as stratification, clustering and unequal selection probabilities.
11.3 Numerical illustrations of capture-recapture sampling (schemes “M₀, M₁, M₄, M₅”)

Tables 11.1-11.4 provide simple numerical illustrations of capture-recapture sampling. The “schemes” shown in the tables are identified following Otis, Burnham, White and Anderson (1978). In each table random (probability) samples are drawn according to the design specified. The population consists of \( N = 100 \) units in all areas.

The schemes or models are as follows.

\[ M₀ \]
“Capture probabilities are constant”

The samples are independent, simple random, and of the same size. This is the basic model. The objective of numerical examples below is to illustrate the effect of departures from this model on overlap between the two samples.

\[ M₁ \]
“Capture probabilities vary with time”

Assumptions are the same as in the basic model, except that the two samples can be of different sizes.

\[ M₄ \]
“Capture probabilities vary by behavioural response to capture”

Selection into the first sample affects the probability of being selected into the second sample.

\[ M₅ \]
“Capture probabilities vary by individual unit”

Each member of the population has its own probability of being selected, independently of all other members of the population.

11.3.1 Table 11.1 (scheme “M₀”)

The two samples are both simple random samples, drawn without replacement and independently of each other. The sample size (\( n \)) for the pair of samples is also the same. Units appearing in both the samples are noted and counted. Two examples have been provided.

In Example 1 with \( n₁ = n₂ = 20 \), \( m₂ = 4 \) units are found to overlap. The standard capture-recapture methodology is used. The proportion of the first sample units observed in the second sample \( (m₂/n₁) \), i.e. \( 4/20 = 0.2 \), provides an estimate of the proportion \( (n₁/N) \) which the first sample forms of the total population. This provides an estimate \( \hat{N} \) if \( N \) was unknown (as is the case in situations where capture-recapture sampling may be appropriate).

This means \( n₁/\hat{N} = 0.2 \) or \( \hat{N} = (20/0.2) = 100 \).

By chance this estimate turns out to be exactly equal to the actual population size. This need not be (and normally would not be) the case in practice. With small sample sizes, in particular with small overlap between the samples, the sampling variability in estimating \( N \) tends to be large. For example, there is a high chance of the overlap between the two samples in the first example coming out as 3 or 5 units, in place
of 4. In the first case we would have $\hat{N} = \frac{20}{(3/20)} = 133$, and in the second case $\hat{N} = \frac{20}{(5/20)} = 80$, in place of the true value of 100.

In Example 2 with $n_1 = n_2 = 50$, $m_2 = 25$ units are found to overlap. The proportion of the first sample units observed in the second sample $(m_2 / n_1)$, i.e. $25/50 = 0.5$, provides an estimate of the proportion $(n_1 / N)$ which the first sample forms of the total population. This provides an estimate $\hat{N}$ as before, as

$$n_1 / \hat{N} = 0.5 \text{ or } \hat{N} = \left( \frac{50}{0.5} \right) = 100.$$

### 11.3.2 Table 11.2 (scheme “M_t”)

This shows variation in sampling rates over time (from one sample to the next). The two samples are both simple random samples, drawn independently and without replacement. The two sample sizes differ: $n_1 = 20$, $n_2 = 50$. From the $m_2 = 11$ recaptures among the 50 units in the second sample, the estimated population size from first sample of 20 units is $\hat{N} = \frac{11}{20/11} = 91$ (against the actual population size of 100). Algebraically, the situation is symmetrical between the two samples: reversing their role, we again have $\hat{N} = \frac{11}{20/11} = 91$.

### 11.3.3 Table 11.3 (scheme “M_b”)

This table shows the impact of selection into the first sample on recapture probabilities in the second sample. Units captured in the first sample may shy away and hence have reduced probabilities of recapture in the second sample. Or, by contrast, they may become more willing to participate, hence have increased probabilities of recapture.

In the first situation, when recapture probabilities are reduced resulting from participation in the first sample, overlap between the two samples is reduced and, consequently, the estimate of the total population size from the first sample is inflated. In the table, consider for example the case where, for a unit in the first sample, the probability of recapture in the second sample is only 50 per cent of the selection probability into that sample of units which were not in the first sample. With such halving of the recapture probabilities, we get $\hat{N} = 200$.

In the second situation, when recapture probabilities are increased resulting from participation in the first sample, overlap between the two samples is increased and, consequently, the estimate of the total population size from the first sample is deflated. Consider for example the case where, for a unit in the first sample, the probability of recapture in the second sample is 1.5 times larger than the selection probability of units which were not in the first sample. With increasing recapture probabilities in this manner, we get $\hat{N} = 80$.

### 11.3.4 Table 11.4 (scheme “M_h”)

This table shows two examples with unit selection probabilities varying in the same manner in the two samples. In so far as this results in a greater concentration of the two samples into the same part of the population, the overlap between the two samples tends to increase. Consequently the population size $N$ is underestimated.
If the unit selection probabilities were uniform throughout, we would expect the estimated population size to be \( \hat{N} = n_1n_2/m_2 = 20 \times 20/4 = 100 \). There are two illustrations in Table 11.4. In the first case, unit selection probabilities are varied among units in the population by factors in the range (1 to 4) – varied in common for the two sample selections. We find the estimate of population size reduced to \( \hat{N} = 80 \).

In the second example, unit selection probabilities have been varied following the same pattern, but with a more extreme range of variation, by factors in the range (1 to 8). Estimate of population size is further reduced to \( \hat{N} = 50 \).

11.3.5 Table 11.5 Control through stratification (scheme “Mh” cont.)

The effect can be controlled through stratification if units with large differences in selection probabilities can be separated out into strata prior to sample selection. This is illustrated in Table 11.5. Stratum 1 consists of 20 units, of which 10 are to be selected for each sample, independently and with SRS. The sampling rate is much lower in Stratum 2, where 10 from the remaining 80 units are selected for each sample as above. Ignoring stratification (which makes it similar to the situation illustrated in Table 11.4), 6 units overlap between the two samples (each of size 20), giving \( \hat{N} = 20 \times (20/6) = 67 \). By dealing with each stratum separately, the situation becomes similar to that illustrated in Table 11.1. The estimate for Stratum 1 is \( \hat{N}_1 = 20 \) (which is identical to actual \( N_1 = 80 \)). For Stratum 2 the sample is extremely small, and the estimate is subject to large variability. We get \( \hat{N}_2 = 100 \) (against the actual size \( N_2 = 80 \)).

11.3.6 Note of the selection method for the simple illustrations

For selecting a simple random sample of fixed size \( n \), each unit in the population is assigned a random number (say in the range (0-1), which is done easily in a spreadsheet such as Excel). Units which happen to be assigned the smallest numbers are taken into the sample.

In Table 11.1 the same procedure is repeated for the second sample, after reassigning random numbers to the units. The two sets of random numbers (say \( r_1 \) and \( r_2 \)) are independent, and hence so are the two samples.

In Table 11.2, units are assigned different random numbers for the two samples as before. The number of the smallest values taken into the sample differ in the two cases, depending on the required sample size, which differs between the two samples.

In Table 11.3, the assigned random numbers \( (r_2') \) for the second sample are modified as:

\[
r_2' = r_2 / (1 + \delta c),
\]

where \( c \) is a constant, and \( \delta = 1 \) if the particular unit was selected in the 1st sample, and \( \delta = 0 \) if the unit was not selected in that sample. The \( n_2 \) smallest values of \( r_2' \) then identify the units selected for the 2nd samples. Thus in the 2nd sample, the selection probability of a unit is altered if the unit was also present in the first sample, but remains unchanged if the unit was not in the first sample. The degree of this direct dependence between the samples is determined by the value of \( c \). A negative value of \( c \) means that units already in the 1st sample have a reduced probability of selection into the 2nd sample; and we have the opposite case when \( c \) is positive. The table shows 2nd samples with \( (1 + c\delta) \) values from 0.5 to 1.5 (corresponding to \( c \) from -0.5 to +0.5).
In Table 11.4, selection probabilities are varied across units by a similar modification to the two (independent) random numbers $r_1$ and $r_2$ associated with each unit, determining its selection into the 1st and the 2nd samples, respectively. In other words, for a given unit the same modification is applied in drawing both the first and the second samples. Thus the two samples are independent, but the form of variation in unit selection probabilities are the same in the two cases.

Mechanically, the procedure can be applied as follows. The population is divided into groups ($k$) of units ($j$). For each selection, a unit’s random number $r_j$ is modified, depending on its group, to $r_{j(k)} = d_k r_j$ where $d_k$ values are determined to obtain the required range of variation in selection probabilities across the groups. Subscript $j(k)$ indicates that the reference is to a unit $j$ belonging to group $k$.

In Table 11.5, units in the population were separated into two strata, with stratum $k = 1$ having units with higher values of the assigned selection probabilities, and stratum $k = 2$ having units with lower values, by factor $d_k$ with $d_1 > d_2$.

### Table 11.1. Independent simple random samples of same size

**Example 1. (Otis, Burnham, White and Anderson, 1978, scheme “M0”)**

| Parameters: | | | | |
| --- | --- | --- | --- | 
| Population size (N) = 100 | | | | 
| Two independent simple random samples (SRS), both of size 20 | | | | 
| Sample sizes: $n_1=20$ | $n_2=20$ | $m_2=4$ | | 
| N (estimate) = 100 | | | | 

| unit | whether selected (1=yes) | | | |
| --- | --- | --- | --- | 
| | 1st sample | 2nd sample | Overlap | 
| 1 | 1 | | | 
| 2 | 1 | | | 
| 3 | 1 | | | 
| 4 | 1 | | | 
| 5 | 1 | | | 
| 6 | 1 | 1 | 1 | 
| 7 | 1 | | | 
| 8 | 1 | | | 
| 9 | 1 | | | 
| 10 | 1 | | | 
| 11 | 1 | | | 
| 12 | | | | 
| 13 | | | | 
| 14 | | | | 
| 15 | | | | 
| 16 | 1 | 1 | 1 | 
| 17 | | | | 
| 18 | | | | 
| 19 | | | | 
| 20 | | | | 
| 21 | | | | 
| 22 | | | | 
| 23 | | | | 
| 24 | 1 | | |
### Example 2

Parameters:

- Population size \( (N) = 100 \)
- Two independent simple random samples (SRS), both of size 50
- Sample sizes: \( n_1 = 50 \) \( n_2 = 50 \) \( m_2 = 25 \) \( N \) (estimate) = 100

<table>
<thead>
<tr>
<th>unit</th>
<th>1st sample</th>
<th>2nd sample</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>unit</th>
<th>1st sample</th>
<th>2nd sample</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>55</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>56</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>63</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>68</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>69</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>71</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>73</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>84</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>87</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>88</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>89</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>92</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>97</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>99</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 11.1 (cont.)

<table>
<thead>
<tr>
<th>unit</th>
<th>1st sample</th>
<th>2nd sample</th>
<th>Overlap</th>
<th>unit</th>
<th>1st sample</th>
<th>2nd sample</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>64</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td></td>
<td></td>
<td>65</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td>66</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td></td>
<td></td>
<td>67</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td>68</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td>69</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td>70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>71</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td>73</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td>74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td>75</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td></td>
<td>1</td>
<td></td>
<td>76</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>77</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td></td>
<td>1</td>
<td></td>
<td>78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td></td>
<td>1</td>
<td></td>
<td>79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td>81</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>32</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>82</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td></td>
<td>1</td>
<td></td>
<td>83</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td></td>
<td>1</td>
<td></td>
<td>84</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td>85</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>36</td>
<td></td>
<td>1</td>
<td></td>
<td>86</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>37</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>87</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td></td>
<td>1</td>
<td></td>
<td>88</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>39</td>
<td></td>
<td>1</td>
<td></td>
<td>89</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td>90</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td>91</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>42</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>92</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>43</td>
<td></td>
<td>1</td>
<td></td>
<td>93</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>44</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>94</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>45</td>
<td></td>
<td></td>
<td></td>
<td>95</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td>96</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>47</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>97</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>1</td>
<td></td>
<td></td>
<td>98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>1</td>
<td></td>
<td></td>
<td>99</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>100</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
## Table 11.2. Independent simple random samples of different sizes
(Otis, Burnham, White and Anderson, 1978, scheme “M,”)

Parameters:
- Population size (N) = 100
- Two independent simple random samples (SRS) of different sizes
- Sample sizes: \( n_1 = 20 \)  \( n_2 = 50 \)  \( m = 11 \)
- \( N \) (estimate) = 91

<table>
<thead>
<tr>
<th>Unit</th>
<th>whether selected (1=yes)</th>
<th>1st sample</th>
<th>2nd sample</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 11.3. Dependent random samples

(Otis, Burnham, White and Anderson, 1978, scheme “M_b”)

<table>
<thead>
<tr>
<th>Parameters:</th>
<th>Population size (N) = 100</th>
<th>Two random samples, both of size 20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st sample: simple random sample</td>
<td>2nd sample: random element sample of same size, but unit selection probability affected by whether the unit was already in the 1st sample</td>
</tr>
<tr>
<td>If captured, the recapture probability changed (reduced or increased) by factor =</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sample size=</td>
<td>n_1=20</td>
<td>n_2=20</td>
</tr>
<tr>
<td>N (estimate)=</td>
<td>0.5</td>
<td>1.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>whether selected (1=yes)</th>
<th>1st sample</th>
<th>2nd sample</th>
<th>Overlap</th>
<th>2nd sample</th>
<th>Overlap</th>
<th>2nd sample</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>unit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 11.3 (cont.)

If captured, the recapture probability changed (reduced or increased) by factor =

<table>
<thead>
<tr>
<th>Sample size</th>
<th>n₁=20</th>
<th>n₂=20</th>
<th>m₂=2</th>
<th>n₂=20</th>
<th>m₂=3</th>
<th>n₂=20</th>
<th>m₂=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>N (estimate)=</td>
<td>200</td>
<td>133</td>
<td>80</td>
<td>200</td>
<td>133</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>Unit</td>
<td>1st sample</td>
<td>2nd sample</td>
<td>Overlap</td>
<td>2nd sample</td>
<td>Overlap</td>
<td>2nd sample</td>
<td>Overlap</td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>40</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>59</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 11.3 (cont.)

If captured, the recapture probability changed (reduced or increased) by factor =

<table>
<thead>
<tr>
<th>Sample size</th>
<th>n&lt;sub&gt;1&lt;/sub&gt;=20</th>
<th>n&lt;sub&gt;2&lt;/sub&gt;=20</th>
<th>m&lt;sub&gt;1&lt;/sub&gt;=2</th>
<th>m&lt;sub&gt;2&lt;/sub&gt;=2</th>
<th>m&lt;sub&gt;2&lt;/sub&gt;=3</th>
<th>m&lt;sub&gt;2&lt;/sub&gt;=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>N (estimate)</td>
<td>200</td>
<td>133</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unit</th>
<th>1st sample</th>
<th>2nd sample</th>
<th>Overlap</th>
<th>2nd sample</th>
<th>Overlap</th>
<th>2nd sample</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>68</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>69</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>71</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>77</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>78</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>79</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>82</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>86</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>87</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>89</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>91</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>92</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>93</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>98</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>99</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Table 11.4. Variable unit selection probabilities (Otis, Burnham, White and Anderson, 1978, scheme “Mₜ”)

**Parameters:**
- Population size (N) = 100
- Two independent random samples, both of size 20
- Relative selection probabilities of units vary in the same way in the two samples (for any unit, relative selection probability is the same in the two samples)

<table>
<thead>
<tr>
<th>Example (A)</th>
<th>Example (B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit relative selection probabilities vary in the range 1-4</td>
<td>Unit relative selection probabilities vary in range 1-8</td>
</tr>
</tbody>
</table>

**Sample size:**
- n₁ = 20
- n₂ = 20
- m₁ = 5
- m₂ = 8

**N (estimate):**
- 80
- 50

<table>
<thead>
<tr>
<th>whether selected (1=yes)</th>
<th>Example (A)</th>
<th>Example (B)</th>
<th>Example (A)</th>
<th>Example (B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>unit</td>
<td>1st sample</td>
<td>2nd sample</td>
<td>Overlap</td>
<td>1st sample</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| unit                     | 1st sample  | 2nd sample  | Overlap     | 1st sample  | 2nd sample  | Overlap     |
| 1                        |             |             |             |             |             |             |
| 2                        |             |             |             |             |             |             |
| 3                        |             |             |             |             |             |             |
| 4                        |             |             |             |             |             |             |
| 5                        |             |             |             |             |             |             |
| 6                        |             |             |             |             |             |             |
| 7                        |             |             |             |             |             |             |
| 8                        |             |             |             |             |             |             |
| 9                        |             |             |             |             |             |             |
| 10                       |             |             |             |             |             |             |
| 11                       |             |             |             |             |             |             |
| 12                       |             |             |             |             |             |             |
| 13                       |             |             |             |             |             |             |
| 14                       |             |             |             |             |             |             |
| 15                       |             |             |             |             |             |             |
| 16                       |             |             |             |             |             |             |
| 17                       |             |             |             |             |             |             |
| 18                       |             |             |             |             |             |             |
| 19                       |             |             |             |             |             |             |
| 20                       |             |             |             |             |             |             |
| 21                       |             |             |             |             |             |             |
| 22                       |             |             |             |             |             |             |
| 23                       |             |             |             |             |             |             |
| 24                       |             |             |             |             |             |             |
| 25                       |             |             |             |             |             |             |
| 26                       |             |             |             |             |             |             |
| 27                       |             |             |             |             |             |             |
| 28                       |             |             |             |             |             |             |
| 29                       |             |             |             |             |             |             |
| 30                       |             |             |             |             |             |             |
| 31                       |             |             |             |             |             |             |
| 32                       |             |             |             |             |             |             |
| 33                       |             |             |             |             |             |             |
| 34                       |             |             |             |             |             |             |
| 35                       |             |             |             |             |             |             |
### Table 11.4 (cont.)

<table>
<thead>
<tr>
<th>whether selected (1=yes)</th>
<th>Example (A)</th>
<th>Example (B)</th>
<th>Example (A)</th>
<th>Example (B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>unit</td>
<td>1st sample</td>
<td>2nd sample</td>
<td>Overlap</td>
<td>1st sample</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td>86</td>
</tr>
<tr>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td>87</td>
</tr>
<tr>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td>88</td>
</tr>
<tr>
<td>39</td>
<td>1</td>
<td></td>
<td></td>
<td>89</td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td>90</td>
</tr>
<tr>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td>91</td>
</tr>
<tr>
<td>42</td>
<td>1</td>
<td>1</td>
<td></td>
<td>92</td>
</tr>
<tr>
<td>43</td>
<td>1</td>
<td></td>
<td></td>
<td>93</td>
</tr>
<tr>
<td>44</td>
<td></td>
<td></td>
<td></td>
<td>94</td>
</tr>
<tr>
<td>45</td>
<td></td>
<td></td>
<td></td>
<td>95</td>
</tr>
<tr>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td>96</td>
</tr>
<tr>
<td>47</td>
<td></td>
<td></td>
<td></td>
<td>97</td>
</tr>
<tr>
<td>48</td>
<td></td>
<td></td>
<td></td>
<td>98</td>
</tr>
<tr>
<td>49</td>
<td>1</td>
<td>1</td>
<td></td>
<td>99</td>
</tr>
<tr>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td>100</td>
</tr>
</tbody>
</table>

### Table 11.5. Variable unit selection probabilities, with stratification (scheme “Mh”)

Parameters:
Random element samples within each of two strata

**Ignoring the stratification**
- sample size: n_1=20 , n_2=20 , m_2=6
- N (estimate)=67

**Separately estimating by stratum**
- Stratum 1: n_1=10 , n_2=10 , m_1=5
- N (estimate)=20
- Stratum 2: n_1=10 , n_2=10 , m_2=1
- N (estimate)=100

<table>
<thead>
<tr>
<th>whether selected (1=yes)</th>
<th>Stratum 1</th>
<th>Stratum 2 (cont.)</th>
<th>Stratum 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>stratum 1</td>
<td>1st sample</td>
<td>2nd sample</td>
<td>Overlap</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### 11.4 Separate estimates for major domains or population subgroups

#### 11.4.1 Need for separate domain estimates

Often the target population is divided into major domains such as domains by gender, age group, geographical location (e.g. areas or work locations of different types), living...
11.4 Separate estimates for major domains or population subgroups

arrangements (e.g. child living with or away from own family), major characteristics by type of activity (e.g. different sectors, day-time versus night work, production or catching versus selling of products), and so on.

Estimates of population size and other characteristics may be produced separately for each major group (‘domain’) for one or both of the following reasons.

(1) Production of separate domain estimates is an objective of the survey. The results may be required for policy, research, and for project design, implementation and evaluation.

(2) By design, or unavoidably, different domains have been sampled at different rates. These differences in unit selection probabilities have to be taken into account at the estimation stage.

A related question concerns the relationship between separate domain estimates and the overall estimates for the total population.

The main technical issue regarding (1) concerns sample size. The sample must be large enough to obtain useful results for each analysis domain. The required sample size increases with the number of domains requiring separate reporting, and also with the variation in domain sizes – especially in the presence of domains which are small in size but still require a certain minimum sample size. In practice, however, the increase in sample size is far from proportionate to the increasing number of domains. Usually a proportionate increase is not affordable, nor is it necessary. Individual domain estimates are generally less important than overall total population estimates. Breakdown by some major categories is normally a part of the basic requirement in any case. Beyond that, a rule of thumb which has been found useful, as a starting point, is to seek a sample size increase roughly in proportion to square root of the number of separate estimates required: $n_D = n_0 \sqrt{D}$; for example, sample size increase by factors 1.4, 1.7, 2.0 and roughly 3.0, respectively, for 2, 3, 4 and 10 domains. This of course has to be refined in view of specific requirements and possibilities.

11.4.2 Analysis by domain

If the sampling rate differs only between but is constant within well-defined subgroups, and there is a large enough sample from each subgroup, the numbers in each subgroup should be estimated separately (Seber 1982, Section 3.2.2).

The following simplified model is often a reasonable approximation to the actual situation with capture-recapture samples.

Unit selection probabilities are variable across the population, but can be taken as uniform within subgroups ($g$). The subgroups may be defined by unit characteristics (sex, age, sector of activity, location, etc.), and/or by sample design (e.g. different sampling rates across strata or other geographical domains). Also, often the subgroup categories are defined in exactly the same way in different samples involved in a capture-recapture study. For simplicity, let us assume that to be the case.

Consider two independent equal probability samples for a group ($g$). The proportion of cases in the second sample which also appeared in the first sample $\left(\frac{m_{2g}}{n_{2g}}\right)$ estimates the sampling rate $\left(p_{1g}\right)$ which must have been applied in selecting the first sample:
$$p_{1g} = \frac{m_{2g}}{n_{2g}}.$$ 

Similarly in reverse: the proportion of the first sample which then also appears in the second sample estimates the sampling rate \(p_{2g}\) for the second sample in group \(g\):

$$p_{2g} = \frac{m_{2g}}{n_{1g}}.$$ 

As usual, the inverse of these probabilities can be used to estimate totals for the group. For example, total \(Y_{1g}\) for a certain variable \(y\) from the first sample is

$$Y_{1g} = \left(\frac{n_{2g}}{m_{2g}}\right) \sum_{j \in \text{ms}_g} y_{1j(g)},$$

where subscript \(j(g)\) refers to individual units \(j\) in group \(g\), and subscript 1 refers to the first sample, and subscript 2 to the second sample.

With group total \(Y_{1g} = \sum_{j \in \text{ms}_g} y_{1j(g)}\), the overall total is

$$Y_1 = \sum_g Y_{1g} = \sum_g \left(\frac{n_{2g}}{m_{2g}}\right) y_{1g}.$$  

(11.14)

Similarly, \(Y_{2g} = \left(\frac{n_{1g}}{m_{2g}}\right) \sum_{j \in \text{ms}_g} y_{2j(g)}\); \(Y_2 = \sum_g Y_{2g} = \sum_g \left(\frac{n_{1g}}{m_{2g}}\right) y_{2g}.

The population size is estimated by simply taking \(y_{1j(g)} = y_{2j(g)} \equiv 1\):

$$N_{1g} = N_{2g} = \frac{n_{1g} n_{2g}}{m_{2g}} = N_g, \text{ say.}$$  

(11.15)

That is, equation (11.1) applies unchanged to each group separately.

Size of the total population is

$$N = \sum_g N_g = \sum_g \left(\frac{n_{1g} n_{2g}}{m_{2g}}\right),$$

which is not the same as from unweighted pooling of the group samples:

$$N \neq \frac{n_1 n_2}{m_2} = \frac{\sum_g n_{1g} \sum_g n_{2g}}{\sum_g m_{2g}}.$$  

Group means are given by, from (11.14) and (11.15):

$$\overline{Y}_{1g} = \frac{Y_{1g}}{N_{1g}} = \left(\frac{1}{n_{1g}}\right) \sum_{j \in \text{ms}_g} y_{1j(g)}, \quad \overline{Y}_{2g} = \frac{Y_{2g}}{N_{2g}} = \left(\frac{1}{n_{2g}}\right) \sum_{j \in \text{ms}_g} y_{2j(g)}.$$  

---

42 Here and in the following, population aggregates such as \(Y\) or \(N\) are meant to be sample estimates. For convenience we have omitted the usual “hat” notation, \(\hat{\cdot}\) etc. for these estimates.
11.4 Separate estimates for major domains or population subgroups

i.e. are estimated correctly by *unweighted* sample means for each group. Overall mean is obtained by weighting these by group sizes in the population (which are the same for both samples):

\[ \bar{Y}_1 = \sum \left( \frac{N_g}{N} \right) \bar{Y}_{1g}, \quad \bar{Y}_2 = \sum \left( \frac{N_g}{N} \right) \bar{Y}_{2g}. \]

**11.4.3 Using sample weights at the level of individual units**

In most situations sample data have to be weighted to produce estimates for the population of interest. When the sample data are to be weighted, it is highly desirable - as a matter of practical convenience - to attach to each individual case or record its weight as a variable in the micro data file (see Section 3.4). Most of the required estimates, such as proportions, means, ratios and rates etc. can then be produced in a very straightforward way without any further reference to the structure of the sample. Variance estimation also becomes simplified: most practical methods of computing sampling errors require weighted aggregates at the level of primary sampling units, along with the identification of PSUs and the strata in which they lie.

Consider, for example, the weighted estimates (11.14) and (11.15) above. The weights vary by group \( g \) (and are uniform for individual units \( j(g) \) within a group). Nevertheless, they can be written in the form of weights of individual units, which can then be conveniently added to the micro-data record of each individual unit:

\[ w_{1j(g)} = \left( \frac{n_{2g}}{m_{2g}} \right), \quad Y_{1g} = \sum_{j \in n_{2g}} (w_{1j(g)} Y_{1j(g)}), \quad Y_1 = \sum_g Y_{1g}, \]

where, as earlier, subscript \( j(g) \) refers to individual units \( j \) in group \( g \). Similarly

\[ w_{2j(g)} = \left( \frac{n_{2g}}{m_{2g}} \right), \quad Y_{2g} = \sum_{j \in n_{2g}} (w_{2j(g)} Y_{2j(g)}), \quad Y_2 = \sum_g Y_{2g}. \]

The population size is estimated by simply taking \( Y_{1j(g)} = Y_{2j(g)} \equiv 1 \):

\[ N_g = \sum_{j \in n_{2g}} w_{1j(g)} = \sum_{j \in n_{2g}} w_{2j(g)}, \quad N = \sum_g N_g. \]

This procedure provides estimates of totals both at the domain level and at the total population level, the latter being simply the sum of the former. Calculation of proportion, means and other ratios is then done separately at each level:

for each domain \( \bar{Y}_g = Y_g / N_g \); \hspace{1cm} (11.18)

for the total population \( \bar{Y} = Y / N = \sum Y_g / \sum N_g \). \hspace{1cm} (11.19)

The above formulation applies also in the more general case, when the weights actually vary at the level of individual units.
11. Capture-recapture sampling

11.5 Weighting for individuals visiting study locations with different frequencies

11.5.1 A simple procedure

Consider a set of locations which are visited by members of the target population, but at different frequencies. For example, over \( I = 5 \) days of the week excluding the weekend, an individual may visit any of the locations during a certain number (\( i = 1 \) to \( I \)) of days, with the particular days for visiting chosen by the individual concerned at random. Individuals never visiting any of the locations (\( i = 0 \)), of course, cannot be included in any survey at those locations. We will use subscript \( i \) to indicate the group of persons visiting the locations \( i \) times per week.

Such a situation was considered in detail in Chapter 10. Here we revisit it from the point of view of the capture-recapture design.

Let \( N_i \) be the number of persons in group \( i \), and \( N = \sum_{i=1}^{I} N_i \) the total numbers in all the groups.

On any arbitrary weekday, the expected number present at the locations from group \( i \) is:

\[
N_i' = \left( \frac{i}{I} \right) N_i,
\]

giving total number present as:

\[
N' = \sum_{i=1}^{I} N_i' = \sum_{i=1}^{I} \left( \frac{i}{I} \right) N_i.
\]

Suppose a first sample at a uniform rate \( p_1 \) is taken on any one day of the week. The expected numbers in the sample by group are:

\[
n_{1i} = p_1 N_i' = \left( \frac{p_1}{I} \right) N_i.
\]

Similarly for a second sample at uniform rate \( p_2 \) on another day:

\[
n_{2i} = p_2 N_i' = \left( \frac{p_2}{I} \right) N_i.
\]

With independence of the two samples, the expected overlap is:

\[
m_{2i} = \left( \frac{p_1}{I} \right) \left( \frac{p_2}{I} \right) N_i = p_1 p_2 \left( \frac{i}{I} \right)^2 N_i.
\]

Note that the proportionate overlap is much smaller for the samples of less frequent visitors (persons with small \( i/I \)).

For each group, population size \( N_i \) is correctly estimated in the usual way from equation (11.1), the expected value of the expression on the left being:

\[
\frac{n_{1i} n_{2i}}{m_{2i}} = \frac{\left( \frac{p_1}{I} \right) N_i \left( \frac{p_2}{I} \right) N_i}{p_1 p_2 \left( \frac{i}{I} \right)^2 N_i} = N_i.
\]

In pooling the micro-data over groups, the relative weight \( w_i \) given to a person in group \( i \) is the inverse of the relative selection probability, i.e. \( w_i = \left( I/i \right) \). With these
Weights, estimates of means and aggregates, etc. are weighted estimates in the form of equations (11.16)-(11.19) above.

11.5.2 Numerical illustration

Table 11.6 provides an illustration. Column (1) is the total number of week-days different groups of persons visit any of the locations in the survey. In column (2) are the assumed numbers (and per cent distribution) of persons according to their frequency of visits. It is assumed that individuals choose their day(s) of visit at random. Column (3) is the number of persons expected to be visiting on any one day. Columns (4) and (5) are two independent samples selected from column (3) at rates, respectively, 20 per cent and 30 per cent, and column (6) is the expected size of their overlap. This overlap forms a reducing proportion of, for example, the second sample with decreasing frequency of visits $i$ (column (7)). This is the effective sampling rate which has been applied for obtaining the first sample $n_1$. It takes into account the chance of finding a person at the location (depending on his/her frequency of visits), and the rate applied for sample selection at that location. The sample size inflated by the inverse of this sampling rate estimates the total population size, shown in column (8). This column simply confirms that, under the simple model and perfect information assumed in the illustration, the standard capture-recapture equation (11.1) correctly predicts the population size for the total population as well as separately for each group defined by $i$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$p_1 = 20%$</th>
<th>$p_2 = 30%$</th>
<th>$p_1p_2 = 6%$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) No. of days visited</td>
<td>(2) $N$</td>
<td>(3) $N' = N_i/5$</td>
<td>(4) $n_1$</td>
</tr>
<tr>
<td>$i = 1$</td>
<td>500</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>900</td>
<td>18</td>
<td>360</td>
</tr>
<tr>
<td>3</td>
<td>1,550</td>
<td>31</td>
<td>930</td>
</tr>
<tr>
<td>4</td>
<td>1,300</td>
<td>26</td>
<td>1,040</td>
</tr>
<tr>
<td>5</td>
<td>750</td>
<td>15</td>
<td>750</td>
</tr>
<tr>
<td>Total</td>
<td>5,000</td>
<td>100</td>
<td>3,180</td>
</tr>
</tbody>
</table>

$N$ total number of persons visiting a location $i$ days per week; $N'$ the number present on any one day.
11.6 Capture-recapture with multistage sampling

For economy, convenience, better operational control, or other reasons, the samples may be selected in multiple stages, with the two samples sharing (some or all) higher stage units. For instance, if the number (say $A$) of locations where members of the target population may be found are too many to be all included in the survey, a sample (of size $a$, say) of those may be selected first. This may be followed by the selection of individuals for both the samples from the common set of locations selected at the first stage. In so far as individuals tend to go to the same locations every time, the observed overlap between the samples, both restricted to the same set of locations, will be inflated.

11.6.1 Observed overlap with samples drawn from the same locations

Let, out of the $n_1$ persons selected in the first sample, $n_{1(s)}$ be the number who are present in any of the $a$ common sample areas at the time of selection of the second sample of size $n_2$, and $m_{2(s)}$ be the observed overlap between the two samples. The quantity $\hat{N}_{(s)} = n_{1(s)}n_2 / m_{2(s)}$ estimates the total population of the $a$ locations in the sample, not the size $N$ of the whole population. The latter would be estimated by the standard equation (11.1), $\hat{N} = n_1n_2 / m_2$, but $m_2$ corresponding to two independent samples is not known. With independent samples, both drawn from the whole population of $A$ locations, the overlap $m_2$ between them would have been generally smaller, so that $\hat{N}_{(s)} < \hat{N}$.

The observed overlap $m_{2(s)}$ depends on the extent to which individuals tend to visit the same location, in contrast to going to different locations on different occasions. The following model, with some simplifying assumptions, explains this in quantitative terms.

Suppose that locations are fairly uniform in size, so that a sample of $a$ out of $A$ locations can be taken to form a proportion $(a/A)$ of the total population. Let $s$ be the average proportion of persons who stay in the same area between the two samples. The remaining proportion $(1 - s)$ are present in some different area at the time of the second sample. Of these, the expected proportion present in the remaining $(a - 1)$ sample areas is $(1 - s)[(A - 1)/(A - 1)]$, and the proportion moving to non-sample areas is $(1 - s)[(A - a)/(A - 1)]$. Hence the number out of the original sample $n_1$ present in one of the sample locations at the time of the second sample is

$$n_{1(s)} = \left( s + (1 - s)\frac{a - 1}{A - 1} \right) n_1.$$  \hspace{1cm} (11.20)

If $m_{2(s)}$ is the observed overlap between the two samples (by definition, confined to the $a$ common sample areas) then, as noted already, the estimated population in the sample areas is\footnote{It is reasonable to assume that movement in and out of sample areas tends to be balanced, so that the total population is unchanged between the two surveys, which is estimated by the equation which follows.}
\[ \hat{N}_{(s)} = \frac{n_1(s) n_2}{m_2(s)} . \]

Estimate of the total population of all areas is approximately
\[ \hat{N} = \left( \frac{A}{a} \right) \hat{N}_{(s)} = \left( \frac{A}{a} \right) \frac{n_1(s) n_2}{m_2(s)} . \]

(11.21)

Given that the correct estimate from equation (11.1) is \( \hat{N} = \frac{n_1 n_2}{m_2} \), equations (11.20) and (11.21) give the following for the expected observed overlap \( m_{2(s)} \) between the two samples:
\[ \frac{m_{2(s)}}{m_2} = \left( \frac{A}{a} \right) \left( s + \frac{(1-s) a - 1}{A - 1} \right) . \]

(11.22)

Special case 1: Every individual always visits the same location. This means
\[ s = 1, \quad n_{(s)} = n_1, \quad \frac{m_{2(s)}}{m_2} = \left( \frac{A}{a} \right) n_1, \quad \hat{N}_{(s)} = \frac{n_1 n_2}{m_{2(s)}}, \quad \hat{N} = \left( \frac{A}{a} \right) \hat{N}_{(s)} . \]

Special case 2: Each individual moves at random between the \( A \) locations in the population, giving:
\[ s = \frac{1}{A}, \quad n_{(s)} = \left( \frac{a}{A} \right) n_1, \quad \frac{m_{2(s)}}{m_2} = 1, \quad \hat{N}_{(s)} = \left( \frac{A}{a} \right) \frac{n_1 n_2}{m_{2(s)}}, \quad \hat{N} = \left( \frac{A}{a} \right) \hat{N}_{(s)} . \]

In general: \( \left( \frac{1}{A} \right) \leq s \leq 1 \), giving
\[ \left( \frac{a}{A} \right) \leq \left( \frac{n_{(s)}}{n_1} \right) \leq 1, \quad 1 \leq \frac{m_{2(s)}}{m_2} \leq \left( \frac{A}{a} \right) \frac{\hat{N}_{(s)}}{\hat{N}} = \left( \frac{a}{A} \right) . \]

Equation (11.22) gives the departure from the standard estimator (11.1) due to dependence between the two samples. The distortion is maximum in case 1 where \( s = 1 \): increase in observed overlap by factor \( (A/a) \), hence underestimation of \( N \) by \( (a/A) \). There is no distortion in case 2, where we have \( s = 1/A \).

11.6.2 Illustrations

A. An example of mobility across locations

The data in the first panel of Table 11.7 are taken from a study by Weir et al. (2003). The objective of the study was to assess the use of the capture-recapture procedure for estimating the size of the bar-based female sex worker population in a sub-Saharan African setting (specifically, in Bulawayo, Zimbabwe). The capture-recapture estimate of the population size was evaluated against an actual count of sex workers in the population of bars. As shown in the selected figures in the table, enumerators counted 6,997 women entering 56 bars apparently for sex-worker activity. The capture-recapture estimate was based on a sample of 15 bars. Two independent samples (of sizes, respectively, \( n_1 = 1,381 \) and \( n_2 = 1,469 \)) were enumerated from those 15 bars on successive Saturday nights. An overlap of \( m_{2(s)} = 521 \) was observed between the two samples. This is considerably larger than the expected overlap \( m_2 = n_1 n_2 / N = 290 \) if the two samples were truly independent and not both confined to the common sample of 15 bars. From equation (11.22), this observed overlap implies an average value of \( s = 0.30 \) approximately, as shown in the second panel of the table. This could be seen as implying that 30 per cent of sex workers always go to the same bar, while the remaining 70 per cent tend to go to any bar chosen at random. Actually, the reality is
likely to be much more complex than such a simple dichotomy. There is likely to be a whole gradation, with different workers having different degrees of mobility between bars, and also non-random patterns in visiting different bars. Nevertheless, \( s = 0.30 \) provides a summary index of the level of mobility of sex workers between bars in the city.

This example is unrealistic in that an estimate of the population size, \( N \), is assumed to be available, and has been used above to estimate an average value of \( s \). Normally, we would need to collect data so that \( s \) can be estimated directly; then the above procedure can be applied in reverse to obtain an estimate of \( m_2 \) from observed \( m_{2(s)} \) using equation (11.22), and then of \( N \) from Equation (11.1).

### Table 11.7. Mobility of sex workers between bars: some illustrative data (Weir et al., 2003)

<table>
<thead>
<tr>
<th>Number of locations in the population</th>
<th>( A )</th>
<th>56</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of locations selected</td>
<td>( a )</td>
<td>15</td>
</tr>
<tr>
<td>Sample size, first survey</td>
<td>( n_1 )</td>
<td>1,381</td>
</tr>
<tr>
<td>Sample size, second survey</td>
<td>( n_2 )</td>
<td>1,469</td>
</tr>
<tr>
<td>Observed overlap</td>
<td>( m_{2(s)} )</td>
<td>521</td>
</tr>
<tr>
<td>Actual population size</td>
<td>( N )</td>
<td>6,997</td>
</tr>
<tr>
<td>Expected overlap</td>
<td>( m_2 = n_1 n_2 / N )</td>
<td>290</td>
</tr>
</tbody>
</table>

\[
s \quad \frac{m_{2(s)}}{m_2} = \left( \frac{A}{a} \right) \left( s + (1 - s) \frac{a - 1}{A - 1} \right)
\]

<table>
<thead>
<tr>
<th>( s )</th>
<th>( \frac{m_{2(s)}}{m_2} )</th>
<th>( m_2 )</th>
<th>( N )</th>
<th>( 100 )</th>
<th>( 50 )</th>
<th>( 30 )</th>
<th>( 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>3.73</td>
<td>1,083</td>
<td>everyone stays in the same location</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>2.34</td>
<td>679</td>
<td>50% stay in the same location</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.30</td>
<td>1.79</td>
<td>518</td>
<td>the observed situation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.02</td>
<td>1.00</td>
<td>290</td>
<td>( s=1/A ), move to any location at random</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**B. The two samples drawn from a common sample of locations**

Table 11.8 summarises some simulated results for a two-stage design imposed on the example in Table 11.6. Data in columns (1) and (2) are from that table. They show the numbers of persons \( (N) \) in the population visiting one of the locations in the scope of the survey for \( i = 1 \) to 5 days during 5 days of a week and, assuming a random pattern, the number \( (N') \) expected to be present at a location during any one day. Column (3) shows assumed percentages \( (s \%) \) of individuals who always visit in the same location. The remaining are assumed to move completely at random between the remaining locations, as was assumed in the simple model developed above.

The illustrative design consisted of the selection of \( a = 15 \) locations out of a total of \( A=36 \) locations in the study population, with probability proportional to the size of the location (the number of individuals visiting it on an average day). At the second stage, a fixed number of individuals were selected from each location. Assuming that the size measures are reasonably accurate, this design gives a uniform probability of selection to all individuals present. (The figures in the table have been worked out with this assumption.) Column (4) shows the expected sample size \( n_1 \) of the 1st sample selected.
11.6 Capture-recapture with multistage sampling

at a constant rate \( p_1 = 20\% \), and the expected number \( n_{1(s)} \) present at the time of the second sample in accordance with equation (11.20). Column (5) shows the 2nd sample selected at a constant rate \( p_2 = 30\% \), with \( n_2 \) being the expected sample size. The figures for \( n_1 \) and \( n_2 \) are the same as those in columns (4) and (5) of Table 11.6. The expected overlap between the samples, \( m_{2(s)} \), in column (6) is estimated from equation (11.22). Comparing columns (6) and (7) with the corresponding columns of Table 11.6, we see that the observed overlap is twice as large here; this results from the two samples coming from the same \( a=15 \) locations. Finally, column (8) shows the predicted population sizes, obtained using equation (11.21). These are exactly the same as the actual sizes in column (1). This is simply an artefact of the consistency in the equations used, as noted in deriving equation (11.22) above.\(^{44}\)

There may be considerable variation among locations in the extent to which they are visited by the same individuals every time. This may depend on the nature and size of the location. Table 11.9 provides an illustration. Details are shown individually for the \( a=15 \) sample locations. The locations were selected with probability proportional to size (PPS), and then a constant number of individuals selected from each location (42 and 64, respectively, for the 1\(^{st} \) and the 2\(^{nd} \) samples, with minor variations for some of the smallest and the largest areas). This gave an approximately equal probability design for samples of individuals.

\[\text{Table 11.8. The two samples drawn from a common sample of locations: variation by individual frequency of visits to locations } \ p_1 = 20\% , \ p_2 = 30\%\]

\begin{tabular}{cccccccc}
\hline
No. of days visited & (1) & (2) & (3) & (4) & (5) & (6) & (7) & (8) \\
\hline
\(i = 1\) & 500 & 100 & 96 & 20 & 20 & 30 & 3 & 0.09 & 500 \\
2 & 900 & 360 & 85 & 72 & 66 & 108 & 19 & 0.17 & 900 \\
3 & 1,550 & 930 & 79 & 186 & 163 & 279 & 70 & 0.25 & 1,550 \\
4 & 1,300 & 1,040 & 69 & 208 & 169 & 312 & 98 & 0.31 & 1,300 \\
5 & 750 & 750 & 64 & 150 & 118 & 225 & 85 & 0.38 & 750 \\
Total & 5,000 & 3,180 & 77 & 636 & 535 & 954 & 274 & 0.29 & 5,000 \\
\hline
\end{tabular}

Columns (1) and (2) are from Table 11.6. \( N \) total number of persons vising a location \( i \) days per week; \( N' \) the number present on any one day.

\[\text{Table 11.9. Illustration of variation among locations in the extent to which the same individuals visit them}\]

\begin{tabular}{cccccccc}
\hline
Sample location & (1) & (2) & (3) & (4) & (5) & (6) & (7) & (8) \\
\hline
1 & 57 & 42 & 57 & 42 & 39 & 0.58 & 0.88 \\
2 & 80 & 42 & 64 & 34 & 33 & 0.52 & 0.97 \\
\hline
\end{tabular}

\(^{44}\) In real situations, the \( m_{2(s)} \) will be empirically observed values. In column (6) we have used their ‘perfect’ or expected values, using Equation (11.22) with \( m_2 \) taken as \( n_1 n_2 / N \) based on the standard Equation (11.1) and using the actual values of population size \( N \) (which in real situations would not be known). This explains the perfect agreement between our estimates in column (8) and the actual values in column (1).
11. Capture-recapture sampling

### Table 11.9 (cont.)

<table>
<thead>
<tr>
<th>Sample location</th>
<th>N</th>
<th>n₁</th>
<th>n₂</th>
<th>m₂(s=1)</th>
<th>m₂(s)</th>
<th>m₂(s)/n₂</th>
<th>s</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>99</td>
<td>42</td>
<td>64</td>
<td>27</td>
<td>24</td>
<td>0.38</td>
<td>0.81</td>
</tr>
<tr>
<td>5</td>
<td>109</td>
<td>42</td>
<td>64</td>
<td>25</td>
<td>22</td>
<td>0.34</td>
<td>0.82</td>
</tr>
<tr>
<td>6</td>
<td>120</td>
<td>42</td>
<td>64</td>
<td>22</td>
<td>20</td>
<td>0.31</td>
<td>0.82</td>
</tr>
<tr>
<td>7</td>
<td>122</td>
<td>42</td>
<td>64</td>
<td>22</td>
<td>19</td>
<td>0.30</td>
<td>0.77</td>
</tr>
<tr>
<td>8</td>
<td>140</td>
<td>42</td>
<td>64</td>
<td>19</td>
<td>16</td>
<td>0.25</td>
<td>0.72</td>
</tr>
<tr>
<td>9</td>
<td>155</td>
<td>42</td>
<td>64</td>
<td>17</td>
<td>13</td>
<td>0.20</td>
<td>0.58</td>
</tr>
<tr>
<td>10</td>
<td>162</td>
<td>42</td>
<td>64</td>
<td>17</td>
<td>14</td>
<td>0.22</td>
<td>0.74</td>
</tr>
<tr>
<td>11</td>
<td>205</td>
<td>42</td>
<td>64</td>
<td>13</td>
<td>12</td>
<td>0.19</td>
<td>0.86</td>
</tr>
<tr>
<td>12</td>
<td>206</td>
<td>42</td>
<td>64</td>
<td>13</td>
<td>8</td>
<td>0.13</td>
<td>0.36</td>
</tr>
<tr>
<td>13</td>
<td>279</td>
<td>42</td>
<td>64</td>
<td>10</td>
<td>9</td>
<td>0.14</td>
<td>0.89</td>
</tr>
<tr>
<td>14</td>
<td>302</td>
<td>42</td>
<td>64</td>
<td>9</td>
<td>8</td>
<td>0.13</td>
<td>0.83</td>
</tr>
<tr>
<td>15</td>
<td>432</td>
<td>48</td>
<td>65</td>
<td>7</td>
<td>6</td>
<td>0.09</td>
<td>0.72</td>
</tr>
<tr>
<td>2,550</td>
<td>636</td>
<td>954</td>
<td>310</td>
<td>274</td>
<td></td>
<td>0.29</td>
<td>0.78</td>
</tr>
</tbody>
</table>

m₂(s): observed overlap between the samples.

m₂(s=1): theoretical overlap if all individuals always visited the same location.

Their ratio provides an estimate of \( S \), the individuals’ propensity to visit the same location.

The figures in Table 11.9 correspond to those in Table 11.8, as can be seen by comparing the last rows of the two tables for corresponding columns. These figures by individual location were in fact adapted from an actual survey. Column (5) shows the expected overlap between the two samples if every individual always visited the same location. This is given by the standard equation (11.1), \( m₂(s=1) = \left( \frac{n₁n₂}{N} \right) \), applied to each location. Column (6) is the observed overlap, \( m₂(s) \), between the two samples in the location concerned. The required proportion \( s \) has been computed from equation (11.22), expressed as:

\[
\frac{m₂(s)}{m₂(s=1)} = \left( s + (1-s)\frac{a-1}{A-1} \right).
\]

The quantity \( s \) is a measure of the extent to which the location is visited by the same set of individuals every time. For example in location “2”, an average of 97 per cent of the visitors remain the same, compared to only 36 per cent in the case of location “12”.

### 11.7 Estimation of sample weights in the more general situation

#### 11.7.1 Reason for weighting sample data

The basic reason for weighting sample data for producing estimates is to compensate for variations in unit selection probabilities. The weight – more strictly what is termed the ‘design weight’ – given to each unit in the sample is the inverse of the unit’s probability of being selected into the sample. Of units with above-average selection probability, too many appear in the sample compared to their proportion in the population, and
11.7 Estimation of sample weights in the more general situation

hence their contribution to the estimated population characteristics has to be weighted down. Similarly the contribution of units selected with below-average probabilities has to be weighted up.45

As regards variation in selection probabilities, in practice the situation is often more complex than that considered so far, for example as in Section 11.4 with variation in weights only across but not within major domains. Variation in unit selection probabilities can be much more widespread, especially in capture-recapture surveys designed for populations which are mobile and lack an adequate sampling frame. Diverse factors contribute to this variation in unit selection probabilities, such as the following.

A. Variations in selection probabilities arising from survey conditions

Problems (1)-(3) below result primarily from the basic survey conditions in a capture-recapture procedure, and less from particular aspects of sample design and implementation.

(1) Conditioning

By this we mean that participation in an earlier survey affects the chance of an individual being included in a later survey. This effect may be positive (individual becoming more keen to participate as a result of previous participation), or it may be negative (individuals avoiding repeated participation). This effect is different from (5) below in that it depends on characteristics and experience at the individual level (and also on the data collection procedures used), rather than being a (predictable) result of the chosen sample design.

(2) Unknown size of the population

Even if none of the other problems discussed here exist, and even if the relative selection probabilities of individuals are uniform or subject to known variation, often the absolute values of the selection probabilities are not known because of unknown size of the target population.

(3) Incomplete frame and haphazard differences in coverage and ‘catchability’

Depending on frame imperfections and/or the particular selection methods used, different groups may be subject to different degrees of under-coverage. Actual selection probability achieved is the product of the selection probabilities applied to the given frame, multiplied by the coverage rate.

Coverage also depends on personal characteristics and situations of the children. Some groups of children may be more difficult to locate, or may shy away from contact with the survey, compared to other groups. There may also be groups which cannot be included at all – thus having a zero chance of being selected into the survey. These problems may result from an incomplete sampling frame, or difficulties in locating and

---

45 Section 3.4 presents practical aspects of sample weighting. The discussion is comprehensive but in general terms. The present section develops specific procedures for application in the context of capture-recapture sampling.
identifying members of the target population. Such haphazard differences in coverage and catchability are often uncontrollable and unknown.

**B. Variations in selection probabilities arising from design and implementation**

**(4) Samples lacking independence by design**

Independent samples means that if, for example, for a unit $j$ in the population the probability of selection into sample 1 is $p_{(1)j}$ and into sample 2 is $p_{(2)j}$, then the probability of the unit appearing in both the samples is the product of the two probabilities: $p_{(1,2)j} = p_{(1)j}p_{(2)j}$. The samples lacking independence means that $p_{(1,2)j} \neq p_{(1)j}p_{(2)j}$ for some units. We can have $p_{(1,2)j} < p_{(1)j}p_{(2)j}$ if, for instance, all or parts of the two samples are forced to come from different areas.

More commonly, probabilities in the samples are positively correlated, and we have $p_{(1,2)j} > p_{(1)j}p_{(2)j}$. An example of this case was discussed in Section 11.6. As illustrated in that example, even if each of the two samples in a capture-recapture design is, in itself, a representative sample of the population, the degree of overlap of the samples is affected (increased) if both are confined to the same sample of higher stage units in a multistage-design.

**(5) Design differences in sampling rates**

We may deliberately sample different domains or groups at different rates depending on the substantive objectives, such as oversampling certain categories of working children such as by sex and age, or according to the type, location and circumstances of their work.

**(6) Inherent differences in unit selection probabilities**

A common situation is the one discussed earlier – concerning the fact that children may engage in activity at different times or with different frequencies, which determines their chance of being present in the survey population at that time of the sample selection.

**(7) Non-response**

Different groups in the sample may be subject to different rates of non-response. In each group, the effective probability of appearing in the achieved sample is reduced: it is the product of the original selection probability applied and the response rate achieved.

**11.7.2 Outline of a methodology**

**A. Absolute versus relative selection probabilities**

A probability sample, by definition, means that every unit in the target population should have non-zero and known probability of being selected into the sample. We may ‘know’ selection probabilities (i) as absolute values for all units in the population, or (ii) only in relative terms – how they vary among units in the sample without knowing their actual values. For example, if a simple random sample of size $n$ is selected from
a population of known size $N$, unit selection probabilities are constant with the actual value as $n/N$. However if $N$ is not known, we may still be able to select a random sample and know that the unit selection probability is a constant, but not know its actual value.

In the procedures below, the focus will be first to determine relative probabilities. The next step would involve estimating $N$ so as to convert them into absolute terms. Note that for certain purposes, such as estimating proportions, means or other ratios, or for more analytic statistics, knowing relative selection probability is sufficient. Values of selection probabilities in absolute terms are required for estimating population totals, which is the special concern of capture-recapture sampling.

B. Basis of the procedure

The fundamental assumption in the following procedure is that sample $n_2$ is a representative independent and random (probability) sample of the given population $N$. It does need to be a simple random sample, nor even an equal probability sample. Being ‘representative’ implies that the sample is large enough to reflect well the characteristics of the population. As such, it can be expected to provide a representative sample of any reasonably large subclass of the population – the representation being better for larger and well-distributed subclasses. The matched sample is obtained by applying the sample $n_2$ selection procedure to the ‘population subclass’ comprising sample $n_1$.

For the overall sample $n_2$, generally the size of the population is unknown. However, its component $m_2$ has the special feature that its population is known: that population is comprised of all the units in sample $n_1$. Hence comparing $m_2$ with $n_1$ gives us the selection probabilities applied to these units as a part of the selection of $n_2$ sample. Note that this gives the actual sampling probabilities (i.e. not simply their relative values). Inverting these actual selection probabilities gives weights which summed over sample units ($m_2$) give the population size (in this case, $n_1$).

In estimating the unit selection probabilities ($p$) and their inverse, sample weights ($w = 1/p$), for units in sample $n_2$, the strategy is to supplement the information available on or from $n_2$ with information which is available only for its component $m_2$.

C. Components of variation in probabilities and sample weights

In order to develop the weighting procedure, let us decompose the variation in unit selection probabilities (and in the corresponding weights) into components corresponding to the sources (1)-(7) described above.

$$P = p_{(1)} p_{(2)} p_{(3)} p_{(4)} p_{(5)} p_{(6)} p_{(7)}$$

$$\frac{1}{P} = W = w_{(1)} w_{(2)} w_{(3)} w_{(4)} w_{(5)} w_{(6)} w_{(7)}. \quad (11.23)$$

Subscript $(q)$ indicates source of variation (1)-(7). Usually probabilities (and weights) are known and needed only in relative terms, rather than absolute terms. This is indicated in the above equations by the use of small letters $p$ (and $w$). The exception is $p_{(2)} (W_{(2)})$ which is in absolute terms, relating to the total population size $N$. This makes the resulting overall probability $P$, and corresponding weight $W = (1/P)$, also in
absolute terms. The weights being in absolute terms means that their sum over units in the sample equals to the population size $N$. Generally we will add subscript $j$ to the above quantities to indicate unit level values. When it is necessary to distinguish between different samples, such as $n_1$ and $m_2$, we will indicate that by appropriate superscript. For instance, $w_{(n_2)}^{(m_2)}$ is the design weight, in relative terms, of unit $j$ in sample $n_2$. Similarly, $w_{(m_2)}^{(n_2)}$ refers to the relative design weight of that unit as a part of sample $m_2$. $W_{(2)}^{(n_2)}$ is the scaling factor which gives the resulting weights $W_j$ values in absolute terms, such that $\sum_{j \in n_2} W_j = N$. (Note that the scaling factor $W_{(2)}$ is a constant value common to all units in the sample.)

Component $p_{(1)}$ (or $w_{(1)}$) arises from conditioning resulting from the unit’s participation in the first survey $n_1$. Hence it applies only to units in the overlap sample $m_2$, and not to other units in $n_2$. For the latter units, $p_{(1)} = w_{(1)} = 1$, by definition.

Similarly, component $p_{(4)}$ (or $w_{(4)}$) applies only to units in sample $m_2$, arising for instance from lack of independence between samples $n_1$ and $n_2$. Hence for $n_2$ units which are not in $m_2$, $p_{(4)} = w_{(4)} = 1$.

For units in $m_2$, these two factors measure how the unit’s weight (or selection probability) when it is seen as a member of sample $m_2$ differs from that when it is seen as a member of sample $n_2$. Hence we may refine the decomposition (11.23) of the weight as follows:

$$W_j^{(n_2)} = W_{(2)} w_{(3)}^{(n_2)} w_{(5)}^{(m_2)} w_{(6)}^{(7)} w_{(7)}^{(n_2)}.$$ (11.24)

for units $j$ taken as part of sample $n_2$ (including its part $m_2$); and

$$W_j^{(m_2)} = w_{(1)} w_{(4)}^{(m_2)} W_j^{(n_2)}.$$ (11.25)

for units $j$ considered parts of sample $m_2$, with $W_j^{(n_2)}$ defined exactly as in (11.24).

**D. Steps involved in the computation of weights**

(1) The weight components $w_{(1)}^{(n_2)}$ and $w_{(4)}^{(m_2)}$ concern only overlap sample $m_2$, and not the whole of the recapture sample $n_2$. They have to be estimated as best as possible from the former sample. Component $w_{(4)}^{(m_2)}$ is determined largely by the relationship between design and implementation of the two samples $n_1$ and $n_2$, and is more tractable. Component $w_{(1)}^{(n_2)}$ depends on complex factors concerning the survey population and condition and patterns of behaviour of individual respondents. These components and possible procedures are discussed in Section 11.7.3.

(2) The weight components $w_{(5)}^{(n_2)}$, $w_{(6)}^{(m_2)}$ and $w_{(7)}^{(m_2)}$ concern sample design and implementation of $n_2$. That sample provides the source for estimating them. These are discussed in Section 11.7.4. Note that these quantities (weights and probabilities) are in relative terms, and can be determined without reference to the usually unknown total population size.

(3) The next step is to estimate the overall sample weight $W_j^{(m_2)}$ for units in sample $m_2$. This can be done on the basis that size and characteristics of the population ($n_1$) of sample $m_2$ are known. By identifying groups with the same or similar
characteristics in \( n_1 \) and \( m_2 \), we can estimate the final probabilities of weights required to get from one to the other. The procedure will be detailed in Section 11.8.

(4) With the quantities estimated in (1)-(3) above, we can compute the remaining factors for units in the overlap sample \( m_2 \):

\[
\{ W_{(2)} w_{(3)j} \} = W_j^{(m_2)}/\{(w_{(1)j} w_{(4)j}) (w_{(5)j} w_{(6)j} w_{(7)j})\}.
\]  

(11.26)

Apart from \( W_{(2)} \), the components are in relative terms with arbitrary scaling. By choosing a suitable scaling for them, the overall scaling factor \( W_{(2)} \) can be isolated and determined from the known population size \( n_1 \), as explained in Section 11.8.2.

(5) Component \( w_{(3)j} \) is described in Section 11.7.5. This component has been determined above only for units in \( m_2 \). The final step is to extrapolate it to the full sample \( n_2 \), and ultimately also to \( n_1 \) if possible. As described in Section 11.8.3, this may be done using a procedure similar to (3) above. By identifying groups with the same or similar characteristics in \( m_2 \) and \( n_2 \), we can carry over grouped \( w_{(3)j} \) values from the former to the latter. In certain cases, it might be possible to consider such carryover in relation to \( n_1 \) as well.

### 11.7.3 A unit’s selection into \( m_2 \) versus selection into \( n_2 \)

**A. Conditioning – problem (1)**

Conditioning – selection into sample \( n_2 \) being affected by a unit’s participation in sample \( n_1 \) – is indeed a basic problem of the capture-recapture method. With only two sources of information – whether samples or administrative lists or a combination of the two – there is no statistical method for adjusting for this effect. With more than two sources, procedures are available to control it, at least in part. These will be briefly described (without full technical details) in Section 11.9. The best practical approach, and certainly when only two sources are being used, is to take steps to try and minimize this problem. For surveys, the recommended strategy is to keep the interview short, confidential and friendly so as not to make children unwilling to participate in the repeat surveys; and to avoid giving substantial gifts or other incentives which can artificially inflate their desire to participate. Asking the same information again in the second survey should be avoided.

Beyond that, attempt should be made to assess whether and to what extent the problem of conditioning exists, and try to take that into consideration in the interpretation of the results. Small-scale intensive studies may help in getting the direction of the effect of conditioning, and perhaps some idea of its magnitude.

**B. Samples lacking independence - problem (4)**

Problems concerning the samples lacking independence are perhaps less common in practice; in any case, they affect the overlap \( m_2 \) and not the main samples \( n_1 \) and \( n_2 \). A detailed example of this problem was discussed in Section 11.6.
If a multistage design is used and samples $n_1$ and $n_2$ share some higher stage units (e.g. both samples are confined to the same sample of clusters), then the amount of the overlap between the samples is increased, compared to the overlap in the case when the two samples are independent. This inflates the observed probability of selection of units into $m_2$, compared to the actual probabilities they received as a part (as just a subclass of) of the selection of sample $n_2$. With such a design, we need information on the chance that a unit selected in the first sample moves out of the common higher stage units by the time the second survey is done.

For a simple illustration of the effect of the above type of design on sample weights, let us consider a sample selected in two stages. The first stage shared by the two samples consists of an equal probability sample of $a$ clusters out of $A$ in the population. For simplicity we also assume the clusters to be of uniform size. This is the same setup as used in the illustration of Section 11.6.

**Case 1:**
If all the $n_1$ units move entirely at random between clusters in the interval between the two surveys, then the expected proportion of $n_1$ units found in the original sample areas (where, by definition, whole of the sample $n_2$ is found) is $r_0 = (a/A)$.

If the two samples were entirely independent (e.g. were drawn from two independent samples of clusters), then the proportion of sample $n_1$ cases expected to be found in the sample areas of $n_2$ would be exactly the same as above, $r_0 = (a/A)$, irrespective of the design of $n_1$.

**Case 2:**
If none of $n_1$ move outside the original $a$ sample areas (their movement within these areas does not make a difference), then their proportion still within the original sample areas is, of course; equal to 1.

**Intermediate case:**
If the movement of $n_1$ units outside the original areas is partial, with proportion $r$, $0 < r < 1$, remaining within the original sample areas then the observed probabilities of units in $m_2$ are inflated by factor $(r/r_0) = r(A/a)$.

**Notes:**
(1) If needed, information for determining parameter $r$ needs to be obtained from the respondents, as noted earlier in the example given in Section 11.6.

(2) If sample areas are selected with probability proportional to area size (PPS), then $r_0$ is defined as the ratio of the sum of size measures of the $a$ selected clusters, to the total size measures of $A$ clusters in the population.

(3) If, as is commonly the case, $n_1$ and $n_2$ are independent samples then the above problem does not exist: $(r/r_0 = 1)$.

### 11.7.4 Variations in selection probabilities arising from design and implementation

**A. Design differences in selection probabilities – problem (5)**

In the absence of information on the population size, the absolute value of selection probabilities are not known. However, they are normally known in relative terms – of each unit in relation to other units in the sample.
The sample weights at the estimation stage are inversely proportional to the design probabilities applied. An important requirement is to properly document variations in the design probabilities and link this information to individual units in the sample.

Typically there can be three types of variation in design probabilities: (1) Variation applying only across a few major sampling domains, with a uniform sampling rate within each domain. (2) In a multistage design, variations across higher stage units (areas, establishments) in the final selection probability of the ultimate units, but uniform probabilities of the ultimate units within higher stage units. (3) General variation at the level of individual ultimate units (children). A common example of the second type is selection of the sample in two stages (for instance, areas, followed by the selection of children in each selected area) with the product of selection probabilities at the two stages (which equals the final selection probability of the ultimate units) being constant within but varying across sample PSUs.

*In so far as the design probabilities are known only in relative terms, the corresponding weights are also scaled arbitrarily.*

It will be shown in Section 11.8 how the scaling factor can be determined using the overlapping part of the two samples in a capture-recapture design.

**B. Inherent variation in selection probabilities - problem (6)**

Considerations similar to the above apply also to inherent variations in selection probabilities resulting from the underlying structure and characteristics of the study population. The main requirement is to collect the necessary information during the survey fieldwork in order to identify these variations. For example, if units differ in the timing and frequency of their presence at the sample location, then the survey should obtain information to quantify the effect of those differences on the individuals’ selection probabilities – asking questions such as how often and when the child comes to any of the locations in the *sampling frame*, in relation to the period(s) when the sample selection is done.

**C. Variation due to differential non-response – problem (7)**

Finally concerning non-response: weighting for non-response has been discussed in Section 3.4.2B. The main aspects include the following. Information concerning non-response is available from survey implementation provided records are kept of the number selected, and among those the number successfully enumerated. It should be noted that some of the units selected may turn out to be ineligible for the survey (e.g. some children selected may turn out to be non-working, or to be otherwise out-of-scope of the target population). In computing the “number selected”, ineligible cases must be excluded.

A practical and common procedure to identify non-response rates is to divide the sample into a (hopefully small) number of groups \(g\) which noticeably differ in response rates, but each group is reasonably homogenous within itself with respect to non-response. Normally the groups will be defined in terms of the main variables affecting response rates (demographic characteristics of the children, their circumstances, type and location of work, etc.). Once the groups have been identified, the computation of response rate for each group is straightforward in principle: it is the ratio of the number \(e\) of
units successfully enumerated, to the number \( s \) of eligible units originally selected: \( R_g = e_g / s_g \). These groups form the weighting classes. The non-response weight given to any enumerated unit \( j(g) \) in group \( g \) is the inverse of the group response rate \( w_{j(g)} = 1 / R_g \).

### 11.7.5 Unknown and inaccessible population

#### A. Unknown size of the population – problem (2)

This of course is the fundamental reason for using the capture-recapture approach, rather than using conventional sampling. When the two samples are completely independent and each is an equal probably sample of the same target population, the basic formula (11.1) gives an estimate of the population size \( N \) as in Equation (11.13):

\[
\frac{m_2}{N} = \left( \frac{n_1}{N} \right) \left( \frac{n_2}{N} \right).
\]

This formula needs to be adjusted in the presence of various sources of variation in selection probabilities. This in fact is the primary motivation for the development of an appropriate weighting system, as in this and the next section.

#### B. Incomplete frame; undercoverage – problem (3)

When administrative lists are used for the capture-recapture method, it is of course taken as given that each list in itself provides an incomplete coverage of the population, including the fact that some units do not appear in any of the lists. Nevertheless, in order to apply the basic procedure, it is assumed that each list is an independent random sample of the whole population. Hence equation (11.13) is applied to obtain estimate of \( N \), and the population not appearing in any of the lists is estimated as in Equation (11.3):

\[
n_0 = \hat{N} - (n_1 + n_2 - m_2) = (n_1 - m_2)(n_2 - m_2)/m_2.
\]

The situation is different when sample surveys are used as the sources. Usually all surveys are subject to the same under-coverage – being selected from the same part, say \( N' < N \) of the population. Equation (11.13) then gives an estimate only of \( N' \), not of the total \( N \):

\[
\frac{m_2}{N'} = \frac{n_1}{N'} \cdot \frac{n_2}{N'} \Rightarrow N' = \frac{n_1 \cdot n_2}{m_2}.
\]

The extent of under-coverage can only be assessed on the basis of some external information, which often is not available. Sometimes it is possible to obtain some idea, even if approximate, of the extent of the under-coverage, and how it differs among different groups in the population. Under-coverage deflates the selection probabilities achieved in comparison with the nominal selection probabilities applied. The effect of under-coverage is often much less marked on relative selection probabilities and hence on estimates of rates and ratios compared to estimates of aggregates from the survey. If there are serious doubts about completeness of coverage, then it may be the best policy
not to publish estimates of total population size and other aggregates, rather than to publish misleading figures.

C. Haphazard variations in coverage and catchability – problem (3) cont.

There can also be the problem of haphazard (unsystematic, unpredictable) variations in coverage and catchability. The problem of haphazard variations in coverage/catchability can be intractable. In principle some information on the nature and extent of the problem may be obtained using special procedures. However, in situations where the capture-recapture sampling procedure is likely to be necessary or useful, typically there are diverse and uncontrollable sources adversely affecting how well the target population can be covered, and can be actually captured in the survey if selected. In the next section a procedure will be described which can help in shedding light on this problem by making use of a special feature of the capture-recapture methodology. This is that $n_2$ forms the known population for the overlap sample $m_2$, that sample having been obtained largely using known $n_2$ sampling procedures.

### 11.8 Weight components derived from the sample overlap $m_2$

#### 11.8.1 Overall weight as a function of unit characteristics

This is a substantive and empirical question. The requirement is to identify a manageable number of variables which account for most of the variability in the selection probabilities. It is also necessary to determine useful categories for each variable. The objective of this exercise is to divide the sample into groups or ‘cells’ such that each cell is homogeneous with respect to the selection probabilities applied, and the variation in selection probabilities is mainly across the cells. The reference is to selection probabilities as applied in $n_2$, while the set of variables and categories defining the cells must be common to $n_1$ and $n_2$ i.e. both surveys should contain the variables, and hence also their overlapping part $m_2$.

A. Simple procedure based on weighting cells

In the simple procedure, groups ($g$) defined in the above manner serve as “weighting cells”. For units in each group the ratio of the number of units in $m_2$ to the number of units in $n_1$ gives the required sampling probability:

$$
\hat{Y}_g = \frac{m_{2g}}{n_{1g}}.
$$

(11.27)

Obviously, groups should be large enough to avoid unstable results. As a rule of thumb, average sample overlap $m_{2g}$ should not fall below 10 units.

B. Use of regression to determine weights

A generally more efficient approach is use some regression model to determine the weights. The dependent variable is a unit's ‘propensity to be selected’, and the
independent variables are a set of covariates - factors which are considered to be most closely related to the unit selection probability. The advantage of the regression approach over simply using weighting cells is that the former can incorporate more variables and categories, and hence capture (explain) more of the variation in the selection probabilities.

In principle one can use data at the individual unit level for regression analysis. However, when the numbers involved are small, there can be too much scatter and it is better to work with grouped data. However, the classification can be much more detailed and the groups sizes much smaller than the ‘weighting cells’ discussed above. As above, we will use subscript ‘\( g \)’ to indicate such groups, though here the references is to generally much smaller groups.

The selection of independent or predictor variables and their appropriate categorisation is perhaps the most difficult problem in regression analysis. As a guiding principle, only a limited number of independent variables can – and should – be used in practical application. The variables included should be easy to measure (e.g. from straightforward interview questions), robust (in the sense of not being subject to large measurement errors), should not be highly correlated with any other variable(s) already included, and should make a significant contribution to explaining the variation in the dependent variable. The covariates should also be substantively meaningful for the purpose. Beyond that, it is not in-scope of the present discussion to go into substantive or statistical details of applied regression; the objective below is simply to outline the procedure.

Let \( X_1, X_2, \ldots \) be a set of covariates on which variations in selection probabilities are expected to depend, and \( x_{1g}, x_{2g} \) be their values for units in group \( g \). As noted, these groups can be and usually are much finer than the groups implied in the ‘weighting cells’ approach, in principle even may refer to individual units. Let \( y_g = \left( m_{1g} / n_{1g} \right) \) denote the observed selection probability, as in equation (11.27). Let \( X_g \) denote a function of values \( x_{ig} \) as

\[
X_g = b_0 + b_1 x_{1g} + b_2 x_{2g} + \ldots ,
\]

where \( b_0, b_1, b_2, \ldots \) are to be determined by regression.

**C. Linear Regression**

The linear regression model is

\[
y_g = X_g + e_g , \tag{11.28}
\]

where \( e_g \) is the “error term”, being the difference of observed \( y_g \) and \( X_g \); \( e_g = y_g - X_g \). The standard linear regression estimates the coefficients \( b_0, b_1, b_2, \ldots \) so as to minimise a certain function of these deviations. The fitted value of \( y_g \) is

\[
\hat{Y}_g = X_g = b_0 + b_1 x_{1g} + b_2 x_{2g} + \ldots . \tag{11.29}
\]

The method used most often for this purpose is least squares or weighted least squares of the deviations \( e_g \), for which standard statistical software is widely available.
D. Logistic Regression

In the present case, the dependent variable is a (0, 1) indicator variable (a unit in sample \( n_1 \) is selected = 1, or is not selected = 0, into sample \( n_2 \), when the \( n_2 \) sampling procedure is applied to the population); or it is a mean \( (Y_g) \) of such indicator variables. Both theoretical and empirical considerations indicate that in such a situation, the shape of the functional relationship will be curvilinear rather than a straight line implied by equation (11.28). In many applications, the functional relationship is found to be shaped like a “tilted-S”, with asymptotes at 0 and 1. Obviously these are the correct limits for \( Y_g \) values, since those values stand for probabilities. The tilted-S shape implies that at very low values as well as at high values of the independent variables, the result is insensitive to change in the actual values of the independent variables. The sensitivity is maximum in the middle range. Such is often the case in real situations.

Logistic regression assumes the relationship to be of the form

\[
\hat{Y}_g = \frac{e^{X_g}}{1 + e^{X_g}} . \tag{11.30}
\]

This relationship is linearized by the so-called logistic or logit transformation

\[
\hat{T}_g = \ln \left( \frac{\hat{Y}_g}{1 - \hat{Y}_g} \right),
\]

giving a form similar to equation (11.29)

\[
\hat{T}_g = X_g + b_0 + b_1 x_{1g} + b_2 x_{2g} + ...... .
\]

We of course begin from observed values \( Y_g \) of \( Y_g \). With the above transformation

\[
t_g = \ln \left( \frac{Y_g}{1 - Y_g} \right),
\]

the estimation equation corresponding to (11.28) is

\[
t_g = X_g + e_g .
\]

Least squares procedures may be used to estimate \( T_g \), and then the required probability \( Y_g \) using equation (11.30). In place ordinary least squares, in theory it is better to use weighted least squares; the appropriate weights are (Neter and Wasserman, 1974, Section 9.8):

\[
s_g = n_g Y_g (1 - Y_g). \tag{11.31}
\]

It is important to ensure that the groupings \( (g) \) are large enough in sample size, so that the group sizes \( n_{1g} \) in sample 1 are not small. In any case, there should be no groups for which the mean value \( Y_g \) of the indicator equals the extreme limits 0 or 1.

E. Unit level overall weight in sample \( m_2 \)

Once estimates of type \( \hat{Y}_g \) have been obtained for groups of units, say \( j(g) \), of sample \( m_2 \), the estimated value is assigned to every unit in the group:

\[
W_{j(g)}^{(m_2)} = \hat{Y}_g . \tag{11.31}
\]
where subscript \( j(g) \) has been used to indicate units belonging to group \( g \).

It may be mentioned that the procedures described in A-D above may also be applied to data weighted by the known components of weights defined in Equation (11.23):

\[
W_{(1)}j \cdot W_{(4)}j \cdot W_{(5)}j \cdot W_{(6)}j \cdot W_{(7)}j = W_{(3)}j, \text{ say.}
\]

In this case the numerator of \( \hat{Y}_g \) in Equation (11.27) would be a weighted quantity, being the sum of weights \( W_{(3)}j \) of the \( m_2g \) units in group \( g \). Equation (11.31) would then give an estimate of the remaining components of the weight, i.e. of \( (W_{(2)} \cdot W_{(3)}j) \) directly. This alternative procedure should be more precise, but is also somewhat more complex.

### 11.8.2 Separating scaling factor \((W_{(2)}j)\) and haphazard component \((W_{(3)}j)\)

Consider the unit weights estimated so far for sample \( m_2 \). Apart from the scaling factor \( W_{(2)}j \), the weights of these units may be scaled such that their average value equals 1.0. (It is a common and convenient practice to scale relative weights in this way.) At the same time the sum of actual weights is meant to equal the size of the population of these units, namely \( n_1 \). This gives the following results.

\[
\sum_{j \in m_2} (W_{(1)}j \cdot W_{(4)}j \cdot W_{(5)}j \cdot W_{(6)}j \cdot W_{(7)}j) = m_2,
\]

\[
\sum_{j \in m_2} W_{(2)}j (W_{(1)}j \cdot W_{(4)}j \cdot W_{(5)}j \cdot W_{(6)}j \cdot W_{(7)}j) = n_1, \text{ hence}
\]

\[
W_{(2)}j = n_1/m_2.
\]

Substituting into Equation (11.26) gives the remaining component \( W_{(3)}j \).

### 11.8.3 Carry over from \( m_2 \) to \( n_2 \); possibly to \( n_1 \)

#### A. Sample \( n_2 \)

In fact, the above-mentioned haphazard variations giving rise to weights \( W_{(3)}j \) are associated with the \( n_2 \) sampling process. The assertion is based on the expectation that on the basis of matching unit characteristics, the estimated haphazard variation in selection probabilities can be imputed from units in sample \( m_2 \) to all the units in sample \( n_2 \).

It is necessary to return to the idea of averaging quantities over groups of units. Let \( g \) refer to groups of units within which the quantities \( W_{(3)}j \) computed above are expected to be reasonably homogeneous, and the main variation in the quantity is across groups. These groups must be defined on the basis of variables available on all units \( n_2 \) (hence on \( m_2 \)) and desirably also over all units in \( n_1 \). These can be the same groups as previously used, e.g. in relation to determining the overall unit weights \( W_{(3);g} \), but can be different if necessary.

As \( m_2 \) is subject to the same sampling procedure as the whole of \( n_2 \), it is reasonable to assume that, for a given \( g \), the already estimated haphazard weighting factors from \( m_2 \) apply also to all the \( n_2 \) units in the same group \( g \). In applying this procedure, individual
unit values are of course replaced by their average over group $g$ to which the unit $j$ belongs.

Thereafter these averaged group values are ascribed to all individual units $j$ in each group $g$ in the whole sample $n_2$.

The value of the scaling factor $W_{(2)} = n_1/m_2$ determined above is of course also carried over from $m_2$ to $n_2$.

With the overall weight for each unit in $n_2$ determined, the important result is that the size of the population is estimated as follows, replacing the basic Equation (11.1):

$$\hat{N} = \sum_{j \in n_2} W_j^{(n_2)}.$$  \hfill (11.33)

It is instructive to examine (11.33) and compare it to (11.1). Consider sample $m_2$. Let $m_2^{(n_2)}$ be its expected size if it were simply a normal part of $n_2$ with no differences in selection probabilities due to conditioning or $(n_1, n_2)$ sample dependence etc. Weights $(w_{(1)}j, w_{(4)}j)$ were introduced in (11.32) to make the weighted sum of $m_2$ the same as those of $m_2^{(n_2)}$ – this is the underlying logic of all sample weighting, to compensate for difference in selection probabilities. This implies that

$$\sum_{j \in m_2} (w_{(1)}j \cdot w_{(4)}j \cdot w_{(3)}j \cdot w_{(5)}j \cdot w_{(6)}j \cdot w_{(7)}j) = \sum_{j \in m_2^{(n_2)}} (w_{(3)}j \cdot w_{(5)}j \cdot w_{(6)}j \cdot w_{(7)}j),$$

giving

$$(m_2/m_2^{(n_2)}) \left[ \sum_{j \in m_2} (w_{(1)}j \cdot w_{(4)}j \cdot w_{(3)}j \cdot w_{(5)}j \cdot w_{(6)}j \cdot w_{(7)}j) / m_2 \right] = 
\sum_{j \in m_2^{(n_2)}} (w_{(3)}j \cdot w_{(5)}j \cdot w_{(6)}j \cdot w_{(7)}j) / m_2^{(n_2)}.$$

Note that from (11.32) the term in square brackets, the mean per unit weight in sample $m_2$, equals 1 by definition. The right hand side is the mean per unit weight of the $m_2$ cases in sample $n_2$. This implies that the mean per unit weight in sample $n_2$ is $(m_2/m_2^{(n_2)})$ times the mean per unit weight (=1) in sample $m_2$.

We can expect this ratio of the means of the two sets of weights to apply more generally to whole of sample $n_2$.

With this (11.33) can be written as

$$\hat{N} = \sum_{j \in n_2} W_j^{(n_2)} = \left( \frac{n_1}{m_2} \right) \sum_{j \in n_2} (w_{(3)}j \cdot w_{(5)}j \cdot w_{(6)}j \cdot w_{(7)}j)$$

$$= n_2 \left( \frac{n_1}{m_2} \right) \left( \frac{m_2}{m_2^{(n_2)}} \right) = n_2 \left( \frac{n_1}{m_2^{(n_2)}} \right).$$  \hfill (11.33a)

This is in the same form as the standard equation (11.1), except that the actual sample size $m_2$ is replaced by $m_2^{(n_2)}$, the sample size which would have been obtained if the overlap sample were not affected by special factors such as conditioning or dependence between the capture and recapture samples. Incidentally, the ratio $(n_1 / m_2)$ can also be seen as the constant weight applied to all $n_2$ cases in the standard equation (11.1).
B. Sample $n_1$

Often both the samples, $n_1$ and $n_2$, are subject to similar haphazard variations. This is quite likely to be the case when the two samples are similar in methodology, timing and size. This is less likely to be the case with administrative lists based on different procedures and sources. When the assumption about $n_1$ and $n_2$ being similar in haphazard variations is reasonable, estimates of haphazard variations in selection probabilities from $m_2$ may be extended also to $n_1$ on the basis of matched unit characteristics.

If $n_1$ is conducted in the same circumstances using the same procedures as $n_2$, the above extrapolation can be applied as well to all units $j(g)$ within each group $g$ in sample $n_1$. That is, we assume that for any group $g$ defined in terms of certain common characteristics in the three samples $(m_2, n_2, n_1)$, the averaged haphazard weights are the same for units in the group in all the three samples.

This is a strong assumption. It is unlikely to be valid if the sources of $(n_2, n_1)$ are of different types, e.g. two different lists, one list and one sample, or two surveys with different methodologies and conditions.

11.9 Use of more than two data sources

Throughout we have assumed that the capture-recapture procedure is applied with two sources of information (samples or lists $n_1$ and $n_2$, with overlap $m_2$). This is likely to be the main realistic option available for application to child labour surveys in developing countries.

However it is also useful to indicate variations from this model, as sometimes they may be feasible and useful. In the following we briefly indicate these possibilities, without going into details.

(1) Many applications of the capture-recapture procedure use more than two data sources.

(2) Certain types of information can also be derived from a single source.

(3) We note in passing that sometimes attempt is made to obtain the capture-recapture information during the course of a single interview – the so called “simulated capture” procedure. Essentially, it seeks hypothetical information on whether the respondent would have been present if a similar visit had been made in the past, or the respondent would be present if such a visit is made in the future.

We will comment briefly only on (1) concerning the use of more than two sources.

Log-linear models are accepted as one of the most useful representations of data in the form of counts (Fienberg 1972; Bishop, Fienberg and Holland 1975; Cormack 1989; also Wittes and Sidel 1968). In the log-linear approach the data are regarded as an incomplete $2^t$ contingency table ($t$ is the number of sources). The missing cell in the table comprises units missing from all the data sources (samples, lists). Log-linear models are fitted to other cells, and the missing cell is estimated by making some
assumption – usually assuming that there is no “interaction” involving all the $t$ data sources.

To indicate the structure of such a model for application to the capture-recapture design, let us begin with the situation in which the capture-recapture design involves only two independent simple random samples, with sampling rates $p_1$ and $p_2$. It is necessary to use a somewhat different notation than the one used hitherto so as to be able to extend the model to the more general case with more than two sources.

Let $m_{ij}$ be the expected number counts in a category defined by $i$ and $j$. The first index $(i)$ indicates the presence of a unit in the first source, say ‘yes present’ as $i = 1$ and ‘not present’ as $i = 0$. Similarly the second index $(j)$ indicates the presence in the second source: present as $j = 1$ and not present as $j = 0$. Thus $m_{11}$ is the expected number of units present in both samples, $m_{10}$ as the expected number present in the first source and not present in second; $m_{01}$ corresponds to the inverse situation: not present in first and present in the second source. Finally, $m_{00}$ is the number not present in either source.

If the two samples are independent, respectively with constant sampling rates $p_1$ and $p_2$, then

$$m_{11} = N p_1 p_2; \quad m_{10} = N p_1 (1 - p_2); \quad m_{01} = N (1 - p_1) p_2; \quad m_{00} = N (1 - p_1)(1 - p_2).$$

Writing $\ln(m_{ij}) = l_{ij}$, the above become

$$l_{11} = \ln(N) + \ln(p_1) + \ln(p_2), \quad l_{01} = \ln(N) + \ln(1 - p_1) + \ln(p_2), \quad l_{10} = \ln(N) + \ln(p_1) + \ln(1 - p_2), \quad l_{00} = \ln(N) + \ln(1 - p_1)(1 - p_2).$$

In summary we have the model

$$l_{11} = u, \quad l_{01} = u + u_1, \quad l_{10} = u + u_2.$$

In the terminology of standard log-linear models of contingency tables, $u$ is the main effect, and $(u_1, u_2)$ are first order effects of the absence from, respectively, the first and the second samples. We have three observations $(l_{11}, l_{01}, l_{10})$ and four parameters $u$, $u_1$, $u_2$ and $N$. Some assumption is required about the cell, $l_{00}$ corresponding to units missing from all the sources. Normally it is in the form

$$l_{00} = u + u_1 + u_2 + u_{12},$$

where the interaction term $u_{12}$ is present if the two sources are not independent (e.g. conditioning effect of the earlier survey on the later survey). The above equation
cannot be solved because both sides contain an unknown \((l_{00}, u_{12})\). When sources are independent, \(u_{12}=0\), and all parameters can be estimated.

*With only two sources, no information can be obtained on \(u_{12}\) – lack of independence between the sources.*

When we have three sources, interaction between all pairs of sources can be identified but not the interaction between all the three sources. The model is in the form

\[
\begin{align*}
I_{111} &= u \\
I_{011} &= u + u_1, \quad I_{101} = u + u_2, \quad I_{110} = u + u_3, \\
I_{001} &= u + u_1 + u_2 + u_{12} \\
I_{010} &= u + u_1 + u_3 + u_{13} \\
I_{100} &= u + u_2 + u_3 + u_{23} \\
I_{000} &= u + (u_1 + u_2 + u_3) + (u_{12} + u_{23} + u_{31}) + u_{123}. 
\end{align*}
\]

Again, the last equation has unknowns on both sides \((l_{000}, u_{123})\); the other terms in this expression can be evaluated from the preceding seven equations. Hence we have 7 observations and 8 parameters to be estimated. Some assumption has to be made about three way interaction \(u_{123}\). Typically we assume that \(u_{123} = 0\).

### 11.10 Some guidelines and recommendations

Below we briefly review main stages of a capture-recapture survey, from identification of the population to estimation from survey data. The objective is to provide some guidelines and recommendations concerning practical aspects.

#### 11.10.1 Identification of the target population

Capture-Recapture sampling can be useful for populations which are relatively small, preferably compact with a high level of concentration, and are well-defined and demarcated.

This is necessary because the capture-recapture procedure requires a sufficient overlap between the samples, which necessitates relatively large sampling rates. Therefore, for the sample sizes to remain within manageable and affordable limits, the population size needs to be small.

Normally the population extent - the physical limits to which the survey results may be generalised - is defined in terms of geographical areas, sites or other workplaces, institutions or facilities, etc., where children in the target population can be found. These form the ‘locations’ from which sampling ‘points’ are to be selected.

The population is also defined in terms of a population ‘time-frame’ such as a season or quarter of the year, or some other similar longer or shorter period. This defines the period to which the survey results are meant to apply. The survey period is a sample in time from this population-time frame.
The target population should not be subject to large flows into or out of the population locations within the population-time frame, other than short term temporary movements. In order to give non-zero selection probabilities to all units in the target population, every individual should be contactable at least at some point/location in the population within the population time-frame.

A fundamental issue is to determine whether the capture-recapture methodology is appropriate for the type and conditions of child labour being studied. It is useful to quote some practical advice from Jensen and Pearson (2002). The capture-recapture methodology

“will not be perfectly suited for every type of the WFCL (worst form of child labour) …. there are certain conditions that are better suited for this strategy. For example, it will work for all situations where children perform the work in specific well-defined places (or, where children pass through specific, well-defined areas). The method will also work best for any place where children are visible, and unlikely to hide what they are doing. Note also though that the place children are observed and interviewed need not be the place of work. For example, if most children gather in a central place unrelated to work, perhaps for entertainment or free food distributed by an organization or body, or if all children working in a mine go to a specific school or centre for benefits, or if children work on a piece-rate basis and bring items gathered to a few central stations (i.e. as with commercial agriculture, mining, or ragpickers/scavengers), we can apply this method.

“Thus, the better settings for implementation include the urban informal sector, ragpickers, porters, and possibly commercial agriculture, fishing and mining. Of course, there is no simple rule, and the method may not be well-suited for some of these cases in a particular country, or there may be other cases where the method works well for other forms of the WFCL (for example, if children engaged in the commercial sex trade are found in the open, for example at beach areas). Exactly which sites are chosen for interviews, and whether these actually capture most of the incidence of a given form of child labour needs to be determined based on local expertise and prior knowledge from key informants. For other forms of the WFCL (like bonded and trafficked children, or child domestic workers), the researcher may have to rely on alternative methods, either instead of, or as a supplement to, this approach”.

11.10.2 Using administrative lists

In applications of the capture-recapture methodology in health, medicine and also some social sectors in many developed countries, use is often made of incomplete lists rather than special interview surveys conducted for the purpose. In applications to child labour studies in developing countries, by contrast, reliance is generally on sample surveys as the source of capture-recapture data. Nevertheless, even in developing countries, there are often lists of working children available – e.g. with welfare and public health agencies – which should be exploited where possible.

In developed country applications, lists may be incomplete, but usually different lists overlap, and each list covers a large proportion of the target population. This is less
likely to be the case in our context in developing countries. Lists, if available, are mostly incomplete, and furthermore, tend to be non-overlapping, each covering a particular segment of the population. The best strategy is likely to be to combine whatever lists are available eliminating duplications, and examine whether the resulting combined list can serve as one of the sources for capture-recapture application – the other source being a special survey representative of the whole target population.

Even if such a list cannot serve as a source for the capture-recapture application, it may still be useful as a supplement to the standard two survey design. The list may contain information on a sub-population of children with special characteristics, hence of special interest to the study objectives. Such a list, or a sample from it, may be used to conduct a special supplementary survey. Children already covered in the list may be removed from the survey-based components, and then the result from the two sources combined.

11.10.3 Survey-based approach

Each of the surveys in the capture-recapture application should be designed to be independent random samples of the whole target population. Independence is required not only in design and selection, but also in application. Survey interviews should be kept as short as possible, and any threatening situations on the one hand, and anything more than token incentives on the other, should both be avoided. The surveys should be close in time, conducted under similar conditions (e.g. on similar days and similar times of the day), and desirably also using the same methodology (questionnaire, mode of interview, type of interviewer, etc.).

One possible model is to conduct the two surveys simultaneously, each by a different team of interviewers, and asking whether the child has been already interviewed by the other team. If an interview has already taken place, no more than identification information for matching need be collected. The interviewing at individual locations should nevertheless be separated, and coordinated so that the two surveys do not interfere with each other.

Sample sizes for the two surveys can be different, but in the just mentioned model at least, the same or similar sizes are preferable.

11.10.4 Choice of sample size

In this section, some recommendations are made of the choice of sample size for capture-recapture surveys in the context of child labour.

As noted, capture-recapture design is suitable for relatively small and reasonably clustered populations. This is necessary to achieve sufficient overlap between samples while keeping the sample sizes modest.
Table 11.10. Illustration of required sample sizes and overlap in a capture-recapture survey

<table>
<thead>
<tr>
<th>Population size N</th>
<th>10,000</th>
<th>9,000</th>
<th>8,000</th>
<th>7,000</th>
<th>6,000</th>
<th>5,000</th>
<th>4,000</th>
<th>3,000</th>
<th>2,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size n₁ = n₂</td>
<td>100</td>
<td>111</td>
<td>125</td>
<td>143</td>
<td>167</td>
<td>200</td>
<td>250</td>
<td>333</td>
<td>500</td>
</tr>
<tr>
<td>Approximate workload involved in the two samples, n₁ + n₂ - m₂</td>
<td>1,000</td>
<td>1,400</td>
<td>1,800</td>
<td>2,200</td>
<td>2,600</td>
<td>3,000</td>
<td>3,400</td>
<td>3,800</td>
<td>4,200</td>
</tr>
</tbody>
</table>

For a given shaded and bordered cell, the value on the x-axis gives a recommended "all purpose sample size" (n₁, n₂) corresponding to the population size N on the y-axis.

In Table 11.10A we consider a range of population sizes 2,000-10,000 – this is merely for illustration. At least two things are required for decision about the sample size: approximate size of the population expected, and what sort of precision is required for the improved estimate of the population size. The table shows various results for a range of sample sizes n₁ = n₂ = n, say from 200-2,000. If n₁ and n₂ are different, then n can be seen as their geometric mean: n² = n₁n₂.

We will assume that the two samples are simple random and independent samples. It is sufficient to take this simple case to indicate the order of the required sample sizes.

Panel (A) of the table shows the expected overlap (m₂) between the two samples under the above assumption of their independence:

\[ m₂ = \frac{n₁n₂}{N} = \frac{n²}{N}. \] (11.34)

The marked cells in the table indicate, for an assumed (a ‘guesstimate’ of the) population size (row), the sort of sample size n which should be adequate for most purposes. The concept ‘adequate for most purposes’ will be explained below.

The row at the bottom of the panel (A) shows roughly the total number of sample cases ‘adequate for most purposes’. This is n₁ + n₂ less the overlapping part m₂ (in order to avoid the overlap being double-counted). It provides an indication of the magnitude of the work involved considering both the samples.

The coefficient of variation of estimated population size is approximately:

\[ cv \left( N \right) = \frac{1}{\sqrt{m₂}}. \] (11.35)
In normal sample surveys, the precision of the estimate depends primarily on sample size, and not on the size of the population other than the secondary effect of ‘finite population correction’. By contrast in a capture-recapture design, the precision depends on the overlap. For achieving a given level of precision, i.e. a required overlap $m_2$, it follows from equations (11.34) and (11.35) that the required sample size $n$ increases as square root of the population size $N$.

This is shown (in per cent terms) in Table 11.10B. Starting from top left corner of the table and proceeding towards the opposite corner, the table has been divided into 5 regions with the following meaning. The criteria used are an adaptation of recommendations in Seber (1982, Section 3.1.5).

<table>
<thead>
<tr>
<th>Population size $N$</th>
<th>(I)</th>
<th>(II)</th>
<th>(III)</th>
<th>(IV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,000</td>
<td>50</td>
<td>25</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td>9,000</td>
<td>47</td>
<td>24</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>8,000</td>
<td>45</td>
<td>22</td>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>7,000</td>
<td>42</td>
<td>21</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>6,000</td>
<td>39</td>
<td>19</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>5,000</td>
<td>35</td>
<td>18</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>4,000</td>
<td>32</td>
<td>16</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>3,000</td>
<td>27</td>
<td>14</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>2,000</td>
<td>22</td>
<td>11</td>
<td>7</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample size $n_1 = n_2$</th>
<th>(I)</th>
<th>(II)</th>
<th>(III)</th>
<th>(IV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>%cv &gt; 25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%cv 15-25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%cv 7-15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%cv 4-7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%cv &lt; 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(I) $cv(N) > 25\%$, not worthwhile

The estimate of $N$ is likely to be too imprecise to serve much useful purpose. It is not useful to conduct such a small survey using capture-recapture methodology. 

(II) $cv(N) 15 – 25\%$, useful for planning

These sample sizes would be generally adequate for preliminary studies or management surveys where only a rough idea of the population size is needed.

(III) $cv(N) 7 – 15\%$, useful for managers

These sizes could be recommended for more accurate management work.

(IV) $cv(N) 4 – 7\%$, useful for analysis

Sample sizes should be suitable for more careful and detailed research into population changes and dynamics.

(V) $cv(N) < 4\%$, not necessary
Such sizes are also needed if estimated $N$ forms an input for computing more complex statistics also involving other estimates such as ratios or measures of change.

Such large sample sizes are unlikely to be necessary in practice for any purpose.

On the margin between (III) and (IV) are the sort of compromise sample sizes which should be able to meet different needs. The expected size of the overlap corresponding to these run along the diagonal from bottom left to top right in Tables 11.10A-C, with cells marked with a border.

Table 11.10C shows the corresponding standard error of $N$: $\text{StErr}(N) = N \cdot \text{cv}(N)$.

<table>
<thead>
<tr>
<th>Population size $N$</th>
<th>StErr(N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,000</td>
<td>5,000</td>
</tr>
<tr>
<td>9,000</td>
<td>4,269</td>
</tr>
<tr>
<td>8,000</td>
<td>3,578</td>
</tr>
<tr>
<td>7,000</td>
<td>2,928</td>
</tr>
<tr>
<td>6,000</td>
<td>2,324</td>
</tr>
<tr>
<td>5,000</td>
<td>1,768</td>
</tr>
<tr>
<td>4,000</td>
<td>1,265</td>
</tr>
<tr>
<td>3,000</td>
<td>822</td>
</tr>
<tr>
<td>2,000</td>
<td>447</td>
</tr>
</tbody>
</table>

Finally, Table 11.10D shows the implied range of the estimated value of population size $N$. This is computed as $\left( N \pm \text{StErr}(N) \right)$, that is, as the “67 per cent confidence interval”. Though it is conventional to use 95 per cent confidence intervals $\left( N \pm 2\text{StErr}(N) \right)$, we feel that in the present application, for most needs of planning, management and policy, a lower level of confidence is sufficient, and possibly also more useful and informative for decision making.
Table 11.10 (cont.)

<table>
<thead>
<tr>
<th>Population size ( N )</th>
<th>( L ) = ( N - StErr(N) ), and ( U ) = ( N + StErr(N) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( L ) 10,000</td>
<td>5,000 5,000 8,333 8,750 9,000 9,167 9,286 9,375 9,444 9,500</td>
</tr>
<tr>
<td>( U ) 10,000</td>
<td>15,000 12,500 11,667 11,250 11,000 10,833 10,714 10,625 10,556 10,500</td>
</tr>
<tr>
<td>( L ) 9,000</td>
<td>4,731 6,865 7,577 7,933 8,146 8,333 8,750 8,466 8,526 8,573</td>
</tr>
<tr>
<td>( U ) 9,000</td>
<td>13,269 11,135 10,423 10,067 9,854 9,712 9,610 9,534 9,474 9,427</td>
</tr>
<tr>
<td>( L ) 8,000</td>
<td>4,422 6,211 6,807 7,106 7,284 7,404 7,489 7,553 7,602 7,642</td>
</tr>
<tr>
<td>( U ) 8,000</td>
<td>11,578 9,789 9,193 8,994 8,716 8,596 8,511 8,447 8,398 8,358</td>
</tr>
<tr>
<td>( L ) 7,000</td>
<td>4,072 5,536 6,024 6,268 6,414 6,512 6,582 6,634 6,675 6,707</td>
</tr>
<tr>
<td>( U ) 7,000</td>
<td>9,928 8,464 7,976 7,372 7,586 7,488 7,418 7,366 7,325 7,293</td>
</tr>
<tr>
<td>( L ) 6,000</td>
<td>3,676 4,838 5,225 5,419 5,535 5,613 5,668 5,710 5,742 5,768</td>
</tr>
<tr>
<td>( U ) 6,000</td>
<td>8,324 7,162 6,775 6,581 6,465 6,387 6,332 6,290 6,258 6,232</td>
</tr>
<tr>
<td>( L ) 5,000</td>
<td>3,232 4,116 4,411 4,558 4,646 4,705 4,747 4,779 4,804 4,823</td>
</tr>
<tr>
<td>( U ) 5,000</td>
<td>6,768 5,884 5,589 5,442 5,354 5,259 5,253 5,221 5,196 5,177</td>
</tr>
<tr>
<td>( L ) 4,000</td>
<td>2,735 3,368 3,578 3,684 3,747 3,789 3,819 3,842 3,859 3,874</td>
</tr>
<tr>
<td>( U ) 4,000</td>
<td>5,265 4,632 4,422 4,316 4,253 4,211 4,181 4,158 4,141 4,126</td>
</tr>
<tr>
<td>( L ) 3,000</td>
<td>2,178 2,589 2,726 2,795 2,836 2,863 2,883 2,897 2,909 2,918</td>
</tr>
<tr>
<td>( U ) 3,000</td>
<td>3,822 3,411 3,274 3,205 3,164 3,137 3,117 3,103 3,091 3,082</td>
</tr>
<tr>
<td>( L ) 2,000</td>
<td>1,553 1,776 1,851 1,888 1,911 1,925 1,936 1,944 1,950 2,000</td>
</tr>
<tr>
<td>( U ) 2,000</td>
<td>2,447 2,224 2,149 2,112 2,089 2,075 2,064 2,056 2,050 2,000</td>
</tr>
</tbody>
</table>

Table 11.10D. Approximate %67% confidence interval*: \( N -/+/ (1 \text{ standard error}) \) corresponding to \( StErr(N) \) in the previous table.

11.10.5 Sample design

When study locations are few in number, it is best to aim at obtaining a simple random sample at each location. A simple random sample throughout would ideally mean a sample size proportional to the population contactable at that site or location, and hence some knowledge of that size. If the sizes significantly differ or cannot be reasonably guessed, one may seek a few (say 2-4) groups of sites which may have very different sampling rates, and use the standard capture-recapture estimator for each separately to estimate its size, \( N_g = \left( n_{1g}n_{2g}/m_{2g} \right) \), and hence the sampling rates \( n_{1g}/N \) and \( n_{2g}/N_g \) which have been applied to the two samples.

The inverse of these sampling rates are then used as (a component of) the weights for estimation. The above estimate of individual \( N_g \) values are meant to be used only for weighting. Usually the interest is mainly in estimating the total population size \( N \).

For substantive reasons, different sampling rates at different sites may also be introduced deliberately.

When there are many survey locations, it could be necessary and even desirable to introduce two-stage sampling, a sample of locations followed by a simple random
sample of children within each location. Children may be selected through their households, establishments or other places where they are found.

Two options are possible for selecting locations for the second (recapture) sample: to take this sample completely independently of the first sample, or to use a common sample of locations for both the samples. Usually, this second option is likely to be more economical and convenient. It will also give a larger overlap between the samples. However, necessary questions must then be asked to determine the probability that a child remains within the common sample areas between the two surveys, so that the effect of dependence between the two samples can be estimated and reduced, as explained earlier.

11.11 Simple illustration of the various procedures discussed

In the preceding sections, we have described the capture-recapture methodology, discussing a full range of issues including the effect of departures from the underlying assumptions, the use of the methodology in general settings where estimates for different domains are required, when sampling is carried out in multiple stages, when there are non-responding units, etc.

The purpose of this section is to illustrate some of these issues with a simple numerical example where the target population comprises 5 working children and capture-recapture sampling is used to estimate the unknown number of working children in the population and their characteristics such as sex and earnings. The distributions of the estimates are obtained for different values of the underlying parameters. Then we examine the effects of departure from the sampling assumptions (closed population, independence, and homogeneity of capture and recapture probabilities), and derive the sampling weights for estimating different characteristics of the working children assuming some non-response, and other design or non-design perturbations.

All the following illustrations in this section come from Mehran (2012).

11.11.1 Children selling flowers at traffic lights

Consider a population of 5 children selling flowers at traffic lights on a street junction of a city neighbourhood every afternoon from 2 to 6 pm. There are three boys and two girls. The presence or absence of children at work is independent of each other. On normal days each child has a probability $p=50$ per cent of being at work in a given afternoon. On sunny days the probability of presence increases to $p=85$ per cent, and on rainy days decreases to $p=15$ per cent.

A. Normal days

During a random visit on a normal day, 3 of the children were seen at work in the junction. In a second visit the next normal day, there were 4 children at work, two of them the same as on the day before. Based on this information, can one find the most likely estimate of the total number of children selling flowers at this street junction?
According to the capture-recapture methodology, the number of children found selling flowers on the first day constitutes the capture \((n_1=3)\) and the number of children found on the second day the recapture \((n_2=4)\). The number of children found working on both days is the overlap between the capture and the recapture \((m_2=2)\). The capture-recapture estimate of the unknown total number of children selling flowers at that junction is then

\[
\frac{n_1 \times n_2}{m_2} = \frac{3 \times 4}{2} = 6.
\]

(The actual total is 5). This particular example is just one among many other possibilities. On the first day we could have found 2 children working at the junction instead of 3, or 4 or just 1, or simply none, or in fact all five of them. Similarly on the second day.

In total, 1024 different patterns of observing children at the street junction could be envisaged on any two days, four for each child: (1) at work on both days; (2) at work on day 1, but not on day 2; (3) at work on day 2, but not on day 1; and (4) not at work in both days. The 1024 possibilities give rise to 11 distinct capture-recapture estimates of the total number of working children at the traffic light. The distinct values and their probabilities are shown in Table 11.11.

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5%</td>
</tr>
<tr>
<td>2</td>
<td>4.9%</td>
</tr>
<tr>
<td>3</td>
<td>12.7%</td>
</tr>
<tr>
<td>4</td>
<td>20.0%</td>
</tr>
<tr>
<td>4.5</td>
<td>5.9%</td>
</tr>
<tr>
<td>5</td>
<td>6.0%</td>
</tr>
<tr>
<td>5.33</td>
<td>2.0%</td>
</tr>
<tr>
<td>6</td>
<td>17.6%</td>
</tr>
<tr>
<td>8</td>
<td>3.9%</td>
</tr>
<tr>
<td>9</td>
<td>2.9%</td>
</tr>
<tr>
<td>?</td>
<td>23.7%</td>
</tr>
<tr>
<td>Total</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

‘?’ corresponds to the situation when there is no recapture \((m_2=0)\). On ‘normal days’ each child has a 50% chance of being at work.

The capture-recapture estimate, 6, could have resulted if any three children were working on the first day and any four on the second as long two of them were also working on the first day. By symmetry, the same estimate, 6, would be obtained if the number of children found working in the first and second day were reversed \((n_1=4\) and \(n_2=3)\). Another pattern (and its reverse) that would also lead to the estimate, 6, is three children found on the first day, two on the second day, one of whom being the same child as on the first day.

In total the probability of any of these patterns to happen is 17.6 per cent as shown in the middle row of the table corresponding to the estimate 6. This value is obtained from
the multinomial distribution with parameters \((p^2, pq, qp, q^2)\), the elements correspond respectively to the probabilities of a child at work on both days \((p^2)\), at work on first day but not on the second \((pq)\), not at work on the first day but at work on the second \((qp)\), and not work in both days \((q^2)\), where \(p=0.5\), is the probability of a child working on a normal day and \(q = 1 - p\).

Among the 1024 possible patterns, there are 243 for which the capture-recapture estimate cannot be calculated with the usual estimation equation \((11.1)\). These correspond to the situations where no child is found at work in both days (that is, \(m_2 = 0\)). The probability of this happening on normal days is 23.7 per cent as shown in the bottom part of the table.

The distribution of the capture-recapture estimates over all possible numeric values is shown graphically in Figure 11.2. The distribution is bi-modal with hollow middle. The mean of the distribution is 4.7 and the standard deviation 1.7. It is instructive to note that although the capture-recapture estimate is a maximum likelihood estimate, it is biased in small samples. In this example, the bias is negative, \(-0.3=4.7-5.0\), corresponding to an underestimation of 6 per cent.

**Figure 11.2. Illustration of a probability distribution of capture-recapture estimates (normal day)**

![Graph showing a probability distribution](image)

**B. Sunny or rainy days**

On sunny days, the chances that the children go selling flowers increase \((p=85\text{ per cent})\) and a larger number of children are likely to be found at the traffic lights on each day. The capture and recapture estimate becomes more substantial and more accurate as shown in Figure 11.3. The probability distribution of the capture-recapture estimate on sunny days when \(p=85\text{ per cent}\) is well centred around the true value 5, and the tails are narrow but somewhat skewed to the right. The mean of the distribution is 5.1 and the standard deviation 0.6.

---

\(^{46}\) A modified formula for the capture-recapture estimate defined also when \(m_2 = 0\) is given by Equation \((11.4)\).

\(^{47}\) See, for instance, Seber (1982, Section 12.8.6).
 CHAPTER 11

11. Capture-recapture sampling

By contrast, on rainy days, the chance that the children go selling flowers diminishes (p=15 per cent) and a smaller number of children are likely to be found at the traffic lights on each day. The capture and recapture estimate becomes much lower and less accurate. Figure 11.4 shows the resulting probability distribution of the estimate. It is heavily skewed to the left with light tail on the right. The mean of the distribution is 2.3 and the standard deviation 1.3.

C. Mixture of days

Now consider the situation where there is mixture of days, that is the weather conditions on the days of capture and recapture are different. There are three possibilities when the weather of the two days is not the same: one of the days is normal (p=50 per cent) the other sunny (p=85 per cent); one of the days normal (p=50 per cent) the other rainy (p=15 per cent); and one of the days sunny (p=85 per cent) the other rainy (p=15 per cent). The corresponding probability distributions of the capture-recapture estimate are shown in Figures 11.5A-C.
It may be observed that the shape of the distribution is to some extent determined by the larger of the probabilities. Thus, if the day of capture was normal and the day of recapture sunny, the capture-recapture estimate (mean = 5.1, standard deviation = 1.1) behaves more or less as if the two days were sunny (mean = 5.1, standard deviation = 0.6).

Figure 11.5A. Illustration of a probability distribution of capture-recapture estimates (normal and sunny days)

Similarly, if the day of capture was normal and the day of recapture rainy, the capture-recapture estimate (mean = 3.6, standard deviation = 1.5) behaves more or less as if the two days were normal (mean = 4.7, standard deviation = 1.7).

Figure 11.5B. Illustration of a probability distribution of capture-recapture estimates (normal and rainy days)

Finally, if the day of capture was sunny and the day of recapture rainy, the capture-recapture estimate (mean = 4.7, standard deviation = 1.0) behaves more or less as if the two days were sunny (mean = 5.1, standard deviation = 0.6).
11.11.2 Departure from sampling assumptions

The capture-recapture methodology assumes a closed population, independence of the capture and recapture samples, and homogeneity of the probability of selection of the sampling units. The effects of deviations from these assumptions are examined in turn, using the child workers example.

A. Turnover

The five children selling flowers at the traffic light of the street juncture do not form a closed population. On some days, new children join in or move out. In fact, on the first day of the capture-recapture survey, a new boy came to the neighbourhood and was equally likely as the others to join the street junction for selling flowers at the traffic lights. Thus, his probability of working on a normal day was also $p=50$ per cent, on sunny days $p=85$ per cent and on rainy days $p=15$ per cent.

On the next visit, this child had gone out of the neighbourhood and therefore had zero probability of being “recaptured” on the second day. But, a new boy had arrived on the second day with the same set of probabilities of selling flowers at the traffic light according to weather. This child had thus a positive probability of being in the “recapture” sample, but had zero probability of being part of the initial “capture” sample.

The effect of the population turnover on the capture-recapture estimate may be examined by considering all possible cases (4096). Ignoring turnover in the calculation of the capture-recapture formula gives the distribution in Table 11.12.
### Table 11.12. Probability distribution of capture-recapture estimates on normal days (p=0.5) ignoring population turnover

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1%</td>
</tr>
<tr>
<td>2</td>
<td>1.5%</td>
</tr>
<tr>
<td>3</td>
<td>4.6%</td>
</tr>
<tr>
<td>4</td>
<td>9.5%</td>
</tr>
<tr>
<td>4.5</td>
<td>3.2%</td>
</tr>
<tr>
<td>5</td>
<td>5.2%</td>
</tr>
<tr>
<td>5.33</td>
<td>1.7%</td>
</tr>
<tr>
<td>6</td>
<td>17.1%</td>
</tr>
<tr>
<td>6.25</td>
<td>0.4%</td>
</tr>
<tr>
<td>6.67</td>
<td>2.4%</td>
</tr>
<tr>
<td>7.2</td>
<td>0.0%</td>
</tr>
<tr>
<td>7.5</td>
<td>3.7%</td>
</tr>
<tr>
<td>8</td>
<td>9.8%</td>
</tr>
<tr>
<td>8.33</td>
<td>0.5%</td>
</tr>
<tr>
<td>9</td>
<td>5.6%</td>
</tr>
<tr>
<td>10</td>
<td>3.7%</td>
</tr>
<tr>
<td>12</td>
<td>5.6%</td>
</tr>
<tr>
<td>15</td>
<td>1.0%</td>
</tr>
<tr>
<td>16</td>
<td>0.7%</td>
</tr>
<tr>
<td>?</td>
<td>23.7%</td>
</tr>
<tr>
<td>Total</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

'?’ corresponds to the situation when there is no recapture ($m_2 = 0$). ‘Turnover’ is as described in the text. But in the presented estimates, it has been ignored, i.e. the closed population formula (11.1) has been used.

The distribution for normal days, when estimated ignoring the turnover, is shown graphically in Figure 11.6, where it is compared with the corresponding distribution under a closed population. It can be observed that with turnover in the population, the distribution of the capture-recapture estimates tends to shift to higher values and cover a wider range of numbers. Generally, ignoring turnover overestimates the population size.

**Figure 11.6. Probability distribution of estimate of number of child workers (normal days)**
With turnover in the population, the mean of the distribution shifts to 6.8 against 4.7 under a closed population. Also, the standard deviation of the distribution, 2.8, indicates a wider range of variation as compared to 1.7 under a closed population.

Similar results are obtained on the shift in the distribution of estimates for sunny or rainy and for mixture of days. Table 11.13 compares the mean and standard deviation of the distributions under closed and turnover populations for different weather conditions.

When there is turnover in the population, it is important to specify the exact nature of the population to be estimated. There are essentially three distinct populations which may be described as follows: (a) core population (i.e. the population present at both capture and recapture periods); (b) average population (i.e. the average of the population present at capture period and the population present at recapture period); and (c) exposed population (i.e. the population present at either capture or recapture period).

In our example, the core population is 5, the average population is 6, and the exposed population is 7. The results for normal days show that the capture-recapture estimate calculated by ignoring the turnover of the population (6.8) tends to grossly overestimate the core population, overestimate the average population, and provide good approximation of the size of the exposed population.

Table 11.13. Mean and standard deviation of the distribution of capture-recapture estimates of number of child workers under closed and turnover populations for different weather conditions

| Weather condition | Probability of working | Mean | | | | Standard deviation |
|-------------------|------------------------|------|----------------|----------------|-----------------|
|                   |                        | Closed population | Pop with turnover | Closed population | Pop with turnover |
| Normal days       | $p_1=0.50, p_2=0.50$   | 4.7  | 6.8            | 1.7  | 2.8            |
| Sunny days        | $p_1=0.85, p_2=0.85$   | 5.1  | 7.4            | 0.6  | 1.4            |
| Rainy days        | $p_1=0.15, p_2=0.15$   | 2.3  | 2.7            | 1.3  | 1.7            |
| Normal, Sunny     | $p_1=0.50, p_2=0.85$   | 5.1  | 7.5            | 1.1  | 2.4            |
| Normal, Rainy     | $p_1=0.50, p_2=0.15$   | 3.6  | 4.7            | 1.5  | 2.3            |
| Sunny, Rainy      | $p_1=0.85, p_2=0.15$   | 4.7  | 6.3            | 1.0  | 2.1            |
| Mean              |                        | 4.3  | 5.9            | 1.2  | 2.1            |

B. Dependency

Another assumption of capture-recapture sampling is the independence of the capture and recapture samples. In some cases, however, this assumption may be violated in practice, for example when the likelihood of working on the second day depends on the amount of sales the child made on the preceding day. Suppose the odds of working the second day are doubled if the preceding day’s sales were low, while the odds are halved if the sales were high.

Let the joint probability of capture and recapture be represented by

\[ p_1 p_2 (1 + \theta), \]
where the parameter $\theta$ represents the degree of dependency between the capture and recapture probabilities. If the odds of working on the second day (recapture) is doubled because of poor sales on the first day, then

$$\theta = \frac{(1 - p_2)}{(1 + p_2)}.$$

Vice versa if the odds of working on the second day is halved because of good sales on the first day, then

$$\theta = \frac{(1 - q_2)}{(1 + q_2)}, \text{ with } q_2 = 1 - p_2.$$

If the odds don’t change, and working on the first and second days are independent of each other, then $\theta = 0$.

Figure 11.7 shows (dashed lines) the distribution of the capture-recapture estimates of the number of working children on normal days under the two levels of dependency. The solid line represents the distribution under independence.

**Figure 11.7. Illustration of dependent capture and recapture probabilities: estimate of number of child workers (normal days)**

In this example, dependence between capture and recapture probabilities has the effect of maintaining the bimodal shape of the distribution, but amplifying the lower tail of the distribution and flattening the upper tail when the dependence is positive and the reverse when it is negative. Table 11.14 compares the mean and standard deviation of the distributions of the number of working children under positive and negative dependence for different weather conditions.
11. Capture-recapture sampling

<table>
<thead>
<tr>
<th>Weather condition</th>
<th>Probability of working</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>θ=1/3</td>
<td>θ=0</td>
<td>θ=1/3</td>
</tr>
<tr>
<td>Normal days</td>
<td>p₁=0.50, p₂=0.50</td>
<td>5.7</td>
<td>4.7</td>
</tr>
<tr>
<td>Sunny days</td>
<td>p₁=0.85, p₂=0.85</td>
<td>7.8</td>
<td>5.1</td>
</tr>
<tr>
<td>Rainy days</td>
<td>p₁=0.15, p₂=0.15</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>Normal, Sunny</td>
<td>p₁=0.50, p₂=0.85</td>
<td>6.9</td>
<td>5.1</td>
</tr>
<tr>
<td>Normal, Rainy</td>
<td>p₁=0.50, p₂=0.15</td>
<td>3.9</td>
<td>3.6</td>
</tr>
<tr>
<td>Sunny, Rainy</td>
<td>p₁=0.85, p₂=0.15</td>
<td>5.3</td>
<td>4.7</td>
</tr>
</tbody>
</table>

The table results indicate that negative dependence has the tendency of increasing the mean value of the capture-recapture estimates, while positive dependence appears to have the reverse effect. The results on standard deviation show a similar pattern. This is because, by definition, negative dependence reduces the overlap between the two days, thus increasing the estimate of population size as can be seen from Equation (11.1). Conversely, positive dependence means increasing the overlap.

Information for determining the dependency parameter \( \theta \) may be obtained from respondents, as noted in an example given in Section 11.6 above.

C. Heterogeneity

The other assumption of capture-recapture sampling is homogeneity of the probability of selection of the sampling units. In the context of our example of child workers, this means that all children have the same probability of going to work on given days. Departure from the homogeneity assumption may occur if for example girls have different probabilities of working than boys, or if by design the capture-recapture study is conducted in the early afternoons (2-4 pm) and then separately in the late afternoons (4-6 pm) because it is thought that children are more likely to go working in daylight than at sunset.

The effect of departure from the homogeneity assumption has been examined on the assumption that boys have different probabilities of going to work on normal days (\( p=0.85 \)) than girls (\( p=0.15 \)). There are three boys and two girls in this example. The capture-recapture estimate assuming homogeneity is given by

\[
\frac{(n_1^{\text{boys}} + n_1^{\text{girls}}) \times (n_2^{\text{boys}} + n_2^{\text{girls}})}{(m_2^{\text{boys}} + m_2^{\text{girls}})},
\]

where \( n_1^{\text{boys}} \) and \( n_1^{\text{girls}} \) are respectively the number of boys and girls in the capture sample, \( n_2^{\text{boys}} \) and \( n_2^{\text{girls}} \) in the recapture sample, and \( m_2^{\text{boys}} \) and \( m_2^{\text{girls}} \) in the overlap. The capture-recapture estimate stratifying for heterogeneity between boys and girls is given by

\[
\frac{(n_1^{\text{boys}} + n_1^{\text{girls}}) \times (n_2^{\text{boys}} + n_2^{\text{girls}})}{(m_2^{\text{boys}} + m_2^{\text{girls}})},
\]
The probability distributions of the two estimates are compared in Figure 11.8. The dashed curve shows the distribution of the estimates calculated assuming homogeneity and the solid curve the distribution based on stratification between boys and girls.

**Figure 11.8.** Probability distribution of capture-recapture estimates of total number of child workers on normal days: effect of heterogeneity among boys and girls

The solid curve is better centred around the true value (5), indicating that stratification among boys and girls improves the capture-recapture estimate, both in terms of bias and variance. The mean of the distribution of estimates of the total number of child workers with stratification is 4.3 (3.0 boys and 1.3 girls), with standard deviation, 0.6.\(^{48}\)

---

\(^{48}\) For methods of dealing with heterogeneity in the case of multiple capture and recaptures, see Chao, A. (1987).
Chapter 12

Controlled selection and balanced sampling

12.1  Introduction

12.1.1  Controlled selection

Controlled selection is a procedure originally proposed by Goodman and Kish (1950) to control the structure of the sample beyond what is possible with ordinary stratification in which the random samples selected from each stratum are independent. Controlled selection permits extra control while conforming to the requirements of probability sampling.

Surveys, in particular of mobile and other difficult-to-access populations, often have to be restricted to a limited area and to a small number of primary units. Controlled selection is a very useful technique when we have to select a small sample of primary units, but at the same time ensure that it is ‘balanced’ and ‘representative’ of the population in terms of many characteristics (or control variables), such as administrative and geographical location, type of place, main economic activity, or level of development. When the sample is to be stratified with multiple variables, crossing all the strata with all the variables may cause the cells to become too small for an independent sample to be selected in each cell. In fact, in some cases the number of cross-stratified cells may actually be larger than the total sample size. In these cases, even a single unit may not be selected in each cell. By contrast to usual stratification, controlled selection permits drawing a sample that is stratified with respect to each variable, without partitioning the population into all the cells of the cross-stratification as the usual practice of stratification requires.

Some procedure of controlled selection is needed when the number of units to be selected falls short of the number of control categories we wish to impose.

In order to convey the basic idea behind this technique, perhaps we can do no better than quote from the original authors (Goodman and Kish, 1950):

“Controlled selection – definition

“As we define it, the expression ‘Controlled Selection’ has a rather broad meaning. This expression is defined to mean any process of selection in which, while maintaining the assigned probability for each unit, the probabilities of selection for some or all preferred combinations of \( n \) out of \( N \) units are larger than in stratified random sampling (and correspondingly the probabilities of selection for at least some non-preferred combinations are smaller than in stratified random sampling).

“It is to be noted that while in stratified random sampling the probabilities of selection for certain combinations are increased, the possibilities of
increasing the probabilities of selection for preferred combinations are by no means exhausted by this process. That is, with controlled selection it is in general possible to go very much further in increasing the probabilities of selection of preferred combinations of units than is done by stratified sampling alone. At the same time the probability of selection for a vast number of other combinations is reduced, generally to zero. It is with the use of control after the possibilities of stratification have been exhausted that this paper is concerned. ...........

“Simple illustration of control beyond stratification

“It is possible to introduce a measure of additional control in the selection process very easily. Suppose we have two strata consisting of first-stage units, from each of which one unit is to be selected. Suppose further that in stratum 1 units B, C, and F lie adjacent to the ocean or other major waterway, that in stratum 2 unit d is similarly located, and that all the other units are located inland.”

The units and the probabilities of selection assigned to them for the present illustration are shown in Table 12.1.

<table>
<thead>
<tr>
<th>Table 12.1. Illustration 1: a simple example of controlled selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illustrative data Re-arranged (see text)</td>
</tr>
<tr>
<td>Stratum 1 Unit Probability Stratum 2 Unit Probability</td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>A 0.10</td>
</tr>
<tr>
<td>B* 0.15</td>
</tr>
<tr>
<td>C* 0.10</td>
</tr>
<tr>
<td>D 0.20</td>
</tr>
<tr>
<td>E 0.25</td>
</tr>
<tr>
<td>F* 0.20</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Stratum 1 Unit Probability Stratum 2 Unit Probability</td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>B* 0.15</td>
</tr>
<tr>
<td>C* 0.10</td>
</tr>
<tr>
<td>F* 0.20</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

*coastal units

Joint probabilities assuming independent selection within strata

<table>
<thead>
<tr>
<th></th>
<th>Stratum 2</th>
<th>Coastal</th>
<th>Inland</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stratum 1</td>
<td>0.20</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>Coastal</td>
<td>0.45</td>
<td>0.09</td>
<td>0.36</td>
</tr>
<tr>
<td>Inland</td>
<td>0.55</td>
<td>0.11</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Source: Goodman and Kish, 1950.
“The sum of probabilities for B, C, F and d is 0.65. It is considered desirable to select one inland and one coastal unit, and undesirable to select two coastal units. The procedure is as follows:

“Re-arrange the units in the first stratum by listing B, C, F first, followed by A, D and E. Re-arrange the units in the second stratum by shifting d to the end, that is, by placing e above d. Then draw a random number from 0 to 100 and let this one number determine the selection in both strata. Thus if the random number is 45 or less, a coastal unit will be selected in stratum 1 and an inland unit in stratum 2, if the number is 46 to 80 an inland unit will be selected in both strata, and if the number is greater than 80 an inland unit will be selected in stratum 1 and a coastal unit in stratum 2.

“By means of the above procedure the probability of selecting one inland and one coastal unit is .65 and the probability of selecting two inland units is .35. With stratified random sampling on the other hand the probabilities are the following: one inland and one coastal unit, .47; two inland units, .44; two coastal units, .09. If one were more interested in securing what appeared to be a balanced sample than a strictly random sample he would be inclined to make sure that he selected one inland unit and one coastal unit to represent the two strata. By a procedure of controlled selection, however, the original assigned probabilities of each unit are rigorously maintained and the probability of securing one inland and one coastal unit is made as large as possible within the limitation of probability sampling.”

Hess and Srikantan (1966) provide a detailed numerical illustration of the procedure of controlled selection, along with the development and illustration of estimation and variance estimation formulae. They also describe a computer program for applying the procedure.

In the following sections, three detailed numerical illustrations are presented and discussed with the objective of explaining the procedure in concrete terms.

12.1.2 Balanced sampling

Controlled selection can also be viewed in the context of the modern theory of balanced sampling elaborated by Deville and Tillé (2004), thus providing the possibility of dealing with a wider range of issues and more efficient sampling algorithms.

The basic idea of balanced sampling is as follows. Let S be a random sample drawn from a population U with known and fixed unit selection probabilities \( \pi_k \). Let \( \{x_k\} = (x_{1k}, x_{2k}, ..., x_{pk}) \) be a vector of \( p \) auxiliary variables defined for all units \( k \) in the population. The sample is said to be balanced on these auxiliary variables if it satisfies the following ‘balancing equations’:

\[
\sum_{k \in s} \left( \frac{x_{ik}}{\pi_k} \right) = \sum_{k \in u} \left( \frac{x_{ik}}{\pi_k} \right) \quad \text{(12.1)}
\]

The right hand side is the actual total of \( x \) variables in the population; the left hand side gives their Horvitz-Thompson estimates from the sample. The objective of balanced sampling is to ensure that the sample estimates the known totals (or means) of the
auxiliary variables without sampling error. Balanced sampling improves the efficiency of the Horvitz-Thompson estimator when the variable of interest is correlated with the balancing auxiliary variables.

Balanced sampling has several potential advantages (see, e.g. Tillé, 2011). Balanced sampling increases the accuracy of the Horvitz-Thompson estimator. It protects against large sampling errors; in fact the most unfavourable samples can be assigned a zero selection probability. The procedure can be used to ensure that small domains receive sufficient and non-zero sample sizes. By controlling the sample characteristics at the design stage itself, balanced sampling reduces the need for introducing large ‘calibration’ weights at the subsequent stage of analysis. In fact, balanced sampling and calibration have similar objectives; the difference is that, while calibration is applied to a given sample at the estimation stage, balanced sampling is applied at the sample selection stage itself. Finally, as will be discussed below, balanced sampling is an elegant method for controlled selection and can be used efficiently in a wide range of applications. The theory also provides for approximate variance calculation, often a difficult task with the procedures of controlled selection.

In the present context, our interest is in the use of balanced sampling procedures as a tool for achieving controlled selection. The problem can be formulated as follows.

A sample is to be selected with specified unit selection probabilities $\pi_k$. The population is stratified according to each of a number of stratification variables $i$. Let subscript $(ij)$ indicate category $j$ of stratification variable $i$. These categories are non-overlapping and exhaustive. For each category of each stratification variable, we define an auxiliary variable

$$x_{ij,k} = \pi_k \delta_{ij,k},$$

where $\delta_{ij,k} = 1$ if unit $k$ is in category $(ij)$; $\delta_{ij,k} = 0$ otherwise. With $x_{ij} = \sum_k x_{ij,k}$ (summed over all units $k$ in the population $U$), the balancing equations require

$$\sum_s (x_{ij,k} / \pi_k) = \sum_k x_{ij,k},$$

leading to

$$n_{ij} = \sum_s \delta_{ij,k} = \sum_s (x_{ij,k} / \pi_k) = \sum_k x_{ij,k} = x_{ij}. \quad (12.3)$$

That is, the balancing equations require the sample size $n_{ij}$ from category $(ij)$ to equal the specified population total $x_{ij}$. This is controlled selection.

In practice, $n_{ij}$ has to be an integer, and may be taken as one of the two nearest integers to $x_{ij}$, for example as:

$$n_{ij} = \text{int}(x_{ij}) + 1, \text{ with probability } x_{ij} - \text{int}(x_{ij}), \text{ and}$$

$$n_{ij} = \text{int}(x_{ij}), \text{ with probability } \{ \text{int}(x_{ij}) + 1 \} - x_{ij}. \quad (12.4)$$

The “cube method” (Deville and Tillé, 2004) provides a general procedure for balanced sampling. It is a sampling algorithm that selects a balanced sample such that

(i) the given unit selection probabilities $\pi_k$ are satisfied exactly for all units $k$, and

(ii) the balancing equations $n_i = x_i$ are satisfied as well as possible for all strata $(ij)$.

The original authors describe the cube method as divided into two phases: a ‘flight phase’ and a ‘landing phase’. The former refers to searching for a balanced sample which satisfies both conditions (i) and (ii). If at the end of this phase no perfectly
balanced sample has been found, the landing phase seeks a sample satisfying condition (i) exactly, but condition (ii) only approximately as well as possible.

The cube algorithm is described in Tillé (2006). Several implementations of the algorithm are available free of charge in the public domain – such as an SAS/IML version (Chauvet and Tillé, 2006), and a package written in R language (Tillé and Matei, 2007). The latter algorithm is available as the function samplecube in R in the CRAN contributed package ‘sampling’ developed by Tillé and Matei (2012). The main inputs required are, for each unit $k$ in the population, the variable giving inclusion probabilities $\pi_k$, and values for a set of $p$ auxiliary variables $[x_{ik} = (x_{i1k}, x_{i2k}, ..., x_{ipk})]$ on which the sample is to be balanced. Available algorithms may differ in the details as to how the procedures are implemented, and generally provide options from which the user may choose.

12.1.3 Numerical illustrations

A simple illustration of controlled selection from Goodman and Kish (1950) has already been provided in Section 12.1.1. Section 12.2 provides an example to illustrate in greater detail implementation of the controlled selection procedure when the sample is stratified in terms of a number of variables. Section 12.3 provides another – a very useful yet quite simple – illustration. These three illustrations draw on the ideas and procedures of controlled selection as originally formulated by Goodman and Kish (1950). Section 12.4 examines the same three examples in-depth and reformulates them in terms of the balanced sampling procedure of Deville and Tillé (2004). Examples of balanced samples generated by the procedure are shown. Finally, Section 12.5 offers some concluding remarks on the relationship between the controlled selection and balanced sampling procedures.

12.2 Illustration 1. Stratification with extra controls using controlled selection procedures

12.2.1 Illustrative data

In this subsection, very detailed illustrations are provided of the application of controlled selection with a small scale numerical example. Our data (Table 12.2A) consists of 52 units. The units have been divided into 3 strata (A, B and C), respectively of 17, 16 and 19 units. Three units are to be selected, one from each stratum, with probability proportional to the unit measure of size (MoS) also shown in the table. We assume that the strata and MoS’s are given, determined by external substantive and statistical considerations. The strata are reasonably uniform in size (in terms of the sum of unit size measures in the stratum); as shown in the last column, the strata sizes vary from 86 to 110.

In practice it may be the case that given strata sizes vary considerably. If some strata are much larger than others, then consideration should be given to splitting the largest ones, and also to merging smaller ones if that is permitted by the survey objectives. If the total number of strata is changed by such reorganisation, then the target number
of units to be selected should also be adjusted. It is most convenient for controlled selection if the number of primary sampling units to be selected is the same as the number of primary strata, one unit being selected per stratum.

An example is provided in Table 12.2B. Here stratum A is nearly twice as large as the other strata B or C. This may be ‘a given’ on the basis of substantive considerations; for instance, the strata may correspond to three types or sectors of child labour activity to be covered in the survey, and one of the strata might be much bigger than the others in terms of the number of children involved. In such a case one option would be to divide A into two strata (A1, A2) of similar size, for instance by dividing the sector geographically, and to increase the number of units to be selected from 3 to 4 – one for each of the new strata. Another option would be to keep the stratification and the number of units to be selected unchanged. But then it should be understood that a unit in the large stratum would receive a lower chance of being selected than a unit of the same size in a smaller stratum. With one selection per stratum, the chance of the unit being selected is equal to its MoS divided by the sum of size measures of all the units in its stratum.

In any case, the unit MoS measures have to be scaled such that the sum is the same in each stratum. This constant stratum size represents one selection under our model of one selection per stratum. In Table 12.2C, the size measures in 12.2A have been rescaled in this manner, the constant stratum size taken as 100.0. (Each figure in 12.2A has been multiplied by 100/total stratum size.)

Incidentally, such rescaling will be required even if the original design involved equal rather than PPS selection of units within each stratum. (The conversion factor from 12.2A to 12.2C would be 100 divided by the number of units in the stratum, 100/17 in stratum A, 100/19 in stratum C.) Hence the original equal probability design has to be treated the same way as a PPS design. It is not necessary therefore to discuss equal probability sampling separately; it is covered by our illustration.

Of course if there were no constraints other than selecting one unit per stratum, the usual selection procedure can be applied separately to each stratum. As noted, the point of introducing controlled selection is that we also need to satisfy additional constraints when the number of strata times the number of additional control categories exceeds the number of selections to be made. For instance, we may have 3 strata each divided into 4 categories by type and size of place, but have only 3 units to select.
12.2 Illustration 1. Stratification with extra controls using controlled selection procedures

Table 12.2. Illustration of controlled selection procedure

Panel A. Data for illustration: strata and unit measures of size (MoS)

Objective is to select one unit per stratum with probability proportional to MoS. Table cells give unit measures of size

| unit id | stratum | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | total |
| A       |         | 9.0 | 1.0 | 2.0 | 2.0 | 7.0 | 2.0 | 8.0 | 7.0 | 10.0 | 2.0 | 8.0 | 8.0 | 5.0 | 7.0 | 4.0 | 9.0 | 6.0 | 97.0 |
| B       |         | 8.0 | 9.0 | 8.0 | 2.0 | 2.0 | 2.0 | 7.0 | 5.0 | 2.0 | 7.0 | 7.0 | 7.0 | 1.0 | 4.0 | 6.0 | 9.0 | 86.0 |
| C       |         | 4.0 | 8.0 | 5.0 | 7.0 | 9.0 | 1.0 | 5.0 | 7.0 | 8.0 | 10.0 | 3.0 | 4.0 | 4.0 | 1.0 | 5.0 | 4.0 | 10.0 | 7.0 | 8.0 | 110.0 |

Panel B. Illustration: Splitting of strata which are significantly larger than others

If some strata are significantly larger than others, they may be split to form more equal strata. This will also require selecting more units, in order to obtain one per stratum.

For example if:

| stratum | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| A       | 18.0 | 2.0 | 4.0 | 4.0 | 14.0 | 4.0 | 16.0 | 14.0 | 20.0 | 4.0 | 16.0 | 16.0 | 10.0 | 14.0 | 8.0 | 18.0 | 12.0 | 194.0 |
| B       | 8.0 | 9.0 | 8.0 | 2.0 | 2.0 | 2.0 | 7.0 | 5.0 | 2.0 | 7.0 | 7.0 | 7.0 | 1.0 | 4.0 | 6.0 | 9.0 | 86.0 |
| C       | 4.0 | 8.0 | 5.0 | 7.0 | 9.0 | 1.0 | 5.0 | 7.0 | 8.0 | 10.0 | 3.0 | 4.0 | 4.0 | 1.0 | 5.0 | 4.0 | 10.0 | 7.0 | 8.0 | 110.0 |

The first stratum may be split into two, e.g. on geographic basis

| New strata | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| A1         | 18.0 | 2.0 | 4.0 | 4.0 | 14.0 | 4.0 | 16.0 | 14.0 | 20.0 | 96.0 |
| A2         | 4.0 | 16.0 | 16.0 | 10.0 | 14.0 | 8.0 | 18.0 | 12.0 | 98.0 |
| B          | 8.0 | 9.0 | 8.0 | 2.0 | 2.0 | 2.0 | 7.0 | 5.0 | 2.0 | 7.0 | 7.0 | 7.0 | 1.0 | 4.0 | 6.0 | 9.0 | 86.0 |
| C          | 4.0 | 8.0 | 5.0 | 7.0 | 9.0 | 1.0 | 5.0 | 7.0 | 8.0 | 10.0 | 3.0 | 4.0 | 4.0 | 1.0 | 5.0 | 4.0 | 10.0 | 7.0 | 8.0 | 110.0 |

Panel C. Unit MoS from panel A adjusted to add up to a constant (100.0) for each stratum

This is to ensure the selection of same constant number of units (=1 in this illustration) from each stratum

| stratum | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| A       | 9.3 | 1.0 | 2.1 | 2.1 | 8.2 | 7.2 | 10.3 | 2.1 | 8.2 | 8.2 | 5.2 | 7.2 | 4.1 | 9.3 | 6.2 | 100.0 |
| B       | 9.3 | 10.5 | 9.3 | 2.3 | 2.3 | 2.3 | 8.1 | 8.1 | 8.1 | 8.1 | 1.2 | 4.7 | 7.0 | 10.5 | 100.0 |
| C       | 3.6 | 7.3 | 4.5 | 6.4 | 8.2 | 0.9 | 4.5 | 6.4 | 7.3 | 9.1 | 2.7 | 3.7 | 3.6 | 0.9 | 4.5 | 3.6 | 9.1 | 6.4 | 7.3 | 100.0 |

Panel D. Controlled selection of units classified by stratum and another control variable or variables

Unit size measures by stratum (A-C) and control group (U1-U4)

| unit id | stratum | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| A       |         | 9.3 | 1.0 | 2.1 | 2.1 | 8.2 | 7.2 | 10.3 | 2.1 | 8.2 | 8.2 | 5.2 | 7.2 | 4.1 | 9.3 | 6.2 | 100.0 |
| B       |         | 9.3 | 10.5 | 9.3 | 2.3 | 2.3 | 2.3 | 8.1 | 8.1 | 8.1 | 8.1 | 1.2 | 4.7 | 7.0 | 10.5 | 100.0 |
| C       |         | 3.6 | 7.3 | 4.5 | 6.4 | 8.2 | 0.9 | 4.5 | 6.4 | 7.3 | 9.1 | 2.7 | 3.7 | 3.6 | 0.9 | 4.5 | 3.6 | 9.1 | 6.4 | 7.3 | 100.0 |

Our final data for the illustration are shown in Table 12.2D. The units are classified by stratum A-C and also by four categories (U1-U4) of another control variable. For instance, the latter may refer to prominent characteristics of child workers, e.g. mainly boys or mainly girls, mainly young or mainly very young. Say in stratum A unit 7 (we will henceforth refer to the unit simply as “A-07”), most of the child workers are very young (say category U2), and in unit B-10 they are mainly older girls (say category U3). Our objective is to ensure, as far as possible, that the 3 units which are selected are (i) not only one from each stratum, but also that (ii) three different control groups
from among the four groups U1-U4 are represented in the sample. It is not desirable to have 2 (and certainly not 3) units belonging to the same group. Independent selection within strata does not ensure this; controlled selection is required.

**12.2.2 Preliminary adjustment to improve control of the sample outcome**

Table 12.3A examines the sizes of the control groups more closely. The panel on the left shows the total MoS for each control group U1-U4 (summed over units belonging to the group as specified in Table 12.2D). Of the total MoS of 300, group U1 has 61.7 and group U3 102.2, for instance. The next columns show the expected number of units to be selected from the group, and the associated probabilities of selecting that number, given that MoS = 100 represents one selection. The expected (average) number \( k_{av} \) to be selected is simply the group size measure divided by 100, for instance \( k_{av} = 61.7/100 = 0.617 \) for U1. Of course the actual number to be selected must be an integer – in fact it is one of the two integers, \( k \) and \( k+1 \), surrounding \( k_{av} \). The probability of selecting \( k \) units is \( (k+1)-k_{av} \), and that of selecting \( (k+1) \) units is \( (k_{av}-k) \), the two adding up to 1.0 of course. Thus for U1, \( k_{av} = 0.617 \), \( k=0 \), so that one unit is selected with probability 0.617, and none is selected with probability 1-0.617 = 0.383. A similar situation applies to groups U2 and U4.

Now we need to pay attention to the situation regarding U3. Here \( k_{av} = 1.022 \), which is very slightly over 1.0. Mostly (probability 0.978) we expect one unit from the group to come into the sample, but rarely (probability 0.022) two units may be selected from the same group – in the latter case only one of other 3 groups will have a sample unit. In practice this is not a not desirable situation. Given that \( k_{av} \) is so close to 1.0 in our example, it would be preferable that the number of units to be selected from U3 is fixed to be exactly 1, so that 2 different groups from the other 3 can be represented in the sample. We can achieve this by making a minor adjustment to the measures of size MoS (i.e. the unit selection probabilities). The given values of the MoS are often approximate and sometimes even subjectively determined. Nevertheless, any adjustment made to those measures at the design stage should be minor.

A possible adjustment of this type is detailed in the table. The adjustment is to shift 2.2 measures of size from U3 to U1 (which happens to be the smallest control group). Hence the adjusted size measure of U3 would become exactly 100 and that of U1 63.9, as shown in the right panel of Table 12.3A. This has to be done such that the total MoS for each stratum remains unchanged at 100, as that is essential to ensure that exactly 1 unit is selected from each stratum. This is done as follows.

1. The size measures of all units (in all strata) in U3 are multiplied by \((100/102.2) = 0.978\).
2. This changes the size measure of U3 in each stratum. For example in stratum A, it changes from the original 41.2 to adjusted 40.3 (by the constant factor 0.978 applied), as shown in Table 12.3B.
3. This difference within each stratum is allocated to U1. Thus in the above example, in stratum A the difference (41.2-40.3 = 0.9) is allocated to U1, changing its size measure from original 21.7 to 22.6. Thus the total of (U3+U1) size measures within each stratum remains unchanged, and hence so does the total stratum MoS.
4. The ratio of the modified to the original MoS of U1 gives an adjustment factor for each stratum – for example 22.6/21.7 = 1.041 for stratum A, similarly 1.070 for stratum C (see table).

5. Size measures of all U1 units in each stratum are multiplied by the stratum-specific adjustment factor in (4).

Panel C of the table shows unit MoS adjusted in the above manner. For each group U1 and U3, these add up to the adjusted figures in panel B. There is no change in the case of units in groups U2 and U4. The total stratum size remains 100 in all cases. Using these adjusted size measures would ensure that exactly 1 unit in the sample will come from U3. As will be seen below, the controlled selection procedure can also try to ensure that the other 2 sample units come from two different groups among (U1, U2, U4).

<table>
<thead>
<tr>
<th>Table 12.3. Illustration of controlled selection procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of units selected per stratum = 1</td>
</tr>
<tr>
<td>(i) Probability of selecting k or (k+1) units for control groups U1-U4</td>
</tr>
<tr>
<td>MoS</td>
</tr>
<tr>
<td>group</td>
</tr>
<tr>
<td>U1</td>
</tr>
<tr>
<td>U2</td>
</tr>
<tr>
<td>U3</td>
</tr>
<tr>
<td>U4</td>
</tr>
<tr>
<td>(ii) Adjusted probabilities after shifting small amount of MoS from U3 to U1</td>
</tr>
<tr>
<td>MoS</td>
</tr>
<tr>
<td>group</td>
</tr>
<tr>
<td>U1</td>
</tr>
<tr>
<td>U2</td>
</tr>
<tr>
<td>U3</td>
</tr>
<tr>
<td>U4</td>
</tr>
</tbody>
</table>

Notes
This is to ensure that exactly one unit is selected from U3.

As illustrated in panels B and C below, it can be achieved for instance by:
multiplying all MoS in U3 by factor (100.0/102.2) = 0.978,
and within each stratum, allocating the difference between the modified and original total U3 size measures proportionately among U1 units in the stratum.

This means that within each stratum, all unit MoS in U3 and U1 have been multiplied by adjustment factors given in panel B below.
No adjustment is made to areas in groups U2 and U4. U1 was chosen for adjustment because its total MoS was the smallest among the groups.

| 12.3B. Calculation of the desired MoS adjustment factors for units in U3 and U1 |
| Size measure by control group |
| original | adjusted | adjustment factor |
| Stratum | U3 | U1 | U3+U1 | U3 | U1 | U3+U1 | U3 | U1 |
| A | 41.2 | 21.7 | 62.9 | 40.3 | 22.6 | 62.9 | 0.978 | 1.041 |
| B | 25.6 | 29.1 | 54.7 | 25.0 | 29.7 | 54.7 | 0.978 | 1.019 |
| C | 35.4 | 10.9 | 46.3 | 34.6 | 11.7 | 46.3 | 0.978 | 1.070 |
| 102.2 | 61.7 | 163.9 | 100.0 | 63.9 | 163.9 |
Table 12.3 (cont.)

<table>
<thead>
<tr>
<th>stratum</th>
<th>unit id</th>
<th>stratum 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>9.7</td>
<td>1.0</td>
<td>2.2</td>
<td>2.2</td>
<td>7.5</td>
<td>2.1</td>
<td>8.2</td>
<td>7.2</td>
<td>10.1</td>
<td>2.1</td>
<td>8.0</td>
<td>8.0</td>
<td>5.1</td>
<td>7.0</td>
<td>4.1</td>
<td>9.3</td>
<td>6.2</td>
<td>100.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>9.5</td>
<td>10.7</td>
<td>9.5</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>8.1</td>
<td>5.8</td>
<td>2.3</td>
<td>7.9</td>
<td>7.9</td>
<td>8.0</td>
<td>1.2</td>
<td>4.7</td>
<td>7.0</td>
<td>10.5</td>
<td>100.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>3.9</td>
<td>7.8</td>
<td>4.5</td>
<td>6.4</td>
<td>8.2</td>
<td>0.9</td>
<td>4.5</td>
<td>6.4</td>
<td>7.1</td>
<td>8.9</td>
<td>2.6</td>
<td>3.6</td>
<td>3.5</td>
<td>0.9</td>
<td>4.4</td>
<td>3.5</td>
<td>9.1</td>
<td>6.4</td>
<td>7.3</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td></td>
</tr>
</tbody>
</table>

Notes
This makes the total size measure for control group U3 exactly 100, as per Table 12.3 panel A(ii) above. This means that exactly one unit will be selected from this group in every pattern. Size measures in control group U1 have been slightly revised to accommodate this change in U3 size measures.

We could use data from Table 12.3C for the illustration to follow in place of Table 12.2D. However, we have used the latter instead, just to explore whether this results in any problems and indicate how they may be handled. Nevertheless, our objective remains as stated above: selecting one unit from U3 as far as possible, and the other two units from two different groups among (U1, U2, U4), in addition to ensuring that exactly one unit comes from each stratum.

12.2.3 Step-by-step illustration of the application of controlled selection

Table 12.4 provides a detailed step-by-step application of controlled selection. In the controlled selection procedure, judgement – even subjective judgement - is permitted without affecting the probability nature of the sample. This is because the procedure only determines probabilities of a set of samples which maintain the given unit selection probabilities unchanged while, as far as possible, the procedure meets the additional controls imposed on the structure of the sample. Hence this illustration is merely one possible choice of the set of samples, determined by the particular criteria and judgements used. It is not a unique set, and different criteria and judgements may produce different, equally valid or even better sets. The degree of arbitrariness in the choices tends to increase towards the end of the procedure when samples with very small probabilities are being defined to complete the operation.

The first panel of Table 12.4 shows the starting point. This is the same data set as in Table 12.2D above.

The requirement is to choose (subjectively, not randomly) a set of 3 units, 1 from each stratum, 1 of them coming from group U3, and 1 each from two of the other three groups. This defines a possible sample, or a controlled selection “pattern”. The probability assigned to this pattern has to be the smallest among the 3 unit probabilities comprising it. The first pattern chosen, as marked in the first panel of the table, consists of units A-09(U3), B-02(U1), and C-17(U4) – the notation indicating the stratum, unit number within the stratum, and the control group to which it belongs. The respective MoS of the above units are 10.3, 10.5 and 9.1. The smallest of these (9.1 per cent) is the probability assigned to the pattern. The final selection of the sample will consist

49 All probabilities are given here are in per cent terms, and hence numerically equal the corresponding unit MoS.
of selecting one of the patterns so defined with its assigned probability, and not a selection of 3 units separately with their own probabilities as in conventional sampling.

Selecting the first pattern defined above with probability 9.1% means that each of the 3 units in it will receive this probability. Unit A-09(U3) still has to receive its remaining probability 10.3-9.1=1.2%; similarly unit B-02(U1) has still to receive 10.5-9.1=1.4%. This will be achieved by including these units in other patterns, hence giving them additional probability of selection. Unit C-17(U4) has been fully dealt with – it has received all its probability, and will not be included in any other patterns.

The sample made of this first pattern meets the constraints listed above. But why have these particular units been chosen to make the first pattern? Many other patterns can also meet these constraints.

The criterion we have used is to take the pattern which meets the constraints, and has the maximum selection probability available at the particular stage in the process. For the first selection, the maximum selection probability for any pattern which meets the constraints can be seen to be 9.1% in our illustration. The pattern A-01(U1), B-16(U4), and C-10(U3) also has a probability of 9.1%. One could have chosen that as the first pattern. But then there are more patterns to come, and indeed we have taken the above as the next pattern to be extracted (see table). The patterns as they are extracted one-by-one are listed in Table 12.5.

Returning to Table 12.4, the second panel shows the situation “after removing pattern 1”. Removing a pattern means that for the units included in the previous pattern, only the remaining unit selection probabilities still to be taken care of are shown. Thus for A-09(U3) we have shown only 1.2 (=10.3-9.1), for B-02(U1) only 1.4, and for C-17(U4) only zero as the unit has already received its full selection probability in the first pattern. These figures are shown in bold in the second panel of the table. But the focus in this panel is on the next step – the extraction of the next pattern from this panel. The bordered cells indicate the next pattern (only this will be indicated in subsequent panels of the table). After removing the first pattern with probability 9.1%, the remaining probabilities total to the same figure in all strata, namely (100-9.1)=90.9%.

The second pattern following the same procedure is A-01(U1), B-16(U4), C-10(U3), also extracted with probability 9.1% as noted above. After removing this pattern, the remaining probability in each stratum is (90.9-9.1)=81.8 as shown in the third panel of the table. The difference between the 2nd and the 3rd panels is only in the cells corresponding to the units removed in the second pattern. The 3rd panel shown is after that extraction. The remaining probability of A-01(U1) is (9.3-9.1)=0.2%, that of B-16(U4) is (10.5-9.1)=1.4%, and of C-10(U3) is zero since all of its 9.1 has been accounted for by pattern 2. The 3rd pattern consists of units A-11(U3), B-01(U1), C-05(U2) with probability 8.2%. And so on.

The final sample will consist of the selection of one pattern with probability proportional to specified pattern measure of size. Each pattern consists of 3 units, one from each stratum. The target is to have 1 of the 3 units from control group U3, and 1 each from 2 among the remaining 3 control groups.
Table 12.4. Detailed example of identifying controlled selection patterns

<table>
<thead>
<tr>
<th>Unit size measures by stratum (A-C) and control group (U1-U4)</th>
<th>Controlled selection of units classified by stratum and another control variable or variables</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unit id</strong></td>
<td><strong>stratum 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 total</strong></td>
</tr>
<tr>
<td><strong>stratum</strong></td>
<td><strong>A</strong></td>
</tr>
<tr>
<td></td>
<td><strong>U1</strong></td>
</tr>
<tr>
<td></td>
<td><strong>B</strong></td>
</tr>
<tr>
<td></td>
<td><strong>U1</strong></td>
</tr>
<tr>
<td></td>
<td><strong>C</strong></td>
</tr>
<tr>
<td></td>
<td><strong>U1</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Cells with border show selection pattern (1). See below.</strong></td>
</tr>
</tbody>
</table>

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (1)

<table>
<thead>
<tr>
<th><strong>Unit id</strong></th>
<th><strong>stratum 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 total</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>stratum</strong></td>
<td><strong>A</strong></td>
</tr>
<tr>
<td></td>
<td><strong>U1</strong></td>
</tr>
<tr>
<td></td>
<td><strong>B</strong></td>
</tr>
<tr>
<td></td>
<td><strong>U1</strong></td>
</tr>
<tr>
<td></td>
<td><strong>C</strong></td>
</tr>
<tr>
<td></td>
<td><strong>U1</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Cells with border show selection pattern (2).</strong></td>
</tr>
</tbody>
</table>

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (2)

<table>
<thead>
<tr>
<th><strong>Unit id</strong></th>
<th><strong>stratum 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 total</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>stratum</strong></td>
<td><strong>A</strong></td>
</tr>
<tr>
<td></td>
<td><strong>U1</strong></td>
</tr>
<tr>
<td></td>
<td><strong>B</strong></td>
</tr>
<tr>
<td></td>
<td><strong>U1</strong></td>
</tr>
<tr>
<td></td>
<td><strong>C</strong></td>
</tr>
<tr>
<td></td>
<td><strong>U1</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Cells with border show selection pattern (3).</strong></td>
</tr>
</tbody>
</table>

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (3)

<table>
<thead>
<tr>
<th><strong>Unit id</strong></th>
<th><strong>stratum 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 total</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>stratum</strong></td>
<td><strong>A</strong></td>
</tr>
<tr>
<td></td>
<td><strong>U1</strong></td>
</tr>
<tr>
<td></td>
<td><strong>B</strong></td>
</tr>
<tr>
<td></td>
<td><strong>U1</strong></td>
</tr>
<tr>
<td></td>
<td><strong>C</strong></td>
</tr>
<tr>
<td></td>
<td><strong>U1</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Cells with border show selection pattern (4).</strong></td>
</tr>
</tbody>
</table>
### Table 12.4 (cont.)

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (4)

<table>
<thead>
<tr>
<th>stratum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.2</td>
<td>1.0</td>
<td>2.1</td>
<td>2.1</td>
<td>7.2</td>
<td>2.1</td>
<td>8.2</td>
<td>7.2</td>
<td>1.2</td>
<td>2.1</td>
<td>0.0</td>
<td>0.9</td>
<td>5.2</td>
<td>7.2</td>
<td>4.1</td>
<td>9.3</td>
<td>6.2</td>
<td></td>
<td>66.3</td>
<td></td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>B</td>
<td>1.1</td>
<td>1.4</td>
<td>2.0</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>8.1</td>
<td>5.8</td>
<td>2.3</td>
<td>8.1</td>
<td>8.1</td>
<td>8.2</td>
<td>1.2</td>
<td>4.7</td>
<td>7.0</td>
<td>1.4</td>
<td></td>
<td></td>
<td>66.3</td>
<td></td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
</tr>
<tr>
<td>C</td>
<td>3.6</td>
<td>7.3</td>
<td>4.5</td>
<td>6.4</td>
<td>0.0</td>
<td>0.9</td>
<td>4.5</td>
<td>6.4</td>
<td>7.3</td>
<td>0.0</td>
<td>2.7</td>
<td>3.7</td>
<td>3.6</td>
<td>0.9</td>
<td>4.5</td>
<td>3.6</td>
<td>0.0</td>
<td>6.4</td>
<td>0.0</td>
<td>66.3</td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (5). 198.9

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (5)

<table>
<thead>
<tr>
<th>stratum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.2</td>
<td>1.0</td>
<td>2.1</td>
<td>2.1</td>
<td>7.2</td>
<td>2.1</td>
<td>8.2</td>
<td>7.2</td>
<td>1.2</td>
<td>2.1</td>
<td>0.0</td>
<td>0.9</td>
<td>5.2</td>
<td>7.2</td>
<td>4.1</td>
<td>2.0</td>
<td>6.2</td>
<td></td>
<td>59.0</td>
<td></td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>B</td>
<td>1.1</td>
<td>1.4</td>
<td>2.0</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>8.1</td>
<td>5.8</td>
<td>2.3</td>
<td>8.1</td>
<td>8.1</td>
<td>8.2</td>
<td>1.2</td>
<td>4.7</td>
<td>7.0</td>
<td>1.4</td>
<td></td>
<td></td>
<td>59.0</td>
<td></td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
</tr>
<tr>
<td>C</td>
<td>3.6</td>
<td>7.3</td>
<td>4.5</td>
<td>6.4</td>
<td>0.0</td>
<td>0.9</td>
<td>4.5</td>
<td>6.4</td>
<td>0.0</td>
<td>0.0</td>
<td>2.7</td>
<td>3.7</td>
<td>3.6</td>
<td>0.9</td>
<td>4.5</td>
<td>3.6</td>
<td>0.0</td>
<td>6.4</td>
<td>0.0</td>
<td>59.0</td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (6). 177.0

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (6)

<table>
<thead>
<tr>
<th>stratum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.2</td>
<td>1.0</td>
<td>2.1</td>
<td>2.1</td>
<td>7.2</td>
<td>2.1</td>
<td>8.2</td>
<td>7.2</td>
<td>1.2</td>
<td>2.1</td>
<td>0.0</td>
<td>0.9</td>
<td>5.2</td>
<td>7.2</td>
<td>4.1</td>
<td>2.0</td>
<td>6.2</td>
<td></td>
<td>51.7</td>
<td></td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
</tr>
<tr>
<td>B</td>
<td>1.1</td>
<td>1.4</td>
<td>2.0</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>8.1</td>
<td>5.8</td>
<td>2.3</td>
<td>8.1</td>
<td>8.1</td>
<td>8.2</td>
<td>1.2</td>
<td>4.7</td>
<td>7.0</td>
<td>1.4</td>
<td></td>
<td></td>
<td>51.7</td>
<td></td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
</tr>
<tr>
<td>C</td>
<td>3.6</td>
<td>7.3</td>
<td>4.5</td>
<td>6.4</td>
<td>0.0</td>
<td>0.9</td>
<td>4.5</td>
<td>6.4</td>
<td>0.0</td>
<td>0.0</td>
<td>2.7</td>
<td>3.7</td>
<td>3.6</td>
<td>0.9</td>
<td>4.5</td>
<td>3.6</td>
<td>0.0</td>
<td>6.4</td>
<td>0.0</td>
<td>51.7</td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (7). 155.1

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (7)

<table>
<thead>
<tr>
<th>stratum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.2</td>
<td>1.0</td>
<td>2.1</td>
<td>2.1</td>
<td>7.2</td>
<td>2.1</td>
<td>8.2</td>
<td>7.2</td>
<td>1.2</td>
<td>2.1</td>
<td>0.0</td>
<td>0.9</td>
<td>5.2</td>
<td>7.2</td>
<td>4.1</td>
<td>2.0</td>
<td>6.2</td>
<td></td>
<td>45.3</td>
<td></td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
</tr>
<tr>
<td>B</td>
<td>1.1</td>
<td>1.4</td>
<td>2.0</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>8.1</td>
<td>5.8</td>
<td>2.3</td>
<td>8.1</td>
<td>8.1</td>
<td>8.2</td>
<td>1.2</td>
<td>4.7</td>
<td>7.0</td>
<td>1.4</td>
<td></td>
<td></td>
<td>45.3</td>
<td></td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
</tr>
<tr>
<td>C</td>
<td>3.6</td>
<td>0.0</td>
<td>4.5</td>
<td>6.4</td>
<td>0.0</td>
<td>0.9</td>
<td>4.5</td>
<td>6.4</td>
<td>0.0</td>
<td>0.0</td>
<td>2.7</td>
<td>3.7</td>
<td>3.6</td>
<td>0.9</td>
<td>4.5</td>
<td>3.6</td>
<td>0.0</td>
<td>6.4</td>
<td>0.0</td>
<td>45.3</td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (8). 135.9
Table 12.4 (cont.)

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (8)

<table>
<thead>
<tr>
<th>stratum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.2</td>
<td>1.0</td>
<td>2.1</td>
<td>2.1</td>
<td>0.8</td>
<td>2.1</td>
<td>0.9</td>
<td>0.8</td>
<td>1.2</td>
<td>2.1</td>
<td>0.0</td>
<td>0.9</td>
<td>5.2</td>
<td>7.2</td>
<td>4.1</td>
<td>2.0</td>
<td>6.2</td>
<td>38.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>1.1</td>
<td>1.4</td>
<td>2.0</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>0.8</td>
<td>5.8</td>
<td>2.3</td>
<td>1.7</td>
<td>1.7</td>
<td>0.9</td>
<td>1.2</td>
<td>4.7</td>
<td>7.0</td>
<td>1.4</td>
<td>38.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>3.6</td>
<td>0.0</td>
<td>4.5</td>
<td>6.4</td>
<td>0.0</td>
<td>0.9</td>
<td>4.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.7</td>
<td>3.7</td>
<td>3.6</td>
<td>0.9</td>
<td>4.5</td>
<td>3.6</td>
<td>0.0</td>
<td>0.0</td>
<td>38.9</td>
<td></td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (9). 116.7

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (9)

<table>
<thead>
<tr>
<th>stratum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.2</td>
<td>1.0</td>
<td>2.1</td>
<td>2.1</td>
<td>0.8</td>
<td>2.1</td>
<td>0.9</td>
<td>0.8</td>
<td>1.2</td>
<td>2.1</td>
<td>0.0</td>
<td>0.9</td>
<td>5.2</td>
<td>0.8</td>
<td>4.1</td>
<td>2.0</td>
<td>6.2</td>
<td>32.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>1.1</td>
<td>1.4</td>
<td>2.0</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>0.8</td>
<td>5.8</td>
<td>2.3</td>
<td>1.7</td>
<td>1.7</td>
<td>0.9</td>
<td>1.2</td>
<td>4.7</td>
<td>0.6</td>
<td>1.4</td>
<td>32.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>3.6</td>
<td>0.0</td>
<td>4.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>4.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.7</td>
<td>3.7</td>
<td>3.6</td>
<td>0.9</td>
<td>4.5</td>
<td>3.6</td>
<td>0.0</td>
<td>0.0</td>
<td>32.5</td>
<td></td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (10). 97.5

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (10)

<table>
<thead>
<tr>
<th>stratum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.2</td>
<td>1.0</td>
<td>2.1</td>
<td>2.1</td>
<td>0.8</td>
<td>2.1</td>
<td>0.9</td>
<td>0.8</td>
<td>1.2</td>
<td>2.1</td>
<td>0.0</td>
<td>0.9</td>
<td>0.7</td>
<td>0.8</td>
<td>4.1</td>
<td>2.0</td>
<td>6.2</td>
<td>28.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>1.1</td>
<td>1.4</td>
<td>2.0</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>0.8</td>
<td>5.8</td>
<td>2.3</td>
<td>1.7</td>
<td>1.7</td>
<td>0.9</td>
<td>1.2</td>
<td>0.2</td>
<td>0.6</td>
<td>1.4</td>
<td>28.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>3.6</td>
<td>0.0</td>
<td>4.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>4.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.7</td>
<td>3.7</td>
<td>3.6</td>
<td>0.9</td>
<td>4.5</td>
<td>3.6</td>
<td>0.0</td>
<td>0.0</td>
<td>28.0</td>
<td></td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (11). 84.0

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (11)

<table>
<thead>
<tr>
<th>stratum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.2</td>
<td>1.0</td>
<td>2.1</td>
<td>2.1</td>
<td>0.8</td>
<td>2.1</td>
<td>0.9</td>
<td>0.8</td>
<td>1.2</td>
<td>2.1</td>
<td>0.0</td>
<td>0.9</td>
<td>0.7</td>
<td>0.8</td>
<td>4.1</td>
<td>2.0</td>
<td>1.7</td>
<td>23.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>1.1</td>
<td>1.4</td>
<td>2.0</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>0.8</td>
<td>1.3</td>
<td>2.3</td>
<td>1.7</td>
<td>1.7</td>
<td>0.9</td>
<td>1.2</td>
<td>0.2</td>
<td>0.6</td>
<td>1.4</td>
<td>23.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>3.6</td>
<td>0.0</td>
<td>4.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>4.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.7</td>
<td>3.7</td>
<td>3.6</td>
<td>0.9</td>
<td>4.5</td>
<td>3.6</td>
<td>0.0</td>
<td>0.0</td>
<td>23.5</td>
<td></td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (12). 70.5
### Table 12.4 (cont.)

#### Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (12)

| unit id | stratum | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | total |
|---------|---------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|------|
| A  | 0.2 | 1.0 | 2.1 | 2.1 | 0.8 | 2.1 | 0.9 | 0.9 | 0.8 | 1.2 | 2.1 | 0.0 | 0.9 | 0.7 | 0.8 | 1.8 | 2.0 | 1.7 | 21.2 |
|   | U1 | U1 | U1 | U1 | U1 | U2 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U4 | U4 | U4 | U4 | U4 | U4 |
| B  | 1.1 | 1.4 | 2.0 | 2.3 | 2.3 | 0.0 | 0.8 | 1.3 | 2.3 | 1.7 | 1.7 | 0.9 | 1.2 | 0.2 | 0.6 | 1.4 |    |    |    | 21.2 |
|   | U1 | U1 | U1 | U2 | U2 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U4 | U4 | U4 | U4 | U4 | U4 | U4 |
| C  | 3.6 | 0.0 | 4.5 | 0.0 | 0.0 | 0.9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 21.2 |
|   | U1 | U1 | U2 | U2 | U2 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U3 | U3 | U4 | U4 | U4 | U4 | U4 |

Cells with border show selection pattern (13). 63.6

#### Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (13)

| unit id | stratum | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | total |
|---------|---------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|------|
| A  | 0.2 | 1.0 | 0.0 | 0.0 | 0.8 | 2.1 | 0.0 | 0.9 | 0.9 | 0.8 | 1.2 | 2.1 | 0.0 | 0.9 | 0.7 | 0.8 | 1.8 | 2.0 | 1.7 | 19.1 |
|   | U1 | U1 | U1 | U1 | U1 | U2 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U4 | U4 | U4 | U4 | U4 | U4 |
| B  | 1.1 | 1.4 | 2.0 | 0.2 | 0.2 | 0.0 | 0.8 | 1.3 | 2.3 | 1.7 | 1.7 | 0.9 | 1.2 | 0.2 | 0.6 | 1.4 |    |    |    | 19.1 |
|   | U1 | U1 | U1 | U1 | U2 | U2 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U4 | U4 | U4 | U4 | U4 | U4 |
| C  | 3.6 | 0.0 | 4.5 | 0.0 | 0.0 | 0.9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 19.1 |
|   | U1 | U1 | U2 | U2 | U2 | U2 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U3 | U3 | U4 | U4 | U4 | U4 |

Cells with border show selection pattern (14). 57.3

#### Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (14)

| unit id | stratum | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | total |
|---------|---------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|------|
| A  | 0.2 | 1.0 | 0.0 | 0.0 | 0.8 | 2.1 | 0.0 | 0.9 | 0.9 | 0.8 | 1.2 | 2.1 | 0.0 | 0.9 | 0.7 | 0.8 | 1.8 | 2.0 | 1.7 | 17.0 |
|   | U1 | U1 | U1 | U1 | U1 | U1 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U3 | U4 | U4 | U4 | U4 | U4 | U4 |
| B  | 1.1 | 1.4 | 2.0 | 0.2 | 0.2 | 0.0 | 0.8 | 1.3 | 2.3 | 1.7 | 1.7 | 0.9 | 1.2 | 0.2 | 0.6 | 1.4 |    |    |    | 17.0 |
|   | U1 | U1 | U1 | U1 | U2 | U2 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U4 | U4 | U4 | U4 | U4 | U4 | U4 |
| C  | 3.6 | 0.0 | 4.5 | 0.0 | 0.0 | 0.9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 17.0 |
|   | U1 | U1 | U2 | U2 | U2 | U2 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U3 | U3 | U4 | U4 | U4 | U4 | U4 |

Cells with border show selection pattern (15). 51.0

#### Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (15)

| unit id | stratum | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | total |
|---------|---------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|------|
| A  | 0.2 | 1.0 | 0.0 | 0.0 | 0.8 | 2.1 | 0.9 | 0.9 | 0.8 | 1.2 | 2.1 | 0.0 | 0.9 | 0.7 | 0.8 | 1.8 | 2.0 | 1.7 | 14.9 |
|   | U1 | U1 | U1 | U1 | U1 | U1 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U3 | U4 | U4 | U4 | U4 | U4 |
| B  | 1.1 | 1.4 | 2.0 | 0.2 | 0.2 | 0.0 | 0.8 | 1.3 | 0.2 | 1.7 | 1.7 | 0.9 | 1.2 | 0.2 | 0.6 | 1.4 |    |    |    | 14.9 |
|   | U1 | U1 | U1 | U1 | U1 | U1 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U3 | U4 | U4 | U4 | U4 | U4 | U4 |
| C  | 1.5 | 0.0 | 4.5 | 0.0 | 0.0 | 0.9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 14.9 |
|   | U1 | U1 | U2 | U2 | U2 | U2 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U3 | U3 | U3 | U4 | U4 | U4 | U4 |

Cells with border show selection pattern (16). 44.7
12. Controlled selection and balanced sampling

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (16)

<table>
<thead>
<tr>
<th>Unit id</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>U1</td>
<td>0.2</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.8</td>
<td>2.1</td>
<td>0.9</td>
<td>0.8</td>
<td>1.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>0.7</td>
<td>0.8</td>
<td>1.8</td>
<td>0.0</td>
<td>1.7</td>
<td></td>
<td>12.9</td>
</tr>
<tr>
<td>B</td>
<td>U1</td>
<td>1.1</td>
<td>1.4</td>
<td>0.0</td>
<td>0.2</td>
<td>0.2</td>
<td>0.0</td>
<td>0.8</td>
<td>1.3</td>
<td>0.2</td>
<td>1.7</td>
<td>1.7</td>
<td>0.9</td>
<td>1.2</td>
<td>0.2</td>
<td>0.6</td>
<td>1.4</td>
<td></td>
<td></td>
<td>12.9</td>
</tr>
<tr>
<td>C</td>
<td>U1</td>
<td>1.5</td>
<td>0.0</td>
<td>4.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td>1.4</td>
<td>1.5</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (17). 38.7

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (17)

<table>
<thead>
<tr>
<th>Unit id</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>U1</td>
<td>0.2</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.8</td>
<td>2.1</td>
<td>0.9</td>
<td>0.8</td>
<td>1.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>0.7</td>
<td>0.8</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td>11.2</td>
</tr>
<tr>
<td>B</td>
<td>U1</td>
<td>1.1</td>
<td>1.4</td>
<td>0.0</td>
<td>0.2</td>
<td>0.2</td>
<td>0.0</td>
<td>0.8</td>
<td>1.3</td>
<td>0.2</td>
<td>1.7</td>
<td>0.9</td>
<td>1.2</td>
<td>0.2</td>
<td>0.6</td>
<td>1.4</td>
<td></td>
<td></td>
<td></td>
<td>11.2</td>
</tr>
<tr>
<td>C</td>
<td>U1</td>
<td>1.5</td>
<td>0.0</td>
<td>2.8</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td>1.4</td>
<td>1.5</td>
<td>0.9</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (18). 33.6

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (18)

<table>
<thead>
<tr>
<th>Unit id</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>U1</td>
<td>0.2</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.8</td>
<td>2.1</td>
<td>0.9</td>
<td>0.8</td>
<td>1.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>0.7</td>
<td>0.8</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td>9.5</td>
</tr>
<tr>
<td>B</td>
<td>U1</td>
<td>1.1</td>
<td>1.4</td>
<td>0.0</td>
<td>0.2</td>
<td>0.2</td>
<td>0.0</td>
<td>0.8</td>
<td>1.3</td>
<td>0.2</td>
<td>1.7</td>
<td>0.9</td>
<td>1.2</td>
<td>0.2</td>
<td>0.6</td>
<td>1.4</td>
<td></td>
<td></td>
<td></td>
<td>9.5</td>
</tr>
<tr>
<td>C</td>
<td>U1</td>
<td>1.5</td>
<td>0.0</td>
<td>1.1</td>
<td>0.0</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td>1.4</td>
<td>1.5</td>
<td>0.9</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (19). 28.5

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (19)

<table>
<thead>
<tr>
<th>Unit id</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>U1</td>
<td>0.2</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.8</td>
<td>0.7</td>
<td>0.9</td>
<td>0.8</td>
<td>1.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>0.7</td>
<td>0.8</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td>8.1</td>
</tr>
<tr>
<td>B</td>
<td>U1</td>
<td>1.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
<td>0.2</td>
<td>0.0</td>
<td>0.8</td>
<td>1.3</td>
<td>0.2</td>
<td>1.7</td>
<td>1.7</td>
<td>0.9</td>
<td>1.2</td>
<td>0.2</td>
<td>0.6</td>
<td>1.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>U1</td>
<td>1.5</td>
<td>0.0</td>
<td>1.1</td>
<td>0.0</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td>1.4</td>
<td>1.5</td>
<td>0.9</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (20). 24.3

Table 12.4 (cont.)
### Table 12.4 (cont.)

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (20)

| unit id | stratum | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | total |
|---------|---------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|-----|
| A       | 0.2     | 1.0 | 0.0 | 0.0 | 0.8 | 0.7 | 0.9 | 0.8 | 0.0 | 0.0 | 0.0 | 0.9 | 0.7 | 0.8 | 0.1 | 0.0 | 0.0 | 6.9 |
| B       | 1.1     | 0.0 | 0.0 | 0.2 | 0.2 | 0.0 | 0.8 | 1.3 | 0.2 | 0.0 | 0.0 | 0.9 | 1.2 | 0.2 | 0.6 | 0.2 | 6.9 |
| C       | 0.3     | 0.0 | 1.1 | 0.0 | 0.0 | 0.9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 | 1.4 | 0.1 | 0.9 | 0.0 | 1.5 | 0.0 | 0.0 | 6.9 |

Cells with border show selection pattern (21).

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (21)

| unit id | stratum | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | total |
|---------|---------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|-----|
| A       | 0.2     | 0.0 | 0.0 | 0.0 | 0.8 | 0.7 | 0.9 | 0.8 | 0.0 | 0.0 | 0.0 | 0.9 | 0.7 | 0.8 | 0.1 | 0.0 | 0.0 | 5.9 |
| B       | 1.1     | 0.0 | 0.0 | 0.2 | 0.2 | 0.0 | 0.8 | 0.3 | 0.2 | 0.0 | 0.0 | 0.9 | 1.2 | 0.2 | 0.6 | 0.2 | 5.9 |
| C       | 0.3     | 0.0 | 1.1 | 0.0 | 0.0 | 0.9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 | 1.4 | 0.1 | 0.9 | 0.0 | 0.5 | 0.0 | 0.0 | 5.9 |

Cells with border show selection pattern (22).

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (22)

| unit id | stratum | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | total |
|---------|---------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|-----|
| A       | 0.2     | 0.0 | 0.0 | 0.0 | 0.8 | 0.7 | 0.9 | 0.8 | 0.0 | 0.0 | 0.0 | 0.9 | 0.7 | 0.8 | 0.1 | 0.0 | 0.0 | 4.2 |
| B       | 0.2     | 0.0 | 0.0 | 0.2 | 0.2 | 0.0 | 0.8 | 0.3 | 0.2 | 0.0 | 0.0 | 0.5 | 1.2 | 0.2 | 0.6 | 0.2 | 4.2 |
| C       | 0.3     | 0.0 | 1.1 | 0.0 | 0.0 | 0.9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 | 1.4 | 0.1 | 0.9 | 0.0 | 1.5 | 0.0 | 0.0 | 4.2 |

Cells with border show selection pattern (23).

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (23)

| unit id | stratum | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | total |
|---------|---------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|-----|
| A       | 0.2     | 0.0 | 0.0 | 0.0 | 0.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.9 | 0.7 | 0.8 | 0.1 | 0.0 | 0.0 | 4.2 |
| B       | 0.2     | 0.0 | 0.0 | 0.2 | 0.2 | 0.0 | 0.8 | 0.3 | 0.2 | 0.0 | 0.0 | 0.9 | 0.4 | 0.2 | 0.6 | 0.2 | 4.2 |
| C       | 0.3     | 0.0 | 0.3 | 0.0 | 0.0 | 0.9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 | 0.5 | 0.1 | 0.9 | 0.0 | 0.5 | 0.0 | 0.0 | 4.2 |

Cells with border show selection pattern (24).
### Table 12.4 (cont.)

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (24)

<table>
<thead>
<tr>
<th>stratum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
</tr>
<tr>
<td>B</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
</tr>
<tr>
<td>C</td>
<td>0.3</td>
<td>0.0</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (25). 10.8

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (25)

<table>
<thead>
<tr>
<th>stratum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>3.3</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
</tr>
<tr>
<td>B</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
<td>0.2</td>
<td>0.0</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>0.4</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>3.3</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>3.3</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (26). 9.9

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (26)

<table>
<thead>
<tr>
<th>stratum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
</tr>
<tr>
<td>B</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
<td>0.2</td>
<td>0.0</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>0.4</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (27). 9.3

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (27)

<table>
<thead>
<tr>
<th>stratum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
</tr>
<tr>
<td>B</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
<td>0.2</td>
<td>0.0</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>0.4</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (28). 8.7
Table 12.4 (cont.)

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (28)

<table>
<thead>
<tr>
<th>unit id</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.1</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.7</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (29).

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (29)

<table>
<thead>
<tr>
<th>unit id</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.6</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>B</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.6</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.1</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (30).

The remaining patterns can have units from groups U2 and U3 only

This means two units from the same control group (mostly from U3)

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (30)

<table>
<thead>
<tr>
<th>unit id</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>B</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.1</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (31).

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (31)

<table>
<thead>
<tr>
<th>unit id</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>B</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (32).
### Table 12.4 (cont.)

| Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (32) |
| unit id | stratum 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | total |
| A | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.3 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| U1 | U1 | U1 | U1 | U1 | U1 | U2 | U2 | U2 | U3 | U3 | U3 | U3 | U3 | U3 | U4 | U4 | U4 | U4 |
| B | 0.0 | 0.0 | 0.0 | 0.2 | 0.2 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| U1 | U1 | U1 | U1 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U3 | U3 | U3 | U4 | U4 | U4 | U4 | U4 |
| C | 0.0 | 0.0 | 0.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 | 0.3 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| U1 | U1 | U1 | U2 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U3 | U3 | U3 | U3 | U4 | U4 | U4 | U4 |

Cells with border show selection pattern (33). 3.0

| Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (33) |
| unit id | stratum 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | total |
| A | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.3 | 0.1 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 |
| U1 | U1 | U1 | U1 | U1 | U1 | U2 | U2 | U2 | U3 | U3 | U3 | U3 | U3 | U3 | U4 | U4 | U4 | U4 |
| B | 0.0 | 0.0 | 0.0 | 0.2 | 0.2 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 |
| U1 | U1 | U1 | U1 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U3 | U3 | U3 | U3 | U4 | U4 | U4 | U4 |
| C | 0.0 | 0.0 | 0.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 | 0.3 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 |
| U1 | U1 | U1 | U2 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U3 | U3 | U3 | U3 | U4 | U4 | U4 | U4 |

Cells with border show selection pattern (34). 2.1

| Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (34) |
| unit id | stratum 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | total |
| A | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.1 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 |
| U1 | U1 | U1 | U1 | U1 | U1 | U2 | U2 | U2 | U3 | U3 | U3 | U3 | U3 | U3 | U4 | U4 | U4 | U4 | U4 |
| B | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.1 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 |
| U1 | U1 | U1 | U1 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U3 | U3 | U3 | U3 | U4 | U4 | U4 | U4 | U4 |
| C | 0.0 | 0.0 | 0.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 | 0.3 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 |
| U1 | U1 | U1 | U2 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U3 | U3 | U3 | U3 | U4 | U4 | U4 | U4 | U4 |

Cells with border show selection pattern (35). 1.5

| Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (35) |
| unit id | stratum 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | total |
| A | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 |
| U1 | U1 | U1 | U1 | U1 | U1 | U2 | U2 | U2 | U3 | U3 | U3 | U3 | U3 | U3 | U4 | U4 | U4 | U4 | U4 |
| B | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 |
| U1 | U1 | U1 | U1 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U3 | U3 | U3 | U3 | U4 | U4 | U4 | U4 | U4 |
| C | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 |
| U1 | U1 | U1 | U2 | U2 | U2 | U2 | U2 | U3 | U3 | U3 | U3 | U3 | U3 | U3 | U4 | U4 | U4 | U4 | U4 |

Cells with border show selection pattern (36). 0.9
12.2 Illustration 1. Stratification with extra controls using controlled selection procedures

Table 12.4 (cont.)

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (36)

<table>
<thead>
<tr>
<th>unit id</th>
<th>stratum 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>B</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (37).

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (37)

<table>
<thead>
<tr>
<th>unit id</th>
<th>stratum 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>B</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (38).

Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (38)

<table>
<thead>
<tr>
<th>unit id</th>
<th>stratum 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>B</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (38).

Note that as we proceed in Table 12.5, the pattern probabilities decline quite rapidly. For instance, that of pattern (16) is only 2.0%, and that of pattern (26) only 0.2%. A part of the reason is that at every step we have sought to extract the pattern with the highest probability available. But the main reason is that as patterns are removed, the probabilities remaining for the units involved are reduced. Further patterns can only be extracted on the basis of those reduced unit selection probabilities, and hence with correspondingly reduced pattern probabilities. This does not mean that these later patterns are less important or less desirable, in so far as they meet the specified constraints. They may be as “good” (as representative) samples as the earlier patterns assigned higher probabilities. The probability a pattern receives depends on the order we bring it in. If taken out earlier in the process, the pattern would have a higher probability assigned. The arbitrariness of order is not a problem so long as unit selection probabilities are accounted for, and the specified constraints in the controlled selection are met. What matters is the selection probabilities individual units ultimately receive. Difficulties can arise towards the end of the process when it becomes increasingly difficult to meet the specified constraints.
The reason for adopting the criterion of extracting, at each stage, the pattern with the highest probability meeting the constraints is twofold. Firstly, it can shorten the procedure by reducing the number of patterns which need to be extracted, but this is not always the case. The main justification is that it takes care of the units with the largest MoS first. The MoS reflects the importance of a unit. A unit is given higher MoS because it is larger, or more important in some other respect. Smaller units are in a sense less important and they can be left to later stages of the operation when, increasingly, it becomes more difficult to extract further patterns while fully meeting the specified constraints. As we proceed, the choice of the next pattern to be extracted tends to become somewhat more arbitrary. Fortunately, the probabilities associated with later patterns are also greatly reduced, as noted.

In the present illustration we find that after extracting 22 patterns, 95 per cent of the overall unit probabilities have been accounted for (see Table 12.5), and after extracting 29 patterns, the figure rises to 97.4 per cent. We also find that after pattern number 29, control groups U1 and U4 are completely exhausted, i.e. all the unit selection probabilities in them have been fully accounted for. Further patterns can have units only from control groups U2 and U3. Two units must be selected from one of them (mainly from U3), and one unit from the other.

The process is completed (i.e. all unit probabilities fully accounted for) after 38 patterns. The last 9 patterns (patterns 30-38), accounting for 2.6% of the total probability, do not meet the criterion of representing 3 of the 4 control groups U1-U4, which the earlier patterns 01-29 met. The problem at later stages of pattern extraction is connected to the fact that we used the data as in Table 12.2D, without preliminary adjustment to the MoS of the type discussed in Section 12.2.2 (Table 12.3C).
12.2 Illustration 1. Stratification with extra controls using controlled selection procedures

Table 12.5. The controlled selection patterns from Table 12.4, with associated probabilities of selection

One pattern may be selected with a random number in the range (0.0–100.0)
Each pattern contains one unit from each stratum A, B, and C
(1)-(29): Patterns with 1 unit from U3, and 1 each from two of the three controls (U1, U2, U4)

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Stratum A unit id</th>
<th>control</th>
<th>Stratum B unit id</th>
<th>control</th>
<th>Stratum C unit id</th>
<th>control</th>
<th>Pattern selection: probability (%)</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>9</td>
<td>U3</td>
<td>2</td>
<td>U1</td>
<td>17</td>
<td>U4</td>
<td>9.1</td>
<td>9.1</td>
</tr>
<tr>
<td>(2)</td>
<td>1</td>
<td>U1</td>
<td>16</td>
<td>U4</td>
<td>10</td>
<td>U3</td>
<td>9.1</td>
<td>18.2</td>
</tr>
<tr>
<td>(3)</td>
<td>11</td>
<td>U3</td>
<td>1</td>
<td>U1</td>
<td>5</td>
<td>U2</td>
<td>8.2</td>
<td>26.4</td>
</tr>
<tr>
<td>(4)</td>
<td>12</td>
<td>U3</td>
<td>3</td>
<td>U1</td>
<td>19</td>
<td>U4</td>
<td>7.3</td>
<td>33.7</td>
</tr>
<tr>
<td>(5)</td>
<td>16</td>
<td>U4</td>
<td>7</td>
<td>U2</td>
<td>9</td>
<td>U3</td>
<td>7.3</td>
<td>41.0</td>
</tr>
<tr>
<td>(6)</td>
<td>7</td>
<td>U2</td>
<td>12</td>
<td>U3</td>
<td>2</td>
<td>U1</td>
<td>7.3</td>
<td>48.3</td>
</tr>
<tr>
<td>(7)</td>
<td>5</td>
<td>U1</td>
<td>10</td>
<td>U3</td>
<td>8</td>
<td>U2</td>
<td>6.4</td>
<td>54.7</td>
</tr>
<tr>
<td>(8)</td>
<td>8</td>
<td>U2</td>
<td>11</td>
<td>U3</td>
<td>18</td>
<td>U4</td>
<td>6.4</td>
<td>61.1</td>
</tr>
<tr>
<td>(9)</td>
<td>14</td>
<td>U3</td>
<td>15</td>
<td>U4</td>
<td>4</td>
<td>U2</td>
<td>6.4</td>
<td>67.5</td>
</tr>
<tr>
<td>(10)</td>
<td>13</td>
<td>U3</td>
<td>14</td>
<td>U4</td>
<td>7</td>
<td>U2</td>
<td>4.5</td>
<td>72.0</td>
</tr>
<tr>
<td>(11)</td>
<td>17</td>
<td>U4</td>
<td>8</td>
<td>U2</td>
<td>15</td>
<td>U3</td>
<td>4.5</td>
<td>76.5</td>
</tr>
<tr>
<td>(12)</td>
<td>15</td>
<td>U4</td>
<td>6</td>
<td>U2</td>
<td>12</td>
<td>U3</td>
<td>2.3</td>
<td>78.8</td>
</tr>
<tr>
<td>(13)</td>
<td>4</td>
<td>U1</td>
<td>4</td>
<td>U2</td>
<td>13</td>
<td>U3</td>
<td>2.1</td>
<td>80.9</td>
</tr>
<tr>
<td>(14)</td>
<td>3</td>
<td>U1</td>
<td>5</td>
<td>U2</td>
<td>16</td>
<td>U3</td>
<td>2.1</td>
<td>83.0</td>
</tr>
<tr>
<td>(15)</td>
<td>10</td>
<td>U3</td>
<td>9</td>
<td>U2</td>
<td>1</td>
<td>U1</td>
<td>2.1</td>
<td>85.1</td>
</tr>
<tr>
<td>(16)</td>
<td>16</td>
<td>U4</td>
<td>3</td>
<td>U1</td>
<td>11</td>
<td>U3</td>
<td>2.0</td>
<td>87.1</td>
</tr>
<tr>
<td>(17)</td>
<td>15</td>
<td>U4</td>
<td>10</td>
<td>U3</td>
<td>3</td>
<td>U2</td>
<td>1.7</td>
<td>88.8</td>
</tr>
<tr>
<td>(18)</td>
<td>17</td>
<td>U4</td>
<td>11</td>
<td>U3</td>
<td>3</td>
<td>U2</td>
<td>1.7</td>
<td>90.5</td>
</tr>
<tr>
<td>(19)</td>
<td>6</td>
<td>U2</td>
<td>2</td>
<td>U1</td>
<td>13</td>
<td>U3</td>
<td>1.4</td>
<td>91.9</td>
</tr>
<tr>
<td>(20)</td>
<td>9</td>
<td>U3</td>
<td>16</td>
<td>U4</td>
<td>1</td>
<td>U1</td>
<td>1.2</td>
<td>93.1</td>
</tr>
<tr>
<td>(21)</td>
<td>2</td>
<td>U1</td>
<td>8</td>
<td>U2</td>
<td>16</td>
<td>U3</td>
<td>1.0</td>
<td>94.1</td>
</tr>
<tr>
<td>(22)</td>
<td>7</td>
<td>U2</td>
<td>1</td>
<td>U1</td>
<td>12</td>
<td>U3</td>
<td>0.9</td>
<td>95.0</td>
</tr>
<tr>
<td>(23)</td>
<td>5</td>
<td>U1</td>
<td>13</td>
<td>U3</td>
<td>3</td>
<td>U2</td>
<td>0.8</td>
<td>95.8</td>
</tr>
<tr>
<td>(24)</td>
<td>8</td>
<td>U2</td>
<td>15</td>
<td>U4</td>
<td>14</td>
<td>U3</td>
<td>0.6</td>
<td>96.4</td>
</tr>
<tr>
<td>(25)</td>
<td>12</td>
<td>U3</td>
<td>7</td>
<td>U2</td>
<td>1</td>
<td>U1</td>
<td>0.3</td>
<td>96.7</td>
</tr>
<tr>
<td>(26)</td>
<td>1</td>
<td>U1</td>
<td>16</td>
<td>U4</td>
<td>11</td>
<td>U3</td>
<td>0.2</td>
<td>96.9</td>
</tr>
<tr>
<td>(27)</td>
<td>6</td>
<td>U2</td>
<td>14</td>
<td>U4</td>
<td>11</td>
<td>U3</td>
<td>0.2</td>
<td>97.1</td>
</tr>
<tr>
<td>(28)</td>
<td>6</td>
<td>U2</td>
<td>1</td>
<td>U1</td>
<td>12</td>
<td>U3</td>
<td>0.2</td>
<td>97.3</td>
</tr>
<tr>
<td>(29)</td>
<td>15</td>
<td>U4</td>
<td>12</td>
<td>U3</td>
<td>6</td>
<td>U2</td>
<td>0.1</td>
<td>97.4</td>
</tr>
<tr>
<td>(30)</td>
<td>14</td>
<td>U3</td>
<td>12</td>
<td>U3</td>
<td>6</td>
<td>U2</td>
<td>0.8</td>
<td>98.2</td>
</tr>
<tr>
<td>(31)</td>
<td>13</td>
<td>U3</td>
<td>7</td>
<td>U2</td>
<td>16</td>
<td>U3</td>
<td>0.5</td>
<td>98.7</td>
</tr>
<tr>
<td>(32)</td>
<td>12</td>
<td>U3</td>
<td>8</td>
<td>U2</td>
<td>14</td>
<td>U3</td>
<td>0.3</td>
<td>99.0</td>
</tr>
<tr>
<td>(33)</td>
<td>6</td>
<td>U2</td>
<td>13</td>
<td>U3</td>
<td>12</td>
<td>U3</td>
<td>0.3</td>
<td>99.3</td>
</tr>
<tr>
<td>(34)</td>
<td>12</td>
<td>U3</td>
<td>4</td>
<td>U2</td>
<td>11</td>
<td>U3</td>
<td>0.2</td>
<td>99.5</td>
</tr>
<tr>
<td>(35)</td>
<td>13</td>
<td>U3</td>
<td>5</td>
<td>U2</td>
<td>3</td>
<td>U2</td>
<td>0.2</td>
<td>99.7</td>
</tr>
<tr>
<td>(36)</td>
<td>8</td>
<td>U2</td>
<td>9</td>
<td>U2</td>
<td>13</td>
<td>U3</td>
<td>0.1</td>
<td>99.8</td>
</tr>
<tr>
<td>(37)</td>
<td>8</td>
<td>U2</td>
<td>9</td>
<td>U2</td>
<td>11</td>
<td>U3</td>
<td>0.1</td>
<td>99.9</td>
</tr>
<tr>
<td>(38)</td>
<td>12</td>
<td>U3</td>
<td>13</td>
<td>U3</td>
<td>3</td>
<td>U2</td>
<td>0.1</td>
<td>100.0</td>
</tr>
</tbody>
</table>
12.2.4 Early termination of the search for controlled selection patterns

The next issue is: what if we stopped the search for patterns before the process is complete, if for instance in our illustration, after pattern 29 we dropped the remaining 9 patterns which do not meet our constraints fully?

The effect is investigated in Table 12.6-12.8.

First considering Table 12.6: panel A of the table shows the given unit MoS from Table 12.2D, and Panel B the 2.6% MoS still not allocated after the first 29 patterns. Panel C is the difference between the above two, showing how the 97.4% allocated so far, after extracting 29 patterns, varies by unit. For comparison with Panel A, these have to be rescaled to add up to 100 by multiplying figures in Panel C by (100/97.4), as shown in Panel D. Panel E shows the ratio, by individual area, between the MoS in Panel D (resulting from stopping after pattern 29, and then recalculating to obtain the given totals), and the original MoS in Panel A. Around 70 per cent of the ratios are slightly higher than 1.0, mostly equalling 1.026. These are simply the units whose probabilities had already been fully accounted for by the first 29 patterns. They do not lose anything by stopping at that point, but gain a small amount as a result of the above-mentioned rescaling. For the remaining units not fully accounted for by the first 29 patterns, the ratio is below 1.0, mostly in the range 0.9-1.0. This is a small difference, in the sense that it makes little difference to the sample to adjust the given unit MoS values by these amounts. For these units, therefore, stopping after 29 patterns is quite acceptable.

For 3 units, the difference is larger, the ratio being 0.68 for units B-13(U3) and C-14(U3), and very low (0.11) for unit C-06(U2) (see figures in bold in Panel E). All these three units are very small, and by always taking care of large units first in constructing the pattern, these small units have not been sufficiently covered by the first 29 patterns. But since all these units are small, the consequences of the above problem are also small. (Note that it is not uncommon in conventional surveys to exclude extremely small units altogether from the sampling frame.) Nevertheless, it is worthwhile to investigate whether some simple adjustment to the procedure can be made to improve representation of the units, in particular of unit C-06(U2) which has a very low ratio in Panel E after extracting the first 29 patterns.

Though any adjustment required is expected to be small, it is desirable to make adjustments only among units in the same stratum and same control group. There are units C-03 to C-08 in U2. One of these units, namely C-05(U2), is quite large and appears in pattern number (3) with probability 8.2%. We can split this pattern into two patterns, (3a) and (3b). The first one has the same composition as the existing pattern (3) – A-11(U3), B-01(U1), C-05(U3) – but is now assigned a slightly lower probability (7.4 in place of the original 8.2). In pattern (3b), we replace C-05(U3) by C-06(U3), which still fully meets our constraints. This pattern – A-11(U3), B-01(U1), C-06(U3) – is with probability 0.8. The combined effect of patterns (3a)+(3b) is identical to that of pattern (3) for the first two units (A-11 and B-01), hence there is no change for them. The probability assigned to C-05 is slightly reduced, but the small amount shifted to C-06 is sufficient to represent that small unit well in the first 29 patterns.
The details of the adjustment are shown in Table 12.7. Panel A shows the units involved in pattern (3) and its split into patterns (3a) and (3b), along with the associated size measures. Panel B shows the remaining MoS after removing patterns, in exactly the same form as in Table 12.4. We start with the situation after extraction of pattern (2) – the figures correspond to the third panel in Table 12.4. Following this, two scenarios are shown: (i) extracting pattern (3) as in Table 12.4; versus (ii) extracting pattern (3a) and then pattern (3b).

Table 12.8 shows the final results after the adjustment. Panels A-E of this table correspond to those of Table 12.6, except that our interest now is only in panel C, where the figures only of the above-mentioned two units, C-05(U2) and C-06(U2), have changed. The ratio for C-05 is reduced from 1.03 in Table 12.6E to 0.93 in Table 12.8E – a small change. But the representation of C-06 is greatly improved, the ratio changing from 0.11 to 1.03.

This example shows that in applying controlled selection, it is worthwhile to pay close attention to detail. One is free to make adjustments to any chosen pattern as long as the constraints are met and the given unit selection probabilities are respected, with the possible exception of small adjustments if needed to improve control of the permitted patterns.
### Table 12.6. The effect of terminating the search for controlled selection patterns

(Example: after 29 patterns worked out in the tables above)

<table>
<thead>
<tr>
<th>Panel A. Original unit size measures by stratum (A-C) and control group (U1-U4) (same as Table 12.3C)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>unit id</strong></td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B. Unit size measures by stratum (A-C) and control group (U1-U4): after removing pattern (29)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>unit id</strong></td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel C. Unit size measures which have been accounted for and controlled up to pattern (29) = Panel A - Panel C</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>unit id</strong></td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel D. Accounted for unit size measures after rescaling (Panel C) to add up to 100 per stratum</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>unit id</strong></td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel E. Implied adjustment of size measures as finally used versus as originally specified (ratio Panel D / Panel A)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>unit id</strong></td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
</tbody>
</table>

Minimum value of ratio 0.88 (disregarding three low values of 0.11, 0.68, and 0.68, in bold)
Maximum value of ratio 1.03
12.2 Illustration 1. Stratification with extra controls using controlled selection procedures

Table 12.7. A simple adjustment to remove any significant effect of termination of the pattern construction after Pattern (29)

Panel A. Splitting a pattern into two parts to change the representation of units

**Notes.** Terminating at pattern (29) means that unit C-06 (U2) gets a very reduced chance of selection. This is in relative terms, as the original selection probability allocated to this unit is very small. This problem is removed by considering pattern (3) to be two patterns, (3a) and (3b) as described below. Pattern (3b) is a minor one (with low probability of 0.8%). In this unit C-05 in the original pattern (3) is replaced by unit C-06 in the new pattern (3b).

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Group</th>
<th>Group</th>
<th>Group</th>
<th>Group</th>
<th>Pattern MoS</th>
</tr>
</thead>
<tbody>
<tr>
<td>(3)</td>
<td>11</td>
<td>U3</td>
<td>1</td>
<td>U1</td>
<td>5</td>
</tr>
<tr>
<td>Split into (3a)</td>
<td>11</td>
<td>U3</td>
<td>1</td>
<td>U1</td>
<td>5</td>
</tr>
<tr>
<td>(3b)</td>
<td>11</td>
<td>U3</td>
<td>1</td>
<td>U1</td>
<td>6</td>
</tr>
</tbody>
</table>

Panel B. Illustration of the effect of this split

**Notes.** Splitting pattern (3) into (3a) and (3b), with no other change, means that the final probability of unit C-5 (U2) is reduced by 0.8, and that of unit C-6 (U2) is increased by the same amount. Note that this split applies across all strata A-C, but the changes made only concern stratum C, and in it only units 5 and 6.

Unit size measures after removing pattern (2)

<table>
<thead>
<tr>
<th>Unit size measures after removing pattern (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>stratum</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
</tbody>
</table>

Cells with border show selection pattern (3).

Unit size measures after removing pattern (3)

<table>
<thead>
<tr>
<th>Unit size measures after removing pattern (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>stratum</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
</tbody>
</table>

Unit size measures after removing pattern (3a)

<table>
<thead>
<tr>
<th>Unit size measures after removing pattern (3a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>stratum</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
</tbody>
</table>

Unit size measures after removing pattern (3b)

<table>
<thead>
<tr>
<th>Unit size measures after removing pattern (3b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>stratum</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
</tbody>
</table>
Table 12.8. Adjusted Table 12.6 after pattern (3) split into two patterns (3a) and (3b)

Panel A. Original unit size measures by stratum (C) and control group (U1-U4)

<table>
<thead>
<tr>
<th>stratum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3.6</td>
<td>7.3</td>
<td>4.5</td>
<td>6.4</td>
<td>8.2</td>
<td>0.9</td>
<td>4.5</td>
<td>6.4</td>
<td>7.3</td>
<td>9.1</td>
<td>2.7</td>
<td>3.7</td>
<td>3.6</td>
<td>0.9</td>
<td>4.5</td>
<td>3.6</td>
<td>9.1</td>
<td>6.4</td>
<td>7.3</td>
<td>100.0</td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
</tr>
</tbody>
</table>

Panel B. Unit size measures by stratum (C): after removing pattern (29), with size measures of units C-05(U2) and C-06(U2) adjusted

<table>
<thead>
<tr>
<th>stratum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.0</td>
<td>0.8</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.6</td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
</tr>
</tbody>
</table>

Panel C. Unit size measures which have been accounted for and controlled up to pattern (29) = Panel A - Panel C

<table>
<thead>
<tr>
<th>stratum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3.6</td>
<td>7.3</td>
<td>4.2</td>
<td>6.6</td>
<td>7.4</td>
<td>0.9</td>
<td>4.5</td>
<td>6.4</td>
<td>7.3</td>
<td>9.1</td>
<td>2.4</td>
<td>3.4</td>
<td>3.5</td>
<td>0.6</td>
<td>4.5</td>
<td>3.1</td>
<td>9.1</td>
<td>6.4</td>
<td>7.3</td>
<td>97.4</td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
</tr>
</tbody>
</table>

Panel D. Accounted for size measures after rescaling Panel (C) to add up to 100 per stratum

<table>
<thead>
<tr>
<th>stratum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3.7</td>
<td>7.5</td>
<td>4.3</td>
<td>6.6</td>
<td>7.6</td>
<td>0.9</td>
<td>4.6</td>
<td>6.6</td>
<td>7.5</td>
<td>9.3</td>
<td>2.5</td>
<td>3.5</td>
<td>3.6</td>
<td>0.6</td>
<td>4.6</td>
<td>3.2</td>
<td>9.3</td>
<td>6.6</td>
<td>7.5</td>
<td>100.0</td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
</tr>
</tbody>
</table>

Panel E. Implied adjustment of size measures as finally used versus as originally specified (ratio Panel D / Panel A)

<table>
<thead>
<tr>
<th>stratum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1.03</td>
<td>1.03</td>
<td>0.96</td>
<td>1.03</td>
<td>0.93</td>
<td>1.03</td>
<td>1.03</td>
<td>1.03</td>
<td>1.03</td>
<td>0.91</td>
<td>0.94</td>
<td>1.00</td>
<td>0.68</td>
<td>1.03</td>
<td>0.88</td>
<td>1.03</td>
<td>1.03</td>
<td>1.03</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>U1</td>
<td>U1</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U2</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U3</td>
<td>U4</td>
<td>U4</td>
<td>U4</td>
</tr>
</tbody>
</table>
12.3 Illustration 2. Controlling sample distribution by location and time segment

In this section we take a very useful yet quite simple illustration of controlled selection. The objective is to control the distribution of the sample simultaneously in space and in time. As elaborated in Chapter 10, such a time-location sampling framework is suitable for surveys of flows. In this framework, the cross of observation points by time segments form the primary sampling units (PSUs). For our illustration, suppose that a survey is to be carried out over 5 days of the week, and each day is divided into 4 observation periods of two hours duration. Thus we have \( n_2 = 5 \times 4 = 20 \) time segments. Let the population consist of \( n_1 \) locations. The \( n_1 \times n_2 \) location by time segments form our PSUs.

Suppose the requirement is to select an equal probability sample of \( n \) PSUs. Controlled selection procedure is to be used to ensure that the maximum possible number of different time segments as well as the maximum possible number of different locations appear in the sample. In relation to the time segments, we have two levels: 5 days of the week, and 4 quarters of each day. The controlled selection procedure should ensure that maximum variety of the day and time-of-day combinations is represented in the sample.

Actually, the controls required are more strict than the above formulation in the following sense.

(1) Representation of locations

The following table illustrates the requirements more precisely for different combinations of \( n_1 \), the number of locations in the population, and \( n \), the number of PSUs to be selected:

| \( n \leq n_1 \) | \( n \) different locations are selected. The number of times any location can be selected is 0 or 1. |
| \( n_1 < n \leq 2n_1 \) | The number of times any location can be selected is 1 or 2. Every location is included in the sample at least once. A random set of \( (n - n_1) \) locations are selected twice. |
| \( 2n_1 < n \leq 3n_1 \) | The number of times any location can be selected is 2 or 3. Every location is included in the sample at least twice. A random set of \( (n - 2n_1) \) locations are selected thrice. |

(2) Representation of time segments

Exactly the same form of specification as above applies here as well. For example, if \( n \leq n_2 \), then \( n \) different time segments should appear in the sample, and no segment should appear more than once. Similarly, if \( n_2 < n \leq 2n_2 \) then \( (n - n_2) \) segments should appear twice in the sample, and each of the remaining \( n_2 \) segments must appear once. And so on.
(3) However, it is also necessary to distinguish between the two dimensions of the time segments: day and time of day. Maximum variation in their combination should be captured in the sample.

(4) The above constraints in terms of representation of locations and of time segments have to be met simultaneously.

Table 12.9 shows a number of examples. Comments on each one in turn follow.

**Case (A).** $n_1 = n_2 = n = 20$

This case is the simplest. We have $n_1 = 20$ locations, $n_2 = 20$ time segments, and $n = 20$ PSUs to select from a total of $n_1 n_2 = 400$.

As shown in the diagram, the diagonal from top left to bottom right, for instance, forms the perfect first pattern. Every location and every time segment are each represented exactly once in the sample. The pattern is marked with “(1)” in the diagram.

Another pattern with the same characteristics is formed by, for instance, shifting a step to the right and running the full length parallel to the diagonal. This pattern is marked with “(2)” in the diagram. Similarly for patterns (3), (4), etc., resulting in 20 patterns. One of these patterns can be selected at random to obtain a sample with the required controls.

**Case (B).** $n > n_1, n_2$

Example: $n_1 = 18$ locations, $n_2 = 20$ time segments, and $n = 25$ sample PSUs.

The number of selections to be made exceeds both the number of locations in the population and the number of time segments. The following provides a simple solution.

(1) Select $(n - n_1) = 7$ locations at random (simple random sampling without replacement), and include them as extra rows in the table. With these duplicates, the table now has $n=25$ rows. One selection per row will be made.

(2) We need to add $(n - n_2) = 5$ columns to the table. The added constraint compared to (1) above is that the additional columns should reflect the maximum possible variety in the day and time-of-day combination. All 5 days can (and should) be represented in the addition, since we have 5 additional columns available. All 4 times of day should be represented, with 5-4=1 period selected at random to be repeated twice.

On the right of panel (B) is shown the 5x4 cross tabulation of day by time-of-day. The above requirement simply means that in selecting the cells for duplication, we spread to the maximum the number of different rows and the number of different columns from which the selections are taken.

(3) Now we have exactly the same situation as case (A): $n=25$ selections are to be made from a 25x25 table. The diagonal from top left to bottom right, for instance, forms the first pattern. Another pattern with the same characteristics is formed by, for instance, shifting a step to the right. And so on. One of these patterns can be selected at random to obtain a sample with the required controls.
12.3 Illustration 2. Controlling sample distribution by location and time segment

Case (C). \( n < n_1, n_2 \)

Example: \( n_1 = 17 \) locations, \( n_2 = 20 \) time segments, and \( n = 16 \) sample PSUs.

The number of selections to be made fall short of both the number of locations in the population and the number of time segments. The following provides a simple solution, which is just the reverse of case (B).

1. Delete \( (n_1 - n) = 1 \) locations at random, removing this row from the table. With this deletion, the table now has \( n=16 \) rows. One selection per row will be made.

2. We need to delete \( (n_2 - n) = 4 \) columns from the table. The added constraint is that the deletions should reflect the maximum possible variety in the day and time-of-day combination. All 4 times-of-day can (and should) be represented in the deletion, since we have 4 columns to delete. Only 4 days can be represented, with 5-4=1 day selected at random to be spared from deletion. See the 5x4 cross tabulation of day by time-of-day on the right of panel (C).

3. Now we have exactly the same situation as the previous cases: \( n=16 \) selections are to be made from a 16x16 table. The diagonal from top left to bottom right, for instance, forms the first pattern. And so on.

Case (D). \( n_1 < n < n_2 \)

Example: \( n_1 = 16 \) locations, \( n_2 = 20 \) time segments, and \( n = 17 \) sample PSUs.

In case (D), we have to add a row for repetition of 1 randomly selected location. We have to delete 3 columns from the 20 time segments. In selecting the columns for deletion, no day and no time-of-day should appear twice. We can simply select a random 3 different rows and 3 different columns in the day by time-of-day cross tabulation, and the 3 intersection cells identify the columns for deletion in the main table.

Case (E). \( n_1 > n > n_2 \)

Example: \( n_1 = 24 \) locations, \( n_2 = 20 \) time segments, and \( n = 22 \) sample PSUs.

Case (E) is the reverse of the above. We delete \( 24-22=2 \) rows (locations) at random.

We have to add 2 columns for the repetition of 2 randomly selected time segments. In selecting the columns for addition, no day and no time-of-day should appear twice. We can simply select a random 2 different rows and 2 different columns in the day by time-of-day cross tabulation, and the 2 intersection cells identify the columns for addition in the main table.
### Table 12.9. Controlled selection: Illustration 3

#### Case (A) \( n_1 = n_2 = n = 20 \)

<table>
<thead>
<tr>
<th>Time segment (day 1-5, time 1-4)</th>
<th>( 1-1 )</th>
<th>( 1-2 )</th>
<th>( 1-3 )</th>
<th>( 1-4 )</th>
<th>( 2-1 )</th>
<th>( 4-1 )</th>
<th>( 5-1 )</th>
<th>( 5-2 )</th>
<th>( 5-3 )</th>
<th>( 5-4 )</th>
<th>( 16 )</th>
<th>( 17 )</th>
<th>( 18 )</th>
<th>( 19 )</th>
<th>( 20 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>(1)</td>
<td>(2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Case (B) \( n_1 = 18; n_2 = 20; n = 25 \)

| Time segment (day 1-5, time 1-4) | \( 1-1 \) | \( 1-2 \) | \( 1-3 \) | \( 1-4 \) | \( 2-1 \) | \( 4-1 \) | \( 5-1 \) | \( 5-2 \) | \( 5-3 \) | \( 5-4 \) | \( 16 \) | \( 17 \) | \( 18 \) | \( 19 \) | \( 20 \) | \( 2 \) | \( 3 \) | \( 4 \) | \( 7 \) | \( 12 \) | \( 13 \) | \( 18 \) |
|---------------------------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| Location                        |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 1                               | (1)      | (2)      | (3)      | (4)      | (5)      |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 2                               | (1)      | (2)      | (3)      | (4)      |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 3                               | (1)      | (2)      | (3)      |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 4                               | (1)      | (2)      |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 5                               |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| ...                             |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 16                              |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 17                              |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 18                              |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 19                              |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 20                              |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| duplicates                      |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 3                               |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 6                               |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 7                               |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 10                              |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 15                              |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 18                              |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 20                              | (2)      | (3)      | (4)      | (5)      |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |

| Columns duplicated             |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 1                               |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 2                               |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 3                               |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 4                               |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 5                               |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
Table 12.9. Controlled selection: Illustration 3

### Case (C) \( n_1=17; n_2=20; n=16 \) \( n<n_1, n_2 \)

<table>
<thead>
<tr>
<th>Time segment (day 1-5, time 1-4)</th>
<th>Columns deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1 1-2 1-3 1-4 2-1 4-4 5-1 5-2 5-3 5-4</td>
<td>1 2 3 4</td>
</tr>
</tbody>
</table>

#### Location

<table>
<thead>
<tr>
<th>Location</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>…</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Case (D) \( n_1=16; n_2=20; n=17 \) \( n_1<n<n_2 \)

<table>
<thead>
<tr>
<th>Time segment (day 1-5, time 1-4)</th>
<th>Columns deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1 1-2 1-3 1-4 2-1 4-4 5-1 5-2 5-3 5-4</td>
<td>1 2 3 4</td>
</tr>
</tbody>
</table>

#### Location

<table>
<thead>
<tr>
<th>Location</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>…</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Case (E) \( n_1=24; n_2=20; n=22 \) \( n_1>n>n_2 \)

<table>
<thead>
<tr>
<th>Time segment (day 1-5, time 1-4)</th>
<th>Columns duplicated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1 1-2 1-3 1-4 2-1 4-4 5-1 5-2 5-3 5-4</td>
<td>1 2 3 4</td>
</tr>
</tbody>
</table>

#### Location

<table>
<thead>
<tr>
<th>Location</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>…</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Deleted

<table>
<thead>
<tr>
<th>Deleted</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Duplicates

<table>
<thead>
<tr>
<th>Duplicates</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(5)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Controlled Sample Distribution by Location and Time Segment

#### Illustration 2

This section discusses the controlled selection of samples based on location and time segment. The table illustrates three cases, each with different sample sizes and distributions:

- **Case (C)**: \( n_1=17; n_2=20; n=16 \) with \( n<n_1, n_2 \).
- **Case (D)**: \( n_1=16; n_2=20; n=17 \) with \( n_1<n<n_2 \).
- **Case (E)**: \( n_1=24; n_2=20; n=22 \) with \( n_1>n>n_2 \).

The table presents the sample distribution for each case, detailing the time segments and locations, with specific columns and locations marked as deleted or duplicated.
12.4 Balanced sampling as a tool for controlled selection

The three illustrations provided in Sections 12.1-12.3 draw on the ideas and procedures of controlled selection as originally formulated by Goodman and Kish (1950). In this section, we examine the same three examples in-depth and reformulate them in terms of the balanced sampling procedure of Deville and Tillé (2004). Examples of balanced samples generated by the procedure are shown.

The illustrations in this section are based on Mehran (2012).

12.4.1 In-depth discussion of illustration in Section 12.1.1

Let us return to the coastal-inland example described in Section 12.1. It is supposed that there are two strata consisting of first-stage units from each of which one unit is to be selected. It is further supposed that in stratum 1, three units lie adjacent to the ocean or other major waterway and in stratum 2 one unit is similarly located. All the other units are located inland.

<table>
<thead>
<tr>
<th>Stratum 2</th>
<th>Coastal</th>
<th>Inland</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stratum 1</td>
<td>0.20</td>
<td>0.80</td>
</tr>
<tr>
<td>Coastal</td>
<td>0.45</td>
<td>0.09</td>
</tr>
<tr>
<td>Inland</td>
<td>0.55</td>
<td>0.11</td>
</tr>
</tbody>
</table>

Let us consider a table showing the joint probabilities assuming independent selection among strata.

<table>
<thead>
<tr>
<th>Inland</th>
<th>Inland</th>
<th>Stratum 2</th>
<th>Coastal</th>
<th>Inland</th>
<th>Unit selection probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
<td>e</td>
<td></td>
</tr>
<tr>
<td>0.015</td>
<td>0.030</td>
<td>0.010</td>
<td>0.020</td>
<td>0.025</td>
<td>0.10</td>
</tr>
<tr>
<td>0.022</td>
<td>0.045</td>
<td>0.015</td>
<td>0.030</td>
<td>0.037</td>
<td>0.15</td>
</tr>
<tr>
<td>0.015</td>
<td>0.030</td>
<td>0.010</td>
<td>0.020</td>
<td>0.025</td>
<td>0.10</td>
</tr>
<tr>
<td>0.030</td>
<td>0.060</td>
<td>0.020</td>
<td>0.040</td>
<td>0.050</td>
<td>0.20</td>
</tr>
<tr>
<td>0.037</td>
<td>0.075</td>
<td>0.025</td>
<td>0.050</td>
<td>0.062</td>
<td>0.25</td>
</tr>
<tr>
<td>0.030</td>
<td>0.060</td>
<td>0.020</td>
<td>0.040</td>
<td>0.050</td>
<td>0.20</td>
</tr>
</tbody>
</table>

It is considered desirable to select one inland and one coastal unit, and avoid selecting two coastal units. The units and the probability of selection assigned to each of them were shown in Table 12.1, and add up by stratum and location as shown in Table 12.10A. The four cross-tabulation cells show joint probabilities assuming independent selection from strata 1 and 2. (For example, in the top left cell, 0.09=0.45x0.20.) The second panel in the table shows joint probabilities of all pairs of units assuming independent selection among strata. (For example, in the top left cell, 0.015=0.10x0.15, respectively the selection probabilities of units A and a in Table 12.1.)
Now, imposing zero probability for selecting two coastal units, one from each stratum, and adjusting the probabilities to be consistent with the marginal probabilities using iterative proportional fitting (Deming, 1943) gives adjusted joint probabilities as shown in Table 12.10B.50

<table>
<thead>
<tr>
<th>Stratum 1</th>
<th>Inland A</th>
<th>Inland B</th>
<th>Inland C</th>
<th>Coastal D</th>
<th>Inland E</th>
<th>Unit selection probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inland</td>
<td>0.0119</td>
<td>0.0239</td>
<td>0.0080</td>
<td>0.0364</td>
<td>0.0199</td>
<td>0.10</td>
</tr>
<tr>
<td>Coastal</td>
<td>0.0281</td>
<td>0.0563</td>
<td>0.0188</td>
<td>-</td>
<td>0.0469</td>
<td>0.15</td>
</tr>
<tr>
<td>Coastal</td>
<td>0.0188</td>
<td>0.0375</td>
<td>0.0125</td>
<td>-</td>
<td>0.0313</td>
<td>0.10</td>
</tr>
<tr>
<td>Inland</td>
<td>0.0239</td>
<td>0.0477</td>
<td>0.0159</td>
<td>0.0727</td>
<td>0.0398</td>
<td>0.20</td>
</tr>
<tr>
<td>Inland</td>
<td>0.0298</td>
<td>0.0597</td>
<td>0.0199</td>
<td>0.0909</td>
<td>0.0497</td>
<td>0.25</td>
</tr>
<tr>
<td>Coastal</td>
<td>0.0375</td>
<td>0.0750</td>
<td>0.0250</td>
<td>-</td>
<td>0.0625</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Unit selection probability

Finally, a draw according to the adjusted joint probabilities gives a sample of two units, one from each stratum, but precluding both being coastal units. By contrast, the probability of getting two coastal units under independent selection by strata is 9%. It can be further verified that the probability of drawing one inland unit and one coastal unit with controlled selection is 65%, higher than under independent selection by stratum (47%). The probability of drawing two inland units (35%) is equivalently lower than under independent selection by stratum (44%).

In order to apply the procedure of balanced sampling for the purpose of achieving the desired controlled selection, we may define the following four auxiliary variables (\(j = 1–4\)) for all units \(k\) in the survey population:

\[
x_{jk} = \pi_k \delta_{jk},
\]

(12.5)

where, as before, \(\pi_k\) is the inclusion probability of unit \(k\). Subscript \(j\) refers to categories for which it is desired to control the sample size. \(\delta_{jk} = 1\) if unit \(k\) belongs to a particular control category, and \(\delta_{jk} = 0\) otherwise. The control categories of interest are:

\(j = 1\): stratum 1
\(j = 2\): stratum 2
\(j = 3\): unit is an coastal unit
\(j = 4\): unit is a inland unit.

For any unit, \(\delta_{jk} = 1\) for one of the two values \(j = 1\) and \(j = 2\), and also for one of the two values \(j = 3\) and \(j = 4\). For example, for a coastal unit from stratum 1, \(\delta_{3k} = \delta_{4k} = 1\) and \(\delta_{2k} = \delta_{sk} = 0\).

50 In this example, the corresponding condition of imposing zero probability for selecting two inland units, one from each stratum, is not possible, being incompatible with the values of the marginal probabilities.
The balancing equations are

$$\sum_{j} \pi_k \delta_{jk} = \sum_{i} \left( \pi_k \delta_{jk} / \pi_k \right) = \sum_{i} \delta_{jk} = n_j,$$

(12.6)

where $n_j$ is the target number of sample units to be selected from control category $j$. With the $\pi_k$ values given in Table 12.1, we have $n_1 = 1$, $n_2 = 1$, $n_3 = 0.65$, and $n_4 = 1.35$, given the total $n = 2$ units to be selected. Since the sample sizes must be integers, the balancing equation may be reformulated as

$$n_1 = n_2 = n_3 = n_4 = 1,$$

with probability 0.65, and

$$n_1 = n_2 = 1, \quad n_3 = 0, \quad n_4 = 2,$$

with probability 0.35.

(12.7)

The above two in combination ensure that the specified unit selection probabilities are respected. Note that no chance has been given to having a sample with both units being coastal units

$$n_1 = n_2 = 1, \quad n_3 = 2, \quad n_4 = 0,$$

probability zero.

Application of the original Goodman-Kish procedure of controlled selection to Illustration 1 (Section 12.1) explicitly takes into account the above formulation and understanding of the problem. The procedure satisfies the marginal distribution and suppresses the combination “both units selected being coastal” with zero probabilities in Table 12.10B. However, neither the controlled selection procedure, nor the application of the balanced sampling procedure, are required to satisfy the full set of joint selection probabilities in individual cells of Table 12.10B. In the controlled selection application in Section 12.1.1, the joint probabilities of selecting pairs of units depended on the joint ordering of the units for systematic selection in the two strata. Recall that units were ordered as shown in the rearranged list in right hand panel of Table 12.1, and a single random number was used to determine the selection in both strata. With the given ordering of units, the unit joint selection probabilities are in fact as shown in Table 12.10C.

<table>
<thead>
<tr>
<th>Stratum 1</th>
<th>Inland</th>
<th>Inland</th>
<th>Inland</th>
<th>Coastal</th>
<th>Inland</th>
<th>Unit selection probability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>Inland A</td>
<td></td>
<td>0.10</td>
<td></td>
<td></td>
<td></td>
<td>0.10</td>
</tr>
<tr>
<td>Coastal B</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.15</td>
</tr>
<tr>
<td>Coastal C</td>
<td>0.15</td>
<td>0.10</td>
<td></td>
<td></td>
<td></td>
<td>0.10</td>
</tr>
<tr>
<td>Inland D</td>
<td></td>
<td>0.20</td>
<td></td>
<td></td>
<td></td>
<td>0.20</td>
</tr>
<tr>
<td>Inland E</td>
<td>0.20</td>
<td>0.20</td>
<td>0.05</td>
<td>0.25</td>
<td></td>
<td>0.25</td>
</tr>
<tr>
<td>Coastal F</td>
<td>0.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.20</td>
</tr>
</tbody>
</table>

**Table 12.10C. Joint probabilities with unit ordering as given in Table 12.10A.**
The joint probabilities of Table 12.10B would be obtained if one considered all possible ordering within the coastal and inland sets of units in each stratum independently. But this is not necessary to meet the specified objectives of the controlled selection.

<table>
<thead>
<tr>
<th></th>
<th>Unit</th>
<th>Stratum</th>
<th>Inland/Coastal</th>
<th>( \pi_k )</th>
<th>Ten balanced samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>Inland</td>
<td>0.10</td>
<td>0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>Coastal</td>
<td>0.15</td>
<td>0 0 0 0 0 0 0 0 0 1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>Coastal</td>
<td>0.10</td>
<td>0 0 0 0 1 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td>Inland</td>
<td>0.20</td>
<td>0 0 0 1 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>E</td>
<td>Inland</td>
<td>0.25</td>
<td>1 1 0 0 0 1 0 1 0 0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>F</td>
<td>Coastal</td>
<td>0.20</td>
<td>0 0 1 0 1 0 1 0 0 0</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>a</td>
<td>Inland</td>
<td>0.15</td>
<td>0 0 0 1 0 0 0 1 0 0</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>b</td>
<td>Inland</td>
<td>0.30</td>
<td>0 0 0 0 0 0 1 0 0 1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>c</td>
<td>Inland</td>
<td>0.10</td>
<td>0 0 1 0 0 1 0 0 1 0</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>d</td>
<td>Coastal</td>
<td>0.20</td>
<td>1 1 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>e</td>
<td>Inland</td>
<td>0.25</td>
<td>0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
</tbody>
</table>

Balanced sample 1-10: Each column shows a balanced sample. The two “1”s in the column identify the particular two units selected into the sample.

Similar results are obtained by using available software implementation of the cube method. One thousand balanced samples were generated, the first ten of which are shown in Table 12.10D.

It can be verified that all ten samples are of size two, one unit from each stratum, and no sample contains two coastal units. In 7 samples, the units are spread, one inland and one coastal. In only 3 samples, both units are inland. With the full 1,000 trials, the sample structure is essentially the same, one unit per stratum and no repeated coastal units. The percentage of samples with one inland and one coastal unit is about 65 per cent, the same as the value derived earlier. Correspondingly, the percentage of samples with two inland units is about 35 per cent.

**12.4.2 Illustration 1. Overlapping strata with insufficient sample size**

In Section 12.2 we presented a detailed numerical illustration of controlled selection with data consisting of 52 units. Let us recall the main points. The units are divided into three strata (A, B and C), respectively of 17, 16 and 19 units. Three units are to be selected, one from each stratum, with probability proportional to a measure of size. In addition, the units are divided into four categories by type and size of place (labelled U1, U2, U3, and U4), but at most only three of them can be represented in a sample.

---

51 The basic inputs are the given unit selection probabilities and the control variables (12.5). These imply constraints “12.6”, interpreted as “12.7”. The objective is to generate samples meeting the specified unit selection probabilities exactly, and the balancing constraints as best as possible. The actual samples generated from the procedures depends on details of implementation of the particular software. Using the R-function, \( \text{samplecube}(X, \pi) \), where \( \pi \) is the vector of inclusion probabilities \( \{\pi_k\} \) and \( X \) is the matrix of balancing variables, \( X = [x_{ik}, x_{2k}, x_{3k}, x_{4k}] \).
of size three. It is desired that, as far as possible, the three units be selected not only one from each stratum, but also from three different of the four control groups U1-U4. It is not desired to have 2 units (and certainly not 3 units) belonging to the same group. In a sense, the goal in this example is to draw a stratified sample according to two sets of overlapping strata (A, B, C, and U1, U2, U3, U4) with only three sample units, an insufficient sample size. A particular sample configuration with the desired properties is shown in Table 12.11A.

<table>
<thead>
<tr>
<th>Table 12.11A. A sample configuration in Illustration 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>U1</td>
</tr>
<tr>
<td>U2</td>
</tr>
<tr>
<td>U3</td>
</tr>
<tr>
<td>U4</td>
</tr>
</tbody>
</table>

The three sample units are shown by “1”. It can be observed that in this configuration each of the three strata is represented in the sample and three of the four U-groups are also represented (here U3 is not represented). Independent selection within strata does not ensure these joint requirements. But, as shown in Section 12.2, controlled selection does. Below, it is shown that balanced sampling procedures can also handle this problem in an efficient manner.

In order to apply the procedure of balanced sampling for the purpose of achieving the desired controlled selection, we may define the following seven auxiliary variables (j=1-7) for all units k in the survey population:

$$x_{jk} = \pi_k \delta_{jk},$$

(12.8)

where, as before, \(\pi_k\) is the inclusion probability of unit k. Subscript j refers to categories for which it is desired to control the sample size. \(\delta_{jk} = 1\) if unit k belongs to a particular control category, and \(\delta_{jk} = 0\) otherwise. The control categories of interest are:

- \(j = 1 - 4\): control groups U1-U4, respectively.
- \(j = 5 - 7\): strata A, B and C, respectively.

The required sample sizes for balanced sampling are given by \(x_j = \Sigma_k x_{jk}\); it can be verified from tables in Section 12.2 that these have the following values, as shown in Table 12.11B. Since in practice the sample sizes must be integral, the table also shows the integral values \(n_j\) on either side of \(x_j\) along with the corresponding probabilities.
12.4 Balanced sampling as a tool for controlled selection

Table 12.11B. Required sample sizes by control groups (U1-U4) and strata (A-C) in Illustration 2

<table>
<thead>
<tr>
<th>$x_j$ = 0.62</th>
<th>0</th>
<th>0.38</th>
<th>1</th>
<th>0.62</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_2 = 0.72$</td>
<td>0</td>
<td>0.28</td>
<td>1</td>
<td>0.72</td>
</tr>
<tr>
<td>$x_3 = 1.02$</td>
<td>1</td>
<td>0.98</td>
<td>2</td>
<td>0.02</td>
</tr>
<tr>
<td>$x_4 = 0.64$</td>
<td>0</td>
<td>0.36</td>
<td>1</td>
<td>0.64</td>
</tr>
<tr>
<td>$x_5 = x_6 = x_7 = 1$</td>
<td>1</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Total sample size $n = (n_1 + n_2 + n_3 + n_4) = (n_3 + n_6 + n_7) = 3$

Table 12.11C gives the basic data for this example and the first ten samples of a run of 1,000 random balanced samples generated with the R-function, `samplecube(X, π)`, where $\pi$ is the vector of inclusion probabilities and $X$ is the matrix of balancing variables, $X = [x_{jk}], j = 1 - 7$.

By design, as can be observed in Table 12.11C, all balanced samples are of fixed size with three units, one from each stratum A, B, and C ($j = 5 - 7$). It can also be verified that for all ten samples, the sample units are well spread over the groups, no sample having more than one unit concentrated in any one of the groups U1, U2, U3 and U4 ($j = 1 - 4$). Overall, the distribution of the units in the ten samples is 7 in U1, 8 in U2, 10 in U3 and 5 in U4. The corresponding percentages in the 1,000 trial samples are 20%, 25%, 34% and 21%, respectively. As expected, except for rounding errors, the values are in proportion to the sum of the balancing variables ($x_1 - x_7$) shown in Table 12.11B.

Table 12.11C. Ten random balanced samples drawn for Illustration 2

<table>
<thead>
<tr>
<th>$k$</th>
<th>Stratum</th>
<th>U-group</th>
<th>$Z_k$</th>
<th>$\pi_k$</th>
<th>Ten balanced samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>U1</td>
<td>9</td>
<td>0.093</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>U1</td>
<td>1</td>
<td>0.010</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>U1</td>
<td>2</td>
<td>0.021</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>U1</td>
<td>2</td>
<td>0.021</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>U1</td>
<td>7</td>
<td>0.072</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>U2</td>
<td>2</td>
<td>0.021</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>U2</td>
<td>8</td>
<td>0.082</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td>U2</td>
<td>7</td>
<td>0.072</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>A</td>
<td>U3</td>
<td>10</td>
<td>0.103</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>A</td>
<td>U3</td>
<td>2</td>
<td>0.021</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>A</td>
<td>U3</td>
<td>8</td>
<td>0.082</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>A</td>
<td>U3</td>
<td>8</td>
<td>0.082</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>A</td>
<td>U3</td>
<td>5</td>
<td>0.052</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>A</td>
<td>U3</td>
<td>7</td>
<td>0.072</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>A</td>
<td>U4</td>
<td>4</td>
<td>0.041</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>A</td>
<td>U4</td>
<td>9</td>
<td>0.093</td>
<td></td>
</tr>
</tbody>
</table>
Now consider the special requirement mentioned in Section 12.2: one of the three units to belong to control group U3 and one each from the other two to be among the remaining control groups U1, U2, and U4. That is, \( n_3 = 1 \) always, and \( (n_1 + n_2 + n_4) = 2 \).

In order to ensure that \( n_3 = 1 \) in all cases, we need to change (rescale) the given unit selection probabilities such that the rescale balancing constraint is

\[
\sum_{k} \pi_k' \delta_{3k} = 1
\]

This would affect the totals \( x_i - x_j \) for the strata (A-C) which contain units belonging to the overlapping control group U3. In order to continue to meet the constraint of exactly

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>A</td>
<td>U4</td>
<td>6</td>
<td>0.062</td>
</tr>
<tr>
<td>18</td>
<td>B</td>
<td>U1</td>
<td>8</td>
<td>0.093</td>
</tr>
<tr>
<td>19</td>
<td>B</td>
<td>U1</td>
<td>9</td>
<td>0.105</td>
</tr>
<tr>
<td>20</td>
<td>B</td>
<td>U1</td>
<td>8</td>
<td>0.093</td>
</tr>
<tr>
<td>21</td>
<td>B</td>
<td>U2</td>
<td>2</td>
<td>0.023</td>
</tr>
<tr>
<td>22</td>
<td>B</td>
<td>U2</td>
<td>2</td>
<td>0.023</td>
</tr>
<tr>
<td>23</td>
<td>B</td>
<td>U2</td>
<td>2</td>
<td>0.023</td>
</tr>
<tr>
<td>24</td>
<td>B</td>
<td>U2</td>
<td>7</td>
<td>0.081</td>
</tr>
<tr>
<td>25</td>
<td>B</td>
<td>U2</td>
<td>5</td>
<td>0.058</td>
</tr>
<tr>
<td>26</td>
<td>B</td>
<td>U2</td>
<td>2</td>
<td>0.023</td>
</tr>
<tr>
<td>27</td>
<td>B</td>
<td>U3</td>
<td>7</td>
<td>0.081</td>
</tr>
<tr>
<td>28</td>
<td>B</td>
<td>U3</td>
<td>7</td>
<td>0.081</td>
</tr>
<tr>
<td>29</td>
<td>B</td>
<td>U3</td>
<td>7</td>
<td>0.081</td>
</tr>
<tr>
<td>30</td>
<td>B</td>
<td>U3</td>
<td>1</td>
<td>0.012</td>
</tr>
<tr>
<td>31</td>
<td>B</td>
<td>U4</td>
<td>4</td>
<td>0.047</td>
</tr>
<tr>
<td>32</td>
<td>B</td>
<td>U4</td>
<td>6</td>
<td>0.070</td>
</tr>
<tr>
<td>33</td>
<td>C</td>
<td>U4</td>
<td>9</td>
<td>0.105</td>
</tr>
<tr>
<td>34</td>
<td>C</td>
<td>U1</td>
<td>4</td>
<td>0.036</td>
</tr>
<tr>
<td>35</td>
<td>C</td>
<td>U1</td>
<td>8</td>
<td>0.073</td>
</tr>
<tr>
<td>36</td>
<td>C</td>
<td>U2</td>
<td>5</td>
<td>0.045</td>
</tr>
<tr>
<td>37</td>
<td>C</td>
<td>U2</td>
<td>7</td>
<td>0.064</td>
</tr>
<tr>
<td>38</td>
<td>C</td>
<td>U2</td>
<td>9</td>
<td>0.082</td>
</tr>
<tr>
<td>39</td>
<td>C</td>
<td>U2</td>
<td>1</td>
<td>0.009</td>
</tr>
<tr>
<td>40</td>
<td>C</td>
<td>U2</td>
<td>5</td>
<td>0.045</td>
</tr>
<tr>
<td>41</td>
<td>C</td>
<td>U2</td>
<td>7</td>
<td>0.064</td>
</tr>
<tr>
<td>42</td>
<td>C</td>
<td>U3</td>
<td>8</td>
<td>0.073</td>
</tr>
<tr>
<td>43</td>
<td>C</td>
<td>U3</td>
<td>10</td>
<td>0.091</td>
</tr>
<tr>
<td>44</td>
<td>C</td>
<td>U1</td>
<td>3</td>
<td>0.027</td>
</tr>
<tr>
<td>45</td>
<td>C</td>
<td>U1</td>
<td>4</td>
<td>0.036</td>
</tr>
<tr>
<td>46</td>
<td>C</td>
<td>U1</td>
<td>4</td>
<td>0.036</td>
</tr>
<tr>
<td>47</td>
<td>C</td>
<td>U1</td>
<td>1</td>
<td>0.009</td>
</tr>
<tr>
<td>48</td>
<td>C</td>
<td>U1</td>
<td>5</td>
<td>0.045</td>
</tr>
<tr>
<td>49</td>
<td>C</td>
<td>U2</td>
<td>4</td>
<td>0.036</td>
</tr>
<tr>
<td>50</td>
<td>C</td>
<td>U4</td>
<td>10</td>
<td>0.091</td>
</tr>
<tr>
<td>51</td>
<td>C</td>
<td>U4</td>
<td>7</td>
<td>0.064</td>
</tr>
<tr>
<td>52</td>
<td>C</td>
<td>U4</td>
<td>8</td>
<td>0.073</td>
</tr>
</tbody>
</table>
one selection per stratum, the selection probabilities of other units will also need to be
adjusted. A simple way to achieve that would be to adjust unit selection probabilities
proportionately in an appropriate manner. Let
\[ S_{j3} = \sum_{k} \pi_k \mid (k \in U_3 \text{ and } k \in j), \quad j = 5 - 7 \]
be the sum of probabilities of all units which belong to control group U_3 and also to stratum j.
Using probabilities of these units redefined according to equation (12.9), we have the
adjusted sum \( S'_{j3} = \sum_{k} \pi'_k \). Now, rescaling selection probabilities of all units in strata
\( j=5-7 \) (meaning, in our original notation, strata A-C) which do not belong to group U3 as
\[
\pi'_k = \left(1 - \frac{S'_{j3}}{1 - S'_{j3}}\right) \pi_k \mid (k \notin U_3 \text{ and } k \in j), \quad j = 5 - 7 \quad (12.10)
\]
ensures that, as required,
\[ \sum \pi'_k = \sum \pi_k = 1 \mid k \in j, \quad j = 5 - 7. \]
For control groups U1, U2, and U4, the required sample sizes are recomputed using the
adjusted unit selection probabilities:
\[ x'_j = \sum \pi'_k \mid k \in j, \quad j = 1, 2, 4. \quad (12.11) \]
This is the same as the small adjustment to unit selection probabilities considered in
Section 12.2 for the same purpose of ensuring exactly one selection from control group
U3, expressed more formally. Of course, once this adjustment to selection probabilities
is made, either of the procedures – controlled selection or balanced sampling –
automatically ensures exactly one selection from U3. The need for early termination
of the search for controlled selection patterns occurred in the example of Section 12.2
only because, in order to construct a more instructive illustration, we chose not to make
the adjustment to begin with.
Table 12.11D presents in a concise manner ten balanced samples randomly drawn with
the R-function \( \text{samplecube}(X, \pi) \) using the control variables modified as above.

<table>
<thead>
<tr>
<th>U1</th>
<th>U2</th>
<th>U3</th>
<th>U4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>1</td>
<td>38</td>
<td>14</td>
<td>50</td>
</tr>
<tr>
<td>26</td>
<td>16</td>
<td>48</td>
<td>12</td>
</tr>
</tbody>
</table>

The cell numbers refer to \( k \), the serial number of the sample unit in the total population,
\( k = 1, \ldots, 52 \). Note that in all cases, there is a sample unit in each strata (A, B, and
C), and a unit representing the control group U3. The other control groups (U1, U2, and U4) are represented evenly, exactly one control group absent in each sample case.

12.4.3 Illustration 2. Controlling sample distribution by location and time segment

The third and last numerical example concerns the problem of the distribution of the sample simultaneously in space and in time in a typical time-location sample survey, discussed in Section 12.3. A survey is to be carried out in \( n_1 \) locations over \( d = 5 \) days of the week, and each day is divided into \( t = 4 \) observation periods of two hours duration, 10:00-12:00, 12:00-14:00, 14:00-16:00 and 16:00-18:00. There are thus \( n_2 = dt = 5 \times 4 = 20 \) time segments. As described earlier, the locations by time segment form the primary sampling units (PSUs) of the survey. There are thus altogether \( N = n_1 n_2 \) primary sampling units from which to select a sample. It is required to sample PSUs with equal probability, and controlled selection procedure is to be used to ensure that the maximum possible number of different time segments (i.e. maximum variety of day and time-of-day combinations) as well as the maximum possible number of different locations appear in the sample. Further detail of the requirements depending on the size of the sample \( n \) relative to the number of locations \( n_1 \) and the number of time segments \( n_2 \) have been given in Section 12.3. PSUs are to be selected with equal probability

\[
\pi = \frac{n}{N} = \frac{n}{n_1 n_2}.
\]  

(12.12)

Concerning time segments, it is necessary to distinguish between \((d)\) different days and \((t)\) different times of the day \((n_2 = dt)\). We will use subscript \(h\) to indicate location and double subscript \((ij)\) to indicate, respectively, day and time of day. These in fact are the stratifying variables. The population is stratified according to each of a number of stratification variables. A sample is to be selected with specified unit selection probabilities \(\pi_k\).

We define \(n_1\) auxiliary variables

\[
x_{h,k} = \pi \delta_{h,k}, \quad \text{with} \quad \delta_{h,k} = 1 \text{ if } k \in h; \quad \delta_{h,k} = 0 \text{ otherwise},
\]  

(12.13)

and \(n_2\) auxiliary variables

\[
y_{i,j,k} = \pi \delta_{i,j,k}, \quad \text{with} \quad \delta_{i,j,k} = 1 \text{ if } k \in (ij); \quad \delta_{i,j,k} = 0 \text{ otherwise}
\]  

(12.14)

and one auxiliary variable for the total fixed sample size \(n\):

\[
z_k = n.
\]  

(12.15)

Summing the above auxiliary variables over all units in the population gives the required marginal sample sizes by location and by time-segment for a balanced sample, namely,

\[
x_h = \Sigma_U x_{h,k}, \quad y_{i,j} = \Sigma_U y_{i,j,k}; \quad \text{of course, it also gives the total sample size} \quad n = \Sigma_U z_k.
\]

Formulating the controlled selection in terms of balanced sampling in this manner requires \((n_1 + n_2 + 1)\) balancing variables. The additional balancing equation is to ensure a fixed sample size \(n\).
In fact, each location has the same number \((n_2)\) of PSUs, and in an equal probability sample, the same number is to be selected, namely \(x_h = \left( \frac{n}{n_1} \right)\). Similarly the required sample size for each time segment is a constant \(y_{ij} = \left( \frac{n}{n_2} \right)\).

In practice, the sample sizes have to be integral, and may be taken as one of the two nearest integers to the above, for example as follows for sample size per location:

\[
\begin{align*}
x_h &= \text{int}\left(\frac{n}{n_1}\right) + 1, \text{ with probability } \left(\frac{n}{n_1}\right) - \text{int}\left(\frac{n}{n_1}\right), \text{ and} \\
x_h &= \text{int}\left(\frac{n}{n_1}\right), \text{ with probability } \text{int}\left(\frac{n}{n_1}\right) + 1 - \left(\frac{n}{n_1}\right).
\end{align*}
\] (12.16)

The same applies for \(y_{ij}\), the constant sample size per time-segment. The overall constraint is the fixed total sample size, \(n\). The relevant numerical values for the five cases considered in Section 12.3 are presented in Table 12.12.

<table>
<thead>
<tr>
<th>Case</th>
<th>Condition</th>
<th>No. of locations (n_1)</th>
<th>No. of time segments (n_2)</th>
<th>Sample size (n)</th>
<th>(N=\frac{n n_2}{n_1 n_2})</th>
<th>Equal selection probability (\pi = n/N)</th>
<th>No. of balancing variables (n_1 + n_2 + 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>(n_1=n_2=n)</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>400</td>
<td>5.0%</td>
<td>41</td>
</tr>
<tr>
<td>B</td>
<td>(n&gt;n_1)</td>
<td>18</td>
<td>20</td>
<td>25</td>
<td>360</td>
<td>6.9%</td>
<td>39</td>
</tr>
<tr>
<td>C</td>
<td>(n&lt;n_1)</td>
<td>17</td>
<td>20</td>
<td>16</td>
<td>340</td>
<td>4.7%</td>
<td>38</td>
</tr>
<tr>
<td>D1</td>
<td>(n_1&lt;n&lt;n_2)</td>
<td>16</td>
<td>20</td>
<td>17</td>
<td>320</td>
<td>5.3%</td>
<td>37</td>
</tr>
<tr>
<td>D2</td>
<td>(n_1&gt;n&gt;n_2)</td>
<td>24</td>
<td>20</td>
<td>22</td>
<td>480</td>
<td>4.6%</td>
<td>45</td>
</tr>
</tbody>
</table>

Specific random samples drawn with the R-function `samplecube` (parameter `method=2`) for different values of \(n_1, n_2\) and \(n\), of the five cases A, B, C, D and E of Section 12.3 are shown in Table 12.13. The spread of the sample over the locations and time-segments may be assessed for each case. It can be verified that in all cases, the detailed requirements of Section 12.3 are indeed satisfied.
### Table 12.13. Balanced samples for controlled selection of time and location

#### Case A: $n_1 = n_2 = n$

| Weekly-day | M | M | M | M | T | T | T | W | W | W | W | T | T | T | T | F | F | F | F | Total |
| Time-segment | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 20 |
| Location | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 20 |

#### Case B: $n > n_1$

| Weekly-day | M | M | M | M | T | T | T | T | W | W | W | W | T | T | T | T | F | F | F | F | Total |
| Time-segment | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 25 |
| Location | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 20 |

### Case C: \( n < n_1 \)

<table>
<thead>
<tr>
<th>Weekly-day</th>
<th>M</th>
<th>M</th>
<th>M</th>
<th>M</th>
<th>T</th>
<th>T</th>
<th>T</th>
<th>T</th>
<th>W</th>
<th>W</th>
<th>W</th>
<th>W</th>
<th>T</th>
<th>T</th>
<th>T</th>
<th>F</th>
<th>F</th>
<th>F</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time-segment</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Location</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

### Case D: \( n_1 > n > n_2 \)

<table>
<thead>
<tr>
<th>Weekly-day</th>
<th>M</th>
<th>M</th>
<th>M</th>
<th>M</th>
<th>T</th>
<th>T</th>
<th>T</th>
<th>T</th>
<th>W</th>
<th>W</th>
<th>W</th>
<th>W</th>
<th>T</th>
<th>T</th>
<th>T</th>
<th>F</th>
<th>F</th>
<th>F</th>
<th>F</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time-segment</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Location</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>16</td>
</tr>
</tbody>
</table>
Case E: $n_1 < n < n_2$

The $(n_1 + n_2 + 1) = (n_1 + dt + 1)$ constraints of equations (12.13)-(12.15) balance the sample in terms of marginal distribution by location and by time-segments. However, in order to control separately for the marginal distributions by day and by time of day – as required in the original formulation of the problem in Section 12.3 – we need to add the following $(d+t)$ balancing equations, giving a total of $n_1 + (d+1)(t+1)$ constraints:

$$d \text{ equations for days: } \sum_j y_{ij} = \left(\frac{n}{n_2}\right) t = \left(\frac{n}{d}\right)$$

$$t \text{ equations for t times of day: } \sum_i y_{ij} = \left(\frac{n}{n_1}\right) d = \left(\frac{n}{t}\right)$$

12.5 Concluding remarks

Balanced sampling is a more general technique than controlled selection. Its objective is to control the distribution of the achieved sample according to some specified set of auxiliary (control) variables. These control variables may include one or more stratification variables, in which case the procedure corresponds to controlled selection as discussed in this chapter.

Both the cube method for balanced sampling of Deville and Tillé (2004) and the controlled selection procedure of Goodman and Kish (1950) require the auxiliary variables (in our context, the stratification variables) to be known for all units in the population. Also, both the procedures are useful when stratification involves multiple variables and is detailed, and consequently, when the stratum sample sizes are small.
For these reasons, the procedures are suitable for selecting primary sampling units in a multistage sampling design. The two approaches also share the initial – and crucial – steps in their application. Both require an understanding and clear specification of the stratification controls required, as is clear from the detailed illustrations presented above.

Once the requirements have been specified, the classical controlled selection procedure involves identifying permissible samples one by one on the basis of subjective judgement. In the process, one may reach a point when no more samples exactly satisfying the conditions concerning (i) the given unit selection probabilities and (ii) the required sample sizes by stratum, can be selected. Judgement has to be made concerning how to proceed from this point and conclude the process. Subjective judgement for carrying out the above tasks requires expertise and can be time consuming. On the other hand this procedure permits greater control and flexibility; for instance, when both conditions (i) and (ii) cannot be satisfied, one may be able to relax both of them a little in order to obtain the best possible samples. The procedure is transparent and does not require specialised software. These are major practical advantages.

Balanced sampling algorithms, based on a procedure like the cube method, require a more formal specification of the balancing equations (see for instance Illustration 3 above). After that, available software can be used to identify balanced samples automatically. But not quite automatically: choices have to be made concerning the precise procedure to be used for identifying balanced samples. Furthermore, when a balanced sample cannot be obtained, some procedures for choosing the ‘best’ (most closely balanced in some defined sense) sample has to be specified. For example, it may involve progressively abandoning the balancing equations one-by-one in a certain order, or seeking the sample with the least departure from balancing in terms of some criterion. With stratified sampling, there is a choice in this respect between ‘global balancing’ and ‘stratified balancing’ (Chauvet, 2009). Some of these options may simply appear as defaults in the software, but they are real choices nevertheless.
V. Reclusive populations
Chapter 13
Snowball sampling

13.1 Introduction

The snowball sampling method belongs to the class of link-trace sampling procedures. It is used for the purpose of obtaining a study sample through referrals made among people eligible for inclusion in the study who have links with or are known to each other. The referrals by those in the initial sample may form ‘chains’: those referred to, themselves become the source of further referrals, and the procedure may be repeated a number of times. For this reason it is often described also as a chain-referral method.

13.1.1 What is snowball sampling?

The term snowball sampling is often used in one of two different senses (see for instance, Handcock and Gile, 2011). It is important to clarify at the outset the sense in which we will use it in the present chapter. The first usage treats snowball sampling as if it is a formal, probability-based procedure. The second takes snowball sampling as a more or less informal method, developed to obtain acceptably representative and large enough samples of difficult-to-access populations. It is this second interpretation which is useful in the practical context of surveying an elusive population of labouring children.

A. Snowball sampling as a formal, probability-based procedure

Snowball sampling may be applied as a formal methodology for making inferences about a population which is difficult to enumerate using normal probability-based methods such as household surveys. It applies a formal sampling method with the purpose either to make inference with regard to a population of individual units, or make inference with regard to the structure of networks or relationships among units in that population. The procedure follows links for the purpose of studying, and therefore sampling, relationships rather than individuals. The use of link-tracing is motivated by the improvement in efficiency allowed by oversampling the relationships involved in the structure being studied.

This type of usage is represented well by the work of Goodman (1961). The Goodman design follows a fixed number of referrals from each participant for a fixed number of waves of snowballing. Variations on this classical design are possible: for instance, one may follow all the links the respondent has (Frank, 1977); or follow exactly one link from each respondent to obtain what is called a random-walk design (Klovdahl, 1989).
B. Snowball sampling in the practical context of surveying elusive populations of labouring children

Many populations of policy concern are hidden populations, often made up of people in stigmatised conditions. These populations pose a range of methodological challenges if we are to understand more about their lives. Practical problems in sampling such populations include its unknown size, the geographical clustering of its members, difficulties in the identification of target groups and establishment of contacts with them, and the inability or unwillingness of individuals to participate in the study. To cope with some of these problems, a frequently applied procedure is to use a link-tracing design, introduced in Section 4.5. Link-tracing methodologies seek to take advantage of the social networks of identified respondents to provide the researcher with an ever-expanding set of potential contacts (Spreen, 1992). This process is based on the assumption that a ‘bond’ or ‘link’ exists between the initial sample and others in the same target population, allowing a series of referrals to be made within a circle of acquaintances. Snowball sampling can be placed within the wider set of link-tracing methodologies (Faugier and Sargeant, 1997). It has been used in studies in many diverse fields involving hidden and reclusive populations.

In this practical context, the term is used for a sampling procedure which is applied as an informal method to reach a target population. Coleman (1958), also drawing on the work of Trow (1957), is perhaps the most commonly quoted reference for the meaning of snowball sampling in this second sense. Current usage of the term has overwhelmingly returned to the idea where the term snowball sampling is taken to refer to a convenience sampling mechanism. It collects a sample from a population in which a standard sampling approach is either impossible or prohibitively expensive, for the purpose of studying characteristics of individuals in the population. Such settings are characterized by the lack of a serviceable sampling frame, so that an initial probability sample is either impossible or impractical.

If the aim of a study is primarily explorative, qualitative and descriptive, snowball sampling offers many practical advantages. Consequently, the method is used most frequently to conduct qualitative research, primarily through interviews. As noted by Atkinson and Flint (2001), “treading an uneasy line between the dictates of replicable and representative research design and the more flowing and theoretically led sampling techniques of qualitative research, snowball sampling lies somewhat at the margins of research practice”. It is in this last-mentioned mode that we find early, landmark examples of the technique in studies such as Street Corner Society by Whyte (1955), and the study A Glasgow Gang Observed by Patrick (1973), which used initial contacts to generate contexts and encounters that could be used to study the gang dynamic.

Over the years there has been a general move away from participant observation of this kind, towards the use of the snowball sampling technique for interview-based quantitative research. Snowball sampling has advanced as a technique toward becoming a more sophisticated method of sampling and data analysis.

Biernacki and Waldorf (1981) review problems and techniques of snowball sampling, and we will draw on this work in the discussion in Section 13.6 below.
13.1.2 Outline of the snowball sampling procedure

With snowball sampling a unit of the target population can enter the sample in one of the following ways.

(1) The unit is directly selected into the initial sample. The collection of units so selected constitutes wave zero of the snowball sample. Only units meeting eligibility criteria for membership of the target population can be included. This applies to sampling for all the snowball waves described below.

*In principle, the initial sample can be a random sample from the population. In practice, usually the initial sample is a (very) restricted convenience sample because of the hidden nature of the population and the lack of a sampling frame.*

(2) Each responding person in the initial sample is asked to ‘name’ (refer) for inclusion in the sample other members of the population he/she is linked to or knows. Some of these individuals may be named by several initial sample respondents. Persons so named, excluding any who were already in the initial sample, are included for the sample. Those meeting the eligibility conditions and responding to the survey constitute wave one of the snowball sample.

(3) Similarly, wave two consists of eligible and responding persons named by any (one or more) wave two individuals, excluding any already included in a previous wave.

(4) The same procedure can be repeated for as many waves as desired.

(5) The snowball sample consists of the initial sample and all the waves successively found around it. A wave is final if none of its members mention any individuals that have not been previously mentioned. Snowball samples often are incomplete in the sense that the sampling is stopped before the final wave so defined.

There are a number of parameters which define the design of a snowball sample.

**Number of waves**

The number of waves in a snowball sample is the largest number of separations (steps) any unit in the final sample has from its seed – that is, the length of the longest chain in the sample. Most chains may terminate before reaching this length. The process ends when the waves cease to produce a predetermined number of new contacts. In small populations, it may take only a few waves before almost no new contacts are obtained in a wave, whereas larger populations may permit more waves.

The ideal number of links in a referral chain will vary depending on the purpose of the study. More links in each chain will generate substantial data about a particular sample, and may also allow access to those most difficult to identify (e.g. those respondents who require the greatest level of trust to be built up before participating). However, it is also more likely that members of such a large single chain sample will share similar and unique characteristics not shared by the wider population. Thus, there may be a case for initiating several discrete chains with fewer links, particularly where any inference about a wider hidden population is considered important. There is in fact a fundamental conflict between the valuable use of long referral chains to move into parts of the population not accessible to ordinary sampling procedures, and the negative
effect of the homogeneity of the clusters created by long chains on the statistical efficiency of the resulting sample. This for instance manifests itself in the form of large design effects often found in applications of the similar technique of respondent-driven sampling (RDS), as will be discussed in Chapter 14.

**Number of contacts to request**

Usually, the number of contacts asked for is kept small (say two to four), partly to minimize the burden on the respondent but also to minimise the potentially biasing impact of participants with very large social networks.

**Criteria for including a participant in the sample**

The initial sample is built from recommendations of experts and informers familiar with the study objectives and the scene, from personal knowledge and observation of the researchers, and from individuals offering themselves as volunteers. Usually individuals are included in the sample for subsequent snowball waves only if (i) they have been referred or named by someone in the sample; (ii) they meet the criteria of eligibility for the study; (iii) self-confirm their eligibility; and (iv) also their willingness to participate in the study. Some studies drop condition (i), i.e. permit inclusion of persons volunteering themselves without being referred or named by someone in the sample.

**13.1.3 Strengths and limitations of snowball sampling**

Snowball sampling is more directed and purposeful than most other non-random sampling techniques, such as convenience sampling that focuses only on the most easily identified and reachable members of a population. When carefully conducted, snowball sampling can provide very useful characterisations of unknown populations.

**A. Potential uses of snowball sampling**

There is a range of potential advantages in favour of the snowball sampling approach.

(1) The primary advantage of snowball sampling is its success in identifying individuals from unknown populations and from small, hidden groups dispersed within a large population. When there is little information available about the population, snowball sampling can help in obtaining a better understanding and more complete characterisation of a population.

Furthermore, snowball methodology provides a means of accessing social groupings which are vulnerable or difficult to penetrate, and a means for obtaining respondents in settings where trust is required to initiate contact. Under these circumstances, the technique may imbue the researcher with characteristics associated with being an insider or group member, and this can aid entry into settings where conventional approaches cannot succeed. The procedure has enabled access to populations whose members may be involved in activities that are considered deviant, or may be vulnerable or stigmatised, making them reluctant to take part in more formalised studies using traditional research methods. Trust may be developed as referrals are made by acquaintances or peers rather than through other more formal methods of identification.
(2) In a study of techniques for assessing the representativeness of snowball samples, van Meter (1990) has elaborated the distinction between ‘descending’ and ‘ascending’ methodologies. Perhaps one of the strongest recommendations for the snowball strategy stems from this distinction. Traditional techniques such as household surveys, as descending procedures, are associated with a largely quantitative tradition of the measurement of social problems that often suffer from non-response, or at best a poor response, from particular groups. Ascending methodologies, such as the use of snowball techniques, can be used to work upwards from individual cases, and locate those hard-to-reach groups on the ground. In this sense snowball sampling can be considered as an alternative complementary to conventional surveys for attaining more comprehensive data on a particular issues, problems, and populations.

(3) While snowball strategies have been used primarily as an aid to accessing the vulnerable or the deviant, the method in fact has a wider applicability in sociological research. Some studies have used it for other purposes – for instance to study exclusive, hard to reach urban elites.

Snowball sampling has been found to be economical and effective in various studies. It has been shown to be capable of producing comparable data on some topics across times, populations, even countries. Snowball sampling can also produce in-depth results and can produce these relatively quickly.

B. Limitations

Snowball samples also have a number of deficiencies. The most important are the following.

(1) The quality of the data, and in particular selection bias which limits the validity of the sample, are a primary concern. Because elements are not randomly drawn but are dependent on the choices of the respondents first accessed, most snowball samples are biased and do not therefore allow valid generalisations from a particular sample. Snowball samples will be biased towards the inclusion of individuals with more inter-relationships, and therefore will over-emphasise cohesiveness in social networks, and will miss isolated individuals who are not connected to any network that the researcher has tapped into.

In addition to selection bias, there is also the issue of ‘gatekeeper bias’ – insiders chosen to obtain access in fact becoming a barrier to achieving that. For instance, in their work Groger, Mayberry and Straker (1999) identified a difficulty when using nursing home staff as go-betweens in obtaining the informed consent of caregivers. These ‘gatekeepers’ were sometimes reticent or protective toward those in their care, and hindered access by the researchers.

(2) The snowball method generally is not an effective tool for producing reliable estimates of total population size or population aggregates of other variables. Sometimes other methods, such as capture-recapture sampling, have been used to supplement it for the purpose. (See Chapter 11; also the example from David and Snijders (2002) described in Section 13.3.1.)
The estimation of statistics such as proportions, means or other ratios for the population requires proportionate representation, or representation with known weights, of different population subgroups in the sample. This generally cannot be ensured with snowball sampling. However, snowball samples can yield valid comparisons between different subgroups of the population. This is because comparisons, such as those between subpopulation means, do not require proportionate samples. Rather, differences between subgroup means are estimated most efficiently by having equal subpopulation sample sizes – which, in principle at least, can be ensured with snowball sampling.

Overall, it is clear that snowball sampling is potentially subject to large bias. However there is some evidence that the problem of selection bias may be partially addressed, (i) firstly through the generation of a large sample (e.g. Atkinson and Flint, 2001); (ii) relying on a variety of sources to develop the seeds for the initial sample; (iii) making special and concerted efforts to reach groups known to be particularly isolated; and (iv) by seeking replication of the results through repeated application in the same as well as in different settings, so as to enhance confidence in the findings.

13.2 Estimation with snowball sampling

Snowball sampling can seem to contradict many of the assumptions underpinning conventional notions of sampling but has a number of special and unique uses as we noted above.

The problem of estimating from a snowball sample arises from the following aspects of the situation. The selection of a unit into the sample depends on two processes, not necessarily independent of each other: (i) the process of being selected into the initial sample; and (ii) the process of being named for inclusion by another unit already selected (or named) into the sample. Either or both these processes may (and often do) lack a random or probability basis.

13.2.1 Non-probability selection into the initial sample

A. Potential problems

As to the selection into the initial sample, there is a series of potential problems.

(1) Exclusion

A unit may have no chance of selection if it is not represented in the sampling frame and cannot be accessed for inclusion through another means. The most common situation is that the selection of the initial sample is confined only to the more accessible parts of the target population. The sample base needs to be made more inclusive. This in fact is the reason for resorting to snowball sampling in the first place.

(2) Non-random selection

Even if the unit is available for inclusion, its selection into the initial sample is usually not random.
In practice, many snowball samples have to begin from whatever eligible units in the target population happen to be known and available for inclusion in the initial sample. The researchers may try to diversify the initial sample to different parts of the target population, but the introduction of random selection procedures is often not possible.

(3) **Unknown selection probabilities**

And even if some measure of randomness can be introduced into the selections, the actual values of the selection probabilities involved may remain unknown.

By definition, there is no frame of the hidden population as long as it is hidden; hence random sampling procedures cannot be designed exclusively for the hidden population. But it may still be possible to introduce randomised selection. For instance, we may first select units from the general population of which the target population is a part, and then identify through screening units of the target population in the random sample of the general population so selected (for instance, first selecting a sample from the population of all small establishments, and then identifying within the sample those establishments which employ child workers). The problem remains that, in the absence of a sampling frame or some other relevant source of information, the size of the population, and hence the sampling rate employed would be unknown. An example is provided by the use of Bernoulli sampling. This design is a way of selecting individual units from a population according to a procedure that decides independently for each unit whether or not the unit should be selected into the sample. For instance, individuals may be selected at random as they enter or leave a location, as discussed under various settings in Chapter 10 above. We may select a random sample of some pre-determined size, but without knowing what the size of the population, and hence what the sampling rate applied, is. The sample may be considered a Bernoulli sample with a common but unknown selection probability. More elaborate schemes could distinguish between different selection probabilities in different strata of the hidden population.

It should be reemphasised, nevertheless, that the above applies only to the part of the target population for which random sampling is possible; it excludes situations (1) and (2) above, namely when units are excluded altogether or, if included, the selection is purposive.

**B. Different sampling situations for the initial sample**

To be more complete, we may view the target population as divided into five components in terms of the sampling situation as follows.

(1) A part in which it is possible to apply normal probability sampling – random selection with known (non-zero) selection probabilities.

(2) A part in which it is possible to apply random sampling, but the sampling probabilities are unknown or are known only in relative terms. The actual (absolute) values of the selection probabilities are not known.

(3) A part from which units can be included into the sample, but it is not possible to apply random sampling for the purpose.
(4) Inaccessible parts of the population from which no unit can be selected or otherwise included in the sample. This may result from the lack of a sampling frame and/or from other barriers preventing access to units.

(5) Units whose eligibility for inclusion in the target population is uncertain, present another problem. This refers to a lack of clarity in the definition of the target population – what is included and what is excluded. While this problem may exist in any survey, it is likely to be a more serious problem when dealing with elusive populations. For instance, as has been often noted in the literature, a standard definition for measurement of the phenomenon of homelessness (or of deprivation, social exclusion, child abuse, hazardous working conditions, etc., for that matter), has not been established. Furthermore, homeless and similar populations can be very variable over time and geographically.

It is instructive to conclude with a quote from Snijders (1992).

“The weak underbelly of the snowball method is the assumption of a random initial sample. Snowball samples are mostly taken from populations for which a sampling frame is not available. There are even cases where the [total population size] is to be estimated from the snowball sample ... Without a sampling frame, how can we draw a simple random initial sample?

“We cannot. The best we can do is to draw the respondents, as much as possible, from independent sources. E.g. if a snowball sample of drug users is to be taken and ‘bars’ is one of the ‘social milieux’ where initial respondents can be sought, not more than one initial respondent is to be sought in one bar or in one small-scale ‘social environment’ of any kind. This physical approximation to independence will hopefully lead to something approaching random sampling in the sense that for individuals to be together in the initial sample is uncorrelated with direct or indirect (larger distance) ties between them. However, practically all ‘ethnographic’ methods to get initial respondents will lead to bias in the sense that the more widely known individuals ... are overrepresented, even in the initial sample. It would be interesting to try some simulation examples to see how badly this affects estimation results, and try to find correction methods for this bias.

“A related point is that there are often separate social sources of initial respondents; e.g. in a study of drug users: bars, police contacts, socio-medical institutions, and educational institutions. The initial sample is then stratified, ... [but] often there is not enough information to determine whether the initial sampling fractions in the subpopulations are anywhere near each other. In such cases, the assumption that the initial sample of the snowball is a stratified random sample may be much closer to reality than the assumption that it is a simple random sample. It could be worthwhile to elaborate estimation methods that are valid under the assumption of stratified random initial sample; or, more generally, under the assumption of known but varying probabilities of inclusion in the initial sample”.
13.2.2 Non-probability inclusion of units in subsequent waves

A. Potential problems

(1) Isolates

Reclusive populations, by definition, are highly likely to contain ‘isolates’ – individuals who are not linked to others in the population to be referred or named by them for inclusion into the sample for subsequent waves. It can be assumed that they are equally or even more unlikely to be available for the initial sample. Therefore, such populations remain unrepresented in the sample – in the initial sample, as well in subsequent additions to it through snowball sampling. Perhaps introducing different, more intensive procedures to supplement the main survey may bring in a part of such isolated subgroups.

(2) Degrees of isolation

Actually, there is not a sharp dichotomy between ‘perfectly linked’ individuals on the one hand, and ‘complete isolates’ on the other. More likely, the population is segmented into groups subject to different degrees of isolation. Isolation among units within a group; isolation between groups; and – what is of particular concern here – isolation of population groups from the sampling process.

(3) Individual differences in network sizes

Among units which are linked, the chance of a unit being named depends on what is called its degree or network size. This means the number of links from other units which lead to it, i.e. the number of units which would name this unit if they were selected.

(4) Links varying in strength

The above is a simplified picture of a more varied situation. In real life, referral links between units are usually more complex. They may not be simple ‘yes-no’ dichotomies. Some links may be stronger or weaker than others, reflecting the chance that the link will actually lead to referral or naming.

By links having ‘different strengths’ we mean that a link between two units may have a certain probability \( L \) of being realised which can differ from 0 or 1, \( 0 \leq L \leq 1 \), indicating the salience of the relationship. For instance, person \( a_1 \) may name person \( a_2 \) with certainty \( L = 1 \), but name person \( a_3 \) (with whom the relationship is weaker) only sometimes \( L < 1 \). To complicate matters further, the salience of the relationship between two individuals may not be simply an objective fact, but something influenced by the survey conditions and procedures. For instance, more probing and careful questioning may identify more links than would casual, less careful survey questioning.

(5) Directional links

This refers to relationships lacking symmetry. A relationship is said to lack symmetry if the presence and salience of the relation from person \( a_1 \) to person \( a_2 \) is different from the presence and salience of the relation from person \( a_2 \) to person \( a_1 \). For instance,
person $a_1$ may count person $a_2$ among ‘friends’, or claim to ‘know’ him, or to have some other form of specified relationship, which determines that $a_1$ would name $a_2$ for inclusion in the survey. It is quite possible that, by contrast, person $a_2$ does not reciprocate these sentiments; he/she may not perceive having a link with person $a_1$, or may see the link as being weaker.

Differences in the strength and lack of symmetry in the links between population elements makes it difficult to measure these in the survey. The questions which need to be asked of the respondent can become too complex to be practicable. For relationships lacking symmetry, what we need is not how many friends person $a_1$ would name if asked, but how many other persons would name $a_1$ as one of their ‘friends’. Person $a_1$ would find it even harder to give the likelihood that each of the other persons would be naming $a_1$.

(6) Non-random selection among the links

It is often assumed that in snowball sampling, each unit names the specified number of links by making the choice at random from his/her total set of links. This is hardly likely to be the case in practice. The process of ‘masking’, for instance, refers to the tendency among people to avoid naming their closest friend for the survey to spare them from the trouble or embarrassment of participation. Or they may artificially promote some friends for the survey to provide them with an opportunity to get the material incentive or the positive experience offered by the participation. There are many other possible scenarios. Individuals can also offer themselves for the survey even without being nominated by anyone.

B. Different sampling situations for the snowball waves

It is instructive to view the progress of snowball sampling from one wave to the next in the following terms. The population may be considered divided according to increasing degree of accessibility.

(1) The most accessible segment of the population is the one which is open to inclusion in the directly selected initial sample.

(2) The next somewhat less accessible is the population segment which, while not available for inclusion in the initial sample, is linked to individuals in segment (1), and therefore is available for inclusion in the first snowball wave on the basis of being named by members of the initial sample.

Individuals in (1) may name others in the same segment, and also name individuals in the next less accessible segment (2). Relative frequencies of naming members from those two levels may vary among different subgroups, and also among individuals in the initial sample.

(3) Similarly, we may think of the next more inaccessible population segment. This segment is not open to selection into the initial sample, nor to naming by individuals in (1). It has direct links to individuals in (2), who may name individuals in segment (3) for inclusion in the second snowball wave. (Of course, they may also name individuals in their own group, or in (1).)
And so on, with increasingly more inaccessible population segments becoming open to inclusion in the snowball sample wave by wave.

The underlying concept in the above scenario is that of distance between individuals or groups, by which we mean the length of the shortest path between them, or the number of barriers between them. The efficacy of a snowball sampling procedure is dependent on how well it moves along the chain

\[(1) \rightarrow (2) \rightarrow (3) \rightarrow \ldots \ldots \ldots \]

in opening up more and more inaccessible parts of the target population for inclusion in the sample. This process will of course involve ‘loops’, going back and forth between different population segments and also within segments.

### 13.2.3 Simplified models for estimation under snowball sampling

#### A. A procedure for estimating size of a hidden population

The following provides an outline and results of a procedure proposed by Frank and Snijders (1994) for estimating size of hidden populations using snowball sampling. We will not go into details of development of the estimators, but simply state the main assumptions and results, and comment on them.

The initial sample is selected from the whole population using Bernoulli sampling with constant selection probability $\alpha$. Thus the variables (say $x_i = (0, 1)$) indicating whether a unit $(i)$ is selected or not are independent identically distributed (IDD) Bernoulli ($\alpha$) variables.\(^{52}\)

The initial sample size $n$ is a binomial $(\theta, \alpha)$, where $\theta$ is the population size.

Similarly, variables $(y_{ij} = (0, 1))$, indicating whether or not a sample unit $i$ is linked to another unit $j$, are assumed to be independent identically distributed Bernoulli ($\beta$) variables, where $\beta$ is constant.

Let us use the original authors’ notation:

- $\theta$ the population size (to be estimated)
- $n$ the initial sample size
- $m$ sample size of the first wave of the snowball
- $r$ the number of links which exist between units both of which are in the initial sample
- $s$ the number of links of units in the initial sample to units in the first wave of the snowball
- $t = (r + s)$ the total number of links between units in the initial sample and other units (both inside or outside of the initial sample).

---

\(^{52}\) That is, variables generated by Bernoulli trials, defined as follows: “Sequences of events, such as those given by coin tossing or dice throwing, in which successive trials are independent and at each trial the probability of appearance of a ‘successful’ event remains constant; the distribution of success is then given by the binomial or Bernoulli distribution.” Kendall and Buckland (1957; 4th edition 1982).
13. Snowball sampling

$k$ similar to $r$, defined as the number of units in the initial sample having one or more links with other units in the initial sample. (The difference from $r$ is that in $k$, multiple links between two initial sample units are counted only once.)

Essentially the basis of the estimation is that the proportion of all the links from the initial sample which are to units within that sample ($r/t$), is the same as the proportion of all units selected into the initial sample ($n/\theta$). This gives an estimate of the unknown population size $\hat{\theta}$ as

$$\hat{\theta} = n \frac{t}{r}. \quad (13.1)$$

A number of slightly different versions of the estimator of population size have been given by the authors. The above is one of them.

The estimators proposed are the following:

$$\hat{\theta}_1 = \frac{nr + (n - 1)s}{r}$$
$$\hat{\theta}_2 = \frac{n(r + s)}{r} = \frac{nt}{r}$$

(the same as the one shown in Equation (13.1)).

$$\hat{\theta}_3$$ computed by iterative solution of the following equation

$$1 - \frac{m}{\theta - n} = \left[1 - \frac{t}{n(\theta - 1)}\right]^n$$

$$\hat{\theta}_4 = \frac{n(k + m)}{k}$$

$$\hat{\theta}_5 = \frac{nk + (n - 1)m}{k}.$$

The first three forms are model-based and the last two design-based.

B. Comment: Limitations of such procedures

The underlying assumptions of constant probability Bernoulli sampling applied on the entire population (to units in the population, as well as to links between units) are clearly unrealistic.

As already discussed, conditions under which snowball sampling is considered for practical application are very different.

In any case, the estimators are likely to provide serious underestimates of the population size for several reasons.

(1) Note that there is no reference to snowball waves beyond the first. Segments of the population which cannot be reached for the initial or the first wave sampling (but some of which may become reachable in later waves) are ignored. The estimation is, at best, of the size of the parts accessible to the initial or wave one sampling.

(2) Secondly it assumes that the relative incidence of links between units in the same population segment (e.g. those between units accessible to initial sampling) is the same as that of links between units in different segments of the population (e.g. links between units accessible to initial sampling and units accessible only to wave 1 of snowball sampling). In reality the former type of link is likely to be more
common. Ignoring this difference is likely to result in underestimation of population size.

An example given later (Section 13.3.1) indicates serious under-estimation with use of above type of estimator.

C. Estimation of means and similar statistics

It is the case that snowball sampling is not suited for estimating size and other population aggregates. It is more likely to yield useful estimates of proportions, means, ratios and similar statistics, than of totals. Snijders (1992), for instance, provides some examples. Expressions are derived for wave-specific probabilities of appearance of units in the sample.

The inverse of these probabilities can be used as weights \( w_i \), for instance in Horvitz-Thompson estimator. The assumptions are still simple random sampling for the initial sample, and random selection of links in moving from one wave to the next of snowballing. Nevertheless, the procedure does take into account variation in units’ probabilities resulting from differences in ‘degree’ (see above).

D. Making snowball estimates more realistic and useful

Directions of possible improvement include the following.

1. Improved statistical models may be developed which are more realistic in the assumptions made about the nature of the data obtained through snowball sampling.

   As an example, Félix-Medina and Thompson (2004) present a scheme combining link-tracing sampling and cluster sampling to estimate the size of a hidden population. The scheme is a variant of link-tracing sampling which is more realistic in that it avoids the assumption of an initial Bernoulli sample of members of the entire target population. Instead, a portion of the target population is assumed to be covered by a sampling frame of accessible sites. The starting point is a simple random sample from this restricted population of clusters and, as in the usual snowballing procedures, links from it to the whole of the target population are followed up.

2. Results of snowball samples should be used for purposes for which they are best suited. For instance, such samples are better suited for estimating means than estimating totals, as noted above. They are better suited for comparison among subgroups than for aggregation over subgroups. They are suited best for ‘digging into’ the target population and reflecting its complexity.

3. The snowball sampling procedure may be improved to reflect more realistically the actual conditions of data collection. For instance, steps may be taken to spread the initial sample to different segments of the population so as to capture the population’s diversity. Such possibilities are discussed more fully in Section 13.6 below.

4. Snowball sampling procedures may be improved so that they become more robust in the face of departures from the model assumptions. Respondent driven
13. Snowball sampling

sampling, discussed in the next chapter, is a development of this type. By using special survey procedures, the method aims to control the effect of biases in the initial sample.

13.3 Examples of use of snowball sampling in surveys

In this section a few examples are provided from surveys using snowball sampling, in order to convey a clearer idea of what the application of this technique may involve. These are not “best practice” examples, but chosen to illustrate some uses and limitations of the approach.

Unfortunately, some of the studies in the examples did not succeed in meeting their objectives. This may be the result of their not following properly the criteria for successful application of snowball sampling listed above (13.2.3D). Nevertheless most examples show how starting from a very restricted and small initial sample, the coverage can be extended to a much bigger and diverse sample. Of course the question remains as to how representative the expanded sample is of the target population.

(1) The first example is from a survey of homeless persons in Budapest, Hungary. It brings out two important problems. The first is that if we begin from a special population which tends to be isolated from the general population, it may prove very difficult to extend the sample through snowballing beyond that special population. The second is that generally the snowball method is not suited for estimating the total size (and other aggregates) of the population. It can seriously underestimate that size. Its primary use is to provide information on population characteristics and diversity.

(2) The second example is from a study of Japanese migrants to Brazil. It also brings out two potential problems. The first again is the difficulty in securing recruits to the sample. Secondly, the snowball sample is expanded on the basis of certain types of link among individuals in the target population. There can be a tendency for over-representation of individuals with stronger links of the type used for snowballing.

(3) The third example is from a major study over several years of migrants from Mexico coming into the U.S. We do not reproduce here technical details of the study, but note its balanced assessment of merits and limitations of the snowball sampling method. While the procedure is not well-suited for producing estimates of population size and aggregates, it can be very valuable in characterising and understanding the population and the situation being studied.

(4) The fourth is a small case-control study in building a snowball sample of a very reclusive population – that of persons with drug abuse and/or related psychiatric disorder – in Brazil. It shows how the snowball criteria for follow-up of the sample may be adapted to control the sample as it develops so as to remain within the confines of the target population. It is possible to be selective in how respondents of different types are requested to make referrals for follow up waves of the snowball.
13.3 Examples of use of snowball sampling in surveys

13.3.1 Snowball estimates of the size of the homeless population in Budapest

This study, by David and Snijders (2002), attempts to estimate the size of the homeless population in Budapest by using two ‘non-standard’ sampling methods: snowball sampling and capture-recapture method. Using the two methods and three different data sets, the study compares the methods as well as the results, and also suggests some further applications. Apart from the practical purpose of the study there is a methodological one as well: to use two relatively unknown methods for the estimation of this very particular kind of population.

A. Background to the study

It is useful to note the background situation being studied, as noted in the study report.

“Before the transition [from the Communist rule], the existence of a special homeless population – mainly young people running away from home – was acknowledged …. In 1989 what took most people – politicians, social workers, scientists and the general public – by surprise was the unexpectedly high number of homeless people emerging out of nowhere. As soon as this phenomenon came to light, the various suggestions as to how to solve this problem (providing shelters, hostels, soup kitchens etc.) were always preceded by the same question: ‘What is the number of homeless? What is the size of the population we have to deal with?’: This question – the basis of all questions – is still unanswered.”

As the definitions vary, so do the estimates of the size of the homeless population. Available estimates included values such as 10-30,000, 25,000, 30-60,000, even up to 200,000 people being under the threat of homelessness, with little official consensus about the number of the homeless.

B. Snowball data

Two methodological problems in the estimation of the number of homeless persons in any population may be noted. One is the definition of homelessness. The homeless, however defined, also constitute a changing population so that the definition has to make a restriction in time and space, and such a limitation has necessarily an arbitrary nature. The second problem is the hidden nature of this population: just like drug addicts for example, the homeless constitute a hidden population in the sense that there is no sampling frame, its members are difficult to locate or to get in touch with, and it may be hard to determine whether a given individual belongs to the population of interest. The problem of sampling hidden populations and some possible sampling methods are discussed in several articles: for example in Sudman, Sirken and Cowan (1988), Watters and Biernacki (1989), and Spreen (1992). Since there is, by definition no sampling frame for hidden populations, these sampling methods are rather unorthodox. These methods, sometimes called ascending methodologies (e.g. van Meter, 1990), include link-tracing sample designs like snowball sampling, targeted sampling, and capture-recapture methods, often supplemented by ratios calculated from secondary data.
In the questionnaire used for the present study, the following snowball question was asked to the homeless participants:

- How many other homeless persons do you know by full name? Please list them!

There were no further restrictions on how the respondent defines this set of persons. The participants were also asked the type of relation they had with such persons: for example friend, relative or acquaintance.

The following are desirable features of a snowball sample aimed at estimating the size of a hidden population according to Snijders (1992):

- In theory the respondents should be a random sample from the population. To approximate a random sample the respondents should be obtained, as much as possible, from several independent sources.
- To obtain an estimate with reasonable precision, the initial sample size should not be much smaller than the square root of the population size, and larger if the average number of nominees per respondent falls below 10.

The base sample was derived from a list of 1,753 homeless people who had been screened for lung disorders, from which a snowball dataset of 1,404 questionnaires was successfully constructed. Since this population was quite segregated not only from the non-homeless population but internally as well, the rate of those who could mention any other homeless person was already low (39 per cent); furthermore, only 29 per cent among them could give full names of the persons they nominated. In the total sample, nearly two-thirds did not give any nominee; the remaining third gave an average of 2.4 nominees per respondent.

Snowball samples don’t always ‘snowball’.

Most people listed by the respondents were from the same place where the respondents themselves usually appeared or spent the night. As the report authors note, this

“probably means that those few relationships that exist between the homeless people are mainly structured around the hostels, shelters and soup kitchens they spend the days and nights in. In principle these accommodations – especially the shelters – are to provide temporal possibilities and not permanent places for the clients. The relatively low rate of fluctuation on one hand can strengthen certain relations but on the other hand it can also hinder the formation of new connections”.

Such clustering of the homeless relationships means that the assumption of independence of the respondent and his nominees belonging to the initial sample is unlikely to be valid. This would tend to restrict the spread of the snowball sample and result in under-estimation of the size of the target population. In order to reduce (but by no means eliminate) this problem, the survey covered 11 different sites and 30 different times throughout a year. The study still could not cover locations and individuals totally unconnected to the places surveyed.
C. Estimate of the total number of homeless people in Budapest

(1) Snowball estimate

The 1,404 people in the original sample mentioned only 426 names in the first wave. This gave an estimate of 3,444 with standard error 167 (using ‘Estimate v5’ proposed by Frank and Snijders, 1994; see Section 13.2.3). In order to reduce the effect of the above-mentioned special structure of relationships among the homeless, the original data set was modified by restructuring it along the 11 different sites of the sampling and adding the corresponding 11 estimates. This gave an estimate of 4,097 (standard error 269). Of course, the snowball estimate is still likely to be an underestimate.

The corresponding capture-recapture estimates were found to be substantially higher.

(2) Comparison with capture-recapture estimates

It is interesting to compare the snowball estimates with those obtained from the capture-recapture method (see Chapter 11) applied during the same study. There were in fact two data sets available for applying the capture-recapture method, producing two different estimates of the total homeless population. These were as follows.

(i) One data set included lists of the homeless people who were screened under “Tuberculosis (TB) programme” in three consecutive years (from 1996 to 1998). The screening sites included shelters, hostels, soup kitchens and public places.

The matrix in Table 13.1 can be regarded as an incomplete three-way contingency table. For each row, codes (1,0) indicate whether (or not) the number shown were screened during each of the three years 1996, 1997 and 1998. (For example, the first row shows that the number of those homeless people who were screened only in 1998 was 1,018.) By fitting log-linear models to this data set, the authors estimated the number of homeless people who were not screened during any of the three years; this estimate is 12,345 (with a wide confidence interval from 8,654 to 17,610). Adding to this the number who were present in at least one list (4,614), it is estimated that the size of the population of persons who were homeless in any period during the years 1996-1998 was about 17,000.


<table>
<thead>
<tr>
<th>Screening pattern</th>
<th>Dates of screening</th>
<th>number</th>
</tr>
</thead>
<tbody>
<tr>
<td>only 1998</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>only 1997</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1997, ’98</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>only 1996</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1996, ’98</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1996, ’97</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>all years</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>never</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total number</td>
<td>2,030</td>
<td>1,753</td>
</tr>
</tbody>
</table>

* Estimated by fitting log-linear models.
Codes (1,0) indicate whether or not screened during a particular year.
(ii) The second data set included three complete lists for the year of 1996. In all cases the numbers refer to the whole year. The three lists included the names of: those who participated in the tuberculosis programme (2,030); those who spent at least one night in a hostel run by the Municipal Social Centre and Institutions (715); and those who were treated with any problem in the biggest emergency hospital in Budapest (174 persons).

For each row in Table 13.2, codes (1,0) indicate whether (or not) the number shown were present in each particular list. In the three 1996 lists the distribution of the homeless people is quite diverse; the overlap is smaller than in the previous data set. By fitting log-linear models to this data set, the authors estimated the number of those homeless people who were not on any of the three lists; this estimate is 3,913 (confidence interval 1,605-9,545). Adding to this the number who were present in at least one list (2,606), the estimated number of homeless in 1996 from this source is about 6,500. Note that the previously given estimate of 17,000 was for persons homeless “any time during the 3 years”. The estimate of around 6,500 in Table 13.2 refers to a single year.

<table>
<thead>
<tr>
<th>List combination</th>
<th>Source of list</th>
<th>number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TB screening</td>
<td>Hostel</td>
</tr>
<tr>
<td>only hospital</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>only hostel</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>hostel &amp; hospital</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>only TB screening</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>TB screening &amp; hospital</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>TB screening &amp; hostel</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>all lists</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>none</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total number</td>
<td>2,030</td>
<td>715</td>
</tr>
</tbody>
</table>

* Estimated by fitting log-linear models.
Codes (1,0) indicate whether or not present in a particular list.

(3) Why is the snowball estimate so low?

As to their best estimate, the authors conclude as follows.

“Considering our estimates and the available statistical data we assume that the number of the homeless population is increasing. If the average rate of this increase is at least 10% per annum from 1996, we are able to give a rough estimation on the size of the homeless population for 1999. According to our calculations this number is between 8000 and 10 000 in Budapest”.

The snowball estimate is similar to the total number (3,578) of places available for the homeless in Budapest in the survey year (1997). Since it is very unlikely that, especially given the situation in the country during the 1990s, there were places available for all – or
even a majority of - the homeless, the estimate of the total number of homeless persons is clearly a gross underestimate. The estimated figures are approximately as follows.

| Total number of places available for the homeless | 3,580 |
| Snowball estimate | 4,100 |
| Capture-recapture estimates | 6,500; & 17,000. |

We can identify at least two reasons for the snowball estimates being low.

Firstly, as the report’s authors note, it is possible that the snowball sample estimate mainly refers to the number of those homeless who have ‘stable’ living facilities, omitting the number of those who are isolated because of their temporary or uncertain accommodation. In order to get a more reliable estimate by using the snowball sampling method, the sites of the survey need to be chosen carefully and widely. It is not only the number of the different sites that is important, but also the type of site included - for example also including hospitals and similar places where the chance for the homeless people being independent of each other is higher than it is in shelters or soup kitchens specifically serving the homeless.

Secondly, we believe that the snowball estimation procedure based on Frank and Snijders (1994) is prone to seriously underestimate the size of the population. The reasons for this were noted in Section 13.2.3.

13.3.2 Survey of migrants in Brazil

We have already discussed in Section 10.4.3 this example of a survey of migrants in Brazil, from McKenzie and Mistiaen (2007). The authors report the results of an experiment designed to compare the performance of three alternative survey methods in collecting data from Japanese-Brazilian families (Nikkei community groups), many of whom send migrants to Japan. The three surveys conducted were: (1) survey of households selected randomly from a door-to-door listing using the Brazilian Census to select census blocks; (2) a location survey in which Nikkei community individuals were sampled during set time periods at a pre-specified set of locations where households in the target group were likely to congregate; and (3) a snowball survey using Nikkei community groups to select the seeds. The objective was to investigate how closely well-designed location and snowball surveys can approach the more expensive conventional household survey method in terms of giving information on the number and characteristics of migrants, the level of remittances received, and the incidence of return migration.

The location survey results were discussed in Section 10.4.3. Here we discuss more fully the snowball component.

A. The snowball survey in Sao Paulo State: methodology

For the snowball survey component, the questionnaire used was the same as the one used for the probability sample based household survey. The survey plan was to begin with a seed list of 75 households, and to aim at reaching a total sample of 300 households through referrals from the initial seed households. Each household surveyed was asked to supply the names of three contacts: (i) a Nikkei household
with a member currently in Japan; (ii) a Nikkei household with a member who had returned from Japan; and (iii) a Nikkei household without any members in Japan or members who had returned from Japan. They were also asked to report the number of other households they knew in each category, which could then be used to weight the sample.

The first step was therefore to select the seed households. One approach commonly followed in snowball surveys is to use ethnic organisations as the source of the seed households. It was decided to use Nikkei associations for the purpose. With the help of a bank which dealt with the Nikkei community, 25 associations throughout the state of Sao Paulo were contacted. The purpose of the survey was explained to each association, and each was asked to supply the names and contact details of three members who could be interviewed for the survey. Twenty of the 25 associations agreed to participate, supplying 67 seed names (several gave more than 3 names). The associations were asked to inform their members about the survey and obtain their consent. However, many of the subsequent participants appeared not to have been informed.

The snowball survey experienced two main problems. The first was that some of the households supplied as seeds by the Nikkei associations refused to answer the survey. The second problem was that among households interviewed, most households did not wish to provide referrals to other Nikkei households. They noted that the length and content of the questionnaire made them reluctant to give the names of friends as potential recruits. In response to these problems, a second phase of the snowballing survey was run. More associations were contacted to provide additional seed names (69 more names were obtained) and, as with the household sample, an adaption of the location survey questionnaire was used when individuals refused to answer the longer questionnaire.

Revising the original sample size target downwards, a decision was made to continue the snowball process until a sample size of 100 households had been achieved. Among them, 75 households received the long survey, and 25 the short survey. Of those receiving the long survey, only 39 per cent provided at least one referral. The mean number of referrals per referral-providing household was 1.5. As a result, an average of only 0.57 referrals per surveyed household was obtained.

The table below provides a summary of the households surveyed using the snowball method. The final sample consists of 60 households who came as seed households from Japanese associations, and 40 households who were chain referrals. The longest chain achieved was 3 links. This means that most of the sample came from the non-probability seed sample; the snowball procedure failed to significantly enlarge or diversify this initial sample.

<table>
<thead>
<tr>
<th>Names on seed list</th>
<th>Interviews on seed list</th>
<th>1st reference</th>
<th>2nd reference</th>
<th>3rd reference</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seed list 1</td>
<td>67</td>
<td>42</td>
<td>19</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Seed list 2</td>
<td>69</td>
<td>18</td>
<td>5</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>136</strong></td>
<td><strong>60</strong></td>
<td><strong>24</strong></td>
<td><strong>9</strong></td>
<td><strong>7</strong></td>
</tr>
</tbody>
</table>

The seed households are drawn from names provided by Nikkei associations, and hence one would expect these households to be more closely connected to Japan
than a randomly chosen Nikkei household. The hope with snowball sampling is that the process of chain-referral will lead to a more representative sample, not as closely connected to Japan. However, as Table 13.3 shows, the snowball seed and referral households turned out to have very similar characteristics. In fact, the only variable where the mean values for the two groups are significantly different is for watching Japanese/Nikkei TV programmes, which was reported to be more common for the referred compared to the seed households. Thus the snowballing does not seem to have succeeded in obtaining households significantly different from the initial seeds.

### Table 13.3. Results from the snowball sample

<table>
<thead>
<tr>
<th>Snowball seeds</th>
<th>Snowball referrals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household Size</td>
<td>3.48</td>
</tr>
<tr>
<td></td>
<td>3.95</td>
</tr>
<tr>
<td>Percentage of households with member who:</td>
<td></td>
</tr>
<tr>
<td>Reads Japanese/Nikkei newspapers</td>
<td>48</td>
</tr>
<tr>
<td>Listens to Japanese/Nikkei radio programs</td>
<td>16</td>
</tr>
<tr>
<td>Watches Japanese/Nikkei TV programs</td>
<td>46</td>
</tr>
<tr>
<td>Reads Japanese/Nikkei books/magazines</td>
<td>52</td>
</tr>
<tr>
<td>Reads newspapers from Nikkei associations</td>
<td>59</td>
</tr>
<tr>
<td>Checks Japanese/Nikkei websites on the internet</td>
<td>21</td>
</tr>
<tr>
<td>% of households which:</td>
<td></td>
</tr>
<tr>
<td>Migration</td>
<td></td>
</tr>
<tr>
<td>Have a Member currently in Japan</td>
<td>28</td>
</tr>
<tr>
<td>Have a Member who has returned from work/study in Japan</td>
<td>49</td>
</tr>
<tr>
<td>Remittances</td>
<td></td>
</tr>
<tr>
<td>Receive remittances from Japan</td>
<td>8</td>
</tr>
<tr>
<td>Refuse to say if they receive remittances</td>
<td>8</td>
</tr>
<tr>
<td>Sample Size</td>
<td>61</td>
</tr>
</tbody>
</table>

**B. Comparison between different survey methods**

In comparing the snowball survey with the location survey component, the following observations were made in Section 10.4.3.

(1) In practice, both the location and the snowball surveys are unlikely to provide a representative sample of the whole population of migrants or migrant-sending families. In particular, they are likely to oversample individuals more closely connected to the community.

While a location-based survey is likely to be more representative than a survey based on snowball sampling, it can be expected that both these surveys oversample individuals who are more connected to Japan and to the Nikkei community in Brazil. This should be the case especially for the seed households in the snowball survey, who are all members of Nikkei associations. Specifically, both the surveys are found to oversample first- and second-generation migrants who are likely to be more strongly connected to Japan, and under-sample third-
and fourth-generation migrants who are likely to be more integrated into Brazil and less likely to attend community events or belong to community associations. Also, both surveys overstate the proportion of households with migrant experience, due to oversampling households with more links to Japan. The above points apply significantly more strongly to the snowball sample. For example, 45 per cent of households in the snowball sample had a member who reads Japanese/Nikkei newspapers, compared to 28 per cent in the time-location survey, and 13 per cent in the household survey. Some more detailed figures were shown in Table 10.7.

(2) The results show that snowball surveys of migrants or their families may be quite ineffective in practice at creating the long referral chains needed for this method to capture the target population.

(3) Furthermore, the snowball method turned out to be only a little cheaper than a representative sampling method.

(4) On the other hand, refusal rates for questions about remittances tended to be higher in the location survey than in the conventional household or the snowball surveys, in which the interviewing took place at a more private location.

(5) Another disadvantage of interviewing in public locations - as is necessary in a location survey, but avoidable in the case of a snowball survey - is that individuals will generally have less time to answer the survey, than during a home visit which the other procedures employ. As a result, the location survey had to use a much shorter questionnaire, thereby collecting less extensive data on the population of interest. By contrast, the snowball survey allowed the use of the same (lengthy) questionnaire as the one used in the probability-based regular household survey.

13.3.3 Mexican Migration Project (MMP)

Durand and Massey (2004) in an essay titled ‘Crossing the Border’, provide the following example of uses and limitations of the data from a snowball sampling based survey of Mexican immigrants interviewed on US side of the border. The survey has been conducted in the context of the Mexican Migration Project (MMP).

“On the U.S. side of the border, the sample of settled out-migrants interviewed in destination areas is representative neither of settled Mexican immigrants nor of settlers from specific sending communities. Samples gathered in the United States are compiled using snowball sampling, ... a serviceable technique imposed by practical constraints but one that ultimately yields data of unknowable representativeness. Although MMP investigators have conducted experiments testing whether representative sampling methods can be applied to select U.S. respondents, their efforts have proved unsuccessful in terms of cost, time, and practicality. Thus, although the MMP’s U.S. interviews provide a snapshot of the characteristics of long-term settlers based on a relatively large number of cases, our data cannot be assumed to be representative and certainly cannot be used to estimate the size of the undocumented Mexican population of the United States.”

“(However,) although MMP data by themselves cannot produce valid estimates of aggregate quantities such as the total number of undocumented
migrants in the United States, the volume of undocumented entries in a given year, or the quantity of “migra-dollars” transmitted to Mexico, they have proved to be extremely useful for characterizing and understanding the social and economic processes that underlie and ultimately produce these aggregate counts ... Data from the MMP can be used to find the value of many important parameters that determine aggregate trends, and they have been used to calculate otherwise unknowable quantities such as the likelihood of undocumented departure from Mexico, the probability of apprehension at the border, the likelihood of return migration, the odds of remigration, the probability of remitting, and the average size of remittances. Such parameter estimates can be combined with published statistics to generate defensible estimates of larger quantities, such as the annual volume of undocumented migration ..., the annual flow of migra-dollars back to Mexico ..., the aggregate effect of remittances on the Mexican economy ..., and the likely size of the future immigrant population ...

“In sum, the MMP data, like all other data, have their strengths and weaknesses and appropriate and inappropriate uses. The point is not that other approaches to data collection should be abandoned in favour of the [snowball] ethno-survey or that the MMP should be used to the exclusion of information from the Bureau of the Census or the Immigration and Naturalization Service but that the data compiled by the MMP using ethno-survey methods provide an important and often crucial complement to standard statistical sources, enabling a clearer interpretation of trends and the more effective use of published statistics”.

13.3.4 A ‘case control study’ using snowball sampling

The investigation, by Lopes, Rodrigues and Sichieri (1996), involves a case-control study in a community in Brazil using the snowball sampling method based on appropriately defined ‘friendship links’ among individuals in the target population.

The objective of the study was to obtain a sufficiently large and diverse sample of a subpopulation subject to high incidence or risk of drug abuse and psychiatric disorder. Individuals in the target population may have been subject to drug abuse (DA) and/or the presence of at least one psychiatric diagnosis (PD). The target population also included individuals who had not actually experienced either of these, but who were closely related to others with DA or PD and hence had a high risk of joining that group.

The basic assumption of the sampling method used was that the target population can be identified and sampled on the basis of suitably defined ‘close friendship’ ties among individuals in the target population.

Boundaries of the target population were somewhat vague, and of course no frame existed for the selection of an even approximately representative sample. Experience had shown that traditional probability sampling procedures – such as those used in a recent high school senior survey and a household survey in the country – had failed to capture reasonable samples of the population of interest. The present study was designed instead to test the feasibility of a snowball sample starting from a very restricted seed sample.
The seed of the snowball sample consisted of only 12 drug abusers, identified with the help of ex-drug abusers, current drug abusers seeking treatment, and councillors located at a drug abuse treatment facility and at a university research centre. Each individual who agreed to co-operate with the study was asked to name (a) a friend who was drug abuser, and (b) a (equally close) friend who had never been involved in the use of drugs. In addition to the existence of friendship ties, non-user referral (b) was also matched by age and gender to the referring individual.

At each wave of the snowball, every person in category (a) – i.e. a referral with drug abuse experience who entered the sample in that wave – was asked to name two friends, one with and another without experience of drug abuse; these formed potential respondents, of category (a) and (b) respectively, for the next wave. By contrast, persons of category (b) were not asked for referrals; their involvement in the survey terminated after the current wave. In either case, only persons not already in the sample in the current or any previous wave were eligible for referral to the next wave. The structure is shown in the diagram below; \((a^{(i)}, b^{(i)})\) refer to the two parts of the sample at wave \(i\).

Figure 13.1. Example of ‘exponential discriminative’ snowball sampling
13.3 Examples of use of snowball sampling in surveys

This structure is what has been termed as “exponential discriminative snowball sampling” shown in Figure 13.1. Linear snowball sampling is when the size of the new sample stays the same from one wave to the next. In the “exponential non-discriminative snowball sampling”, all units in a wave generate multiple referrals; but with “exponential discriminative snowball sampling”, referrals from only particularly categories of units are followed up.

Presumably, the assumption behind choosing this structure of snowballing was that, compared to those in (b), individuals in category (a) would be more likely to name referrals falling within the target population of interest, i.e. within the population with experience or high risk of drug abuse and/or psychiatric problems; and furthermore, that they would be more capable of naming persons from both groups (a) and (b). By contrast, persons in group (b) – those without personal drug abuse experience – may be more ‘open’ in the sense of having close friendship links with individuals outside the population of interest. Referrals from them may, therefore, fall outside the target population of the survey more often.

This referral structure required information on the part of the referring person on drug abuse status of persons being referred to, for the purpose of classification of the latter into the two groups (a) and (b). Such behavioural information concerning drug abuse is likely to be shared among individuals linked through close friendship ties. However, even with such ties, individuals are less likely to be knowledgeable about their friends’ experience of psychiatric disorder. The latter criterion therefore has not been used in selecting referrals in the snowball procedure.

Starting with a seed sample of 12 individuals – all with drug abuse experience – the final achieved sample consisted of 185 pairs of matched individuals (370 persons), each pair consisting of one person with and the other without drug abuse experience, all believed to be members of the target population defined above. Experience of psychiatric disorder among the two parts of the sample was correlated as follows.

### Table 13.4. Illustration of relationship between drug abuse and psychiatric disorder experience

<table>
<thead>
<tr>
<th></th>
<th>Sample (a) – all with DA experience</th>
<th>Sample (b) – all without DA experience</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All (1)</td>
<td>All (2)</td>
</tr>
<tr>
<td></td>
<td>had PD? yes</td>
<td>had PD? no</td>
</tr>
<tr>
<td>had PD? yes</td>
<td>120  65%</td>
<td>96  52%</td>
</tr>
<tr>
<td>had PD? no</td>
<td>65  35%</td>
<td>89  48%</td>
</tr>
<tr>
<td>All</td>
<td>185  100%</td>
<td>185  100%</td>
</tr>
</tbody>
</table>

DA - drug abuse  
PD - psychiatric disorder

Around two-thirds (65 per cent) of persons with drug abuse had experienced psychiatric disorder (PD), the remaining one-third (35 per cent) being without PD (column (1)). Among persons with no drug abuse experience, nearly equal proportions did and did not have psychiatric disorder experience (column (2)). The two variables are closely related.

However, the proportion with a psychiatric disorder turned out to be close to one-half (48-52 per cent) among non-drug abusers in group (b), irrespective of the psychiatric disorder experience of individuals with drug abuse experience in group (a) who named
the former for inclusion in the sample (columns (3) and (4)). The neglect of psychiatric disorder as a criterion in selecting referrals for the snowball sampling procedure therefore seems justified. This “absence of bias” as the report’s authors term it, is certainly a convenient feature of the situation because of the greater complexity of the information required for assessing the psychiatric disorder experience of individuals, compared to their drug abuse status.

### 13.4 Simulation 1: estimating a proportion in the population

In normal practical situations, snowball sampling has to start from a restricted and unrepresentative sample. The objective (and hope) of snowball sampling is to obtain a less restricted (larger, spread out to different parts of the population) and a more representative sample as the number of snowball sampling waves increases. Whether, and the speed with which, these objectives are achieved depends on the nature of the links existing among members of the target population, and on how these links are realised and exploited by the survey procedures chosen. The type of links may differ greatly from one population subgroup to another. The surveyor’s skill lies in devising sampling and field procedures which can identify and control those differences towards the objective of obtaining a more representative sample of the target population. Apart from inherent differences among different population groups, there are also factors which are more under the control of the statistician, such as the number of waves of snowball sampling and the number of referrals per unit allowed at each wave of sampling.

In this and the next sections we develop a number of numerical illustrations in order to give a feel of how snowball sampling may work in practice.

#### 13.4.1 Assumed model for the illustrations

The first set of illustrations is concerned with constructing a snowball sample with the objective of estimating a proportion in the population. These illustrations assume the following simple statistical model about the target population.

The population consists of two groups, identified by subscripts $a$ and $b$. The objective is to estimate $P_a$, the proportion belonging to group $a$. The target population may, for instance, be children working in some specified hazardous or otherwise difficult to locate and survey occupations or working conditions. Among these working children, group $a$ may refer to those who are subject to exceptionally harsh working conditions, difficulties, dangers, abuses, etc., at work. No sampling frame is available for selecting a reasonably representative sample of the target population for estimating proportion $P_a$. However, children in the target population are assumed to have social relations (links) among themselves. The objective of snowball sampling is to make use of those links in order to obtain a more representative sample starting from a very restricted initial (‘seed’) sample.

Suppose that an initial sample of size $n^{(0)}$ is selected from lists or similar sources available for the purpose. An investigation is conducted to identify the two subgroups
a and b in the sample: \( n^{(0)} = n^{(0)}_a + n^{(0)}_b \). The proportion \( p^{(0)}_a = n^{(0)}_a / n^{(0)} \) in this initial sample may be quite different from the actual proportion \( P_a \) in the target population which is to be estimated.

Each person in the initial sample \( n^{(0)}_a \) is asked to name \( x_a \) eligible persons not already in the sample. Similarly, each person in the initial sample \( n^{(0)}_b \) is asked to name \( x_b \) eligible persons not already in the sample. By ‘eligible’ is meant persons belonging to the target population, for instance children working in the specified sectors or circumstances.

The naming for inclusion in the snowball sample is based on friendship or other links among the children, and it is very likely that such links are stronger or more numerous among children with similar characteristics or children in similar circumstances, than among children with different characteristics or circumstances. In other words, children may have a preference to name children from their own group. The strength of such preference may differ from one group to another; that is, children in different groups may be differently disposed to naming children from outside their own group.

In these illustrations, we will assume that the following simple model reflects how the sample expands with waves of the snowball sampling.

\[
\begin{align*}
    n^{(i+1)}_a &= a n^{(i+1)}_a + b n^{(i+1)}_a \\
    a n^{(i+1)}_a &= x_a n^{(i)}_a (f_a + (1 - f_a) P_a) \\
    b n^{(i+1)}_a &= x_b n^{(i)}_b (1 - f_b) P_a
\end{align*}
\]

(13.2a) (13.2b) (13.2c)

where \( i = 0, 1, 2 \ldots \) refers to waves of snowballing.

Quantity \( a n^{(i+1)}_a \) is the number of sample cases in group \( a \) in wave \( (i+1) \) coming from individuals in the same group (a) in the preceding wave(i)

Quantity \( b n^{(i+1)}_a \) is the number of sample cases in group \( a \) coming from the different group (b) in the preceding wave; \( x_a \) is the average number (assumed fixed) of new cases brought into the sample through referral by a group \( a \) individual in the preceding wave.

Quantity \( x_b \) is the similarly defined quantity for group \( b \). By ‘new’ is meant cases not already included in the current or any preceding wave.

The basic assumption of the model is that a certain proportion \( f_a \) of referrals made by group \( a \) persons are automatically within the same group (a), while the remainder proportion \( 1-f_a \) are made at random from each group (a and b) in proportion to the size of the group in the population (the relative sizes being \( P_a \) and \( 1-P_a \) for the two groups, respectively).

Thus \( n^{(i)}_a \) persons at wave \( i \) make \( x_a n^{(i)}_a \) referrals for the next wave; of these a proportion \( f_a \) are automatically to group (a) at wave \( (i+1) \), while among the remaining, proportion \( (1-f_a) P_a \) are assigned at random to the same group, and proportion \( (1-f_a) (1-P_a) \) are to group b.\[53\]

\[53\] The last-mentioned quantity appears in (13.3c) giving the composition of \( n^{(i+1)}_b \), while its complement appears in (13.2c) giving the composition of \( n^{(i+1)}_a \).
Similar breakdown applies to the $n_b^{(i)}$ persons in group $b$ at wave $i$. Equations similar to (13.2) apply to the expansion of group $b$ over waves of the sample:

\begin{align}
    n_b^{(i+1)} &= b n_b^{(i+1)} + a n_b^{(i+1)} \\
    b n_b^{(i+1)} &= x_b n_b^{(i)} (f_b + (1 - f_b)(1 - P_a)) \\
    a n_b^{(i+1)} &= x_a n_a^{(i)} (1 - f_a)(1 - P_a)
\end{align}

The proportion of sample cases in group $a$ at wave $(i+1)$ is

$$m_a^{(i+1)} = n_a^{(i+1)}/(n_a^{(i+1)} + n_b^{(i+1)}).$$

And the cumulative proportion for all waves up to $(i+1)$ is

$$p_a^{(i+1)} = \frac{\sum_{j=0}^{i+1} n_a^{(j)}}{\sum_{j=0}^{i+1} (n_a^{(j)} + n_b^{(j)})}.$$

The point of interest is how close $p_a$ becomes to $P_a$ with increasing $i$ (increasing waves of the snowball sampling).

### 13.4.2 Case (1): snowball sampling assuming identical structure of referrals

This illustration is based on the assumption of identical structure of referrals in the two groups, $a$ and $b$. By this we mean that parameters $x$ and $f$ are the same for the two groups. Parameter $x$ determines how rapidly the sample is allowed to expand. This is primarily a design decision under the control of the researcher, in so far as the population is large enough and co-operative enough for a chosen value of $x$ to be maintained over waves of the snowball sampling. (In practice, it often happens that small or non-cooperating populations get progressively ‘exhausted’, so that the desired number of new referrals cannot be maintained.)

Parameter $f$ is a measure of the tendency of individuals in the sample to name individuals from their own group for the following wave of the snowball sample. This parameter is a characteristic of the given population group. Approximate information on the parameter may be available from previous experience or pilot studies; also, the value of the parameter may be modified to a limited extent through careful design and implementation of survey procedures. However, generally this parameter is not known to, nor can be much manipulated by the survey researcher.

Table 13.5 applies the model described above to nine snowball waves following the initial sample.

In Table 13.5A, we take a population with $P_a = 0.25$, i.e. group $a$ comprises 25 per cent of the total population. *This parameter is of course not known and the objective of the exercise is to estimate it from the snowball sample.* Not knowing this proportion, let us assume that we take an equal number of cases $n_a^{(0)} = n_b^{(0)}=10$ from the two groups for the initial or base sample, i.e. group $a$ forms proportion $p_a^{(0)}=0.50$ of the total sample.
Taking fixed and equal values \( f_a = f_b = 0.50 \), the table shows how the snowball sample would develop with different (but equal) values of \((x_a, x_b)\), varying from 2.00 to 0.50. Large values of \( x \) mean that the sample size expands rapidly with snowball waves. For example, with \((x_a = x_b = 1)\), i.e. on the average with one referral per case in the sample, the total sample size expands by 6 times (120/20) after five waves, and by 11 times (220/20) after 10 waves. With \((x_a = x_b = 2)\), the expansion is by nearly 5,000 times after 10 waves! (This last one is of course not a realistic scenario; such large expansion cannot be sustained over so many waves.)

| Table 13.5. Snowball sampling under the assumption of an identical structure of referrals by different population groups |
| A. For different rates \((x)\) of spread of the sample |
| Panel (1) | Panel (2) | Panel (3) |
| Class \(a\) | Class \(b\) | Class \(a\) | Class \(b\) | Class \(a\) | Class \(b\) |
| \(p\) | 0.25 | 0.75 | 0.25 | 0.75 | 0.25 | 0.75 |
| \(n\) | 10 | 10 | 10 | 10 | 10 | 10 |
| \(x\) | 2.00 | 2.00 | 1.00 | 1.00 | 0.50 | 0.50 |
| \(f\) | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |

Composition of wave \(i\):

| \(i=0\) | 10 | 10 | 0.50 | 10 | 10 | 0.50 | 10 | 10 | 0.50 |
| \(i=1\) | 15 | 25 | 0.38 | 8 | 13 | 0.38 | 4 | 6 | 0.38 |
| \(i=2\) | 25 | 55 | 0.31 | 6 | 14 | 0.31 | 2 | 3 | 0.31 |
| \(i=3\) | 45 | 115 | 0.28 | 6 | 14 | 0.28 | 1 | 2 | 0.28 |
| \(i=4\) | 85 | 235 | 0.27 | 5 | 15 | 0.27 | 0 | 1 | 0.27 |
| \(i=5\) | 165 | 475 | 0.26 | 5 | 15 | 0.26 | 0 | 0 | 0.26 |
| \(i=10\) | 5,125 | 15,355 | 0.25 | 5 | 15 | 0.25 | 0 | 0 | 0.25 |

Cumulative composition after \(i\) waves of snowball sampling:

| \(p_a\) | \(p_b\) | \(p_a\) |
| 0 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| 5 | 0.27 | 0.30 | 0.33 | 0.30 | 0.30 |
| 10 | 0.25 | 0.30 | 0.33 | 0.30 | 0.30 |

"Composition of wave \(i\)" shows the development of the sample size for groups \(a\) and \(b\).

- \(m_a\) - proportion of the sample population in class \(a\), at a certain wave
- \(p_a\) - proportion of the sample population in class \(a\), cumulated up to a certain wave

In any case, with the chosen value of \(f\) in the illustration, the structure of the additional sample in each wave - measured by the share of class \(a\) in the total, namely parameter \(m_a\) - rapidly converges to the population value \(P_a = 0.25\) for the whole range of \(x\) values shown. However, with small \(x\), say \(x < 1\), there is little expansion of the sample after a few waves and the effect of imbalance in the initial sample tends to linger on in the total (cumulative) sample. For example, even after 10 waves \(p_a^{(10)}\) does not fall below 0.30 for \(x = 1\) and below 0.42 for \(x = 0.5\), compared to the starting value \(p_a^{(0)} = 0.50\) and the actual value \(P_a = 0.25\).
As a concluding remark we may note that in the illustrations, \( x \) refers to the average value of referrals per case in the sample. Fractional average values can occur due to non-response or the respondents’ inability to find the required number of referrals. They can also be achieved by design, for example by assigning values \( \text{int}(x) \) and \( \text{int}(x) + 1 \) at random to cases in appropriate proportion to achieve the required average value of \( x \). The respective proportions will in fact be \( (\text{int}(x) + 1 - x) \) and \( (x - \text{int}(x)) \).

In Table 13.5B, we take a constant value \( x_a = x_b = 1.50 \) throughout, and examine the effect of varying \( f \), shown here in the range \( f_a = f_b = 0.8 \) to 0.2. Large values of \( f \) mean that individuals in the group concerned have a strong tendency to name individuals from their own group for the next wave. This means that any imbalance in the initial sample tends to be more persistent. For example, while with \( f = 0.2 \) proportion \( p_a \) become 0.27 from a starting value of 0.50 after 5 waves, and becomes 0.25 after 10 waves, with \( f = 0.8 \) on the other hand, the corresponding values remain high more persistently, the values being 0.37 after 5 and 0.30 after 10 waves, compared to 0.25 for the population. Note that with fixed value of \( x \), the total sample size at any wave is the same for different \( f \) values; only its proportions in classes \( a \) and \( b \) are different.

<table>
<thead>
<tr>
<th>Table 13.5 (cont.)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B. For different degrees of preference (( f )) for referral within own group</strong></td>
</tr>
<tr>
<td>Panel (1)</td>
</tr>
<tr>
<td>Class ( a )</td>
</tr>
<tr>
<td>( p )</td>
</tr>
<tr>
<td>( n )</td>
</tr>
<tr>
<td>( x )</td>
</tr>
<tr>
<td>( f )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Composition of wave ( i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m_a )</td>
</tr>
<tr>
<td>( i=0 )</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cumulative composition after ( i ) waves of snowball sampling:</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_a )</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>

Table 13.5C still assumes an identical structure of referral in the two groups \( a \) and \( b \), but examines the effect of different extents of imbalance in the original (seed) sample. In other words, though the unknown \( P_a \) value is 0.25, results are shown with the initial samples departing from this by various amounts, with \( P_a^{(0)} \) values taken as 0.20, 0.50 and 0.80. As before, we have defined \( P_a^{(0)} = p_a^{(0)} = n_a^{(0)}/(n_a^{(0)} + n_b^{(0)}) \).
Parameter \( f \) is taken as 0.5 throughout, and parameter \( x \), the average number of referrals per case in each wave of snowballing, is taken as a constant, \( =1.5 \). With this fairly large value of \( x \), the effect of the initial imbalance in the sample is rapidly corrected: in the table, \( p_a^{(5)} \) (sample proportion of class a after 5 waves) is in the range 0.25-0.35, and \( p_a^{(10)} \) is practically identical to the actual value \( p_a = 0.25 \) in all cases.

This remarkable result is, of course, a consequence of the assumed simple and identical structure of referrals in the two groups. In reality, when dealing with heterogeneous and reclusive populations, often the situation is more complex.

### Table 13.5 (cont.)

C. For different extents of imbalance in the seed sample: moderately large number of referrals per case (\( x=1.5 \))

<table>
<thead>
<tr>
<th></th>
<th>Panel (1)</th>
<th>Panel (2)</th>
<th>Panel (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class a</td>
<td>Class b</td>
<td>Class a</td>
</tr>
<tr>
<td>( P )</td>
<td>0.25</td>
<td>0.75</td>
<td>0.25</td>
</tr>
<tr>
<td>( n )</td>
<td>4</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>( x )</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
</tr>
<tr>
<td>( f )</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
</tr>
</tbody>
</table>

**Composition of wave \( i \)**

<table>
<thead>
<tr>
<th></th>
<th>( m_a )</th>
<th>( m_b )</th>
<th>( m_a )</th>
<th>( m_b )</th>
<th>( m_a )</th>
<th>( m_b )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( i=0 )</td>
<td>4</td>
<td>16</td>
<td>10</td>
<td>10</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>23</td>
<td>11</td>
<td>19</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>34</td>
<td>14</td>
<td>31</td>
<td>17</td>
<td>28</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>51</td>
<td>19</td>
<td>49</td>
<td>22</td>
<td>46</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>76</td>
<td>27</td>
<td>74</td>
<td>29</td>
<td>72</td>
</tr>
<tr>
<td>5</td>
<td>38</td>
<td>114</td>
<td>39</td>
<td>113</td>
<td>41</td>
<td>111</td>
</tr>
<tr>
<td>10</td>
<td>288</td>
<td>865</td>
<td>289</td>
<td>865</td>
<td>289</td>
<td>864</td>
</tr>
</tbody>
</table>

**Cumulative composition after \( i \) waves of snowball sampling:**

<table>
<thead>
<tr>
<th></th>
<th>( p_a )</th>
<th>( p_b )</th>
<th>( p_a )</th>
<th>( p_b )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( i=0 )</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>101</td>
<td>315</td>
<td>120</td>
<td>295</td>
</tr>
<tr>
<td>10</td>
<td>851</td>
<td>2,569</td>
<td>874</td>
<td>2,546</td>
</tr>
</tbody>
</table>

Even with these favourable assumptions, the effect of imbalances in the seed sample persists if the expansion of the sample is very restricted, i.e. if the number of referrals per unit in the sample (\( x \)) is small – e.g. for \( x = 0.5 \) as shown in Table 13.5D. Such a small value of \( x \) means that the sample does not expand sufficiently in size to be able to compensate for the initial imbalance. In fact with \( x \) as low as 0.5, there is practically no expansion of the sample after 3 or 4 waves. With \( p_a^{(3)} = 0.20 \) for example, the \( p_a \) value is seen to quickly stabilise at \( p_a = 0.22 < 0.25 \); with \( p_a^{(10)} = 0.80 \) it persists in being very high, at \( p_a = 0.62 > 0.25 \). This follows from the fact that most of the sample comes from earlier waves where, in most of our illustrations, group a is overrepresented.
In summary, the following are the main conclusions from this illustration, based on the assumption of identical referral patterns in the two groups.

(1) Moderate to rapid expansion \((X \geq 1)\) reduces the effect of imbalance in the initial sample. But contraction \((X < 1)\) makes any initial imbalance persist; in this case the sample is not self-correcting.

(2) Initial imbalance tends to persist when parameter \(f\) is large, i.e. respondents have a strong tendency to refer (name) persons from their own group for the next wave.

### 13.4.3 Case (2): snowball sampling with differing structure of referrals

The following illustrations reflect more realistic conditions. Different groups and individuals can have quite a different structure of referrals.

Consider for instance a population of children engaged in some hazardous work. Suppose they are divided into two groups. Children in subpopulation \(a\) (say a group subject to particularly harsh working conditions) have closer ties with each other, and hence have a greater tendency to provide names for snowballing from within their own group, compared to children in subgroup \(b\) who work in less harsh conditions. The former group has a larger value of parameter \(f_a\), compared to the value of parameter \(f_b\) for the latter group.
It is possible for the situation to be just the opposite. Children in harsher conditions (group \(a\)) may be more open to children in either group, compared to children working in less harsh conditions (group \(b\)) who may have a tendency to avoid the more unfortunate children in the former group. In this case \(f_a < f_b\).

Some results from simulation of this situation are shown in the set of tables 13.6A-C.

In Table 13.6A, the effect of different combinations of \(f\) values is examined, ranging from \((f_a = 0.25, f_b = 0.75)\), to the reverse situation with \((f_a = 0.75, f_b = 0.25)\). The initial sample size \((n_a^{(0)} = n_b^{(0)} = 10)\) and the average number of referrals per case in each wave \((x_a = x_b = 1.5)\) have been kept constant throughout. As before, in the seed sample, \(p_a^{(0)} = 0.50 > p_a = 0.25\).

<table>
<thead>
<tr>
<th>A. Different values of parameters (f_a) and (f_b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel (1)</td>
</tr>
<tr>
<td>(P)</td>
</tr>
<tr>
<td>0.25</td>
</tr>
<tr>
<td>0.75</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Composition of wave \(i\)

<table>
<thead>
<tr>
<th>(m_i)</th>
<th>(m_i)</th>
<th>(m_i)</th>
<th>(m_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i=0)</td>
<td>10</td>
<td>10</td>
<td>0.50</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>23</td>
<td>0.25</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>38</td>
<td>0.16</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>59</td>
<td>0.12</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>90</td>
<td>0.11</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>136</td>
<td>0.10</td>
</tr>
<tr>
<td>10</td>
<td>115</td>
<td>1,038</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Cumulative composition after \(i\) waves of snowball sampling:

<table>
<thead>
<tr>
<th>(p_i)</th>
<th>(p_i)</th>
<th>(p_i)</th>
<th>(p_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
<td>10</td>
<td>0.50</td>
</tr>
<tr>
<td>5</td>
<td>59</td>
<td>356</td>
<td>0.14</td>
</tr>
<tr>
<td>10</td>
<td>360</td>
<td>3,060</td>
<td>0.11</td>
</tr>
</tbody>
</table>

In this situation, when \(f_a < f_b\), expansion of the sample is slower in group \(a\) and faster in group \(b\). This is because, in relative terms, individuals in group \(a\) are more likely to name (bring into the snowball sample) persons from group \(b\), than the other way round.

In the illustration, the initial imbalance in the sample, \(p_a^{(0)} > p_a\), is overcome quickly and is then overshot, resulting in \(p_a^{(i)} < p_a\) after a few waves. For example, with \((f_a = 0.25, f_b = 0.75)\), \(p_a\) stabilises at a very low value of 0.11; with \((f_a = 0.40, f_b = 0.60)\) it stabilises at 0.19 – still below the actual value \(P_a = 0.25\).

The situation is reversed with \(f_a > f_b\). The imbalance \(p_a^{(0)} > p_a\) in the initial sample tends to persist. For example with \((f_a = 0.60, f_b = 0.40)\), \(p_a\) does not fall below 0.34 even after 10 waves of the snowball sample. Case \((f_a = 0.75, f_b = 0.25)\) is an ‘extreme case’...
in the sense that \( P_a \) stays unchanged at the initial value \( p_a^{(0)}=0.50 \), grossly above the actual value \( P_a =0.25 \).

With an even greater difference in favour of \( f_a \), \( f_a > f_b \), snowballing would increase the imbalance in the initial sample which began with an over-representation of group \( a \) (\( p_a^{(0)} > p_a \)).

The general pattern is that, under the assumed model, while snowballing usually helps to reduce bias in the initial sample, the correction may overshoot, or alternatively the correction may be only partial, depending on values of the parameters involved. In extreme cases, the imbalance can even increase.

<table>
<thead>
<tr>
<th>Panel (1)</th>
<th>Panel (2)</th>
<th>Panel (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class a</td>
<td>Class b</td>
<td>Class a</td>
</tr>
<tr>
<td>( P )</td>
<td>0.25</td>
<td>0.75</td>
</tr>
<tr>
<td>( n )</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>( x )</td>
<td>2.00</td>
<td>1.50</td>
</tr>
<tr>
<td>( f )</td>
<td>0.50</td>
<td>0.50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Composition of wave</th>
<th>( m_a )</th>
<th>( m_b )</th>
<th>( m_a )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( i=0 )</td>
<td>10</td>
<td>10</td>
<td>0.50</td>
</tr>
<tr>
<td>1</td>
<td>14</td>
<td>16</td>
<td>0.46</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>25</td>
<td>0.44</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>36</td>
<td>0.43</td>
</tr>
<tr>
<td>4</td>
<td>38</td>
<td>52</td>
<td>0.43</td>
</tr>
<tr>
<td>5</td>
<td>54</td>
<td>74</td>
<td>0.42</td>
</tr>
<tr>
<td>10</td>
<td>315</td>
<td>432</td>
<td>0.42</td>
</tr>
</tbody>
</table>

<p>| Cumulative composition after ( i ) waves of snowball sampling: |</p>
<table>
<thead>
<tr>
<th>( p_a )</th>
<th>( p_b )</th>
<th>( p_a )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>163</td>
<td>212</td>
</tr>
<tr>
<td>10</td>
<td>1,043</td>
<td>1,418</td>
</tr>
</tbody>
</table>

Table 13.6B illustrates the effect of difference among groups in the number of referrals (\( x_a, x_b \)) per respondent allowed each wave. The other parameters are taken as constants as before: \( P_a =0.25, \ p_a^{(0)} =0.50, \ f_a =f_b =0.50 \). Results are shown for (\( x_a=2.0, \ x_b=1.0 \)) versus (\( x_a=1.0, \ x_b=2.0 \)).

With (\( x_a>x_b \)), sample size for group \( a \) is permitted to expand at a faster rate than the sample for group \( b \), so that the initial imbalance in the composition (\( p_a^{(0)}>p_a \)) persists. For example, with (\( x_a=2.0, \ x_b=1.0 \)), \( p_a^{(i)} \) stabilises at 0.42, well above the population value \( P_a =0.25 \).

The reverse is the situation with (\( x_a<x_b \)). Sample size for group \( a \) expands too slowly, so that the initial imbalance in the composition (\( p_a^{(0)}>p_a \)) is quickly overcome and
then reversed with waves of the snowball. For instance, with \((x_a=1.0, x_b=2.0)\), \(P_a^{(i)}\) quickly stabilises at 0.17, below the population value \(P_a=0.25\).

### 13.4.4 An important practical implication

We saw from Table 13.6B that with \((x_a>x_b)\), the sample for group \(a\) expands with snowballing more rapidly than that for group \(b\), and the sample proportion \(p_a\) increases with waves of the sample, and may even end up overestimating the population value \(P_a\). Conversely, with \((x_a<x_b)\), the sample for group \(a\) is not allowed to expand as much as that for group \(b\), and \(p_a\) falls with waves of the sample, and may even end up underestimating \(P_a\).

#### Table 13.6 (cont.)

<table>
<thead>
<tr>
<th>Panel (1)</th>
<th>Panel (2)</th>
<th>Panel (3)</th>
<th>Panel (4)</th>
<th>Panel (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(P)</td>
<td>(x)</td>
<td>(f)</td>
<td>(x*f)</td>
<td>(P)</td>
</tr>
<tr>
<td>0.25</td>
<td>2.10</td>
<td>0.30</td>
<td>0.63</td>
<td>0.25</td>
</tr>
<tr>
<td>0.75</td>
<td>0.90</td>
<td>0.70</td>
<td>0.63</td>
<td>0.75</td>
</tr>
</tbody>
</table>

#### Composition of wave \(i\)

<table>
<thead>
<tr>
<th>(i)</th>
<th>(m_a)</th>
<th>(m_b)</th>
<th>(m_a)</th>
<th>(m_b)</th>
<th>(m_a)</th>
<th>(m_b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
<td>10.50</td>
<td>10</td>
<td>10</td>
<td>10.50</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>11.36</td>
<td>11</td>
<td>11</td>
<td>11.36</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>12.93</td>
<td>13</td>
<td>12.39</td>
<td>13</td>
<td>12.39</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>14.27</td>
<td>17</td>
<td>16.42</td>
<td>17</td>
<td>16.42</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>16.26</td>
<td>22</td>
<td>17.59</td>
<td>22</td>
<td>17.59</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>19.26</td>
<td>28</td>
<td>20.82</td>
<td>28</td>
<td>20.82</td>
</tr>
<tr>
<td>10</td>
<td>47</td>
<td>47.25</td>
<td>125</td>
<td>125</td>
<td>125</td>
<td>125</td>
</tr>
</tbody>
</table>

Cumulative composition after \(i\) waves of snowball sampling:

<table>
<thead>
<tr>
<th>(i)</th>
<th>(\rho_a)</th>
<th>(\rho_b)</th>
<th>(\rho_a)</th>
<th>(\rho_b)</th>
<th>(\rho_a)</th>
<th>(\rho_b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
<td>10.50</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>82</td>
<td>195.30</td>
<td>101</td>
<td>242.29</td>
<td>120</td>
<td>295.29</td>
</tr>
<tr>
<td>10</td>
<td>252</td>
<td>701.26</td>
<td>476</td>
<td>1,357.26</td>
<td>874</td>
<td>2,546.26</td>
</tr>
</tbody>
</table>

Similar is the situation concerning \((f_a, f_b)\). We saw from Table 13.6A that with \(f_a>f_b\), a higher proportion of referrals from group \(a\) remain within that group, compared to the situation in group \(b\). Consequently, the sample for group \(a\) tends to expand more rapidly, resulting in increasing \(p_a\) with waves of the sample, which may even end up overestimating the population value \(P_a\).

With \(f_a<f_b\), the situation is reversed.
We can expect $p_a$ to better reflect $P_a$ – i.e. expect the sample bias to be reduced, reducing the imbalance we began with in the seed sample – if some information is available on the relative f-values so that values of parameter $x$ can be varied inversely to those as a compensation.

In the illustration of Table 13.6C, we take $x \propto (1/f)$, i.e. $x_a f_a = x_b f_b$ for a wide range of values of the parameters. In all examples, bias in the seed sample disappears with waves of the snowball, and the sample proportions $m_a$ (or $p_a$) closely approximate the population value $P_a$.

It is interesting to derive the results of Tables 13.6(A)-(C) analytically (algebraically) under the assumed model. Below we derive an expression for the stable value to which the sample proportion $m_a$ (or $p_a$) tends with increasing waves of the snowball, as a function of parameter values $(x_a/x_b, f_b/f_a)$. Table 13.6C is a special case with $(x_a/x_b = f_b/f_a)$, resulting in the sample estimate $m_a$ tending to the true population proportion $P_a$.

### 13.4.5 Algebraic derivation of the procedure

Defining

$$k = \frac{n_a^{(i)}}{n_b^{(i)}} = \frac{n_a^{(i+1)}}{n_b^{(i+1)}}$$

(13.4)

for some sufficiently large value of snowball wave $i$ as the required limiting value of the relative sample sizes of the two groups, it can be easily derived that $k$ is the positive root of the following equation:

$$ak^2 + bk + c = 0$$

(13.5)

with

$$a = g(1 - f_a)(1 - P_a)$$

$$b = 1 - P_a(1 + g) - gf_a(1 - P_a) + P_a f_b$$

$$c = -((1 - f_b)P_a$$

$$g = (x_a/x_b).$$

This gives the limiting value, for sufficiently large $i$, as

$$m_a = \frac{n_a^{(i)}}{n_a^{(i)} + n_b^{(i)}} = \frac{k}{k+1}$$

(13.6)

Value $P_a$, the overall proportion in the sample cumulated over $i$ waves, is generally close to the above value $m_a$ if the sample is allowed to expand sufficiently over waves of the snowball.

Result (13.5) follows from the observation that, using equations (13.2) and (13.3), equation (13.4) can be written as:

$$k = \frac{k x_a}{x_b} [f_a + (1 - f_a)P_a] + (1 - f_b) P_a$$

$$[f_b + (1 - f_b)(1 - P_a) + k \left( \frac{x_a}{x_b} \right) (1 - f_a)(1 - P_a)]$$

(13.7)
Table 13.7. Verification of the numerical results

<table>
<thead>
<tr>
<th>Panel of Table 13.6A</th>
<th>Panel of Table 13.6B</th>
<th>Panel of Table 13.6C</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>( f_a )</td>
<td>0.25</td>
<td>0.40</td>
</tr>
<tr>
<td>( f_b )</td>
<td>0.75</td>
<td>0.60</td>
</tr>
<tr>
<td>( f_a / f_b )</td>
<td>3.00</td>
<td>1.50</td>
</tr>
<tr>
<td>( g = x_a / x_b )</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel of Table 13.6A</th>
<th>Panel of Table 13.6B</th>
<th>Panel of Table 13.6C</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>( a )</td>
<td>2.25</td>
<td>1.80</td>
</tr>
<tr>
<td>( b )</td>
<td>2.00</td>
<td>1.40</td>
</tr>
<tr>
<td>( c )</td>
<td>-0.25</td>
<td>-0.40</td>
</tr>
<tr>
<td>( k )</td>
<td>0.11</td>
<td>0.22</td>
</tr>
<tr>
<td>( k / (1+k) )</td>
<td>0.10</td>
<td>0.18</td>
</tr>
<tr>
<td>(*) ( m_a )</td>
<td>0.10</td>
<td>0.18</td>
</tr>
</tbody>
</table>

(*) As computed in the numerical illustrations in Tables 113.6A-C.

In our numerical illustrations, \( P_a = 0.25 \) and the above equations take the following simplified form (since only the relative magnitudes of the parameters matter, the values below have all been multiplied by 4):

\[
\begin{align*}
  a &= 3g(1 - f_a) \\
  b &= (3 - g) - 3gf_a + f_b \\
  c &= -(1 - f_b) \\
  g &= (x_a / x_b).
\end{align*}
\]

In Table 13.6A we had \( g = 1 \) throughout, and the left hand part of Table 13.7 shows the limiting values of \( m_a \) for different combinations of \( (f_a, f_b) \) computed using the above formulae. These are identical to the numerically obtained values in the detailed Table 13.6A in its corresponding panels (1)-(5), and shown in the last row of Table 13.7.

Similar results are presented for Table 13.6B in Table 13.7. Here we had \( f_a = f_b = 0.50 \) throughout, and three different values of \( g = x_a / x_b \) (2.0, 1.0, 0.5) for the three panels (1)-(3).

Results for different combinations of \( (x_a f_a = x_b f_b) \) of Table 13.6C appear in right hand part of Table 13.7, which are the same as the numerically obtained results in panel (1)-(5) of Table 13.6C.

In all cases, the values of \( m_a \) from Equations (13.5)-(13.6) are identical to those from the simulations in Table 13.6. And when \( x_a f_a = x_b f_b \), \( m_a \) always equals the population value \( P_a \).

The observation that - under our assumed model of the structure of snowball referrals – with \( (x_a f_a = x_b f_b) \) the sample proportion \( m_a \) (or \( P_a \)) converges to the population value \( P_a \), is an important one. Below we note its algebraic verification in fuller detail.

After some manipulation, it can be seen that the above equations with \( g = x_a / x_b = f_b / f_a \) (and given \( P_a = 0.25 \)) become:

\[
\begin{align*}
  a &= 3g(1 - f_a); \\
  b &= (3 - g) - 2gf_a; \\
  c &= -(1 - gf_a),
\end{align*}
\]

giving
and hence
\[ k = \frac{-b + \sqrt{b^2 - 4ac}}{2a} = \frac{-(3-g) + 2gf_a + g(1-f_a) + 3}{6g(1-f_a)} = \frac{2g(1-f_a)}{6g(1-f_a)} = \frac{1}{3} \]
irrespective of the particular \((x_a f_a = x_b f_b)\) value.

Finally, the limiting sample proportion after a sufficient number of snowball waves is
\[ m_a = \frac{n_a^{(i)}}{n_a^{(i)} + n_b^{(i)}} = \frac{k}{k+1} = \frac{1}{4} = P_a. \]

In practice, parameter \(f\) is determined by the nature of the existing referral links between individuals, and cannot be freely chosen by the survey investigator. Perhaps with appropriate selection of the methodology and careful design and implementation of survey procedures, the \(f\) values can be influenced to a limited extent. Variations in \(x\) are much more likely to be under the control of the survey investigator, unless the target population is small and the required numbers of new referrals become increasing difficult (finally impossible) to find with increasing number of survey waves.

The problem still remains that the \(f\) values for population groups of interest may not be known a priori. Previous experience and small-scale pilot surveys may provide useful estimates.

The first wave snowball sample resulting from referrals from the initial sample cases can be decomposed into four classes according to the group \(a\) or \(b\) from where the referral came, and the group \(a\) or \(b\) to which the referred individual belongs.

<table>
<thead>
<tr>
<th>Referring persons at wave 0</th>
<th>Referred persons at wave 1:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class (a)</td>
<td>Class (b)</td>
</tr>
<tr>
<td>Class (a) (n_a^{(0)})</td>
<td>(a n_a^{(1)} = n_a \times \frac{x_a + (1 - f_a)P_a}{x_a})</td>
<td>(a n_b^{(1)} = n_a \times \frac{x_b + (1 - f_b)(1 - P_a)}{x_b})</td>
</tr>
<tr>
<td>Class (b) (n_b^{(0)})</td>
<td>(b n_a^{(1)} = n_b \times \frac{x_a + (1 - f_b)P_a}{x_b})</td>
<td>(b n_b^{(1)} = n_b \times \frac{x_b + (1 - f_b)(1 - P_a)}{x_b})</td>
</tr>
</tbody>
</table>

The above expressions are the same as those given in Equations (13.2) and (13.3), except that here we allow for the possibility of the sample expansion factors \((x_a, x_b)\) being wave-dependent. As before, symbol \(y n_x^{(i+1)}\) means the number of sample cases referred from group \(y\) at wave \(i\) to group \(x\) at wave \((i+1)\), and \(n_x^{(i+1)}\) is the total sample size of group \(x\) at wave \((i+1)\). With the following quantities defined from the results obtained at wave 1
\[ r = \frac{b n_b^{(1)}}{b n_a^{(1)}}, \quad s = \frac{a n_a^{(1)}}{a n_b^{(1)}} \]

it can be shown that
13.5 Simulation 2: spreading the sample to less accessible strata

13.5.1 Conditions, objectives, and options

The objective of the following is to numerically illustrate some aspects of a commonly encountered situation when applying the snowball sampling procedure. The situation being considered has the following characteristics.

(1) The target population is heterogeneous in terms of its visibility and accessibility to the sampling process. Suppose the population consists of children working in a set of hazardous activities. Some of the activities may be more hazardous, threatening, shameful, illegal, or hidden, while some others may be less so. Also suppose that, on this basis, we can divide the population into strata in accordance with the degree of hazard of the child workers face – say from the least hazardous to the most hazardous. The least hazardous may be the most visible activities, in the sense that a frame exists for the selection of a sample of children engaged in them and also that the children involved are relatively accessible for the survey. The most hazardous activities may be hidden, with no sampling frame or at best only a very incomplete sampling frame available.

(2) An initial sample of child workers is taken from individual strata. The initial sample may include cases from all strata, or only from one or more of the more accessible strata with less hazardous activities, strata for which some sort of sampling frames are available for the purpose. We leave aside the question of how and what type of initial sample is selected within each stratum. The focus of our discussion here is the distribution of the final sample across the strata defined according to the degree of hazard involved in the children’s work.

(3) The initial sample constitutes wave zero. After being interviewed in wave zero, each person in the initial sample is asked to refer (‘name’) one or more eligible persons for inclusion in the sample. These referrals (excluding those which were already included in the initial sample) constitute the first wave. The procedure is repeated with the first wave respondents in turn, to obtain a sample for the second wave, and so on. The snowball sample is terminated after a certain number of waves, as determined by the chosen survey design.

\[
f_a = \max \left( 0; 1 - \frac{1}{(1+s)(1-P_a)} \right), \quad f_b = \max \left( 0; 1 - \frac{1}{(1+r)P_a} \right).\]

We have here two independent equations from the observed sample components at wave 1, but three unknowns \((P_a, f_a, f_b)\). With \(r\) and \(s\) obtained from the results of snowball wave 1, the above equations can be seen to provide a relationship between \(P_a\) and \((f_b / f_a)\). We may be able to obtain an indication of parameter \((f_b / f_a)\) from an idea of the plausible \(P_a\) values for our survey population, and thus choose an improved value for \((x_a / x_b)\) for referrals from wave 1 onwards.
(4) It is assumed that individuals in each ‘degree of hazard’ stratum have a certain propensity to refer (name) for the next wave other persons belonging to the same stratum as their own, and also certain propensities to refer persons in the neighbouring strata some less hazardous and others more hazardous than their own stratum. This allows the initial sample to spread out from the starting stratum or strata. These propensities may have certain fixed average values for each stratum, but these averages may differ from one stratum to another. We shall distinguish these three types of referral as:

- type (i): referrals from the same degree of hazard stratum
- type (ii): referrals from less hazardous strata
- type (iii): referrals from more hazardous strata

(5) Starting with an initial sample, the objective of the exercise is to seek – at least approximately – a certain specified distribution of the final sample across the strata. The type of distribution desired would depend on the survey objectives. From one end to the other, we may wish the final sample: (i) to be concentrated on the less hazardous strata, with only a small number of cases coming from strata with high hazard activities; or (ii) to be proportionate to the strata sizes; or (iii) to have an equal sample size in each stratum; or (iv) to be concentrated in the more hazardous strata, with only a small number of cases coming from strata with low hazard activities; or (v) have some other form of distribution, for instance concentrated on strata with middle hazard levels. The required distribution is determined by the survey objectives.

We take aspects (1)-(5) above as given, determined by the survey conditions (including the availability of sampling frames), and also by the particular objectives of the survey.

Most importantly, we also assume that some information is available on the referral propensities mentioned in (4) above concerning the referral by respondents of persons within their own stratum, versus the referral of persons outside their stratum. The information may be available from pilot studies, past surveys, field observations, external sources, etc.

There are two parameters which can be manipulated in the sample design procedures so as to meet objective (5), namely the objective of obtaining a final sample with some desired distribution across the strata. These parameters to control the size and distribution of the snowball sample are as follows.

(6) The average number of referrals per respondent from wave \( \omega \) which we seek for inclusion in the follow-up wave \( \omega+1 \) may be varied to control how the size and distribution of the sample develops with snowball waves. In principle, this average number of referrals per respondent may be varied across strata, and possibly also from one wave to another. In practice, there can be limits on how far the average number of referrals can be increased because of limits in the size of individuals’ social networks; that is, respondents may be simply unable to make more than a certain number of new referrals.

(7) At any wave, we may decide to follow-up all the referrals, or only a sample of the referrals made by the respondents in that wave. Again, the sampling rate may be varied across strata, and possibly also across waves of the snowball sampling. This
also includes the number of waves after which no more referrals are sought for follow-up, and the snowball sampling is terminated.

### 13.5.2 Models for the numerical illustrations

To be specific, in the numerical illustration of snowball sampling below, we assume the following feature concerning the population and sample design. These correspond to the aspects (1)-(7) noted above.

1. **The target population of children engaged in certain hazardous activities has been divided into 5 strata, \( i = 1 \cdots 5 \), according to the degree of hazard involved and the related accessibility of the population to the sampling process. Stratum \((i = 1)\) is the most accessible, and stratum \((i = 5)\) the least accessible.**

2. **We consider a range of distributions of the initial sample by stratum – from the whole of the initial sample coming only from the most accessible (the least hazardous) stratum, to the initial sample being uniformly distributed across the five strata. The first-mentioned distribution, where the availability of the sampling frame and hence the initial sample is confined to the most accessible stratum, represents quite a realistic situation, and a majority of the illustrations below assume this scenario.**

3. **It is assumed that the snowball sample terminates after 9 follow-up waves after the initial sample. In each example, the sample structure and parameters remain unchanged over the waves, though these aspects are chosen to differ from one example to another.**

4. **We assume throughout a constant pattern of the propensities to refer (name) individuals from the refereeing person’s own stratum, and also from other strata involving child labour activities with levels of hazard lower or higher than own stratum. It is assumed that on the average, 60 per cent of the referrals are to persons within the refereeing person’s own stratum; 20 per cent are to persons in the neighbouring stratum with less hazardous activities; and the remaining 20 per cent are to persons in the neighbouring stratum with more hazardous activities.**

There are no referrals to more distant strata in terms of the degree of hazard.

5. **It is assumed that substantive considerations require a final sample which is more or less uniformly distributed across the strata. This of course is only a working assumption, with no particular significance beyond that. There are also some examples with different distributions, as determined by the choice of parameters in particular cases.**

As concerns the design choices (6) and (7), the following combinations are assumed.

6. **The expansion of the sample from one wave to the next and the subsampling rate of the referrals to be followed-up are adjusted jointly such that the wave-specific sample size remains unchanged over the waves. Thus starting with an initial sample size of \( n (=40 \text{ in the examples}) \), the same amount of new sample is**

---

\[54\] For the least hazardous stratum, these figures are taken as 80% (same stratum) and 20% (next more hazardous stratum); similarly, for the most hazardous stratum, these are taken as 80% (same stratum) and 20% (next less hazardous stratum)
added at each wave, so that after follow-up waves the cumulative sample size is \( n(w+1) \) (=400 with the initial sample, plus 9 follow-up waves in the examples).

(7) The subsampling is applied at the same rate, say \( f < 1 \), to the first two of the three types of referral: the 60 per cent referrals in the refereeing person's own stratum, and the 20 per cent referrals in the neighbouring less hazardous stratum. The third type of referrals, those to individuals in the neighbouring more hazardous stratum, are all included (without any subsampling, \( f = 1 \)) into the next wave's sample. This appears to be quite a realistic model, since generally the initial sample lacks cases from the less accessible, more hazardous strata, and it is desirable to retain a higher proportion of such cases from referrals made for subsequent waves.

The above specific design choices are by no means required features of the model, but they are convenient and useful for the purpose of illustration.

The algebra of the balance between the sample expansion factor and the subsampling rate in order to retain a constant sample size over waves as required in (6) is as follows.

Let \( X \) be the average number of (actual, realised) referrals per respondent from one wave to the next, and \( p_{(i+1)} \) be the propensity of the respondents to refer (name) persons of the neighbouring more hazardous stratum (in our examples, \( p_{(i+1)} \) equals 0.2 throughout). These are what we called above as type (iii) referrals, and assumed that no subsampling of this type of referrals is involved. With these assumptions, it can be easily shown that, in order to keep the size of the new sample the same in each wave, the common sampling rate \( f \) to be applied to referrals of types (i) and (ii) – i.e. to referrals to the same and the neighbouring less hazardous strata - has to be

\[
f = \left( \frac{1}{X} - p_{(i+1)} \right) / \left( 1 - p_{(i+1)} \right).
\]

(13.8)

Note that no subsampling is involved \( (f = 1) \) if \( X = 1 \), that is, when we require only one (effective, achieved) referral per respondent. The upper limit for \( X \) in this model is \( X = 1/p_{(i+1)} = 5 \) in our example, which corresponds to \( (f = 0) \) meaning that no referrals of type (i) or (ii) are followed up. Only referrals of type (iii) – and all of them – are followed-up.

Slight adjustment is required to the reporting propensities for respondents in the least hazardous \( (i = 1) \) and the most hazardous \( (i = 5) \) strata, as already implied in the last footnote. Denote any stratum by \( i \), the neighbouring less hazardous stratum by \( (i - 1) \), and the neighbouring more hazardous stratum by \( (i + 1) \). The assumed referral propensities of respondents in stratum \( i \) are \( p_{(i-1)} = p_{(i+1)} = 0.2, p_{(i)} = 0.6 \). For \( i - 1 \), \( p_{(i-1)} \) is not relevant (there is no less hazardous stratum than \( i = 1 \)), and we simply take \( p_{(i)} = 0.8, p_{(i+1)} = 0.2 \). Similarly for \( i = 5 \), \( p_{(i+1)} \) is not relevant (there is no more hazardous stratum than \( i = 5 \)), and we simply take \( p_{(i-1)} = 0.2, p_{(i)} = 0.8 \).

### 13.5.3 Illustrative results

**A. Table 13.8. Spread of the snowball sample**

This table provides numerical illustration of the snowball sample resulting from different values of the model parameter. Table 13.8A has the following columns.
13.5 Simulation 2: spreading the sample to less accessible strata

| Column (1) | Initial sample size (also the sample size as it remains in each wave), \( a = 40 \). |
| Column (2) | Average number of referrals \((X)\) per respondent from any wave \(w\) to the next wave \((w + 1)\). |
| Column (3) | Total number of referrals from wave \(w\) to the next wave \((w + 1)\), \( A = aX = 40X \). |
| Column (4) | Propensity of a person in wave \(w\) to refer ('name') for the sample of next wave \((w + 1)\) someone belonging to: \(-1\) neighbouring less hazardous (more accessible) stratum, \( p_{(-1)} = 0.2 \). \(0\) referring person’s own stratum, \( p_{(0)} = 0.6 \). \(+1\) neighbouring more hazardous (less accessible) stratum, \( p_{(+1)} = 0.2 \). |
| Column (5) | Total number of referrals from a respondent in wave \(w\) to persons in different strata of wave \((w + 1)\): \(-\) in neighbouring less hazardous (more accessible) stratum, \( A_{(-1)} = \Delta p_{(-1)} \). \(-\) in respondent’s own stratum, \( A_{(0)} = \Delta p_{(0)} \). \(-\) in neighbouring more hazardous (less accessible) stratum, \( A_{(+1)} = \Delta p_{(+1)} \). |
| Column (6) | Subsampling rate to be applied to referrals to the respondent’s own and the neighbouring less hazardous strata \((f < 1)\), and to the neighbouring more hazardous stratum \((f = 1)\), determined so as to keep the sample size constant from one wave to the next. From (13.8), we have \( f = (5 - X)/4X \). |
| Column (7) | Resulting number of referrals from a respondent in wave \(w\) to different strata of wave \((w + 1)\): \(-\) to neighbouring less hazardous (more accessible) stratum, \( a_{(-1)} = f \cdot A_{(-1)} \). \(-\) to respondent’s own stratum, \( a_{(0)} = f \cdot A_{(0)} \). \(-\) to neighbouring more hazardous (less accessible) stratum, \( a_{(+1)} = A_{(+1)} \). |
| Column (8) | \( = \) Column (7) / Column (1). This shows how referrals (after subsampling according to (6) above) are distributed among the three – less, same and more hazardous - strata |

The panel at the bottom shows the special case of the least hazardous (most accessible) stratum, \( i = 1 \), with \( a_{(-1)} = 0 \), \( a_{(0)} = 0.8 \), and \( a_{(+1)} = 0.2 \). Similarly for the most hazardous (least accessible) stratum, \( i = 5 \), we will have \( a_{(-1)} = 0.2 \), \( a_{(0)} = 0.8 \) and \( a_{(+1)} = 0 \). The range of values of \( X \) shown is \( X = 1.0 \) to \( X = 5.0 \), in steps of 1.0. Note that with \( X = 1 \), the sample size automatically remains unchanged from one wave to the next - no subsampling is involved \((f = 1)\). By contrast, with \( X = 5 \), referrals to the neighbouring more hazardous stratum are fully retained \((f = 1)\), while referrals to the respondent’s own and the neighbouring less hazardous strata are completely dropped \((f = 0)\).
Table 13.8. Spread of the snowball sample to hidden strata
A. Numerical illustrations with different values of model parameters

<table>
<thead>
<tr>
<th>Level</th>
<th>i &gt; 1</th>
<th>a=</th>
<th>X=</th>
<th>A=X*a</th>
<th>stratum</th>
<th>p(i)=A(i)*p(i)</th>
<th>f(i)=A(i)*f(i)</th>
<th>s(i)=a(i)/a</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
<td>(7)</td>
</tr>
<tr>
<td>(-1)</td>
<td>0.2</td>
<td>8</td>
<td>1.00</td>
<td>8</td>
<td>0.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(+1)</td>
<td>0.2</td>
<td>8</td>
<td>1.00</td>
<td>8</td>
<td>0.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(-1)</td>
<td>0.2</td>
<td>16</td>
<td>0.38</td>
<td>6</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(+1)</td>
<td>0.2</td>
<td>16</td>
<td>1.00</td>
<td>16</td>
<td>0.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>40</td>
<td>80</td>
<td>80</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(-1)</td>
<td>0.2</td>
<td>32</td>
<td>0.60</td>
<td>2</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(+1)</td>
<td>0.2</td>
<td>32</td>
<td>1.00</td>
<td>32</td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>40</td>
<td>160</td>
<td>160</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(-1)</td>
<td>0.2</td>
<td>40</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(+1)</td>
<td>0.2</td>
<td>40</td>
<td>1.00</td>
<td>40</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>40</td>
<td>200</td>
<td>200</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level</th>
<th>i = 1</th>
<th>a=</th>
<th>X=</th>
<th>A=X*a</th>
<th>stratum</th>
<th>p(i)=A(i)*p(i)</th>
<th>f(i)=A(i)*f(i)</th>
<th>s(i)=a(i)/a</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
<td>(7)</td>
</tr>
<tr>
<td>40</td>
<td>1</td>
<td>40</td>
<td>40</td>
<td>(0)</td>
<td>0.8</td>
<td>32</td>
<td>1.00</td>
<td>32</td>
</tr>
<tr>
<td>(+1)</td>
<td>0.2</td>
<td>8</td>
<td>1.00</td>
<td>8</td>
<td>0.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>2</td>
<td>80</td>
<td>80</td>
<td>(0)</td>
<td>0.8</td>
<td>64</td>
<td>0.38</td>
<td>24</td>
</tr>
<tr>
<td>(+1)</td>
<td>0.2</td>
<td>16</td>
<td>1.00</td>
<td>16</td>
<td>0.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>3</td>
<td>120</td>
<td>120</td>
<td>(0)</td>
<td>0.8</td>
<td>96</td>
<td>0.17</td>
<td>16</td>
</tr>
<tr>
<td>(+1)</td>
<td>0.2</td>
<td>24</td>
<td>1.00</td>
<td>24</td>
<td>0.60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>4</td>
<td>160</td>
<td>160</td>
<td>(0)</td>
<td>0.8</td>
<td>128</td>
<td>0.06</td>
<td>8</td>
</tr>
<tr>
<td>(+1)</td>
<td>0.2</td>
<td>32</td>
<td>1.00</td>
<td>32</td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>5</td>
<td>200</td>
<td>200</td>
<td>(0)</td>
<td>0.8</td>
<td>160</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>(+1)</td>
<td>0.2</td>
<td>40</td>
<td>1.00</td>
<td>40</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Similarly for Level $i = 5$: the only change is that neighbouring stratum (+1) is replaced by stratum (-1)
### Table 13.8B. Composition of the resulting snowball sample by strata defined according to level of hazard of child labour activity

<table>
<thead>
<tr>
<th>Initial sample confined to the least hazardous stratum ($i = 1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a = 40$</td>
</tr>
<tr>
<td>$X = 1.0$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>stratum</th>
<th>&quot;Spread&quot;**</th>
<th>Wave</th>
<th>total</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i = 1$</td>
<td>$s(i-1)$</td>
<td>$s(i)$</td>
<td>$s(i+1)$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.2</td>
<td>40</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.6</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.8</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>total</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>stratum</th>
<th>&quot;Spread&quot;**</th>
<th>Wave</th>
<th>total</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i = 1$</td>
<td>$s(i-1)$</td>
<td>$s(i)$</td>
<td>$s(i+1)$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.6</td>
<td>40</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.3</td>
<td>0.6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.3</td>
<td>0.6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.3</td>
<td>0.6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.9</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>total</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>stratum</th>
<th>&quot;Spread&quot;**</th>
<th>Wave</th>
<th>total</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i = 1$</td>
<td>$s(i-1)$</td>
<td>$s(i)$</td>
<td>$s(i+1)$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>1.0</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0</td>
</tr>
<tr>
<td>total</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
</tbody>
</table>

* Spread: it refers to column (8) of Table 13.8A. These figures are used to compute wave-by-wave development of the snowball sample. A figure in row $i$ (stratum) column $w$ (wave) is spread out to rows $(i-1)$, 1 and $(i+1)$ at wave $(w+1)$ according to the $s(.)$ values.

Throughout this illustration we assume that any non-response or other loss in the sample has been compensated by drawing extra cases.

Table 13.8B shows the corresponding composition of the resulting snowball sample by wave, and the overall (cumulative) picture after 10 waves. Because of the manner noted earlier in which the parameters ($X$ and $f$) have been chosen, the sample size is constant in each wave, $n = 40$, and the cumulative sample size after 10 waves is 400.

Panels of the table have been constructed with different values of parameter $X$, the average number of referrals per respondent ($X = 1, 3, 5$). In each case, the initial sample has been confined to the least hazardous (most accessible) stratum, $i = 1$.  

---

**Table 13.8B. Composition of the resulting snowball sample by strata defined according to level of hazard of child labour activity**

For different values of number of referrals per respondent

- Initial sample confined to the least hazardous stratum ($i = 1$)
- $a = 40$, $X = 1.0$
- $a = 40$, $X = 3.0$
- $a = 40$, $X = 5.0$
After 10 waves, nearly 60 per cent of the final sample still belongs to this stratum when $X = 1$, i.e. when no sample expansion or subsampling from one wave to the next is involved. In the extreme case with $X = 5$, 60 per cent of the final sample after 10 waves belongs, by contrast, to the most hazardous (least accessible) stratum, $i = 5$, and a uniform 10 per cent to each of the other four strata ($i = 1 – 4$).

The stratum sample sizes of the final sample are approximately equal with $X = 3$, which we have assumed for the purpose of this illustration to be the desired pattern for the given survey objectives.

B. Table 13.9. Starting from different spreads of the initial sample

This table shows the effect of spread of the initial sample across the strata on the composition of the final sample. Throughout we take $X = 1$, so that there is no sample expansion or subsampling from one wave to the next ($f = 1$). In the first panel, the initial sample is confined entirely to the least hazardous (most accessible) stratum, $i = 1$, as was the case in all the panels of the previous illustration. As already noted, in this case nearly 60 per cent of the sample after 10 waves still comes from this stratum.

In the second panel we show the results when the initial sample is spread equally between the two least hazardous (most accessible) strata ($i = 1, 2$, with an initial sample of size 20 in each of these strata). Here, just over 40 per cent of the final sample comes from the most accessible stratum ($i = 1$), compared with 60 per cent in the previous panel. This figure drops to around 30%, then to 25% and finally to 20%, as the initial sample is spread to the most accessible three, four and finally all the five strata.

The gradual spread of the sample to more hazardous strata will be faster and more marked with larger $X$ values. This is because with $X > 1$ and subsampling of the strata with less and the same hazard level, a higher proportion of the followed-up sample comes from the more hazardous stratum. This effect can be seen, for example, in Table 13.10 below, which also has another illustrative objective as noted there.
13.5 Simulation 2: spreading the sample to less accessible strata

Table 13.9. Composition of the snowball sample with different spreads of the initial sample across strata (Parameters throughout. $\alpha = 40; X = 1$).

<table>
<thead>
<tr>
<th>Initial sample all from stratum 1 (least hazardous stratum)</th>
<th>Wave</th>
<th>total</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>stratum &quot;Spread&quot;**</td>
<td>s(i-1)</td>
<td>s(i)</td>
<td>s(i+1)</td>
</tr>
<tr>
<td>1</td>
<td>0.8</td>
<td>0.2</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Initial sample equally from strata 1 and 2</th>
<th>Wave</th>
<th>total</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>stratum &quot;Spread&quot;**</td>
<td>s(i-1)</td>
<td>s(i)</td>
<td>s(i+1)</td>
</tr>
<tr>
<td>1</td>
<td>0.8</td>
<td>0.2</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Initial sample equally from strata 1, 2 and 3</th>
<th>Wave</th>
<th>total</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>stratum &quot;Spread&quot;**</td>
<td>s(i-1)</td>
<td>s(i)</td>
<td>s(i+1)</td>
</tr>
<tr>
<td>1</td>
<td>0.8</td>
<td>0.2</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Initial sample equally from strata 1 to 4</th>
<th>Wave</th>
<th>total</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>stratum &quot;Spread&quot;**</td>
<td>s(i-1)</td>
<td>s(i)</td>
<td>s(i+1)</td>
</tr>
<tr>
<td>1</td>
<td>0.8</td>
<td>0.2</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Initial sample equally from all five strata</th>
<th>Wave</th>
<th>total</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>stratum &quot;Spread&quot;**</td>
<td>s(i-1)</td>
<td>s(i)</td>
<td>s(i+1)</td>
</tr>
<tr>
<td>1</td>
<td>0.8</td>
<td>0.2</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
</tbody>
</table>

* Spread: see note to Table 13.8B.
C. Table 13.10. Controlling stratum sample sizes

This table shows the effect of simultaneous adjustment of (i) decreasing spread of the initial sample, from equally over all the 5 strata to the initial sample being confined to stratum \( i = 1 \); and (ii) the average number of referrals \( (X) \) per respondent from one wave to the next. The combination of these parameters has been chosen such that the final sample is approximately equally distributed over the five strata after 10 waves of the snowball sampling. As in other examples, there are 10 waves and referrals at each wave are subsampled such that, for a given \( X \), the sample size from one wave to the next remains the same. This amounts to choosing the subsampling rate \( f \) to balance the chosen \( X \) in each case. The following are the \( (X, f) \) combinations to achieve this desired result in the table. The figures of course follow from Equation (13.8).

<table>
<thead>
<tr>
<th>Initial sample uniformly distributed over:</th>
<th>( X )</th>
<th>corresponding ( f )</th>
</tr>
</thead>
<tbody>
<tr>
<td>all 5 strata ( (i = 1 - 5) )</td>
<td>1.0</td>
<td>1.00</td>
</tr>
<tr>
<td>the 4 most accessible strata ( (i = 1 - 4) )</td>
<td>1.4</td>
<td>0.64</td>
</tr>
<tr>
<td>the 3 most accessible strata ( (i = 1 - 3) )</td>
<td>1.8</td>
<td>0.44</td>
</tr>
<tr>
<td>the 2 most accessible strata ( (i = 1, 2) )</td>
<td>2.2</td>
<td>0.32</td>
</tr>
<tr>
<td>only the most accessible stratum ( (i = 1) )</td>
<td>2.8</td>
<td>0.20</td>
</tr>
</tbody>
</table>

As noted earlier, in the present illustration the above subsampling rate \( f \) is applied to referrals to the neighbouring less hazardous stratum, \( (i - 1) \), and the refereeing respondent’s own stratum \( (i) \). The is no subsampling of referrals to the neighbouring more hazardous (less accessible) stratum, \( (i + 1) \).

Table 13.10. Balancing different spreads of initial sample across strata and the wave-to-wave sample expansion to obtain snowball sample roughly equally distributed across strata

<table>
<thead>
<tr>
<th>Initial sample equally from all five strata</th>
<th>( a = 40 )</th>
<th>( X = 1.0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>stratum &quot;Spread&quot;</td>
<td>Wave</td>
<td></td>
</tr>
<tr>
<td>( i = )</td>
<td>( s(i-1) )</td>
<td>( s(i) )</td>
</tr>
<tr>
<td>1</td>
<td>0.80</td>
<td>0.20</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>0.60</td>
</tr>
<tr>
<td>3</td>
<td>0.20</td>
<td>0.60</td>
</tr>
<tr>
<td>4</td>
<td>0.20</td>
<td>0.60</td>
</tr>
<tr>
<td>5</td>
<td>0.20</td>
<td>0.80</td>
</tr>
<tr>
<td>total</td>
<td>40</td>
<td>40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Initial sample equally from strata 1 to 4</th>
<th>( a = 40 )</th>
<th>( X = 1.4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>stratum &quot;Spread&quot;</td>
<td>Wave</td>
<td></td>
</tr>
<tr>
<td>( i = )</td>
<td>( s(i-1) )</td>
<td>( s(i) )</td>
</tr>
<tr>
<td>1</td>
<td>0.72</td>
<td>0.28</td>
</tr>
<tr>
<td>2</td>
<td>0.18</td>
<td>0.54</td>
</tr>
<tr>
<td>3</td>
<td>0.18</td>
<td>0.54</td>
</tr>
<tr>
<td>4</td>
<td>0.18</td>
<td>0.54</td>
</tr>
<tr>
<td>5</td>
<td>0.18</td>
<td>0.82</td>
</tr>
<tr>
<td>total</td>
<td>40</td>
<td>40</td>
</tr>
</tbody>
</table>
13.6 Snowball sampling: issues in planning and implementation

The purpose of this section is to describe practical aspects of planning and implementing the snowball sampling procedure: to bring out some common problems and suggest some possible solutions.

A major problem with many of the studies which have utilised the snowball sampling procedure is

"[that they] contain only a sparse account of the sampling method. It is as if the use of snowball method entails little more than to start it rolling"
through a personal contact or through an informant and then simply to sit back and allow the resulting chain to follow its own course. ... In spite of the widespread use of [the method], little has been written specifically about it. ... This apparent neglect may be a result of the [mistaken] assumption that the problems that may arise through the use of this sampling method are so simple or self-evident that they require little or no explanation. ... Through omission, the existing methodological literature suggests that [this] method of sampling is a self-contained and self-propelled phenomenon, in that once it is started it somehow magically proceeds on its own. This, however, is simply not the case; rather, the researcher must actively and deliberately develop the sample's initiation, progress, and termination.” (Biernacki and Waldorf, 1981).

We may identify the following major problem areas in the implementation of the snowball sampling procedure. In this discussion we draw on the excellent and widely referred to article by Biernacki and Waldorf quoted above.

### 13.6.1 Some issues in survey planning

**A. Defining the study objectives**

As in all statistical enquiries, the approach and procedures which are most appropriate depend on the nature of the study objectives. Those objectives determine the scope and type of the sample which would be acceptable.

In a sample survey under normal conditions, when it is feasible to capture a representative sample of the target population, the study aim may include all of the following objectives.

- **Objective (1):** merely exploratory, in order to assess the main features of the phenomenon.
- **Objective (2):** description of characteristics of the target population, without aiming at precise quantification.
- **Objective (3):** exploration and analysis of the process that worked to bring about the phenomenon being studied.
- **Objective (4):** quantification of the structure of the target population, for instance by estimating the relative sizes of population subgroups by age, sex, other personal characteristics, work location, sector of activity, etc.
- **Objective (5):** quantitative estimation of the size of the target population.

These objectives range from the most basic (1) to the most ambitious (5). A more ambitious objective in this list generally subsumes any preceding ones. In going from (1) to (5), generally increasingly large and representative samples are needed. This is reflected in the increasingly stringent requirements concerning control and development of the snowball sample. In each particular situation, a judgement has to be made about the type of sample which would be adequate to meet particular objectives, and on that basis, the type of snowball procedure required.
B. Defining the population and identifying subjects

In order to minimize sampling bias in studies of difficult-to-reach populations, it is necessary to begin with a clear conceptual definition of the population in question, followed by a plan to operationally identify potential respondents.

The explicit specification of a conceptual and operational definition of the study population (in terms of its content, extent, and time reference – see Section 4.1) is important so as to maintain consistency over the implementation of the study, and to control the sampling process. In particular, the choice and precise definition of the target population determines the type of sampling frame needed.

The above applies certainly to quantitative studies, aimed at all the objectives (1)-(5) listed above. Qualitative studies, concerned primarily with objectives (1)-(3), may begin with a looser conceptual definition that may evolve to guide more precisely the subsequent sampling as the study advances. A maxim commonly applied to sampling in qualitative research - maximize the phenomenon, i.e. capture as many and as varied members of the target population as possible - is a critical consideration (Penrod, Preston, Cain and Starks, 2003). What are the most appropriate sources of participants who have experienced the phenomenon or concept of interest? Although answers to this query may emerge only gradually during the study, thinking through this issue of scope early in the study can help to direct sampling efforts efficiently.

C. Choice of sample size

In qualitative studies, purposive sampling techniques may be used in conjunction with link-trace sampling to attain the necessary criteria of appropriateness and adequacy of the sample for objectives (1)-(3). The sample size and structure needs to be directed and adapted to explore the situation as new insights emerge.

For the purpose of quantitative estimation, a common procedure is the use of power analysis in the determination of the number of respondents necessary to control error (Kraemer and Thiemann, 1987). If multiple segments of a population are conceptually or operationally distinct, power analysis should be conducted for each segment of interest to enable statistical analysis of each segment.

But in any case, whether the objectives are mainly qualitative or more precise and quantitative, by establishing the minimum number of subjects necessary to control error and strengthen findings, researchers can plan and control the pacing and monitoring of the referral chains.

D. Selecting and assessing settings for the study and gaining access

The selection of the initial setting or settings as potential sources of respondents is usually a critical step in studies of reclusive and other special populations. Settings directly affect the representativeness of the sample. Care must be taken to select a diversity of settings that allows access to a wide range of characteristics of the study population.

The number of settings selected is also related to the required size of the sample. Through a strategic assessment of different sources of information, including facilities and organisations serving the target population groups, settings through which snowball
Sampling may be initiated can be identified. A review of related literature along with personal and professional insights establishes a baseline for starting the process. Careful consideration of potential biases introduced due to the restricted variety of the populations covered in the setting must be made and will help determine the array of settings to be included in the initial wave of sampling.

By their nature, members of a hidden population are difficult to locate. Often studies require some previous ‘knowledge of insiders’ in order to identify initial respondents. Such prior knowledge may not be readily available to researchers and it may be very time consuming and labour intensive to acquire. Under these circumstances, it is possible that people in positions of relative authority or proximity may provide a route into the required population. For example, staff of organisations working with children, especially with child labourers, might be able to introduce the researcher to children engaged in sectors of activities or in working conditions which are of interest to the study. It should be stressed that there are clear ethical implications for such work and that informed consent should be considered a prerequisite.

Developing a research agenda on sensitive topics requires entry into communities that are often hostile to people attempting to collect information. Entry can be influenced positively or negatively by the characteristics of the researcher (race, gender, age, cultural background, socio-economic status). Sensitivity of research topics on special populations has cultural connotations that will vary among groups. Thus, researchers must be culturally sensitive to perceiving and respecting the risks to and sensitivities of the target group. If potential respondents tend to react negatively to particular characteristics of data collectors, then a different type of field worker may be sought.

Once desired settings have been identified, the researcher’s entry must be established through leaders or members within each setting. ‘Gatekeepers’ are individuals who can help provide entry and access to the setting and help create the necessary link between researchers and the target population. Gatekeepers may facilitate processes of obtaining approval, if required, from authorities in the study area, and of recruiting potential respondents and disseminating information about the study to them. For example, local organizations may agree to contact potential participants with whom they are in contact.

In certain circumstances, it is necessary to minimise direct contact between researchers and respondents. In such cases, the protocol for involvement in the study must be explicitly stated in language and terms understood by the gatekeeper. In studies involving interviews or face-to-face data collection, initial contacts with the researcher are made through self-referral that is stimulated by information received through the gatekeeper. If anonymity of respondents is desired, the gatekeeper can help distribute data collection instruments (e.g. a written questionnaire) directly to eligible participants so that contact between researchers and respondents is minimised or avoided.

13.6.2 Practical aspects of developing a snowball sample

A. The initial sample

Once the survey objectives and the type of population being targeted are identified, the next step is to select and access the initial sample from which the snowballing procedure can be started and the sample developed in waves.
In some descriptions, applications or modelling of the snowball procedure, it is assumed that one can begin with a probability (random, representative) initial sample. However, as a rule, the type of population for which snowball sampling is of interest is a population dispersed, or rather ‘lost’, in a much larger general population. A representative sample of such a target population may sometimes be obtained by drawing a probability sample of the general population. Then, assuming the efficacy and honesty of the screening process, one may screen the general population sample to locate members of the target population in the sample. However, such an approach has to be generally excluded because, among many other possible reasons, the costs involved in drawing a large enough sample of the target population can be prohibitive (see discussion in Chapter 6, especially Section 6.7). A large scale screening is even less justified if the study objective is a limited one in the range of objectives (1)-(5) identified above.

A more realistic context for snowball sampling is when it is unavoidable but still considered acceptable to begin from a restricted and less than representative initial sample. Nevertheless, it is advantageous to begin from as representative a sample as possible. That would facilitate obtaining a more representative final sample after snowballing, and obtaining it more quickly (within fewer waves). What can facilitate obtaining a good initial sample?

In many situations, a reasonably good frame covering a part of the target population can be found. For instance, organisations or facilities providing services to some groups in the target population may be able to provide formal or informal assistance in locating, or providing referrals to, individuals in those groups. This is facilitated if concerns of confidentiality can be respected.

Another helpful feature can be that, for a part of the target population, while there is no sampling frame available, some information is available concerning locations where members of that group may be found in significant concentrations. In such situations, it may be feasible to undertake special operations to prepare a sampling frame or otherwise capture a sample of such persons.

There can also be more general sources of information which can direct us to members of the target population not covered in the sampling frame. General knowledge of the situation, previous studies, individuals already identified for the sample, etc. may suggest potential sources for locating cases from other categories of the target population.

Sometimes finding recruits to the initial sample may seem to have come about as if simply by good fortune. However, as Biernacki and Waldorf (1981) note,

“this is not entirely a process of chance but results from an increasing sensitivity and attentiveness to information related to the study’s focus that develops as the researcher becomes steeped in the research area. In a sense, it is the prepared mind that both knows and can take maximum advantage of opportunity”.

For instance, someone employed as a domestic worker by an acquaintance of the researcher may narrate something which results in contact with one or more child workers in the target population of interest.
Going beyond such accessible groups is likely to be more difficult. Sometimes a big hindrance comes from a mistaken belief among sources, whose help is being sought for finding sample cases, that the target population simply does not have any persons who are not in contact with at least some relevant organisation or facility, or a mistaken belief that it is simply impossible for any eligible cases to be present in certain groups, at certain locations, etc.

B. Sample of respondent chains

Starting with an initial or ‘seed’ sample, the snowball sample consists of respondents linked to other respondents. Thus what we have is a sample of respondent chains. The suitability of the sample for particular survey objectives, its representativeness and other characteristics, depend on how and what type of chains have been constructed, and how many and how diverse they are for the purpose of capturing different parts of the target population. The length of the chains which we are able to construct is also a very important feature. This is because longer chains tend to take us to other parts of the population, parts which may be quite different from the limited base of the initial sample from which we were able to start—thus making the final sample more diverse and representative. Often chains can lead to dead-ends when no new recruits can be found, or they close (‘loop’) quickly to simply return to their starting point.

The type of respondent chains which can be developed of course depends on the nature of the target population at hand and the conditions under which the survey is being conducted. But much more importantly for us, it also depends on the survey design, the procedures applied, and how they are implemented in practice, factors which can be manipulated by the survey researchers.

Practical implementation involves not only controlling the type, number and length of respondent chains, but also pacing and monitoring how the chains develop. Referrals, through which chains develop, are largely determined by the subjective perceptions of initial respondents about the involvement of others in the same activity. Thus particular individuals (the most popular ones, group leaders, those with wider social networks) are more likely to be identified than others. Much of snowball sampling rests on the assumption that social networks consist of groups with relatively homogenous social traits. In the case of a particular type of work or working conditions, for example, it is assumed that someone in this group would know others to whom a researcher could be directed. However, some groups may themselves consist of highly atomised and isolated individuals whose social networks are relatively impaired. Young unemployed men (and similarly older working children) constitute a prime case in this respect and may therefore be very difficult to identify or to initiate contact with. It is therefore apparent that snowball samples are both time consuming and labour intensive.

C. Verifying the eligibility of potential respondents

Verification of the criteria for membership of a sample will depend on the nature of the research question being posed. In the case, for instance, of very specific child labour activities, it may be that some referrals for the initial sample will not be accurate. This can also happen more generally, during later waves of the snowball sample.

It should be emphasised that any potential respondents must meet the eligibility condition required for their inclusion in the survey. Verification of eligibility is especially...
important before using any recruits as seeds for generating referral chains in search for additional respondents.

A possible source is the information for verification provided by others in the snowball sample. One strategy which can be useful for corroborating a respondent’s claims is made possible through a feature of the chain referral method itself. The respondents often end up discussing their own experiences as well as the experiences of others, who are closely related or known to them and are themselves likely to be in the snowball sample. Such voluntary information may serve as an additional source for verification of the information provided by different respondents.

The above applies in the case of respondents who entered the sample through the regular referral process. However, there can also be respondents who enter the survey through self-referral, that is by themselves taking the initiative to be included in the study sample. It can be more difficult to verify the eligibility of such cases. In all cases of doubt, it is prudent to assume a sceptical or doubtful stance towards persons volunteering to participate and seek verification concerning their claim to be eligible for the study.

D. Capturing diversity

The next consideration is to diversify the sample to capture different features and parts of the target population – at least in qualitative terms even if not in rigorous statistical terms at this stage. How diverse the sample needs to be depends, of course, on the nature of the target population and survey objectives. Sometimes the target population is limited and defined by very specific characteristics, such as child workers engaged in some very particular type of activity or at a very particular place. Normally, however, the target population comprises of individuals with different characteristics and in different circumstances which need to be captured in the survey. For instance, the target population may include child workers of different ages, both boys and girls, different family backgrounds, socio-economic situation, ethnic or racial characteristics, different behaviours in terms of school enrolment and attendance, etc. Also the target population may cover different areas or locations, and establishments, work places, activity, or working conditions of different types.

It is worthwhile to develop a check-list of population characteristics which appear important to represent in the snowball sample, and then take specific steps to develop the sample in order to capture at least some of that diversity. It is desirable that such effort is explicitly recognised as an aspect of the development of the snowball sample.

E. Broadening the initial sample base

It often becomes necessary to go further towards broadening the initial sample base. This may require moving into areas (locations or social groups) where there are no pre-established contacts. This can increase recruitment difficulties and bring in new problems. For one thing, it may become more difficult to verify that the persons appearing as potential respondents meet the eligibility condition required for their inclusion in the survey. Also, it becomes more likely that the effort of obtaining additional sample cases comes to nothing. In any case, finding new respondents often becomes more and more costly and difficult. “Verification of eligibility, as well as the accounts provided by
respondents, become increasingly problematic as the sources used to initiate referral chains become more distant and the knowledge about the sources less personal” (ibid).

In certain instances it might be possible to obtain additional cases and hopefully stimulate new chains through open, more public channels, such as soliciting potential respondents by asking members of the public, asking for volunteers, through some sort of ‘advertising’. However, it is recommended that such strategies “be used only as a last resort and, when they are used, that special care be given not to reveal all the criteria of eligibility of the study nor too few of them. Revealing too many of the eligibility criteria can result in problems related to verification, while revealing too few details can create management problems related to screening and perhaps difficulties in turning away non-eligible but willing study participants” (ibid).

A common problem is that the initiation of referral chains is restricted to areas of which those conducting the survey are aware of and which are accessible to them. It can be helpful to use paid and/or unpaid persons to help locate new respondents and start additional referral chains. Such locators can serve two interrelated purposes: first, because of their particular pasts, occupations, social positions and/or life styles, they can have relatively easy access to certain data sources and, as a result, can make contacts for possible interviews more efficiently than could the researchers. Second, because the locators often know the persons referred in the study, they can verify the respondents’ accounts. “The use of locators in snowball sampling is akin to the use of significant [that is, key] informants in field studies. Their use assumes that knowledge is differentially distributed and that certain persons, as a result of their past or present situations, have greater accessibility and knowledge about a specific area of life than do others. These persons can more easily develop referral chains because they may already be aware of potential respondents or may be more likely to have others reveal their potential to them. ... When the snowball sampling method is used and study respondents are enlisted to help find other potential respondents, they become de facto research assistants [members of the survey team]. Although all of a study’s respondents might be asked to refer others, all of them cannot and should not be engaged to assist the research on a regular basis. The characteristics of the study respondents differ and so do their abilities to help with the research effort” (ibid).

**F. More selective development**

As the sample diversifies, it usually becomes necessary to guide its development more selectively. The objective gradually shifts from simply reflecting the diversity of the target population in the sample, to capturing that diversity in a more proportionate, balanced manner. The sample has to be steered to avoid obvious bias, such as gross over-representation of some parts or characteristics of the population, or gross under-representation of others. For instance, some chains may be ‘stuck in grooves’- may lead to too many cases of the same type, never entering other parts of the population. If it is not possible to redirect such chains, it may become necessary to terminate them and start new ones with different characteristics. “Another problem that must be addressed and controlled when using the chain referral sampling method is that of limiting the number of cases within any subgroup in the sample. The researcher must continually ask: How many more cases should be collected and in what direction should the referral chain be guided? The decision here should be based on at least two
13.6 Snowball sampling: issues in planning and implementation

considerations: representativeness of the sample and repetition of the data. ... The number of cases provided through any type of referral chain should also be limited when the data become repetitious. At this point the researcher should be confident that the possible variations extant in that particular subgroup have been exhausted” (ibid).

The ability to control selectively the development of respondent chains depends on at least two factors. Firstly, we need to know something concerning the structure and characteristics of the target population towards which the snowball sample is to be steered. Such knowledge can be critical, even if it can only be approximate. It may be derived from pilot studies, past surveys, experts in the field, key informants, media, even knowledge of a purely general kind about the situation – ideally one should explore as many of such sources as possible. Some simple examples were provided in the simulated numerical illustrations in the preceding sections, where we discussed the choice of model parameters for capturing certain features of the target population structure, such as its composition and distribution across strata. As the illustrations show, knowledge of certain general characteristics of the target population can greatly help in steering the development of the snowball sample.

G. Data analysis

The second type of information source for controlling and steering development of the snowball sample is continuing analysis of the data as they come in from the survey implementation. The analysis may have many objectives and uses, but an important one among them is to direct and guide existing and future referral chains, particularly at early stages of the survey.

The data should be analysed on an on-going basis. The analysis should ideally involve both statistical and theoretical aspects. By statistical aspects we mean analysis of the structure (size, distribution, and other characteristics) of the sample and its comparison with whatever is known/expected regarding the structure of the target population. The objective of the analysis is to assess, as far as possible, whether the sample is developing towards being more representative of the target population. This does not mean that one is necessarily seeking the sample to be proportionately distributed across strata or domains of the population. Different distributions of the sample may be aimed at depending on the specific objectives – such as equal domain sample sizes to facilitate comparisons among domains, or proportionate sample allocation to facilitate aggregation over domains.

By theoretical aspects of the analysis we mean looking for new aspects, relationships, processes, etc., which might be present in the target population – and which might require reassessment of the study orientation, objectives, methods, including possibly increasing the size and targeting of the sample to the new aspects identified.

H. Controlling the pace and quality of data collection

In a conventional sample survey situation, it is usually easy to determine the pace of data collection as required, for instance to maximise its efficiency or quality, or to meet other constraints the survey organisation may be faced with. It needs to be recognised that in snowball sampling there are special practical pressures to speed-up the data collection process, or conversely, to slow it down from its ‘natural rhythm’.
“It is important to make a distinction here between the natural evolution and dissipation of a referral chain as it builds and exhausts itself through a social network, and the deliberate regulation of the speed with which new contacts and interviews are developed and conducted” (ibid).

A serious problem can be that enough referrals cannot be generated to develop, or even continue, the snowball sample.

By contrast we may also have the reverse situation. Certain pressures such as the following can combine to create a sense of urgency and to rush the development of new contacts and the completion of interviews. (i) There is, above all, a commonly existing fear that if referrals are not followed up immediately, the leads will go ‘cold’ and possibly lost. (ii) Pressure to meet predetermined sample quotas can be another problem, more serious in the present case than the normal survey situation because of the sensitivity of the snowball procedures to many difficult-to-control factors. (iii) Another possible problem arises from the nature of the snowball method itself: it is that referrals sometimes appear suddenly, in large groups, thus introducing unpredicted surges in the workload. (iv) Finally we should also note the problem resulting from the need to analyse the information quickly, almost simultaneously with its collection, if it is to be used to guide the direction in which the sample should be developed.

Two important signals of possible problems are: that enough new cases are not being found; and/or that the recruits being found are too homogeneous to adequately reflect the diversity of the target population.

The process of controlling and steering development of the snowball sample is not as simple as it may appear. It might be extremely difficult to start referral chains that will yield respondents of certain types. The target population may be composed of subgroups which are reclusive to different degrees and lack connections with each other, connections which may be used to link new units onto the chains already being developed. For instance, some worst forms of child labour occur in extremely isolated conditions and/or the children involved are very reclusive. Accessing such isolated groups would require starting new chains directly from within those groups.

In contrast with the conventional sampling situation, the snowball sample is not pre-selected, nor is the sample wholly determined by objective criteria as for instance is the case with adaptive sampling. The snowball sample often has to be developed through deliberate intervention.
Chapter 14
Respondent-driven sampling

14.1 Methodology of respondent-driven sampling

14.1.1 Introduction

Respondent-driven sampling (RDS) is the name given to a set of special procedures based on the snowball sampling method (Heckathorn, 1997, 2002). It is a form of link trace sampling that has been used to identify hidden population groups like injection drug users, jazz musicians, and young adult ecstasy users (Félix-Medina and Thompson, 2004; Frank and Snijders, 1994; Heckathorn, Broadhead and Sergeyev, 2001; Heckathorn and Jeffri, 2001; Heckathorn, Semaan, Broadhead and Hughes, 2002; Salganik and Heckathorn, 2004). The method is suitable for populations where no sampling frame exists and such a frame is impossible to establish.

Since it was first proposed, RDS has been widely used in various countries in studies of most-at-risk population subgroups for HIV. A 2008 review of studies in the literature that used RDS in hard-to-reach populations identified 128 HIV surveillance studies performed in developing and ex-Communist countries during 2003-2007 (Johnston et al., 2008). Likewise, another review the same year identified 123 studies, 59 of which were in Europe, 40 in Asia/Pacific, 14 in Latin America, 7 in Africa, and 3 in Oceania.

As in snowball sampling generally, the sample is collected with a chain-referral design. As described in Chapter 13, snowball sampling is useful in settings where a network of social relations links members of the target population, such that previously sampled individuals can facilitate the sampling of others in the population. The sampling begins with a set of initial participants who serve as ‘seeds’, and expands in waves. The procedure relies on personal contacts of the people in the sample to gather information about other prospective respondents, and is often very effective at recruiting large samples from hard-to-reach populations. Respondent-driven sampling is also based on such a design, but with a number of special feature as described in this chapter.

The initial samples usually cannot be drawn at random. The RDS procedure attempts to take steps to get away from the consequences of this common weakness in chain-referral sampling arising from the initial sample being restricted.

Furthermore, usual chain referral sampling methods, relying on referrals from initial subjects to generate additional subjects, produce biased samples because respondents who have a large number of social connections are able to provide large numbers of respondents. The result is a final sample that is overrepresented in characteristics of respondents who have many social connections, and underrepresented in characteristics of respondents with few social connections. The latter usually are the more hidden groups. This limitation produces a sample whereby statistical inference from the sample to the larger target population cannot be made reliably (Van Meter, 1990).
Bias can arise from two sources: the initial sample; and the subsequent recruitment process.

(1) Firstly, the sample may be biased because it is likely to attract volunteering, more co-operative participants, while subjects less co-operative are underrepresented.

(2) Secondly, bias is also introduced as the initial respondents recruit their friends. Since referrals occur through social network connections, socially isolated individuals tend to be excluded from the sample. Participants may have a preference to recruit friends who share the same characteristics as them (Erickson, 1979); but on the contrary, they may also avoid recruiting friends in order to protect them.

Salganik and Heckathorn (2004) have argued that the problem of bias in the ordinary application of chain-referral procedures can be solved on the following lines. Previous chain-referral sampling techniques have treated members of the hidden population as atomized units, not recognising that hidden populations are made of real people connected in a network of relationships. Rather than attempting to estimate directly from the sample to the population – as in traditional sampling and estimation – RDS uses an indirect method. First, the sample is used to make estimates about the social networks connecting the population. This information provides sample weights, which are then used to derive the proportion of the population in different groups.

**Figure 14.1. RDS estimation framework**

Schematic representations showing the differences between traditional and respondent-driven sampling and estimation. By not estimating directly from the sample to the population, but through social networks, respondent-driven sampling avoids some estimation problems common in chain-referral sampling.

*Source: Salganik and Heckathorn (2004).*

The special procedures introduced in RDS to reduce these biases include the following (see for instance Fashola, 2010). We will elaborate some of these subsequently.
(1) While the initial sample is generally non-representative, even non-random, attempt is made to follow it up with long chains of additional recruitment, so that the resulting sample becomes more and more free of the bias in the initial sample.

(2) The procedure requires direct recruitment of peers by their peers. RDS is based on the recognition that peers are better able than researchers to locate and recruit peers whom they know as members of a hard-to-reach population.

(3) Furthermore, the researchers keep track of who recruited whom into the sample.

(4) Information is obtained on the degree of each respondent, that is the number of people each participant reports having in their social network.

(5) Estimation and analysis of survey data is based on a mathematical model aimed at reducing the biases typically associated with snowball sampling. The mathematical model combines principles from Markov chain theory and biased network theory into a common data analysis framework.

- To account for the non-random selection of individuals and the possible over-representation of individuals with given characteristics in the study population, the recruitment is modelled by a Markov process.

- Biased network theory is based on the assumption that social network connections are formed randomly, through a non-deterministic process, where friendships tend to form among those who are similar. This phenomenon is termed homophily. By contrast, connections can also be formed based on how different groups complement each other – this has been termed heterophily. The theory also defines the parameters for how to calculate the percentage of time in-group connections are formed and the percentage of time out-group connections are formed in the social networks of a given sample.

To summarise, RDS presents two main innovations:

- The sampling design relies on the respondents at each wave to select (‘drive’) the next wave of sampling through their selection of other members of the target population. This is typically achieved through the distribution of coupons by respondents to other members of the target population.

- The second main innovation is that, while an RDS sample begins with a convenience sample of individuals, through many waves of sampling the dependence of the final sample on the initial convenience sample is reduced.

Improvements offered by the RDS procedure are summarised in Table 14.1.
### Table 14.1. Biases in usual chain-referral (snowball) sampling, and RDS ‘solutions’

<table>
<thead>
<tr>
<th>Usual snowball sampling</th>
<th>Bias or problem</th>
<th>Respondent-driven sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respondents may refer to an unlimited number of peers</td>
<td>1. Differential recruitment: those with larger network sizes can recruit more peers, who are likely to have similar traits 2. Clustering: leads to lower effective sample size</td>
<td>Limiting recruitment coupons to individuals limits clusters, thereby reducing recruitment bias and excessive homophily (see Section 14.2)</td>
</tr>
<tr>
<td>Social network properties are ignored</td>
<td>1. Clustering by network traits cannot be measured 2. Size of social networks affect probability of selection</td>
<td>1. Coded coupons permit linking respondent with recruiter 2. Analysis is weighted to account for measurable network properties</td>
</tr>
<tr>
<td>Respondent contact and recruitment may involve researchers too closely</td>
<td>1. Only members accessible to ‘outsiders’ participate 2. Concerns of intrusion and lack of confidentiality for the (often reclusive) respondents</td>
<td>1. Peers recruit peers, and can exert social influence to facilitate participation; surveyors can retain greater distance 2. Strategy reduces intrusion and confidentiality concerns</td>
</tr>
<tr>
<td>Convenience sample – analysis cannot be generalised to the population</td>
<td>Probability of selection is unknown</td>
<td>Attempts to move the sample closer to probability sampling. 1. Collects network size to estimate selection probabilities. 2. Uses network properties to account for clustering</td>
</tr>
</tbody>
</table>

**Source:** Johnston and Sabin, (2010).

### 14.1.2 Underlying concepts: homophily, equilibrium, degree, network structure

The following are among the basic concepts in RDS methodology.

**A. Homophily and heterophily**

Homophily refers to the tendency to recruit peers with characteristics similar to those of the recruiters. The concept of homophily is important because there is a need to weight the sample to compensate quantitatively for differences between groups because of the tendency to recruit similar peers within groups. The limited number of recruits that recruiters are allowed in RDS helps to increase the social distance between initial seeds and recruits who enrol in later waves. Respondents can also have a tendency to protect their friends, people like themselves, and hence avoid recruiting them. This is heterophily. These tendencies depend on whether individuals see participation in the survey as embarrassing and burdensome, or as beneficial, rewarding and interesting.

**B. Equilibrium**

Studies have shown that, with proper procedures, specific features resulting from the initial respondents are progressively weakened with each recruitment wave. Equilibrium refers to the stage where bias from the seeds disappears when the number of waves is
large (Heckathorn, 1997). It identifies the point at which sample characteristics seem to become independent of the seeds, thereby indicating control of bias introduced by the non-random selection of seeds. The concept of equilibrium helps in determining the number of waves needed to reach the desired sample size that is representative of the target population. Sampling may be continued after reaching equilibrium, if necessary, and terminated when the required sample size is reached.

A methodological point should be emphasised. The fact that, after a sufficient number of waves of the survey, the sample may reach ‘equilibrium’ in terms of stability of some characteristics does not necessarily mean that those stabilised characteristics are representative of the target population. Certain very isolated or hidden subgroups in the population may remain totally or in large part unrepresented in the sample. Nevertheless, reaching equilibrium in terms of sample characteristics has two important implications. Firstly, reaching an equilibrium does provide a positive (even if not a definitive) indication of the sample characteristics moving closer to those of the population. Secondly, reaching an equilibrium means that no further improvements in the structure of the sample can be expected by adding more waves to it. That can only augment the sample size. Sample characteristics can be influenced only by seeking new sources and new methods for recruitment.

Figure 14.2 provides an example of equilibrium being reached after a small number of RDS waves, taken from a study by Kajubi et al. (2008). The methodology of the study will be described in more detail in Section 14.3.3 as one of the examples of surveys using RDS. The main aim of the study was to describe the demographic characteristics of gay and bisexual men in Kampala, Uganda. The target population was men aged 18 years or older who self-identified as gay or bisexual. The study used respondent-driven sampling to recruit 224 participants for the survey.

**Figure 14.2. Examples of equilibrium being reached following RDS waves**
The following comments are provided in the study report concerning Figure 14.2.

"[The figure] shows equilibrium by recruitment wave and ... population estimates for education and having sex with a male partner from a country outside of Africa—two key variables we used to monitor the survey. Despite actively seeking seeds of lower education level, the majority had completed secondary education (Figure 14.2a, wave 0). As recruitment progressed, the level of education decreased and stabilized near 50% having completed secondary education by the second to third wave. We could not identify [any seeds having male sex partners only from within an African country - Figure 14.2b, wave 0]. However, the proportion of respondents reporting foreign male sex partners from countries outside Africa dropped from 100% with successive waves, stabilizing by the third to fourth wave with 44% having any foreign male sex partner from outside Africa”.

C. Degree

In practical application of the procedure, all recruits to the sample at each stage are requested to report their degree. Degree refers to the multiplicity of the person. In terms of the RDS framework, it is defined as the number of individuals who, if they had the opportunity, would be ready to pass a coupon to the respondent concerned. In terms of practical measurement, the concept has to be interpreted somewhat differently. The respondent cannot be expected to know and report what another might do. Hence the degree, as measured in the survey, is based on assessment by the respondent of the total number of persons to whom the respondent would be willing to pass a coupon. On the assumption of reciprocity of relationships, this is taken to be the number of individuals who, if they had the opportunity, would be ready to pass a coupon to the respondent concerned. In this way the measured degree is used as a surrogate for the true degree for determining the individual's probability of selection into the sample.
D. Network structure

While RDS does not call for a priori characterizing the network structure of the target population, such a structure can affect the success of using RDS on a given population. The following account draws on Semaan (2010).

Network characteristics are usually described in the network literature and are measured by several network related variables:

- **Density** (i.e. total number of connections divided by the possible number of connections) represents the extent of communication within the network.
- **Centralization** measures the extent to which the majority of the connections relate to a small set of nodes and indicates the degree to which communication is centralized around a small number of participants.
- **Clique count**, the average size of maximally connected subgroups, provides a measure of social integration and network cohesion that predominates in the population.
- **Effective network size**, the average number of effective personal links (disregarding redundant links), measures connections to the same person through more than one link.

It is often the network structure that affects the success or otherwise of a RDS, as reflected by the density of the network that characterizes the target population. Network size is usually based on reported data, it is measured by asking participants to recall how many peers they know and have seen during a set period of time (e.g. one month). Network sizes are an essential part of RDS data analysis as they are used to weight the sample data to compensate for the oversampling of participants who have larger than average network sizes and who thus have more recruitment paths leading to them. Furthermore, RDS requires additional analytical steps which take into account a combination of differential recruitment effectiveness across groups. Differential recruitment is assessed by collecting data on who recruits whom.

14.2 Aspects of implementation

14.2.1 Development of design and procedures

A. Formative research

RDS require *formative research* – preparatory research prior to implementation of the survey - to acquire relevant information about the target population, and to provide the basis and rationale for the choice of a particular sampling strategy.

First of all, adequate formative research is needed to characterize the target population and to develop procedures to use RDS with that population. RDS is best used after formative research has provided relevant data, including data on the density of networks and on other network-related variables for the purpose of characterising the target population and assessing the extent to which RDS can be successful. Data on the network size and connections of the target population provide variables which affect the
ability to reach the desired sample size. Adequate formative research or ethnographic evaluation can provide information on whether the target population is sufficiently connected or too fragmented to support the use of RDS. Indicator data should also be analysed to characterize the network connections of the target population because recruitment is dependent on peers who have a reciprocal relationship with peers in the target population.

Also, formative research with the target population, including people in different economic circumstances, can help in understanding the level of incentives that may or may not motivate the target population to participate in the study.

B. Developing procedures and protocols

Addressing logistical, regulatory, and ethical dimensions of sampling is important to ensure that studies being planned for implementation are feasible, provide statistically valid data, and are ethically sound.

Basic steps in preparing to use RDS include developing and implementing protocols for training seeds and participants to serve as recruiters of their peers, pilot-testing remuneration levels for participant-driven recruitment, pilot-testing questions on screening and eligibility criteria, training staff members in using the coupon management system, and establishing stable, safe, and accessible interview sites (Semaan, Lauby and Liebman, 2002).

There are general procedures and safeguards that can be used irrespective of the sampling strategy chosen. Safeguards include field experience in working with the target population, knowledge of guidelines and regulations for protecting participants, and awareness of local factors that can influence the field work. Equally important is the need to share information about the process and results with representatives of the target population, community leaders and authorities, and with the scientific community.

As noted by Semaan (2010), in order to avoid disappointment or conflict, investigators may share information with the participants about the conditions under which the coupons will not be accepted (e.g. when target sample size has been reached, when the coupon has been tampered with, when the coupon receiver has already been interviewed). Quality control measures, such as the extent to which the data in the coupon management system are collected and evaluated, are important in RDS implementation. Training staff in relevant regulations and ethical considerations to protect the participants' rights and welfare and the integrity of the survey is also valuable.

C. Ethical Considerations

In terms of ethical considerations, procedures for protecting participants have to be developed and refined. These procedures would specify how sensitive information must be guarded and how remuneration for the time and effort in survey participation or in participant-driven recruitment must be balanced to avoid undue inducement or coercion. In working with any target population, it is important that recruiters and recruits do not feel coerced or pressured to participate, and that investigators protect against such occurrences. Respondents need to know that participant-driven recruitment is voluntary and that there is no penalty should participants decide not to recruit their peers.
14.2 Aspects of implementation

14.2.2 Seeds

RDS begins with a set of participants, or seeds, purposively selected from the target population. Seeds are treated as any other participant (i.e. screened, interviewed, etc.) and must meet the eligibility requirements for inclusion in the study. However, seeds must have special attributes that will facilitate effective recruitment by them.

Ideally, seeds should have large social network sizes, be respected by members of the target population, be able to convince others to participate in the study and have some interest in the study goals (Heckathorn, 1997; 2002). If seeds are well-connected and well-motivated members of the target population, they can more easily jump-start the recruitment process and reduce bias by speeding up the attainment of equilibrium (Semaan, 2010).

Studies often begin with 6 to 8 seeds, though a larger number (e.g. 20) is preferable. In many settings, individuals in the target population are separated by type (e.g. children by the nature, conditions or location of their work) into sub-networks across which communication and contact are limited. These ‘cleavages’ could hinder, even preclude, recruitment across type, resulting in a sample of two or three distinct groups rather than one complete social network component. In any case, there should be a sufficient number of seeds to reach different parts of the target population and reflect its diversity.

Figure 14.3. Example of a recruitment graph starting from seeds

Seeds are represented with larger symbols.

Symbols used in the diagram to indicate different subgroup types in the target population of female sex workers:

- Street/park/port = shaded circle within a square
- Karaoke bar/restaurant/disco = solid (black) square
- Hotel/guesthouse = circle
- Agent/telephone/internet = triangle

Source: Johnston and Sabin, 2010.
Figure 14.3 from a study by Johnston and Sabin (2010) of RDS application to female sex workers (FSW) presents an example of eight well-integrated recruitment chains of FSW types based on the question ‘where do you most often solicit your clients?’ Note that, in the end, all types of FSW were recruited within one of several recruitment chains, overriding the seed type with which they started. Seeds that represent important characteristics can help to ensure that the sample reaches into different sub-networks of the target population. Ideally, formative assessment in advance of data collection can identify potential cleavages and assist in the selection of effective seeds.

Figure 14.4 provides another example of the development of an RDS sample from a small initial sample of seeds, taken from the already mentioned study by Kajubi et al. (2008) of gay and bisexual men in Kampala, Uganda. The study is described in more detail in Section 14.3 as one of the examples of surveys using RDS.

Figure 14.4. Another example of RDS sample development from a small initial sample of seeds

Source: Kajubi et al. (2008).

The following comments are provided in the study report concerning Figure 14.4.

“[The figure] shows the recruitment chains for our RDS survey of gay and bisexual men in Kampala. Eight seeds germinated and their chains grew to different lengths. The longest referral chain extended to 13 waves; the average number of waves was four. As the chains grew, 650 coupons were distributed; 230 men presented themselves for the survey; 224 men were screened as eligible and all of these completed the survey. …. We did not ask participants how many persons they attempted to recruit but
were unsuccessful. Only one person said that the person who gave them the study coupon was a stranger; he is included in the analysis. The sample of 224 was achieved in less than two months. Recruitment started slow but accelerated up to the moment that we stopped the study. The median reported number of gay and bisexual men that participants said they knew by name, nickname, or face and had seen in the last six months was 30 (interquartile range [IQR] 20–100). The median number of gay or bisexual men considered as close friends was 5 (IQR 2–10). When asked to estimate how many gay and bisexual men lived in Kampala, the median response was 1,000 (IQR 300 to 2,600).

14.2.3 Recruitment by previous respondents

For recruitment to work, respondents must know one another as members of the target population. ‘Knowing’ someone involves general recognition, allowing both acquaintances (weaker ties) and friends (stronger ties) to be recruited. As already noted, and explained in more detail in Section 14.2.4 below, recruiters receive a predetermined number of non-transferable, hard-to-duplicate coupons to recruit peers from their social network. Following referral by recruiters and arrival at the study site, staff members explain the study to recruits, screen them, and if they are found to be eligible, invite them to participate in the study.

Once each seed’s recruits have participated in the survey, they in turn are given a set number of coupons with which to recruit peers from their social networks to participate. This process normally results in the sample expanding from wave to wave to form recruitment chains, and continues until the final sample is obtained. Figure 14.5 from Johnston and Sabin (2010) provides a (hypothetical) example.

RDS’s recruitment method has the potential for rapid recruitment because every participant is meant to be a recruiter. This is particularly true when participants have large social networks and strong ties within those networks; therefore RDS can be especially successful at rapid recruitment in dense urban environments (Abdul-Quader, Heckathorn, Sabin and Tobi, 2006).

This diagram shows only those links which lead to a new recruit, not already in the sample in the current or any previous wave. It also assumes a target population large enough to sustain expansion of the sample over survey waves. The recruitment of new cases can dry out in small populations, in the sense that increasingly, the networks of respondents contain only persons who have been already included in the survey and do not yield new respondents. Recruitment paths tend to form closed loops in this case.
Hence, while the potential for rapid recruitment is one of the advantages to using RDS, there is the possibility that recruitment may be very slow if participants are not recruiting their peers. There are a variety of reasons why recruitment may be sluggish, or even fail: small network size, lack of connections among members of the target population, privacy concerns, or a high level of stigma associated with membership of the target population.

In any case, often recruitment rates are unpredictable. Practical steps which will be effective in sustaining recruitment at the required level and tempo tend to be very situation specific. The only guideline that can be given is to closely monitor the development of the sample case by case, in order to pinpoint to the extent possible why recruitment may have failed, succeeded as required, or overshot the target.

An important assumption in the recruitment process is that recruitment occurs symmetrically, that ties are reciprocal, so that any recruit, selected by a recruiter would have been just as willing to recruit his/her recruiter if that person had not already participated in the survey. Process and analytical mechanisms should be established to prevent repeated participation in the survey. Biometric measurements and even fingerprints have been used to prevent repeated enrolment in other studies, but it is doubtful whether child workers can be subject to such procedures without causing offence or other problems.
14.2.4 Incentives and coupons

A. Incentives

The RDS method also draws on the strategy of giving incentives to the participants. Its implementation relies on a variety of incentives to encourage participation. One incentive is simply peer pressure by the recruiters on their peers. Another incentive may be the opportunity to receive specific benefits, such as medical care for some problem the person may be suffering from when such problems come to light in the process of participation in the study. This can be especially effective in situations where alternative sources for the same benefits do not exist, are not affordable, or the subject wishes to keep the problem private, or to remain hidden from others.

The main incentive used to promote the recruitment and thus the growth of the networks is, of course, material reward. Remuneration for the time and effort for participation and for participant-driven recruitment is a core element in the use RDS. Remuneration is an important element in studies that depend on voluntary participation because although altruism is a necessary force, it is normally insufficient. Choice of the type and amount of incentive depends on the target group’s characteristics. The fact that there is a dual incentive structure, in which respondents are rewarded for being interviewed and for recruiting new respondents, is a key element of the procedure. The first incentive, called the primary incentive, is given to participants when they complete their participation in the study. They are then asked to recruit peers to respond, and are rewarded according to the number of peers responding to the criteria of inclusion in the study. This is the secondary incentive.

The reward may be in cash, or in the form of items useful in daily life or in work for the respondent. One difficulty about determining incentives is that if they are too high, it could lead to the bartering or selling of coupons and encourage non-eligible persons to pretend to be part of the eligible population. When incentives are too low, recruitment may be slow and, in some cases, attractive only to poorer members of the population (Kendall et al., 2008). Both of these situations can lead to selection bias.

In relation to the study venue, it is important that the study be carried out in an easy-to-reach place for the participants. A special, private but not isolated or threatening, ‘interview office’ may be helpful in winning the confidence of respondents, especially of child respondents. For instance, the interview place should be sheltered from public view, but should not be claustrophobic or dark. Providing a convenient and private place for recruits to present themselves would be an effective incentive, just as a threatening or exposed place would be a disincentive.

B. Coupons

The number of coupons distributed to participants is usually fixed (usually at 2 to 4) to allow each participant to recruit roughly the same number of peers. A coupon quota ensures that everyone has an equal chance to recruit peers – this is referred to as the recruitment quota.

To facilitate and control this procedure, it is usual to give each recruitment coupon a unique identification number. To participate in the study, the invited individuals must present their coupons as a proof that they have come through network recruitment. In
addition, the unique serial numbers of the coupons are used later to identify network structure and recruitment patterns. Normally, the coupons are redeemed at a fixed interview location within a set period of time (e.g. 10 days). After participating in the study these new sample members are also provided with the same fixed number of coupons, which they then use to recruit others. This sampling process continues until the desired sample size is reached, or a lack of recruits terminates it.

In principle, RDS peer-to-peer recruitment avoids selection bias of survey staff, and the coupon quota minimises bias associated with the over-representation of participants with large networks. In addition, RDS requires that recruitment continue far beyond the seed and his/her recruits. The sequence of recruitment from one respondent to another is continued until the final sample comprises long recruitment chains made up of several waves of participants (sometimes as long as 20 waves). It reduces the influence of the seeds on the final sample, increasing the sample’s power to capture hidden individuals in the study population. Long recruitment chains allow for deeper penetration into the target population networks and help to ensure that the sample meets theoretical assumptions behind the claim of it being representative.

14.2.5 Sample size

The most common approach for establishing the sample size for any sampling project is to do a power analysis to determine the size of the sample that can detect the variable of interest. The usual approach for a power analysis assumes a simple random sample which, by definition, has a design effect of one. However, in order to measure the effect of the sampling plan on the variance of the survey estimates, one has to use a proper value of the design effect (see Section 3.6).

RDS sample designs usually have large design effects. To estimate the sample size required in the case of samples collected with RDS, Salganik (2006) proposes selecting a sample size for RDS that is twice as large as the sample size that would be needed under simple random sampling; this amounts to assuming a design effect (ratio of variances in the actual sample and a simple random sample of the same size) of 2.0. Later, simulation models that took into account the homophilia effect and dependence of observations showed much larger design effects, greater than 4 (Goel and Salganik, 2009). It is recommended to do a standard power analysis to determine the equivalent simple random sample size required, and then increase the sample size so determined by a factor of 2-4 in calculating the desired sample size. If similar RDS studies have already been conducted, it may be possible to use those studies to get a more precise idea of the magnitude of the design effects involved.

14.3 Examples of use of respondent-driven sampling

In this section examples are provided from surveys using the RDS method. The examples bring out some practical issues such as issues concerning the formative research needed to set up the study, the initial sample, the number of referrals required from each respondent, controls to avoid the same child being included more than once,
and the structure and amount of reward offered to respondents for participation and making referrals for the next wave.

The examples discussed include the following. It is noteworthy that all studies begin from extremely small (but generally carefully selected) initial samples.

(1) A survey of street children in Bamako and Accra

The study uses two techniques in combination: capture-recapture for estimating the size of the population of street children, and RDS with the aim to capture a proportionate sample to measure the structure and characteristics of that population. The study brings out problems concerning difficulties in moving across boundaries of relatively segregated subpopulations and in achieving a balanced development of the sample (in terms, for example, of distribution by location, gender, or age of the children).

(2) A study of female sex workers in Brazil

This study illustrates how a reclusive population may be penetrated by using the RDS method, in this example achieving a 25-fold increase from the initial sample of 100. It also shows the crucial role of incentive payment in determining the size and tempo of recruitment to the sample. The study also found that the size of personal networks can vary a great deal among different segments of the study population.

(3) A survey of gay and bisexual men, Uganda

The study highlights the considerable amount of formative research needed to set up the study for an extremely reclusive population. The study is exceptional among RDS studies in not offering any incentive, either for participating in the survey or for recruiting for the survey.

(4) Study into community impact of hate-crime, Canada

This is a somewhat unusual application of the RDS method. It seeks to enumerate a sample from a community to study a certain experience of the community, an experience not necessarily personal to the individuals involved. The RDS method is used to encourage reluctant community members to come forward and respond to the survey interviewing concerning the impact of hate crimes in the locality.

14.3.1 A survey of street children in Bamako and Accra

Street children are difficult to study because, as a group, they are rare, mobile and reclusive. Their way of living means that, unlike other children, they are not registered through the household or the school. It is particularly difficult to estimate the total number of street children within a given area.

The description and discussion of the study below is based on Hatløy and Huser (2005).

A. Study method

In this study, two sampling techniques were used and compared to collect information on the street children in two cities, Bamako (Mali) and Accra (Ghana): capture-
14. Respondent-driven sampling

recapture (CR), mainly used to estimate the number of street children; and respondent-driven sampling, mainly to study the characteristics of this population group. The idea was to combine the two sampling techniques applied during a short period of time. After an initial mapping of each city, it was planned that locations would be selected to conduct interviews using RDS, representing the ‘capture’; then a second round of interviews using RDS at the same locations would constitute the ‘recapture’. NGOs that circulate in the cities during night time and worked directly with street children informed that there were fixed places in the city where the street children usually spent the nights. In the RDS, the children were asked about their exact sleeping location the previous night. In order to maximise coverage of possible locations of children, the information collected during RDS was compared with and integrated into information from the other available sources.

The main reasons for the use of the RDS method to reach the street children were to get a sample of street children proportional to the population, and to give accurate characteristics of the street child population. This method is good at drawing representative samples, but it is more difficult to use it as the basis for estimating the total number in the population.

B. Study population and recruitment process

The study population for the RDS was street children, defined as children less than 18 years of age who spend all their time in the streets on their own or with peers. The recruitment process started with some initial subjects, the seeds. The children were told that they would get a reward for responding to the questionnaire, and that they would get an additional reward for recruiting each eligible child for the next wave. The choice and administration of rewards is an important practical factor determining the success of implementation of the RDS sampling procedure. The choice of rewards was made after long discussions with NGO’s and field workers in the study. It was important to find rewards with a certain value for the children, rewards that also increased with an increase in the number of children recruited, but which were not too expensive for the study. Also, logistical aspects had to be taken into account: the rewards had to be easy to store outdoors in a hot and rainy climate, and the stock easy to refill whenever needed. It turned out that the reward for being interviewed was equivalent to half a day’s income for a street child, and that the reward for recruiting six children was equivalent to two day’s income.

Due to time constraints, the survey was limited to four waves (the initial seed, plus three additional waves). For the same reason, the researchers planned to start with a relatively big group of seeds (10 street children), and then try to expand the sample rapidly by asking every street child responding in a wave to recruit (a maximum of) 6 eligible street children for the next wave of the survey. If these numbers could be realised, one would get a sample of nearly 2,600 street children after 4 waves. In practice, it turned out to be difficult to find enough children willing or able to bring in the hoped for number of recruits. The following tables show the sample sizes actually achieved in the two study cities.
C. Achieved sample size

(1) Bamako, Mali

The initial subjects for the survey were selected in a centre for street children that was open only during daytime, and did not offer food to the children. Ten tickets with a unique stamp for the project were given to ten children known to be street children. The total sample size achieved was merely 132 children as follows.

<table>
<thead>
<tr>
<th>Initially recruited</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seed sample: number successfully interviewed</td>
<td>8</td>
</tr>
<tr>
<td>Of these, number who brought at least 1 recruit</td>
<td>5</td>
</tr>
<tr>
<td>Wave 1: number successfully interviewed</td>
<td>25 (5.0 per recruiter)</td>
</tr>
<tr>
<td>Of these, number who brought at least 1 recruit</td>
<td>8</td>
</tr>
<tr>
<td>Wave 2: number successfully interviewed</td>
<td>41 (5.1 per recruiter)</td>
</tr>
<tr>
<td>Of these, number who brought at least 1 recruit</td>
<td>15</td>
</tr>
<tr>
<td>Wave 3: number successfully interviewed</td>
<td>58 (3.9 per recruiter)</td>
</tr>
<tr>
<td>Total achieved sample size</td>
<td>132</td>
</tr>
</tbody>
</table>

Of the total of 132 children, on the first day 8 children were interviewed, the second day 33 children, the third day 65 children, the fourth day 21 children, and then the number subsided, with only 5 new children showing up on the fifth day. This indicated that the study had reached most of the street children that were willing to be interviewed in that area.

Plotting the respondents on a map of Bamako showed that the recruitment had been from the centre of Bamako. Even though the site for interviewing was only a few meters from the bridge crossing the Niger River, practically none of the respondents came from the other side of the river. It was known from other sources that there were quite a number of street children on the other side of the river as well, primarily centred around a bus station. The researchers moved to the bus station to carry out a new RDS. However, due to various practical difficulties, this attempt did not succeed in recruiting any street children. For one thing, there was a misunderstanding in the recruitment procedure. The children were not given clear enough information on which children they should recruit, which caused the children to recruit not only street children according to the survey definition, but also street beggars and school children. This experience shows how extremely important it is to be clear right from the first wave of the recruitment procedure, so that the children understand what kind of children they can recruit.

(2) Accra, Ghana

The street children in Accra were located at many different places around the city, usually near a market, bus station, or train station. In order to reach different groups of children, it was considered desirable to spread out the sample geographically in different parts of the city. Local NGOs working with street children helped to map out the main areas where the children lived, on the basis of which four different locations were selected.
The implementation appeared problematic, as the following results on development of the sample show. The achieved samples turned out to be very imbalanced: samples from two of the four sites contained overwhelmingly only street girls, and one site overwhelmingly only street boys.

D. Observations

The study report offers the following observations.

“Findings from our survey, combined with the experience of local NGOs, suggest that the community of street children in Accra is quite segregated. Children of the same ethnicity and gender, originating from the same regions, stay together in groups. They sleep and socialize in groups; they look after each other and often have a group leader to protect them. The children we interviewed in the first wave tended to recruit other children that were mainly from their own group. At the first location, all four girls in the first wave, were Mole Dagbon from the Northern region and, in the total sample, we got almost solely this group of children. The same pattern seemed to repeat itself at the other locations, although varying somewhat depending on the size of the ethnic group in the area. Only when there were no more children from their own ethnic group did the children start recruiting from other groups. … [Also], the items we offered as a reward became very popular and were perceived as luxury goods, which the children would rarely buy themselves. This was probably the reason that the children kept the opportunity to participate in the survey within their own group.

Another important practical issue is to ensure that the same children are not interviewed more than once. When applying respondent-driven sampling, there is a risk that a child is interviewed more than once. Although it is not possible to eliminate this problem, the study tried to minimize it in several ways. In Bamako, the sample size of children was limited, thus making it possible for the field workers to remember the children and, when in doubt, compare questionnaires. In Accra, the supervisors worked together, each supervisor talked to each child two times, except for the children in the fourth wave – at the child own recruitment, and at the time the child brought in his/her recruits. Another control mechanism can be the children themselves: children often try to make sure that everybody plays by the rules, and if someone tried to cheat, they get upset and try to inform the field workers”.

<table>
<thead>
<tr>
<th>Site</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original target seed sample size</td>
<td>Boys</td>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Girls</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Achieved seed sample size</td>
<td>Boys</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Girls</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Final total achieved sample size</td>
<td>Boys</td>
<td>3</td>
<td>278</td>
<td>3</td>
</tr>
<tr>
<td>Girls</td>
<td>296</td>
<td>33</td>
<td>562</td>
<td>98</td>
</tr>
<tr>
<td>Total</td>
<td>299</td>
<td>311</td>
<td>565</td>
<td>152</td>
</tr>
</tbody>
</table>
14.3.2 A survey of female sex workers in Brazil

In relation to sexual transmission of HIV infection, female sex workers as a group are recognized as a most-at-risk population. In Brazil, the population size of female sex workers has been estimated at 1 per cent of the Brazilian female population 15-49 years of age, that is, more than half a million women. The use of conventional sampling strategies in studies of most-at-risk subgroups for HIV is generally problematic, since such subgroups form a small proportion of the total population, and are associated with stigmatized behaviour or illegal activity.

Respondent-driven sampling was used in the present study involving 2,523 female sex workers. According to the study report, RDS proved appropriate for recruiting sex workers, allowing the selection of a sort of probabilistic sample and the collection of previously missing information on this group in Brazil.

The description and discussion of the study below is based on Damacena, Szwarcwald and Júnior (2011).

A. Implementation of RDS among female sex workers in Brazil

The study, called the ‘Health Chain’, was carried out in Brazil from August 2008 to July 2009, aimed at estimating the prevalence rates for HIV and syphilis and to identify knowledge, attitudes, and practices related to HIV infection and other sexually transmitted diseases in female sex workers. Ten Brazilian municipalities (counties) based on the size and importance of the local AIDS epidemic were chosen for the study. The eligibility criteria were the following: (1) age 18 years or older; (2) being a woman; (3) working as a sex worker in the municipality where the study was being conducted; (4) having had sex in exchange for money at least once in the previous four months; (5) accepting to participate under the study conditions and agreeing to sign the informed consent form; (6) presenting a valid invitation to participate in the study; (7) not having participated in the study previously; and (8) not being under the influence of illegal drugs or alcohol at the time of the interview.

Before the data collection phase, in each municipality a preparatory survey was done to facilitate the study’s implementation at each site according to the characteristics of prostitution in each city. The following issues were addressed: identification of the main prostitution venues, and diversity of the type of female sex workers present in the city, choice of the health service centres for conducting the study, and above all, identification of possible seeds. In the preparatory survey, focus groups and in-depth interviews were held with sex workers and their local representatives, using a previously prepared script and qualitative methodology.

The study consisted of a preliminary interview to verify the eligibility criteria and characterise the respondent’s network and work venue, obtain a self-completed computer questionnaire, and conduct rapid tests for syphilis and HIV. In all ten municipalities, fieldwork was done in health service centres, the interviewers working in teams. The research teams were also responsible for monitoring the networks and encouraging participation by the sex workers. For this purpose, promotional material for the study was prepared and distributed in the main prostitution venues in the cities included in the study.
In each municipality, five to ten initial participants, the seeds, were selected non-randomly. Seeds were chosen with differences in age group, race, socioeconomic class, schooling, and work venues such as the street, nightclubs, saunas, hotels, call services, and others. In addition, this initial focus was on sex workers who were well-connected in the community and had reported extensive social networks during the preparatory survey.

Each seed received three coupons to give to her friends or acquaintances. The women who were invited by the seeds and participated in the study comprised the first wave. After participating in the interview, they received three coupons themselves to invite their friends and acquaintances. This process was repeated until the target sample size was reached in each site.

The coupons were numbered in order to prevent counterfeiting, and to allow identification of the recruitment patterns in the population. The participants were only identified by the identification code on the coupon, consisting of two letters (the municipality’s initials) and a set of numbers indicating the recruits and the wave. During the research project’s implementation, each participant received the primary incentive when she completed the interview, plus an additional reward (secondary incentive) in the form of cash for recruiting each peer that successfully completed the study.

B. Data analysis

The data were weighted using the following question to measure the network size: “How many female sex workers do you know personally?”; the estimation weights were taken to be inversely proportional to the network size so measured.

In addition, since the study was done in 10 municipalities, in order to produce results for the entire sample, calibration was performed according to the relative number of women 15-49 years of age in each municipality, assuming a constant proportion of female sex workers in all the sites studied. This is a major assumption, but no more precise information on the variation of this proportion across different cities was available.

Concerning the possible selection bias described by Heckathorn (2002), in the case of the Chain of Health project, the homophilia effect occurred according to HIV serological status, since participants with HIV infection tended to recruit other HIV-positive participants. These findings indicate a great dependency in the observations, which needs to be taken into consideration in data analysis.

C. Discussion

(1) Incentives for participation and recruitment

In relation to incentives, the primary incentive (for own participation) was used from the very beginning of the study in all the sites. The secondary incentive was not used at the beginning of the study, on the assumption that there would be spontaneous interest in participating, based on the concepts of solidarity and the social networks among the female sex workers. However, in the majority of the participating municipalities, it proved necessary to use the secondary incentive in order to bolster participation in the
study, since the networks were developing too slowly and there were constraints on the time and resources available.

According to the original developers of RDS, while the primary incentive motivates individual participation based on autonomous decisions in the concept of cooperation, the secondary incentive brings the social influence to the surface, which is taken advantage of in the sampling process. However, provision of the secondary incentive has generated controversy. On the one hand, the dual incentive stimulates both participation and recruitment in the study; on the other it becomes a kind of parallel currency in groups with low socioeconomic status, generating duplications in the sample, participation of non-eligible individuals, and even clashes among individuals in the population competing to be nominated by a participating peer for inclusion in the study. In the case of the Health Chain study, the use of the secondary incentive functioned to stimulate and speed up the inclusion of participants, and had the merit of obtaining the participation of more hidden and socially excluded sex workers – many of whom were without the slightest awareness of the importance of their participation, but were only interested in the incentives they were going to receive. However, although the process fostered the use of health services (even if temporarily), individuals’ desire to participate without meeting the eligibility criteria and to improperly receive the incentive payment caused problems, leading in some cases to physical assault against the research team.

(2) Wider social implications

In the application of RDS among female sex workers in Brazil, a link was also made with the ‘social movement’ in the community. In each of the municipalities in the study, focal points, i.e. persons with influence in the community of sex workers, were chosen. These persons helped to publicise the study and accompany the participants with invitations to the health service. The use of focal points helped overcome the distrust and fear among sex workers, who are routinely exposed to violence and social discrimination.

(3) Sample weighting

As for the network size, the study showed that some seeds take a long time to ‘germinate’, while some other types germinate faster, interfering in the sample’s diversity. In addition, due to the homophilia effect, a long network does not always guarantee capturing the diversity of type of individuals in the population segment, something which is essential for the sample to be representative of the target population. Data need to be weighted to compensate for differences in network size.

In the current study the size of the network among sex workers varied according to the work venue, and was smaller among women that worked in the streets as compared to other places (nightclubs, saunas, hotels). The weighting method incorporated stratification by work area.

As noted, a tendency was observed among participants with HIV infection to recruit HIV-positive peers. Considering that the great majority (75 per cent) of HIV-positive participants were unaware of their serological status, this tendency is probably explained by common characteristics in their peer network (rather than by serological status itself), characteristics like age, prostitution site, type of client, social condition, price
charged for sex, drug use, and other factors. From the epidemiological point of view, while dependence of observations helps in understanding the risk factors associated with social networks, from the statistical point of view the structure of dependence complicates the picture and requires special statistical techniques for data analysis.

D. Concluding remark

In short, the use of RDS proved appropriate for studying female sex workers in Brazil, since it facilitated recruitment and allowed the selection of an apparently representative sample in this population subgroup. Also, it succeeded in expanding the initial sample substantially – from 100 seeds to 2,500 achieved interviews. On the other hand, the experience with the implementation of RDS leaves some challenges to be explored, like “improvement of the methodology for statistical analysis of data collected with RDS, adequate data calibration, and interpretation of the social network”, as noted in the study report.

14.3.3 A survey of gay and bisexual men in Kampala, Uganda

HIV/AIDS disproportionately affects gay and bisexual men around the world; however, little is known about this population in sub-Saharan Africa. Surveying HIV/AIDS has been the field with most wide-spread application of the RDS method.

Respondent-driven sampling was used in the study described below to survey a sample of 224 gay and bisexual men in Kampala, Uganda. We have already referred to this study to illustrate in Figure 14.2 the concept of equilibrium achieved after a sufficient number of waves of the survey, and also in Figure 14.4 in the description of the development of the RDS sample in the form of chains from the original sample of seeds. Below we give some further details of the study methodology as an illustration of application of the RDS procedure.

The description and discussion of the study below is based on Kajubi et al. (2008).

A. Study objectives and approach

According to the authors, the aims of the study were to describe the demographic characteristics of gay and bisexual men in Kampala, gauge awareness and level of sexual risk behaviour, foster their inclusion in HIV/AIDS prevention and care programme planning, and also field test survey methods that might serve as a model for epidemiological research and HIV surveillance among MSM in sub-Saharan Africa. The survey was conducted in Kampala, the capital city of Uganda and the country’s largest urban centre with a population of well over one million. The target population was men aged 18 years or older who self-identified as gay or bisexual and were residents of Kampala.

The procedure used to recruit participants for the survey was respondent-driven sampling. The study demonstrates that gay and bisexual men in Uganda are willing to identify themselves and participate in research. The sample was able to capture diverse subgroups. About 60 per cent of these men reported themselves as being ‘gay’ and 40 per cent as ‘bisexual’. Over 90 per cent of them were Ugandan from all parts of society; 37 per cent had had unprotected receptive anal sex in the last six months, 27
14.3 Examples of use of respondent-driven sampling

per cent were paid for sex, 18 per cent paid for sex, 11 per cent had history of urethral discharge. Perception that gay and bisexual men are at risk for HIV infection was low among the participants.

B. Formative phase

Sometimes it has been claimed as a strength of RDS that a lot of formative research is not needed to be able to access diverse, hidden networks of the target population because members of the population themselves recruit each other for the study. Such is not the case, however. A focused formative phase is needed to assess the feasibility of RDS and guide the conduct of the survey (see Section 14.2.1A). In the present study, the formative phase goals of the study were therefore seen as gaining a basic understanding of the diverse networks of the population in terms of demographic characteristics, assessing the willingness of gay and bisexual men to participate in the survey, identifying initial seeds to start recruitment chains, recruit gay and bisexual men to be trained as interviewers, and deciding on practical issues such as the use of incentives, where to field test the questionnaire, and where to conduct interviews.

The researchers gained initial contact through websites targeting gay and bisexual men in Uganda. They also contacted a local member of the clergy whose name appeared in the popular media in connection with providing counselling services to gay and bisexual men. They discussed issues of gay and bisexual men with persons providing HIV prevention and care services to the general population. Key informants included seven AIDS service providers. Appointments were made with key informants at a location of their choice and semi-structured interviews covering the study goals were conducted.

According to the study report, “the key informants played a really key role!”. There was an initial debate on whether to use the respondent-driven sampling method, or the time-location sampling method which entails mapping the venues where gay and bisexual men congregate and sampling random time periods for recruitment at the venues (see Chapter 10). It is noteworthy that the researchers selected the RDS design during the formative phase only after consulting key informants on whether they, the key informants, preferred RDS or time-location sampling, after the two basic approaches were explained to them. Key informants indicated they preferred RDS for several reasons: (i) gay and bisexual men would be better at knowing who are the other gay and bisexual men; (ii) the researchers would be able to reach men whom they would not be able to reach at venues; (iii) better participation can be expected with RDS; and (iv) because the RDS method would better preserve confidentiality of the few venues in Kampala where individuals in the target population felt safe. Key informants were also asked for other persons and members of organizations connected to gay and bisexual men in Kampala to serve as key informants or seeds; they also identified a few locations (e.g. bars, restaurants, dance clubs) where gay and bisexual men and women congregated. The researchers conducted informal observations of these venues and interviews with venue owners, employees and patrons. Altogether, 20 key informants were interviewed, including gay and bisexual businessmen, professionals, students, unemployed youth, and Africans living in Kampala who were non-Ugandan nationals.
The formative phase of the study also succeeded in identifying seeds in basic demographic groups and in hiring gay and bisexual interviewers for the study.

Other decisions made from the formative phase information included not offering an explicit monetary incentive for participation, using a mobile phone number to have participants call for appointments, conducting interviews at any location of the participant's choice, not conducting HIV testing in this first study, keeping the length of the questionnaire to a minimum, and choosing which variables to track to monitor when the sample achieved equilibrium.

C. The initial sample

Six self-identified gay men and two bisexual men among persons referred by key informants were recruited to serve as seeds. These eight seeds were chosen to cover different ethnic groups and socio-economic and educational levels. All eight identified seeds succeeded in starting recruitment chains. In the present study, seeds were removed from the data for the adjusted analysis.

D. Follow-up recruitment to the sample

After completing the study questionnaire, the initial seeds and subsequent participants were given three recruitment coupons and instructed to give them to other men they knew personally and knew to be gay or bisexual. Persons who presented themselves with a coupon and who were eligible and consented to participate, were enrolled. After their participation, they in turn were given three recruitment coupons.

Recruitment was flexible to the needs of the potential participants. Potential participants called a mobile phone number printed on the coupon to schedule their screening at any location comfortable to them. This was often a restaurant or bar well known by gay and bisexual men, but participants also designated intersections or neighbourhoods to walk and talk. At or near the location, a private area was found for interviewing. Men wishing to participate were screened for eligibility by age (18 years and above), being gay or bisexual (in response to the question: “Would you describe yourself as gay, bisexual, straight, or heterosexual?”—the latter two were excluded), and being in possession of a coupon indicating they were recruited by another participant with whom they were acquainted.

No personal identifying information was collected in order to protect confidentiality.

The study did not offer an explicit ‘primary’ incentive for survey participation. Nor did it provide an ‘secondary’ incentive to recruit others, which is common in other RDS studies.

While it may have been impossible to eliminate every person who was faking being gay or bisexual, the following additional precautionary steps were taken.

First, as mentioned, the study did not advertise or offer monetary incentive to participate. Second, the researchers carefully instructed the initial seeds and each subsequent participant to recruit only men they personally knew to be gay or bisexual and explained why this was important. Third, the interview team included gay and bisexual men whose experience provided insight as to who was or was not gay or bisexual, including gauging the potential participants’ knowledge of and comfort with
terms in common usage by gay and bisexual persons. Finally, to reduce the potential that persons would participate in the survey multiple times, all potential participants were screened by the principal investigator.

14.3.4 A survey to research into the community impact of hate-crimes in Canada

This study, by Fashola (2010), represents a different type of application of the RDS method. Here the target was not a group in the population with certain characteristics or conditions, but the community which had experienced some trauma or undesirable incidence. This application shows the versatility of method.

The main purpose of this research was to understand the impact of hate crimes on different communities—geographic as well as ethnic, racial or cultural ‘identity communities’. The research design involved two case studies at sites where an allegedly hate-motivated crime had occurred.

The first case study was a violent attack on a Sudanese refugee by a group of about 10 men at Victoria Park in Kitchener, Ontario, in 2006. The second case study involved the assault, by two men, on a Chinese-Canadian male victim who was fishing near the Mossington Bridge on the Black River in Sutton, Georgina. This particular incident was one of a series of attacks against Asian-Canadian anglers in the area.

Data were collected at the two sites where the incidents took place. At each site, two main communities were selected for data gathering. At the first site, data were collected from the “African identity community” (individuals from the racial/ethnic community of the victim) and the “Kitchener geographic community” (individuals living in the Kitchener, Ontario, region). At the second site, data were collected from the “Chinese identity community” (individuals from the racial/ethnic community of the victim) and the “Georgina geographic community” (individuals living in the study region).

A survey was administered to the geographic and identity communities. After describing the allegedly hate-motivated incident, the survey asked a number of questions about the impact of the event on the community. RDS was chosen as a sampling method for this study because the racial/ethnic identity communities were groups for which no exhaustive list of all their members was available for the purpose of random sampling. Stratified sampling was used to generate a statistically reliable sample for the geographic communities. The RDSAT (RDS Analytical Tool) software specially developed for analysing RDS-based data was employed for analysis because “it allows researchers to make population prevalence estimates with confidence intervals” (RDSAT, 2008).

Using RDS recruitment, the study began with five seeds in each of the two identity communities — the African identity community and the Chinese identity community. Each seed was given four coupons with which to recruit four new participants from his/her social network. The four referred participants received four coupons each and were expected to refer four more participants with the coupons. There were a total of five recruitment waves in the sample. As a reward for their participation, respondents were given the option to enter into a raffle draw for a prize.
Generally, a 95 per cent confidence level is considered statistically reliable. In most cases, this would entail collecting a sample of about 400 for each identity community. Using RDS to collect a sample of 400 for each identity community proved to be a challenge because many participants were unwilling to provide contact information for their friends. There also appeared to be little motivation for them to contact their friends on behalf of the researchers. In total, there were 196 survey respondents from the African identity community in Kitchener and 288 survey respondents from the Chinese identity community of the Greater Toronto Area.

The study report notes the following in conclusion:

“RDS is a sampling method utilized in instances where researchers are attempting to study hard-to-reach populations. RDS combines “snowball sampling” with a mathematical model that weights the sample in such a way that eliminates some of the biases that may have been introduced into the sample by the non-random choice of initial recruits. Less biased prevalence estimates can then be produced and confidence intervals can be constructed around those estimates.

RDS is a relatively new sampling method, and as its use increases, so will researchers’ familiarity with its possibilities and limitations. In this study, RDS did not eliminate some of the challenges in recruiting victims of crime; there remained issues where potential participants did not trust the researchers or just did not want to talk about the incident. As such, recruitment was difficult. Using RDS, however, contributed to increasing the statistical reliability of the sample compared to using a [normal] snowball sample. It also made possible to draw conclusions from the sample about the impact of hate crime on communities of identity with a greater degree of statistical reliability. This, in and of itself, suggests that RDS merits further attention as a sampling method when trying to reach hard-to-reach populations”.

14.4 Statistical modeling of the RDS procedure

14.4.1 Basic estimation procedure

A. Basis: Hansen-Hurwitz estimator

The estimation procedure involves the following basic steps. The procedure is elaborated in Volz and Heckathorn (2008). According to those authors, estimation from respondent-driven sampling has as its basis the Hansen-Hurwitz (1943) estimator, written as

\[
\bar{y} = \frac{\sum_i (y_i/p_i)}{\sum_i (1/p_i)}.
\]  

(14.1)

where

\( \bar{y} \) Mean of variable \( y_i \) over selections \( j \) comprising the sample

\( y_i \) The value of some variable for selection \( j \)
14.4 Statistical modeling of the RDS procedure

The draw-wise selection probability of \( j \)

\[
\sum_j
\]

Sum over all \( j \) in the sample

The Hansen-Hurwitz estimator is based on the draw-wise selection probabilities, i.e. on the probability of selection of a unit at a particular draw \( j \), assuming with-replacement sampling. For a given unit, this probability is assumed to remain the same across selections. The number of links a unit has with other units in the target population is its degree \( (m_j) \). The procedure assumes that the (draw-wise) probability of selection of a unit is proportional to its degree:

\[
p_j \propto m_j. \tag{14.2}
\]

This implies assuming that, apart from the effect of the unit’s degree, the unit selection probabilities are uniform. Equations (14.1) and (14.2) give:

\[
\bar{y} = \frac{\sum_j(y_j/m_j)}{\sum_j(1/m_j)}. \tag{14.3}
\]

This estimation procedure is based on a set of assumptions about the sample selection procedure of RDS. In particular, the use of equation (14.1) assumes that the sample is selected with replacement, or that the sampling rate is small so that the chance of the same unit being selected more than once is negligible.

In practice, this assumption is frequently violated. One reason is that often the RDS procedure is applied for the study of special, relatively small populations. Secondly, in order to ‘go deep’ into the population, it is necessary to carry on the RDS procedure over many waves. This process may be terminated only when it has ‘exhausted’ the population: when additional waves bring in no (or very few) new cases. This is likely to happen when the sample already forms a significant part of the population.

B. In the form of multiplicity estimator

It should be noted that estimator (14.1)-(14.3) is simply the multiplicity estimator which underlies several of the sampling procedures described in other chapters – such as in multiplicity sampling, multi-frame sampling, adaptive cluster sampling, and time-location sampling. For this reason, it is useful to examine the estimator a little further.

Equation (4.3) is not affected by the scale of measurement of \( m_j \). However, the scale has to be taken into account in using its numerator and denominator separately. Let \( f \) be the constant of proportionality in Equation (14.2):

\[
p_j = f m_j. \tag{14.4}
\]

where \( p_j \) is the actual selection probability of \( j \), and \( m_j \) is its degree – which is the same thing as its multiplicity, introduced in Chapter 7.

A population aggregate is estimated by

\[
\hat{Y} = \sum_j(\frac{y_j}{p_j}) = \sum_j(\frac{y_j/f}{m_j}) \tag{14.5}
\]

and population size by
\[ \hat{N} = \sum_j (1/p_j) = \sum_j \left( \frac{1/f}{m_j} \right) \]  \hspace{1cm} (14.6)

Parameter \( f \) is the constant sampling rate applied to all units, except for the units’ multiplicity \( m_j \). Equation (14.5) is identical to the multiplicity estimator (4.13), except for being a special case with \( f_j = f \), a constant. As before, index \( j \) stands for a particular selection of a given unit. Multiplicity \( m_j \) is in principle the same for all selections of a given unit. However, if it differs between selections of the same unit due to measurement error then, in line with the usual procedure, for each selection the measured value corresponding to that particular selection can be used.

### 14.4.2 Assumptions of the RDS model

The estimation of unit selection probabilities \( (p_j) \) from the units’ degrees or multiplicities \( (m_j) \) is the foundation of the estimation procedure. This assumes a number of characteristics of the RDS model and its application. Volz and Heckathorn (2008), the authors of the procedure, list the following.

1. **Degree.** Respondents accurately report their degree in the network.

2. **Recruitment is random.** When recruiting others, respondents select uniformly at random from their personal network.

3. **Reciprocity.** Network connections are reciprocal. Respondents recruit those with whom they have a pre-existing relationship, such as acquaintances, friends, and those closer than friends. Such connections are reciprocal, e.g. my friends and acquaintances consider me to be a friend or acquaintance. Consequently, in network theoretic terms, the potential recruitment network is undirected, so if respondent \( a \) can recruit \( b \), then \( b \) can also recruit \( a \). …

4. **Convergence.** Recruitment is modelled as a Markov process (MP), where the state of the MP is the last individual recruited. …. We assume that the MP is irreducible and that each state has a finite return time. Therefore, a unique equilibrium to the MP exists and recruitment rapidly converges to this equilibrium. The implication is that after a modest number of steps, the sample composition becomes independent of the initial respondents who initiated the chain-referral process. The irreducibility condition is equivalent to the condition that the social network is well-connected; that is to say, every node [unit] can be reached by a finite path from any other node [unit]. Furthermore, our social networks are assumed to be finite (though very large), so the expected return time must be finite as well”.

The authors continue:

“On the surface, the irreducibility assumption [that the social network is well-connected] may seem unrealistic, especially for large populations, where it is most likely that some units will be isolated from the network as a whole. This, however, is usually not a cause for concern. It is known from random network theory that most networks possess a so-called giant component, a subset of nodes such that a network path exists between any two and which occupies a non-vanishing fraction of the network as the population size goes
14.4  Statistical modeling of the RDS procedure

to infinity. The giant component usually encompasses the vast majority of the population, so long as some basic conditions are met. For instance, in pure random graphs, the giant component will consist of 99% of the population if nodes have just 5 links on average. RDS studies have typically exceeded this margin comfortably. … With that said, field RDS studies should come with the caveat that statistical inference is limited to the giant component, rather than the total population. But provided the giant component is very large, this is usually a minor distinction.

Furthermore, research on the small-world problem (Milgram, 1967; Watts, 1999) has led to the observation that almost all social networks have very short mean path length. Consequently, there are relatively few intermediaries between any two randomly selected individuals in most social networks. … [It is] likely that the selection probability for any individual in the network will stabilize after just a few recruitments, as almost anyone in the population can be reached in a small number of steps.

Another assumption commonly called into question is that respondents recruit uniformly at random from their network neighbours. Indeed, it is difficult or impossible to enforce random recruitment among respondents, and in many cases respondents may have special reasons for selecting a particular recruit. However, non-random recruitment, if it occurs, will not necessarily bias our estimator. As long as recruitment is not correlated with any variable important for estimation (e.g. the study-variable or degree), the aggregate effect is for recruitment to appear uniform-random. …"

The above assertions, while justified to an extent, are over-optimistic. See discussion in the examples given in Section 14.7 below.

Furthermore, RDS requires rigorous adherence to implementation and analysis requirements which are sometimes ignored, resulting in studies claiming to use RDS when in fact they are merely ordinary snowball samples (Johnston et al., 2008).

14.4.3 Estimation of unit selection probabilities and sample weights

With RDS field procedures, correct analysis can account for certain known biases found in most snowball type of sampling. To this end, two essential pieces of data must be acquired during data collection: (1) the social network size of each participant; and, (2) connections between recruiters and their recruits. Social network size data are used to weight estimates to account for over representation of those with larger social network sizes (ability to recruit more persons) and vice versa. Recruiter-recruit connections are used in a mathematical modelling of the recruitment process to generate relative inclusion probabilities (Heckathorn, 1997; 2002).

Theoretically, the suggested number of waves for reaching equilibrium is six, since according to Watts (1999), all the individuals are connected to each other through only half a dozen links. However, if the homophilia effect is large, the sampling process needs to consist of a larger number of waves in order to reach Markov equilibrium (for given variables) and the minimum estimated sample size.
According to Markov chain theory, if there are a fixed number of states or conditions, one can calculate the probabilities of moving from one state to another. In this way, future states depend only on the present state, and are independent of the history of the states through which the unit has passed.

Furthermore, according to the ‘law of large numbers’, the probability of the process in any state over the course of a large number of steps will be independent of the state in which it began. In the case of recruitment in the RDS method, this means that memory of the recruitment occurs wave-to-wave, that is, the characteristics in the recruited individual depend only on the characteristics of his/her direct recruiter, and not on those of the recruiter’s recruiter or of any participant in previous waves. After a sufficient number of waves, the characteristics of the individuals in the final sample are independent of the seed’s characteristics. In other words, Markov chain theory suggests that if peer recruitment occurs through a sufficiently large number of recruitment waves (called long chains), the representation of the population within the sample will stabilize and further recruitment waves will not change the sample’s representativeness by a significant amount. This process is called reaching equilibrium, as noted earlier. The procedure can be used to calculate weights for estimation from the sample, aiming at overcoming bias in the initial sample introduced by the non-random choice of seeds from which recruitment began. Within this framework, supposedly unbiased prevalence estimates for the population of interest can be produced and confidence intervals can be constructed around these estimates (Salganik 2006).

Data collected with RDS can be weighted according to an expansion factor based on the inverse of the selection probability. Under the hypothesis that (i) the seed is selected with probability proportional to her/his network size, and that (ii) participants randomly recruit their peers from their network of friends and acquaintances, Salganick and Heckathorn (2004) demonstrated that the probability of selecting each participant is also proportional to his/her network size, i.e. to the number of acquaintances known in the target population. Subsequently, the authors are able to relax assumption (i) on the basis that using a Markov chain argument it can be shown that “the estimates are asymptotically unbiased no matter how the seeds are chosen”.

It is helpful to clarify the above model in concrete, albeit in simplified, terms by the following example.

Figure 14.6 shows a hypothetical network structure consisting of points or ‘nodes’ (A-M ... X-Z), with links (‘paths’, ‘arcs’) as shown in the diagram.

(1) Some destination, say X, may be reached from a starting point, say A, by several different routes, in this case as shown in the figure.55

(2) Suppose that each participant is asked to choose one person for inclusion in the follow-up wave, then the model assumes that the choice is made at random, with equal probability, from the existing links in the chooser’s network. For instance, starting from A, one of the 4 links (A-B), (A-C), (A-F) and (A-X) will be chosen at random with equal probability (1/4). Similarly, the choice from C is (C-A), (C-B), (C-D), (C-E) and (C-F), chosen at random with equal probability (1/5); the choice from B is between (B-A) and (B-C) with equal probability (1/2).

55 By definition, a path cannot pass through a point more than once.
14.4 Statistical modeling of the RDS procedure

Figure 14.6. Illustration of a simple network

(3) In practice, repeated recruitment of the same units is not allowed, even though that is assumed in the theoretical model. For instance, if one arrives at A from B (rather than starting from A as a seed), then there are only 3 available links to proceed further from A: (A-C), (A-F) and (A-X). Similarly, arriving at C from A via B, the available links at C are only (C-D), (C-E) and (C-F).

(4) In the diagram, there are two clusters, (A,B,C,D,E,F,X) and (G,H,I,J,K,L,M,Y,Z), with many connections among units within each cluster, but very limited connections between the clusters (only D-Y in this illustration). Having all the seeds confined to only one of the clusters makes it unlikely (but not impossible) that the sample will reach to include any units from the other cluster. Hence it is very desirable in practice to identify and try to represent at least the main clusters in the initial sample of seeds.

The theoretical model behind the method makes the following three important assumptions.

(5) The network is connected (just as in the illustration on Figure 14.6), meaning that one can reach any point starting from any other point in the network.
With a sufficiently large number of steps (waves), the probability of reaching any particular point is stabilised at a value independent of where the process began, and of the details of the path followed in reaching that point.

This probability is proportional to the number of paths leading to the point being reached (its degree or multiplicity). Assuming symmetry in relationships (links), this equals the number of paths leading out from the point concerned. For instance, for A, the probability (in relative terms) is 4, for B it is 2, and for C it is 5.

### 14.4.4 Analysis tools

Since RDS is a complex and relatively new sampling method, the appropriate statistical techniques for analysing the collected data are still being developed. For example, over time, two distinct estimators have been proposed for estimating prevalence rates, which have been discussed in the recent literature on RDS (Volz and Heckathorn, 2008). While the point estimators are quite similar, some questions remain about the best way to estimate variances. The procedure tends to be subject to very large design effects (for example, see illustration in Section 14.7.3 below). It is necessary therefore to have more information on design effects.

RDS is a method based on various statistical assumptions about the probabilities of reaching a representative sample of the study population. Despite its potential to produce probabilistic samples, new applications of this relatively recent procedure still require reconsideration and re-evaluation of details of the analysis procedure.

The analysis tools which have been developed for RDS-based data are an important advantage of the procedure. However, the complexities in some cases can make their application difficult, especially for researchers who may not be well-versed in advanced statistical methods. We may note that there is an RDS statistical software package available for analysis of RDS-based survey data, called the RDS Analytical Tool (RDSAT, 2008). However, at its current state of development, the application presents important limitations, like its inability to perform multivariate analyses. Also, it needs to be updated for the new estimation procedure developed in Volz and Heckathorn (2008).

### 14.5 Note on alternative approaches

This section briefly describes two alternatives to or variants of ordinary snowball and respondent-driven sampling. These various approaches to link-trace sampling all have their merits and limitations. Their relative success depends not only on the objectives and circumstances of their application but also on the details of how they are actually applied.

#### 14.5.1 Targeted Sampling

Targeted sampling calls for the collection and use of quantitative data (e.g. survey or administrative data) and qualitative data (e.g. ethnographic interviewing) in selecting the sample. For instance, existing data from public or private agencies may be used to
delineate geographical boundaries and to describe the target population. These data, supplemented by qualitative data, may be used to develop the sampling frame of locations where the target population is found, and to characterize the population in terms of its characteristics. Specific locations for sampling can then be selected randomly, and interviewers sent to recruit and interview eligible respondents. We draw on Watters and Biernacki (1989) for the following characterisation of the target sampling method.

Targeted sampling is a procedure requiring extensive formative research and complex ethnographic work. In order to identify the locations where the study participants will be recruited, it is necessary that professional outreach workers and experienced ethnographers map the areas based on field observations and interviews with key informants from inside and outside the subgroups which will be studied.

The ethnographic data collected is one of the strengths of targeted sampling. It provides a characterisation of the population, a better knowledge of the socioeconomic determinants of the situation, and key information about personal situations and motivational factors associated with the variables studied. In this context, recruitment is not only a matter of selecting members to participate in the study; it also involves a deeper exploration of the individuals, their networks and their communities. Although this method does not produce unbiased samples and its external validity is very low because generalization of the results is limited to the population actually reached through the recruitment process, targeted sampling could be the choice of preference for research aimed to design and develop social interventions.

The quality of targeted sampling depends on the accuracy and comprehensiveness of ethnographic mapping and the availability of quantitative data. Targeted sampling has several limitations. It requires experienced staff. A prolonged period of data collection and analysis may be needed before recruiting and interviewing can begin. Because administrative data are collected for other aims, they may not be sufficiently precise and pertinent for characterising the target population. For example, poor and minority populations tend to be overrepresented in police statistics as a result of prejudice on the part of the authorities. Also, data may be reported for different geographic or administrative areas than those of the study. The sample obtained is difficult to replicate, and interviewer and self-selection biases are common. For example, security concerns may exclude the recruitment of areas considered unsafe. Additionally, there are often practical and logistic problems. For instance, the presence of project staff may cause curiosity as well as defensive reaction or hostility among members of the community in the data collection area. Poor weather conditions affect recruitment negatively, and the fact that recruitment and interviews are conducted in the field makes it difficult to keep privacy and avoid interferences. Properly addressing these situations usually implies extra costs (such as the need to hire private transport facilities, or interviewers working in teams simply for reasons to do with security). Sometimes locations initially identified as high yield recruitment areas fail to provide the expected number of eligible participants, which results in having to spend more time to recruit the planned number of subjects or the moving to other previously selected areas of recruitment. Generally, such practical difficulties, present in any survey, are more common and serious in surveys using the targeted sampling methodology. Finally, it is not usually possible to represent geographical areas and other population groups proportionately in the
sample, or to know what the proportions are. Hence probabilities of selection generally remain unknown. All of these factors influence the validity or generalizability of the data. Nevertheless, targeted sampling is clearly better than convenience or other forms of nonprobability sampling because it ensures the inclusion of persons with varied characteristics.

Aldana and Quintero (2008) provide the following summary of relative advantages and limitations of targeted sampling.

<table>
<thead>
<tr>
<th>Probability sampling</th>
<th>No. It does not provide unbiased population estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>External Validity</td>
<td>Limited</td>
</tr>
<tr>
<td>Ethnographic and formative research</td>
<td>Extensive and complex. Can provide valuable information on socioeconomic determinants of risk behaviour.</td>
</tr>
<tr>
<td>Recruitment</td>
<td>Professional outreach workers familiar with the target population are recommended.</td>
</tr>
<tr>
<td>Working conditions</td>
<td>Normally interviewing is conducted in the field (outdoors, in the street). Lack of privacy. Safety concerns. Weather conditions can affect recruitment and interviewing.</td>
</tr>
<tr>
<td>Cost efficiency</td>
<td>Can give useful – sometimes unique - results, but it requires more staff-time per interview than alternative methods.</td>
</tr>
<tr>
<td>Need for further research</td>
<td>Yes. Current experience is limited and tends to be situation-specific.</td>
</tr>
</tbody>
</table>

### 14.5.2 Indigenous fieldworker sampling

The following remarks are based on Platt et al. (2006).

Over the last 15 years, the ‘indigenous fieldworker sampling’ (IFS) procedure has become the established sampling method for recruiting hidden populations of injection drug users (IDUs) and sex workers internationally. Typically, such studies have employed ‘privileged access interviewers’. These *indigenous field workers are interviewers who are either current or former drug users or individuals who have experience working with drug users and have privileged access to IDU networks*.

The IFS recruitment method uses a standard chain referral approach. Indigenous field workers undergo training covering the aims of the study, fieldwork protocols, ethics, informed consent, interview skills and safety procedures. Field workers (FWs) identify individuals known to them from IDU networks, recruit them, and then interview them in community settings, separately from the rest of the research team. Eligible participants are given an incentive to take part and also asked to introduce their peers to the FW. The use of multiple site and network recruitment ensures a wide coverage of the population. There is some evidence that the use of FWs with direct access to IDU social networks facilitates recruitment and reduces masking (under-sampling reclusive respondents), volunteer bias (oversampling cooperative respondents) and underreporting of socially undesirable behaviour. Basic features of the IFS and RDS approaches are compared in the table below.
14.6 Assessment: comparisons with alternative sampling approaches

The primary difference between IFS and RDS is that in the IFS the onus for a host of activities - recruitment, testing eligibility of the recruits, follow-up of recruits in the field for the interview, and controlling the data collection procedure – is on the fieldworkers. The role of researchers is to work out and specify the study parameters, provide guidance and advice on its implementation, and undertake quality checking and analysis of the data. By contrast in RDS, the researchers select the seeds. The seeds and subsequently previous respondents do the recruiting, generally without involving the researchers or fieldworkers. Then at each wave the researchers themselves test for eligibility, and often also conduct the interview. The interviews are conducted at chosen locations where the recruits present themselves.

### 14.6 Assessment: comparisons with alternative sampling approaches

In the following, we will present some experience aimed at assessing the performance of the RDS method. Two modes of assessment are considered. First, in this section we describe case studies which used RDS as one among several sampling methods employed. This permits a comparison of the performance of RDS against alternative procedures. Then, in Section 14.7 we provide some examples of how far implementations of the RDS method have conformed to the theoretical assumptions on which the approach is founded.

In this section the following two examples are discussed.

**(1) Comparison of three approaches in a survey of “men who have sex with men” (MSM) in Brazil.**

The empirical study compares time-location sampling of venues, ordinary snowball sampling, and RDS. Both the snowball and RDS are chain-referral procedures; the difference is in how the chains are developed and implemented. The study experience concludes RDS to be the clear favourite in the particular case.

**(2) A comparison of RDS with ‘indigenous fieldworker sampling’ (IFS), in Russia and Estonia.**

As noted in the previous section, IFS is an alternative chain-referral method, but using a different strategy to develop and implement the sample. In this method, carefully selected local fieldworkers play an active role in recruitment to the sample (supplementing referrals by past respondents with their own referrals), and conduct the interviewing in the field. The study concludes that neither of the two methods, RDS...
and IFS, is clearly superior overall, so that a preferred approach would be to adopt the best aspects of both methodologies.

**14.6.1. Comparison of different approaches in a survey in Brazil**

Below is an evaluation of RDS compared to some alternatives, drawn from a study by Kendall *et al.* (2008). The study compares snowball sampling, time location sampling (TLS), and respondent-driven sampling (RDS) applications to surveys of MSM in Fortaleza, Brazil.

**A. Background**

The focus of the survey was on measuring the socioeconomic status of MSM and their risk-related behaviour to others with AIDS/HIV and to the general population. The HIV behavioural surveillance methods used to sample MSM in the study changed over a short period of time from snowball sampling, to TLS, and then to RDS. The date, sample size and the methodology used in the four rounds conducted over a decade were as follows.

<table>
<thead>
<tr>
<th>Round</th>
<th>Survey year</th>
<th>Sample size</th>
<th>Method used</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>1995</td>
<td>400</td>
<td>snowball</td>
</tr>
<tr>
<td>(2)</td>
<td>1998</td>
<td>200</td>
<td>snowball</td>
</tr>
<tr>
<td>(3)</td>
<td>2002</td>
<td>127</td>
<td>snowball</td>
</tr>
<tr>
<td></td>
<td></td>
<td>274</td>
<td>TLS</td>
</tr>
<tr>
<td>(4)</td>
<td>2005</td>
<td>406</td>
<td>RDS</td>
</tr>
</tbody>
</table>

For all surveys, the eligibility criteria were the same: male, aged 14 years and older, and reporting oral or anal sex with other men during the past 6 months. Interviewers were selected from self-identified MSM and trained by the investigators.

**B. Sampling methods**

**(1) Conventional snowball sampling**

Through formative research, venues around the city where MSM meet were mapped. These included bars, cinemas, streets, beaches, public squares, and bathhouses. Some members of the MSM population at these venues were approached, their consent obtained, and then interviewed. They were asked to identify others to be recruited to the survey. Participants among the latter were asked in turn to identify other members, and so on, until reaching the target sample size.

**(2) Time-location sampling**

TLS in Fortaleza began with an ethnographic mapping of the different venues and establishments targeting or serving the MSM population and where MSM were known to congregate. The 2002 survey round updated maps and attendance patterns of venues prepared for the earlier MSM snowball surveys. The mapping generated a
sampling frame of venues-and-time-periods of MSM attendance. The sampling frame was divided into ‘venue-day-time’ (VDT) increments that formed the unit of random sampling. The VDT roster took into account the volume of patrons by the time of the day, day of the week, and type of venue and hours of operation. To identify appropriate time divisions in these venues, the researchers visited these venues and interviewed key informants recruited from two collaborating NGOS and their contacts. They identified venues which were more crowded and attracted a wider range of clientele during the weekends as opposed to weekdays. The final list of venues was classified into day/night, weekday/weekend, and holiday/non-holiday types. These time-venues were then selected randomly from the list. A final roster of 54 venues and their associated VDTs was catalogued for inclusion in the survey and included bars, nightclubs, cinemas, bathhouses, and other ‘cruising’ areas (e.g. streets, beaches, public squares). To achieve a sample size of 400, 40 VDTs and 10 MSMs per VDT were randomly selected. In bars and nightclubs, patrons were given pieces of paper sequentially numbered as they entered the venue. In bathhouses, the lockers in use were assigned sequential numbers. At the end of the first hour in the field, ten numbers were selected at random using a random number table. Persons who held those numbers were invited to participate in the survey. Interviews were conducted in a private area in or near the premises. If there was a lack of privacy or other obstacle (e.g. loud noise), the participant could make an appointment to complete the questionnaire at another location of his choice. Twenty-four interviewers were trained and worked in pairs.

(3) Respondent-driven sampling

In 2005, RDS was used to recruit 406 MSM for behavioural surveillance. Common details of the RDS methodology have been described in Section 14.1; specific points relating to the present study are noted below.

A brief formative research was conducted for one week prior to initiation of RDS, to explore the appropriate incentive level, the selection of a brief HIV prevention intervention that would be delivered at the interview site, and the acceptability of being interviewed in the study offices at the project site.

MSM were chosen from the clients and staff of two local NGOs serving the MSM community to serve as the initial seeds. A total of 10 MSM seeds were enlisted - five from each NGO to represent five different levels by socio-economic status (SES). In contrast to conventional snowball sampling, recruitment was limited to up to two MSM per participant in order to limit any individual’s influence on sample accrual. Each seed received two uniquely coded coupons, identified by a different colour for each NGO, to be used to recruit other MSM. Eligible MSM who presented themselves with a coupon were enrolled, and in turn given two recruitment coupons until the sample size was reached, targeting approximately 200 for each NGO. By way of incentive, each individual received US$5 for his own interview (US$7 in the last wave when further recruitment stopped), and a further US$5 for each participant interviewed whom the person had recruited. Interview sites were located in two buildings, with easy access using public transportation. Participants could drop-in or phone to make an appointment.

As a variation to the usual RDS practice, the study included two different approaches in instructing participants how to recruit. At one NGO, MSM were recruited by participants
in the traditional RDS format, that is, by allowing them to recruit any MSM in their social network. At the other NGO, participants wrote a list of names of MSM who were part of their circle of relationships. These names were placed inside of an envelope, the interviewer randomly drew two, and the participant was instructed to invite these two MSM to participate in the study. Examination of the data found no substantial differences in the makeup of the samples using these two techniques, supporting the assumptions of the traditional RDS format of leaving the selection of the recruits to the participants.

Additionally, information on linkages between recruiters and recruits and the size of participants’ social MSM networks were collected to statistically adjust for recruitment biases. Relative social network sizes affect a person’s probability of being recruited into the study and their ability to recruit others. In the study, the question on social network size translates from Portuguese approximately as follows:

“How many men do you know in person, who have had sex with other men and who know you and you have seen at least once in the past 6 months?”

C. Relative performance of RDS procedure

This comparative empirical study of the three sampling methods applied to MSM in Fortaleza, Brazil, highlights one advantage of RDS: the long-chain recruitment method can be effective in reaching very inaccessible segments of a hidden population. A major difference in the samples obtained by the different recruitment methods was the socio-economic level of the participants. The snowball and TLS samples included over 50 per cent of MSM in the highest social classes, compared to only 3-4 per cent in the RDS sample. As the study report notes,

“the RDS survey was closer to the social class structure of Brazil, Fortaleza and MSM AIDS cases of Ceara’. Identifying and recruiting at venues where risky sex takes place may be advantageous from a prevention or intervention point of view, but for surveillance purposes and understanding the full impact of the HIV epidemic, the major concern is reaching all segments of the target population. Given the observed disparities of AIDS cases in Brazil, it is particularly important to reach MSM in the lower SES classes”.

Hence it is fairly clear that data collected through snowball sampling and TLS at venues overrepresented men having the financial resources to attend these venues. The primary reason why RDS achieved a more diverse sample with respect to SES is, perhaps, that the long-chain recruitment is geared toward expanding recruitment across many interconnected networks, and is not limited to the most visible parts of the MSM population.

But there are also other factors which may operate to introduce bias in the other direction:

“The location of the interviews for the RDS round, two NGOs sites located in Fortaleza’s inner city, may have been a barrier to higher SES MSM. This area near the central market has congested traffic, limited parking and increased crime rates compared to other areas of the city. In contrast, the TLS and
snowball survey interviews were conducted by appointment at a site of the respondents’ choosing.”

In addition,

“the monetary incentives for participation in RDS may have had a much stronger appeal to the lower SES MSM than higher SES MSM. … Tellingly, our key informants expressed concern about the overrepresentation of the upper social classes in earlier survey rounds [rounds which used different methodologies], but also about their under-representation in the RDS round”.

There are trade-offs in methods of sampling hard to reach populations. Efficiency in recruitment is often pitted against scientific rigor, such as achieved through probability-based sampling. RDS provides a theoretical basis for a representative sample; however, many assumptions must be met, and putting theory into practice may also introduce biases. In the absence of a gold standard, it is difficult to judge the representativeness of the different samples produced. In any case, the overall conclusion of the study is clearly in favour of the RDS method:

“.. we favour a more diverse sample over one that is more restrictive. … the RDS sample was not only more diverse, it also more closely matched the makeup of known MSM AIDS cases and the larger society in terms of SES. … RDS produced a sample with wider inclusion of lower socioeconomic status (SES) than snowball sampling or TLS – a significant finding given that the majority of AIDS cases reported among MSM in the state were low SES. RDS also achieved the sample size faster and at lower cost. For reasons of inclusion and cost-efficiency, RDS is the sampling methodology of choice for HIV surveillance of MSM in Fortaleza” (ibid).

### 14.6.2 Comparison of RDS with indigenous fieldworkers sampling (IFS)

Evidence suggests diffusion of injecting drug use and associated HIV infection in the Russian Federation since 1996. Approximately 60 per cent of HIV cases have been associated with injecting drug use. It has been documented that HIV in much of the Russian Federation and former Soviet Union is a concentrated epidemic, with prevalence consistently above 5 per cent in high risk groups (e.g. IDUs – injecting drug users) compared to less than 1 per cent in the general population.

It is usually possible to sample persons who have received treatment at a public, private or community-based facility. There remains a need for research among non-treatment and community-recruited samples of IDUs – i.e. IDUs outside institutions, among the general population - to better estimate the dynamics of HIV transmission and to improve treatment and health services access. Many surveillance studies of IDUs conducted in the 1990s relied on non-probability sampling to recruit members of the target group.

The study described below is based on the work of Platt et al. (2006).
A. Study objectives

The study compares two sampling methodologies for surveying such populations - respondent-driven sampling, and another chain referral sampling procedure using indigenous fieldworkers (see Section 14.5.2 for a description of the IFS procedures.) Both methods work on the assumption that peers are better able to recruit members of a hidden population than researchers.

The aim is to investigate the relative effectiveness of RDS to reach more marginal and hard-to-reach groups. It compares the two sampling methodologies in terms of cost effectiveness, duration of fieldwork and effect on the demographic composition of the sample. Specifically, it evaluates the relative efficiency of RDS to recruit the sample at a lower cost in comparison to IFS, and also provides a theoretical comparison of the two approaches. The study draws upon nine community-recruited surveys of IDUs that used either IFS or RDS, undertaken in the Russian Federation and Estonia between 2001 and 2005. Sampling effects on the demographic composition and injecting risk behaviours of the samples generated are compared using multivariate analysis.

B. Data collection

Between 2001 and 2005, nine community surveys of IDUs were undertaken in the Russian Federation and Estonia. Four studies used IFS to recruit IDUs, and five used RDS. All IDUs were recruited from community settings. All studies collected some standardised indicators and defined current IDUs as individuals who had injected drugs for non-medical purposes in the past 4 weeks.

For each of the IFS studies, IDUs were recruited using a team of 10-12 field workers at each site. Settings included street locations and respondents’ homes but excluded drug treatment centres and clinics. Volunteers and outreach workers at ‘local harm reduction’ non-governmental organizations (NGOs) were employed as FWs.

For studies of both types, at each location two to three trained research staff from a local university, two experienced supervisors from Moscow and a researcher from the UK provided technical expertise and management.

In all studies FWs recorded in detail their observations on the drug scene, progress of the fieldwork and any difficulties arising from the research. These observations provide a useful additional comparison between the two sampling methods.

A total of 3,771 IDUs were recruited into the nine surveys across four cities in the Russian Federation and two cities in Estonia. A total of 2,049 (54 per cent) participants were recruited through IFS (4 studies), and 1,722 (46 per cent) through RDS (5 studies).

C. Evaluation

Platt et al. (2006) note the following points.

(1) With IFS, the responsibility for selecting the right target group is placed with the field workers, and its success depends on establishing a trusting relationship between the researchers and the fieldwork team. With RDS, issues of trust are less important, as researchers undertake the interview themselves.
(2) The problems of establishing whether respondents are genuine members of the target group remain. Although measures - such as using indigenous field workers to screen participants or recording biometric measurements to avoid the same respondent being interviewed twice - can be put in place which might reduce the incidence of this happening, it is very difficult to measure how much fabricated data may be entering a survey. A disadvantage of both methods is that study participants who are not members of the target group may lie about their membership in order to receive a reward.

(3) Determining the most suitable type and amount of material incentive for participation is difficult, and has many implications for the study, especially for RDS studies where the secondary incentive is crucial to recruitment success. Recruitment can be inhibited, especial in RDS, if the incentives are too small, and also because of certain properties of the social network.

(4) Adjusting the RDS sample to obtain population estimates depends on the ability to recruit a random population within a subject’s social networks and a positive probability of recruiting everyone in the network. The possibility that the network is highly dependent on the incentive raises the question whether the latter condition obtains. This is particularly relevant when the definition of the population of study is fluid or artificially constructed by the research, as with IDUs and sex workers. It should also be noted that the collection of information describing network characteristics which allows RDS analysis to produce population estimates requires the respondent to recall detailed information on the composition of their network, including its size and each member’s relationship with the recruiter. This process carries a large potential for error.

(5) There are also important issues relating to personal safety. There are safety considerations that favour RDS as respondents attend a fixed site for an interview at which a minimum number of staff is always present. In the IFS method, interviewers may find themselves travelling to an area with which they are unfamiliar, and unintentionally put themselves in danger, especially if it becomes known that they are carrying financial rewards or gift packs with them.

(6) It is often assumed that, after adjusting for network sizes and homophily, RDS is more successful at gaining a more representative sample of the population than is IFS. However, the study being discussed concluded that “to date there is no evidence to suggest that this is the case”, and that further research is needed to assess how the practicalities of implementing RDS in the field compromise the requirements mandated by the theoretical assumptions of RDS for adjusting the sample data to obtain estimates of the wider population.

(7) The findings suggest that RDS does not appear to succeed any better than IFS in recruiting more marginalised sections of the IDU community nor those engaging in riskier behaviour. However, the RDS method does have practical advantages over IFS in the implementation of fieldwork in terms of greater recruitment efficiency and safety of field workers, but at a higher cost.

The conclusion is that, in the meantime, pending demonstration of statistical superiority of the RDS approach, a preferred approach may be to adopt the best aspects of both methodologies, depending on the resources available. A combination could include
the use of coupons for recruitment, but also training indigenous field workers to work alongside researchers to undertake interviews, serving to increase their capacity in research skills whilst ensuring that the correct target group is being reached.

14.7 Assessing validity of the model assumptions

The theoretical assumptions of RDS, noted in Section 4.4, can be tested in field studies to examine the extent to which RDS is used properly and appropriately. There are three main areas for testing.

(1) The first assumption relates to respondents' tendency to recruit those with whom they have a pre-existing, reciprocal and sufficiently close relationship. The reciprocity assumption can be tested by asking respondents about their relationships to their recruiters, and by computing a reciprocity index, indicating the extent to which both recruiters and recruits know each other.

(2) The second assumption relates to respondents' tendency to recruit peers as though recruiters are sampling randomly from their personal networks. This assumption can be tested by asking respondents about the composition of their personal networks and by comparing these data to recruitment patterns, and by asking about the characteristics of those who refused to be recruited and by comparing their characteristics to those who were recruited.

(3) Important analytical activities include using statistical procedures to control for differential network size and recruitment patterns, conducting validity comparisons to assess the success of the sampling process, and evaluating the sampling process and its outcomes for different groups of analytic interest (e.g. children by age, sex, type of activity, working conditions).

The following provides examples of studies undertaken for assessing validity of the RDS model assumptions – as to how far implementations of RDS method have conformed to the theoretical assumptions on which the approach is founded. The examples include the following.

(1) A review of RDS field experience in an international setting, involving a large number of studies in developing countries.

(2) An example of common difficulties in developing long referral chains quickly enough.

(3) A large-scale simulation study to access the impact of network structure on the performance of the RDS method.

(4) Simulations to assess sensitivity of the performance of the RDS method to conditions of implementation.

From a review of the rich literature on RDS applications, it is quite clear that the procedure has had (and continues to have) some remarkable success in dealing with very reclusive populations, albeit less striking than what is sometimes claimed by originators and supporters of the procedure. There are also shortcomings in the method; it is those that the following studies aim to bring out.
14.7 Assessing validity of the model assumptions

14.7.1 Review of RDS field experience in an international setting

A. Introduction

Below we summarise findings from a wide-ranging review by Johnston et al. (2008) of experience of RDS applications in developing countries. The review provides valuable insights into the requirements and problems faced in these studies, and into the achievements of this sampling methodology so far.

This review gathered data from 128 HIV surveillance studies using respondent-driven sampling, conducted in more than 28 developing and transition countries up to October 1, 2007. The studies measured HIV and other sexually transmitted infection (STI) prevalence and their associated risk factors. More than 32,000 injecting drug users (IDUs), men who have sex with men (MSM), sex workers (SWs) and high-risk heterosexual (HRH) men were enrolled in these studies. Many more such studies are being conducted in many different settings.

The review was focused on RDS-based surveys of injecting drug users (IDUs). It examined predictors of poor study outcomes and the operational, design and analytical challenges associated with conducting RDS in international settings, and offered recommendations to improve HIV surveillance.

B. Main findings of the review

1) Defining eligibility and measuring social network sizes

Of the 128 studies identified, 66 (51 per cent) had insufficiently defined eligibility criteria, which included not clearly defining parameters such as eligible age range, sex, geographic area, or specific types of behaviour or activity of the study population. These omissions matter in RDS studies because the eligibility criteria are used to build the question on social network size, which is necessary for analysing RDS data. Some studies also neglected to include the social network question during interviewing, therefore ignoring a standard requirement in RDS methodology and analysis. The social network size question requires careful construction if RDS is to be successfully conducted.

2) Designating appropriate incentives

Incentives that are too high can result in coupon bartering and selling. In some countries, some IDUs waited in front of the survey site to acquire a coupon from an IDU who was coming out, thereby increasing the chances that coupons were being distributed to strangers rather than someone known to the recruiter and was in his/her social network. Some individuals misrepresented themselves as MSM, and some individuals misrepresented themselves as IDUs to participate in the study and receive an incentive. High incentives also can result in logistical challenges. In some studies, incentives were perceived as being so attractive that interview staff became overwhelmed with the number of referrals wanting to redeem their coupons. When some IDUs were told to return another day, sometimes they became disruptive and threatened staff. Furthermore, the high incentive increased the possibility that IDUs would try to participate in the study more than once.
Low incentives, on the other hand, can result in slow recruitment. During the first three months of an RDS study, for instance, no IDUs participated due to the perceived low incentives.

(3) Social networks

RDS will not function if the population of interest is not socially networked. Some surveys did not attain their calculated sample size, most likely due to insufficient social networks among members of the population.

As examples of studies that looked at diversity in social network connections, sex workers (SWs) were found to form diverse (inter-connected) social network associations by type of sex work, while other studies have found that some populations do not form diverse social network connections for certain characteristics, but separate and disconnected networks. In a few cases, efforts to increase crossover among distinct groups within a social network have relied on identifying seeds that can recruit across those groups.

(4) Target sample size and equilibrium

In the review, 18 of the 118 studies which could be analysed on this variable did not reach 90 per cent of their calculated sample size or equilibrium, or both.

To assess whether certain RDS study characteristics might affect participant recruitment, the researchers defined sample size ratio as the number of reported subjects that actually was recruited through RDS (recruited sample size) divided by the number of subjects that was pre-specified during the study design (calculated sample size). They then examined factors in the implementation, design and analysis phases of the studies that might influence the sample size ratio. These factors included the type of study population (IDUs, MSM, SWs or HRH men), whether formative work was conducted before initial recruitment, the use of mobile as opposed to fixed venues, and the use of monetary primary and secondary incentives for recruitment rather than combinations of other incentives. Other variables were the number of sites being used for recruitment of subjects (one versus more than one), use of a coupon expiration date to limit referral and enrolment time (limited versus no limit), and use of a realistic (sufficient large, at least exceeding 1.5) design effect for calculating the required sample size.

C. Conclusions

According to the authors of the review, it appears that most of the studies reviewed were able to follow fairly high standards. Relatively more successful outcomes among IDUs seem appropriate given that RDS was originally developed as a method to sample and provide peer education among IDUs.

However, results from specific study experiences demonstrate numerous operational, design and analytical challenges. Of most importance, many studies failed to use clearly defined eligibility criteria, which are essential in structuring the question about social network size.

A certain proportion of studies failed to reach their target sample size, or to achieve equilibrium in terms of sample characteristics.
Some studies had difficulty determining an appropriate incentive level, thereby introducing logistical and sampling problems during data collection. In response, some studies increased or reduced their incentive amount during data collection. Selecting an appropriate incentive is not an exact science and is often one of the more difficult RDS components to plan, especially in light of differing country-specific socioeconomic factors. It is unknown to what extent those studies used formative research to determine which incentives and in what amounts were likely to be appropriate. The authors conclude that input from target population and key informants, as well as a thorough understanding of the study’s specific economic and political context, would improve the chances of selecting an appropriate incentive.

### 14.7.2 A common problem: sluggishness in recruitment

The following brief example, based on the study by Wang *et al.* (2005), is provided here to illustrate a common problem in implementing link-trace sampling such as RDS: namely the difficulty in getting sufficient numbers of recruits with sufficient speed. However, it also shows that this difficulty by no means precludes getting satisfactory results from the sampling process in the end.

This study provides a methodological assessment of the application of RDS among young adult MDMA/ecstasy users in Ohio, USA. RDS was applied to recruit a sample of 402 active, young adult MDMA/ecstasy users to participate in a study examining patterns of drug use over time, risky sexual behaviour, and adverse consequences related to drug use. In this study, sample analysis was conducted to evaluate how RDS works in different geographical locations with different hidden populations.

#### A. Positive assessment of the overall results

According to the authors, the results showed that the sample compositions in their study converged to equilibrium within a limited number of recruitment waves, independent of the characteristics of the initial recruits (i.e. seeds). The sample compositions approximated the theoretical equilibrium compositions, and were not significantly different from the estimated population compositions—with the exception that White respondents were over-sampled and Black respondents were under-sampled. The effect of volunteerism and masking on the sampling process was found not to be significant. Though identifying productive seeds and improving the referral rate are significant challenges when implementing RDS, the findings demonstrate that RDS is a flexible and robust sampling method. RDS has the potential to be widely employed at least in studies of illicit drug-using populations.

The recruitment quota system was implemented by giving each respondent three recruitment coupons to pass on to other potential respondents. Each respondent was paid US$50 for his/her time spent completing the baseline interview, which lasted between 2–3 hours. Initially, US$10 was given to a respondent for each eligible peer referred to the project. This reward was limited to a modest level to prevent the exercise of social influence and peer pressure from becoming coercive. To speed up the recruitment process, compensation for a referral was increased to US$15 about 2 months after the initiation of the sampling process. Additional recruitment coupons
were given when the recruitment process slowed down. However, no financial reward was given for additional referrals beyond the third recruit.

B. Slow speed and low level of recruitment

The initial recruitment of seeds lasted 1-2 months. The ethnographer and consultants identified 28 initial respondents, or seeds. Of the 28 seeds, 11 did not refer anyone to the project; thus, they were defined as ‘infertile’ seeds. During the 12-month recruitment period, a total of 374 respondents were recruited by 17 seeds. This gives a total of 402 respondents, or 391 excluding the infertile seeds.

Initially, the recruitment process was slower than expected. The number of recruits was about 25 in each of the first 2 months when about 40 new recruits were expected. After increasing the referral incentive from US$10 to US$15 per referral, recruitment improved, and the average number of new recruits increased to about 32 per month.

It is noteworthy that among the 17 seeds, only two (one in eight) fulfilled their quotas, i.e. had three recruits. The majority of the referral chains began with only one ‘branch’ (i.e. began with a single successful referral). Among the 17 referral chains, only nine (little over half) had a length of four or more recruitment waves. However, these nine referral chains produced 349 recruits, accounting for 93 per cent of the total recruits. In addition, there were five lengthy referral chains, each of which had between 7 and 11 waves, producing a total of 296 recruits, representing almost 80 per cent of the total recruits. An example of a recruitment tree that began with a productive seed is shown in Figure 14.7.

A substantial number (n = 196) of the respondents did not make any referrals and produced no recruits. As a result, the ratio of referrals (i.e. the ratio of total referrals to the total number of respondents) was only 50 per cent. Among those who produced recruits (n = 195), 44 per cent had a single recruit, 32 per cent had two, 16 per cent had three, and 8 per cent had more than three recruits. The actual growth curve of recruitment did not approach the theoretically expected growth curve.

C. Reaching diverse parts of the population

Since it was not possible to select the seeds randomly from the target hidden population, the initial sample compositions, or the characteristics of the seeds, were considered biased. For example, no Black MDMA users were identified among the initial recruits (i.e. among the seeds). However, the sample compositions of each wave changed and gradually stabilized over waves. In the final sample, the composition had become quite diverse and probably also representative: the corresponding shares were 63 per cent men and 37 per cent women; 82 per cent White, 11 per cent Black, and 7 per cent Others. In addition, 37 per cent were younger than age 20, 50 per cent were between 20 and 24, and 13 per cent were aged 25 or older. These figures appear quite representative of the population.

In the stochastic process of RDS, the characteristics of the recruiters theoretically affect the characteristics of their recruits. As such, recruiters in different groups generated different structures of recruits. It appears that an in-group recruiting tendency existed among men, Whites, Blacks, and younger age groups (<20 and 20–24), but cross-group recruiting also occurred. The in-group and out-group selection probabilities for female recruiters were almost the same (51 per cent versus 49 per cent), while out-
group recruiting occurred preponderantly among respondents in ethnic group classified as “Others” (than Whites or Blacks), and in the “25+” age group.

Figure 14.7. Recruitment tree beginning with a single productive seed

Source: Wang et al., 2005.

The study illustrates two noteworthy points.

(1) The level and form of incentives is an important determinant of the success of recruitment.

(2) Even if the average number of referrals per respondent remains disappointing low and the pace of recruitment appears slow, sometimes the development of a few very large chains can ameliorate the situation and help in the sample reaching equilibrium, as well as diversity, as shown by the experience of this study.
14.7.3 Simulations to assess the impact of network structure

The study by Goel and Salganik (2010) described below investigates the performance of RDS by simulating sampling from 85 known network populations.

A. Background

RDS uses the standard multiplicity estimator (14.3). The accuracy of RDS estimates is affected by the structure of the underlying social network, the distribution of traits within the network, and the recruitment dynamics. In particular, RDS can perform poorly when traits cluster in cohesive subpopulations – such as in the case of infectious diseases (Goel and Salganik, 2009). The authors note that gauging the combined effect of these factors has proven difficult, and previous attempts to assess the performance of RDS have been largely inconclusive.

There is a variety of approaches for assessing the quality of RDS.

(1) First, an RDS has been carried out on a population with known characteristics: undergraduates at a large residential university. Because only a single RDS sample was taken, however, it is difficult to ascertain sampling variability of estimates.

(2) Second, RDS estimates have been compared to estimates derived from alternative sampling methods, as in the examples below. These comparisons have not yielded consistent patterns and are hindered by the fact that true population values remain unknown.

<table>
<thead>
<tr>
<th>Study population</th>
<th>Location</th>
<th>Alternative methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDU</td>
<td>Barnaul, Russia</td>
<td>Indigenous fieldworker sampling</td>
</tr>
<tr>
<td>IDU</td>
<td>Volgograd, Russia</td>
<td>Indigenous fieldworker sampling</td>
</tr>
<tr>
<td>IDU</td>
<td>Detroit, USA</td>
<td>Time-location sampling</td>
</tr>
<tr>
<td>IDU</td>
<td>Houston, USA</td>
<td>Time-location sampling</td>
</tr>
<tr>
<td>IDU</td>
<td>New Orleans, USA</td>
<td>Time-location sampling</td>
</tr>
<tr>
<td>IDU</td>
<td>Seattle, USA</td>
<td>Targeted and venue-based sampling</td>
</tr>
<tr>
<td>DU</td>
<td>New York City, USA</td>
<td>Targeted and venue-based sampling</td>
</tr>
<tr>
<td>MSM</td>
<td>Fortaleza, Brazil</td>
<td>Snowball and time-location sampling</td>
</tr>
<tr>
<td>MSM</td>
<td>Tallinn, Estonia</td>
<td>Internet convenience sampling</td>
</tr>
<tr>
<td>Latino gay men</td>
<td>Chicago, USA</td>
<td>Simulated time-location sampling</td>
</tr>
<tr>
<td>Latino gay men</td>
<td>San Francisco, USA</td>
<td>Simulated time-location sampling</td>
</tr>
</tbody>
</table>

Abbreviations: IDU, injection drug users; MSM, men who have sex with men; DU, drug users. Source: Goel and Salganik (2010).

(3) The stability of repeated cross-sectional RDS estimates has been examined, again yielding ambiguous results. For example, in studies of men who have sex with men (MSM) in Beijing, China in over three years 2004-2006, year-to-year RDS estimates for age and employment status were stable, whereas estimates of education and sexual orientation were suspiciously volatile.

(4) There have been several, but generally not satisfactory, simulation studies on ‘synthetic networks’. Recently, Gile and Handcock (2010) have evaluated the bias of RDS estimates on more realistic synthetic networks (see Section 14.7.4 below).
In contrast to the above approaches, performance of the RDS method may be evaluated by simulating the sampling and estimation process on real populations mapped in previous studies. The following is a review of a study of the last-mentioned type.

B. Data and methods of the study

The study by Goel and Salganik (2010) evaluated the performance of RDS by simulating the sampling and estimation process on 85 real populations mapped previously. In all cases, both the network structure of the population and demographic traits for each individual were available. The study was thus able to directly compare empirical RDS estimates to true population values and, in particular, to measure the variability of estimates.

The study used two data sources. The first source of data was a large, multi-year study that began in 1987 as a prospective examination of the influence of network structure on the propagation of infectious disease. The second data source, the National Longitudinal Study of Adolescent Health, mapped the friendship networks in 84 middle and high schools in the United States.

The simulations for the study assume the same, idealized sampling conditions considered in the theoretical RDS literature. Specifically, the assumptions are that: (i) initial sample members are chosen independently and proportional to network degree; (ii) relationships within the population are symmetrical, i.e. if A is a contact of B, then B is also a contact of A; (iii) participants recruit uniformly at random from their contacts; (iv) those who are recruited always participate in the study; (v) individuals can be recruited into the sample more than once; (vi) the number of recruits per participant does not depend on individual traits; and (vii) respondents accurately report their social network degree.

Starting from ten initial seeds, each participant recruits between 0 and 3 other individuals. The exact recruitment distribution assumed is: 1/3 of participants recruited no one, 1/6 recruited one other participant, 1/6 recruited two other participants, and 1/3 recruited three other participants, the maximum allowed. The simulated recruitment procedure continues until a sample size of 500 is reached. The entire sampling process was repeated 10,000 times on each network to generate replicate estimates.

C. The problem of large design effects

From each simulated sample, the RDS multiplicity estimator was used to infer the population proportion of a given trait. It was found that RDS generates approximately unbiased estimates, but that the variability of RDS estimates is significantly larger than generally acknowledged.

This variability can be quantified in terms of design effect, which benchmarks the performance (variance) of RDS against that of simple random sampling (see Section 3.6 above). A design effect of 10, for example, effectively reduces an RDS sample of nominal size 500 to an SRS sample of size 50. Consistently large design effects are seen in both the data sets used in the simulations.
However, a review of 91 studies found that more than half assumed a design effect less than 1.5, and all assumed a design effect less than 2.5. Given that the study finds typical design effects greater than 5 – well above for example the rule-of-thumb design effect of 2 that has been suggested by Salganik (2006) - it is likely that, even under optimistic sampling conditions, many RDS studies do not have sufficiently large sample sizes to meet their stated study objectives. Compounding these large design effects, the standard RDS confidence intervals developed in earlier studies are often misleadingly narrow.

To put the results in context, the study compares the RDS estimator - which weights simple members inversely proportional to their network degree - to the unweighted mean of the RDS sample, an estimation method that mimics traditional snowball sampling and that is widely considered problematic. However, though the simple mean generally has larger bias than the RDS estimator, the two estimators are comparable in terms of their standard error and their overall performance, as quantified by root mean square error (RMSE).

Large design effects arise primarily because of the ‘clustering’ inherently imposed by the link-trace design: chains formed from the same seed tend to have a high degree of homogeneity, because people choosing (nominating for the survey) each other tend to be similar.

D. Conclusion

Past work has emphasized that RDS in theory generates approximately unbiased estimates - and the study indeed finds this to be the case from the simulations. However, by neglecting to consider the variance, this result has been widely interpreted as indicating that RDS has low error.

The study argues that, because the study models a best-case scenario in which the theoretical RDS sampling assumptions hold exactly, it is unlikely that RDS performs any better in practice than in its simulations.

Furthermore, it seems very likely that actual RDS sampling dynamics only exacerbate the problems indicated by the results. Specifically, initial participants are generally a convenience sample and are almost certainly not chosen in the judicious, independent manner of the simulations in the study. There is also evidence of non-random recruitment of peers and of differential participation and recruitment rates. In a study for example of MSM in Brazil (see Section 14.6.1), participants were more likely to recruit those who they thought engaged in riskier behaviour and who would therefore most benefit from HIV testing; this same study found that some individuals refused to participate for fear of disclosing their sexual orientation. Evidence of differential recruitment rates was seen in a study of jazz musicians in New York, with women on average recruiting over 60 per cent more participants than men.

Furthermore, ascertaining an individual’s network degree is a challenging problem - particularly in the context of RDS - and so self-reported network size represents a possibly significant source of non-sampling error absent from the above simulations.
It should be mentioned, nevertheless, that for the kind of surveys being considered, no reasonable samples could have been obtained at all without resort to some form of link-trace sampling strategy, of which RDS provides a fine example.

14.7.4 Simulations to assess sensitivity to conditions of implementation

The following discussion is taken from the study by Gile and Handcock (2010). It is a simulation study to assess the performance of existing RDS estimators in three areas concerning sensitivity of the method to deviations from the random walk model: sensitivity to the convenience sampling procedure used to select seeds for the initial sample; sensitivity to respondent behaviour in bringing in new respondents for subsequent waves; and sensitivity to the assumption that the sampling is with replacement.

A. Bias introduced by seed selection

The simulations in the study illustrate “that the rate of reduction of seed bias is sensitive to the level of homophily and the number of sampling waves, and that the number of waves typically sampled in RDS studies may not be adequate for removing the bias induced by seed selection, especially in highly clustered populations”. In terms of costs, there is a trade-off between many broadly chosen seeds and many waves of sampling. Interest in increasing the number of waves has to be balanced against the desirability of accessing the diversity of subgroups of the target population.

Seeds should be selected to be as representative of the target population as possible. “Clearly, when the underlying network is disconnected, the disconnected subgroups represented will be determined by the seed selection, and seed bias cannot be removed by increasing the length of sample chains ... In cases of extreme homophily in a connected network, the resulting RDS estimators can be highly uncertain”.

B. Selection of referrals and response quality

The survey data are subject to biases in the process of referral through which respondents bring new respondents into the survey – in the specific context of RDS, how coupons are passed from respondents to new recruits. Other sorts of biases may be also present. These arise, for example, from the respondents’ inability to report their number of links (their ‘degree’), or a lack of correspondence between these links and the actual groups to which the coupons are distributed, or a lack of reciprocity in the relationships along which coupons are passed.
C. High sampling rates

The authors also point to difficulties in using the existing RDS estimators in the presence of substantial sampling fractions. The presence of large sampling fractions negates the assumption of ‘sampling with replacement’. When there are differential activity levels between the groups of interest (the case which RDS is designed to address), and in the presence of homophily, larger sampling fractions lead to significant biases in the resulting estimators. This bias is larger for populations with greater differentials in activity levels. The simulations indicate that for very large sampling fractions, the ordinary (unweighted) sample mean out-performs the RDS estimators. “In short, we find the random walk model to be an inadequate approximation to RDS sampling whenever there is a sizable sample fraction, differential activity levels by characteristic of interest, and homophily on that characteristic, circumstances which are not uncommon among populations studied via RDS”.

In conclusion, the authors sound a cautionary note. They argue that “the much-lauded allegedly ‘asymptotically unbiased’ property of the estimators is not a panacea. … and it is not at all clear that RDS samples reach anywhere near the depth necessary to assure … convergence. Furthermore, any convergence properties are heavily dependent on a large number of assumptions. Practitioners should be aware of the strict conditions necessary for these desirable properties”.

Their overall conclusion is more positive, however. “Current RDS methodology is powerful and innovative. As a sampling strategy, it effectively reaches large and varied samples of many hard-to-reach populations. As tools for inference, the existing estimators leverage the waves of sampling to reduce the inevitable bias in convenience sampling from such populations”.
Annexes
Annex A

Illustrations from studies of the worst forms of child labour

In this book we have described sampling procedures appropriate for special groups, situations and types of child labour. As was noted in Chapter 2, *a good understanding of the situation to be studied is absolutely essential for the choice of an appropriate sampling strategy, and more generally of the survey methodology*. Child labour situations and problems are very diverse and very specific to the particular situation. Hence it is necessary and useful to provide a wide range of illustrations from national practices, covering different types of activity in different settings. The objective of this annex is to identify and illustrate the variety of special characteristics, situations and types of child labour in order to provide the necessary background and context for the diverse sampling techniques discussed.

Illustrations from national studies are provided for each main sector of hazardous child labour. The sectors are:

(A.1) child domestic work  
(A.2) agriculture including commercial crops  
(A.3) fishing and aquaculture  
(A.4) mining and quarrying  
(A.5) manufacturing including handicrafts  
(A.6) construction  
(A.7) street work and the informal sector  
(A.8) unconditional worst forms of child labour: including (i) child trafficking; (ii) commercial sexual exploitation of children; (iii) children in forced or bonded labour; (iv) children engaged or living in armed conflict; and (v) children involved in illicit activities, in particular in drug trafficking.

Reference has been made of all the examples below in the body of Chapter 2. The chapter provides the background and general features concerning each sector of hazardous child labour illustrated with examples here.

A.1 Child domestic work

Example A. Child domestic work in South-east and East Asia

The following illustration is summarized from Matsuno and Blagbrough (n.d.), *Child domestic labour in South-east and East Asia: emerging good practices to combat it.*

Child domestic labour has to be analysed from different angles: education, culture/tradition, gender equality, migration and trafficking, violence against children, minimum age for employment and the worst forms of child labour.
Education

Both the education status and achievement level of child domestic workers are far from desirable. Though many children and possibly parents start off believing it is possible to combine the two, many surveyed young people report that it is difficult for them to keep up the household work and schooling. In urban Mongolia, for example, one third ... of child domestic workers surveyed had dropped out of school. In Hanoi, a third of children surveyed had dropped out of school somewhere between grades 1 and 5 and were forced to find work. Accessing lower secondary education is a challenge, especially for girls. Despite all the difficulties in continuing education however, the willingness of surveyed child domestic workers to resume their formal education is high. Given the opportunity, some 72 per cent of them in the Indonesia study said they would like to resume their formal education.

Culture/tradition

Seeking work in a household is a practice that has long been culturally accepted as a coping mechanism for poor families in Asia. In many societies, placing children of poor families in the care of a wealthier relative is an established tradition and rarely, if ever, questioned.

Gender equality

Children in domestic labour are predominantly girls. For example, 98 per cent of child domestic workers aged 12–17 years in Hanoi are girls. In Indonesia, 93 per cent are girls. Even if they want to find employment elsewhere, very few options are available to uneducated children, especially girls. Once employed in households, the experiences of girls and boys differ. While girls tend to remain within the employers’ household, boys are given tasks outside, such as fetching water, shopping, washing cars, herding livestock, etc. The common problem of being isolated and/or confined that many child domestic workers endure is experienced less by boys. In addition, their future prospects are not the same. While boys tend to have more options other than domestic labour when they get older, girls tend to have little choice but to remain as domestic workers.

Migration and trafficking

Rapid urbanization and rise of the middle class in urban areas has affected the demand for domestic workers in general. In Thailand, for instance, domestic work is the second largest work sector for migrant workers after agriculture. Human traffickers often use household employment as an entry point. For children who have low educational attainment, especially girls, a housekeeping or nanny job is appealing. It does not require prior education or training, and it seems like a safe work. It is, however, not uncommon to hear that children recruited as child domestic workers actually end up in commercial sexual exploitation and/or very exploitative/abusive child domestic labour situations. This situation happens not only within a country, but also in a context of cross-border migration.
The worst forms of child labour

Excessive working hours, no rest time or rest day, no or limited remuneration, exposure to safety and health hazards, abuse and exploitation, bonded labour and trafficking are among the conditions making some domestic work as worst form of child labour. The impact of a poor working environment may manifest in a child's health condition, which is closely linked with emotional development. In Cambodia, 23 per cent of child domestic workers complained of exhaustion, 21 per cent complained of constant fear and 20 per cent suffered from insomnia. In Indonesia, 31 per cent of child domestic workers also stated that they could not sleep and had no appetite (29 per cent).

Example B. Child domestic work in Côte d’Ivoire

The following illustration is summarized from Diallo (2009), *Children’s Work, Child Domestic Labour, and Child Trafficking in Côte d’Ivoire*. The essay presents a broad overview of the nature and prevalence of children’s work in Côte d’Ivoire, and examines two especially problematic and often interrelated aspects of child labour in the country: child domestic work and child trafficking. Child domestic work and child trafficking are particularly difficult to track, largely due to their informal nature and the fact that they occur in private and hidden spaces.

“Historically, from the colonial period to now, child domestic labour has taken different forms and includes trafficking or work that by its nature or the circumstances in which it is carried out is hazardous and likely harms the health, safety, or morals of children. Also historically, there have been three waves of the phenomenon of domestic labour in Côte d’Ivoire. From the colonial period to the 1970s, domestics were mostly mature girls, the majority from two sociolinguistic areas of the country. Most were illiterate and from poor households. Then, during the 1980s, increasing numbers of children became involved in domestic labour due to exclusion from school. Contrary to the earlier domestics, these children were not from any specific rural areas of the country. Finally, in the 1990s, the phenomenon of child domestic labour became extensive, with the development of employment agencies and increasing numbers of children excluded from school. …”

The 2005 SIMPOC child labour survey shows that there were about 50,000 children between the ages of six and fourteen years doing domestic work on a full-time basis outside the homes of their biological parents in the regions covered by the survey. These children account for about 2.2 per cent of all children ages six to fourteen years old. They perform a variety of tasks. Girls are slightly more involved than boys in domestic work (52% girls against 48% boys). While child labour in Côte d’Ivoire is mostly a rural phenomenon, child domestic labour is more concentrated in the cities, almost one-third of child domestic workers being in Abidjan. Studies highlight the difficult working conditions of child domestic workers: long and tiring working days, handling dangerous items such as knives, insufficient or inadequate food and accommodation, and degrading treatment. “Half of the children work twenty hours per week or more. Only 2 per cent of child domestics are paid [in cash] for their work. Many do not receive the wages they have been promised. For others, the employment arrangements do not involve the exchange of money.”
One main concern regarding child domestic workers concerns their schooling. Data show that over 20% of child domestic workers have never attended school.

There is a strong correlation between child-fostering arrangements and child domestic labour. “More than 90% of children in domestic work are relatives of the head of the household in which they work. This means that many child domestics have been given by parents or guardians to another person to be fostered, but in reality the child becomes an unpaid servant for the host family. Child fostering is a temporary and reversible transfer of child-rearing responsibilities to people other than the natural parents. In most countries in sub-Saharan Africa it is common to see natural parents sending their children to relatives. The phenomenon is more diffused in Côte d’Ivoire. Traditionally, the institution of fostering has favoured the movement of children to the houses of relatives, where they become domestic workers. In this way, the distinction between child placement and child domestic employment is sometimes hard to pin down. However, some child domestic workers are recruited through recruitment agents or have sought out a placement on their own initiative. In addition, children may be trafficked for domestic labour. It has been reported that children are being trafficked for domestic labour, agriculture, the informal sector, and mining in Côte d’Ivoire”.

Example C. “Small maids” in Morocco

The following illustration is summarized from Sommerfelt (2009), Small Maids in Morocco.

Household work in urban Moroccan homes is to a large extent performed by paid maids, girls aged 5-15 years nicknamed ‘small maids’. It has been estimated that these small maids make up 2.3-3.0 per cent of the total number of girls in their age-group, thus constituting a significant proportion of the child labour in Morocco, outnumbering girls in urban areas who work outside of private homes. The small maids “perform household tasks, like washing and cleaning, taking care of children, and running errands. They live in their employers' homes, and contact with their own parents is limited. Young girls' labour as maids in Morocco is locally defined as work, not as ‘caretaking’ (as is often done farther south on the African continent). Correspondingly, many who employ girls make a point of emphasizing that girls are hired for work and are not ‘daughters of the house’. … In spite of the fact that the domestic service of small maids is discussed in terms of wage labour and employment, terms of work are vague and are not discussed when parents and employers meet. Salary, however, is determined upon the girl’s appointment. … In the majority of cases, the salary is given directly to the parents …. For many of the girls, parents' monthly visits to collect their daughters' salaries are the only occasions for parental contact. Some of the little maids are allowed to visit their families for special occasions”.

In general, small maids do not attend school. In a study carried out by Fafo and Save the Children UK (Sommerfelt (ed.), 2001), parents described the ideal employer as one that ensures their daughter’s welfare, providing her with food and clothes, and refrains from beating. “Parents also expect employers to ensure that their daughters are kept under surveillance, preventing them from ‘running around’ (also implying engagement in premarital sex). …Parents have few possibilities of checking on their
daughters’ well-being, and abuse of little maids, including sexual abuse, is slowly becoming a subject of public discussion”.

Small maids are, by and large, an urban phenomenon, and the young maids make up more than 5 per cent of the population of girls in urban areas. The vast majority come from rural areas or urban outskirts. “For rural families, placing daughters as maids is a response to a desperate economic situation. Daughters’ incomes contribute significantly to the parents’ household income. The gender structures in this response to poverty are clear. Girls are sent to work as maids because housework is a female domain. Boys are given priority for education, and when they do work, it is mostly in nondomestic settings, and they generally stay with kin”. Small maids’ employment is often ended when they grow older; the presence of young, unmarried women in the employer’s home is seen as morally problematic or indecent.

The small maids are recruited through informal middlemen or women. Occasionally, young maids themselves recruit relatives from their home areas. “Recruitment processes appear to take the shape of ‘chain migration’. Rural girls … occasionally claim to want to join their sisters to work in towns, imagining urban centres as places of opportunity”.

Example D. Child domestic work in the Philippines

The following illustration is summarized from Kitada (2009), Child Domestic Work in the Philippines.

Invisible behind the walls of private homes, child domestic workers are vulnerable to abuse. Hiring domestic workers is a common practice not only among the upper class but also in middle-class households. According to the 1995 labour force survey over 300,000 domestic workers were nineteen years old or younger, “but the actual number of child domestic workers today is estimated to be more than three times that figure” (Flores-Oebanda, Pacis and Alcantara, 2004). There are both female and male child domestic workers, but the vast majority are female.

“In the face of economic hardship, some families, especially in the rural areas, are willing to let their children work as domestics in the provincial cities or Manila in exchange for money (advance payment or salary), or for the children’s education plus room and board. Children themselves often take up such work willingly, as they aspire to get educational opportunities in the urban centres while they work. Many regard working as domestics their only chance to go to school. However, once within the compound of a private home, away from their friends and families, children are easily taken advantage of. They are on duty twenty-four hours a day, and some do not receive the education, pay, holiday, food, or bed that they were promised. Worse still, some children are subjected to psychological, physical, and sexual abuse”.

While there are children who are forced into work, other children actively look for work to help their families. “As domestic work is seen as an accessible and appropriate job for females, sometimes children seek a part-time job as domestic help at a better-off family in their own neighbourhood, combined with schooling …. In these situations where children live with their own families and work close to home as part-time domestic workers, children’s social contacts can quickly intervene in the event of mistreatment. However, when children work as live-in domestic workers in families
other than their own, they are extremely vulnerable to abuse. They are isolated from
sources of help, and their welfare is completely dependent on the employing family”.

Child domestic work has links with child trafficking. Children are often trafficked by
recruiters whose promises often involve deception, whether it is about the conditions,
type, or exact location of work. Fortunately “The Philippines is one country where
there is in-depth documentation of the plight of child domestic workers and where
some cutting-edge work is under way to address this potentially very harmful form
of child labour”.

A.2 Agriculture; commercial agriculture

Example A. The size and trend of child labour in some countries of South Asia

The following is based on de Groot (2009), *Rural Child Labour in South Asia*.

Throughout South Asia child labour remains a serious issue, especially in Nepal
and Bangladesh, with a labour market participation rate in 1999 of 29 per cent in
Bangladesh and 43 per cent in Nepal for children aged 10-14. Table A.1 provides
rough indications of the differences between countries and of the longitudinal trends.
Care has to be taken in interpreting the data as there are significant variations in
the estimates of child labour in the subcontinent and the figures in the table are
generally understood to underestimate child labour. One of the reasons has to do with
the problem of defining child labour. The concept of ‘nowhere children’ is especially
crucial in discussing rural child labour: it relates to the millions of children who are
not enrolled in school and who are not enumerated as ‘economically active’ in the
periodic census or labour surveys. The focus on economic activities, however, has the
disadvantage of missing many children who may work invisibly in the household or
in the informal sector of the economy. In rural areas, many children work as ‘unseen
hands’ - unpaid and unacknowledged – that facilitate the work of adult men and
women and who must be classed as child workers, whether or not they are formally
recognized as such.

Table A.1 shows some figures on the incidence of child labour in South Asia. The
figures in the table lack comparability across countries, but are perhaps more
compared over time for a given country.

| Table A.1. Incidence of Child Labour in South Asia, 1950–2000 |
|-----------------------------|--------|--------|--------|--------|--------|--------|
| Bangladesh            | 37    | 36    | 35    | 34    | 31    | 27    |
| India                 | 29    | 26    | 24    | 21    | 16    | 11    |
| Nepal                 | 68    | 66    | 57    | 49    | 41    | 38    |
| Pakistan              | 14    | 15    | 15    | 15    | 13    | 12    |
| Sri Lanka             | 12    | 7     | 4     | 3     | 2     | 2     |

Source: Chaudhri, Nyland, and O’Rourke (2004).
Example B. Pakistan Baseline Survey 1996

The description is based on an unpublished report of Pakistan 1996 survey on child labour (5-14 years old).

The survey was designed as a conventional household-based survey. It used a two-stage design with census enumeration blocks in urban areas and villages in rural areas as the PSUs. From each area selected with probability proportional to population size within each sampling stratum, one cluster of approximately 75 households was selected. Thus 1,865 clusters in the sample formed the effective PSUs. In these clusters, a total of 140,298 households were listed, obtaining information on whether or not the household contained at least one economically active child aged 5-14. A total of 10,438 households (7.4 per cent of all households listed) contained an economically active child, giving a sample of 13,962 such children. As to the sample clusters, nearly 75 per cent of the originally selected contained an economically active child. Compared to the 1993-94 labour force survey, the 1996 child labour survey is reported to under-estimate the child economic activity rate by nearly 10 per cent.

The sample was allocated very disproportionately between urban and rural areas (and also across provinces): two-thirds of the sample clusters were allocated to urban and only one-third to rural areas. By contrast, the urban sector accounts for only 11 per cent of the estimated number of economically active children aged 5-14, and the rural sector for 89 per cent.

Figures in Table A.2 of course reflect the predominance of agriculture and of sectors related to agriculture, in child labour – sectors which are essentially rural. The table summarises some figures reported or derived from the survey report. Among children 5-14, the reported activity rate is 8.3 per cent, with a large differential between rural and urban areas (10.0 per cent rural versus 3.2 per cent urban), and between boys and girls (11.5 per cent for boys versus 4.4 per cent for girls).

<table>
<thead>
<tr>
<th>Table A.2. Pakistan baseline survey 1996: selected figures</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>% distribution of all children 5-14</strong></td>
</tr>
<tr>
<td>Rural</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>72.0</td>
</tr>
<tr>
<td><strong>% of children 5-14 economically active</strong></td>
</tr>
<tr>
<td>Rural</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>10.0</td>
</tr>
<tr>
<td>Boys</td>
</tr>
<tr>
<td>11.5</td>
</tr>
<tr>
<td><strong>% distribution of economically active children 5-14</strong></td>
</tr>
<tr>
<td>Rural</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>89.0</td>
</tr>
</tbody>
</table>

Of economically active children 5-14:
- **% distribution by sector, in rural and urban areas**

<table>
<thead>
<tr>
<th></th>
<th>Rural</th>
<th>Urban</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>% in agriculture</td>
<td>74.0</td>
<td>4.0</td>
<td>67.0</td>
</tr>
<tr>
<td>% in manufacturing</td>
<td>9.0</td>
<td>31.0</td>
<td>11.0</td>
</tr>
<tr>
<td>% other</td>
<td>17.0</td>
<td>65.0</td>
<td>22.0</td>
</tr>
<tr>
<td>Total</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>% literate</td>
<td>32.0</td>
<td>41.0</td>
<td>33.0</td>
</tr>
</tbody>
</table>
### Table A.2 (cont.)

#### Of economically active children 5-14:

- % by age and gender (all child workers aged 5-14 = 100%)

<table>
<thead>
<tr>
<th></th>
<th>Boys</th>
<th>Girls</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>aged 5-14</td>
<td>73.0</td>
<td>27.0</td>
<td>100.0</td>
</tr>
<tr>
<td>aged 5-9</td>
<td>9.0</td>
<td>7.0</td>
<td>16.0</td>
</tr>
<tr>
<td>aged 10-14</td>
<td>64.0</td>
<td>20.0</td>
<td>84.0</td>
</tr>
<tr>
<td>% literate</td>
<td>40.0</td>
<td>11.0</td>
<td>33.0</td>
</tr>
</tbody>
</table>

#### - % in elementary (unskilled) occupations relating to agriculture

<table>
<thead>
<tr>
<th></th>
<th>Boys</th>
<th>Girls</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>% in agriculture and related *</td>
<td>68.0</td>
<td>80.0</td>
<td>71.0</td>
</tr>
<tr>
<td>% in agriculture</td>
<td>63.0</td>
<td>77.0</td>
<td>67.0</td>
</tr>
</tbody>
</table>

* including sales+services, mining, manufacturing, transport sectors where farm activity dominates

#### - % distribution by employment status

<table>
<thead>
<tr>
<th></th>
<th>Boys</th>
<th>Girls</th>
<th>Rural</th>
<th>Urban</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>unpaid family helpers</td>
<td>67.0</td>
<td>78.0</td>
<td>75.0</td>
<td>30.0</td>
<td>70.0</td>
</tr>
<tr>
<td>employees</td>
<td>25.0</td>
<td>17.0</td>
<td>18.0</td>
<td>62.0</td>
<td>23.0</td>
</tr>
<tr>
<td>self-employed</td>
<td>8.0</td>
<td>5.0</td>
<td>7.0</td>
<td>8.0</td>
<td>7.0</td>
</tr>
<tr>
<td>Total</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

#### - % distribution by normal hours worked per week

<table>
<thead>
<tr>
<th></th>
<th>Boys</th>
<th>Girls</th>
<th>Rural</th>
<th>Urban</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;=35</td>
<td>52.0</td>
<td>67.0</td>
<td>58.0</td>
<td>27.0</td>
<td>54.0</td>
</tr>
<tr>
<td>36-55</td>
<td>34.0</td>
<td>25.0</td>
<td>31.0</td>
<td>48.0</td>
<td>33.0</td>
</tr>
<tr>
<td>56+</td>
<td>14.0</td>
<td>8.0</td>
<td>11.0</td>
<td>25.0</td>
<td>13.0</td>
</tr>
<tr>
<td>Total</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Agriculture accounts for 67 per cent of the child labour, or for 71 per cent if we include parts of sectors (services, mining, manufacturing and transport) where farm activity dominates.

Among the economically active children, 70 per cent are 'unpaid family helpers'. This is obviously associated with the predominance of agriculture in child labour.

#### Example C. Children in India’s tea industry

The following is paraphrased from Ashraf (2009a), *Children in India’s Tea Industry*.

There were almost 130,000 tea estates in India in 2006, nearly all of them smallholdings of less than ten hectares. These smallholders typically sell their output to middlemen, to other larger plantations, or to tea factories. Child labour in the tea industry is concentrated on tea plantations where tea is grown and plucked, and not in the factories where tea is processed and packed for sale.

*Tea plantations need large areas of land and a large labour force.* Workers are employed as permanent, temporary, or casual labour on the plantations. Of them, around 50 per cent are women, and 5 to 10 per cent are children, though no reliable estimates of child labour on tea plantations exist. Among permanent workers, child labour is minimal; children are mainly needed for temporary work during peak plucking season. With scarce employment opportunities in the plantation regions,
workers are forced to bring their children to work with them. Plantation owners also encourage family-based employment as it saves on recruitment and housing costs.

*Most plantation workers live in unhygienic conditions within the tea estates.* Unlike other agricultural operations, tea plantations are considered to be part of industry. The employment conditions of plantation workers are protected just like those of industrial workers. Casual workers receive the same wages as regular workers, but they are not entitled to any benefits. For example, casual women workers have no maternity benefits, and their children cannot study in the plantation schools.

*Workers are needed in the field for plucking, weeding, pesticide spraying, pruning, clearing the garden, and making pits for new saplings.* Plucking and weeding are generally done by women and their daughters. Tea plucking is a tough job and is physically debilitating. The heavy baskets workers carry concentrate the weight on the head and neck, leading to severe back pain. Spraying of pesticides in the tea gardens is done by male adolescents and children, leading to exposure to toxic materials.

*Work begins at sunrise.* Working hours for adults are limited to no more than eight hours per day for a six-day week. Adolescent and child workers are limited to no more than four and a half hours per day for a six-day week. Wages are of three kinds. Workers are paid a time-rated wage (daily wage), a price-rated wage (extra leaf price), and are given food grains at subsidized rates.

*In conclusion, many children in poor families are sent to work without getting an opportunity to educate themselves and learn skills.* Child labour has been prevalent in different industries for historical reasons, as parents passed on traditional skills to their children, but on tea plantations the reasons seem to be the low wages received by adult workers and the lack of other viable employment opportunities.

**Example D. Child labour on sugarcane plantations in the Philippines**

The following is based on de Boer (2009), *Child Labour on Sugarcane Plantations in the Philippines*.

Sugarcane provides an income to more than 6 per cent of the Philippine people. It is one of the main commercial crops of the archipelago. With agriculture and fishing being the sectors where the bulk of working children can be found, sugarcane alone is responsible for 1.5 to 5 per cent of all Filipino child workers. Nowadays most sugarcane is produced on big commercial haciendas. Workers are hired on a seasonal or daily basis, and pay is low. Accommodation for the workers and their families is provided on the hacienda. Children of the sugarcane workers are thus an accessible workforce that can be hired during the busy seasons of planting, weeding, and harvesting. They are pushed into the fields by family poverty, dependency on the hacienda for income, and high costs of education. At the same time, they are pulled by the demand for simple manual labour. Plantation owners prefer child workers since they cannot be trade union members.

Children from the age of six are involved in clearing fields, planting, weeding, and fertilizing. Older children are also involved in the hard labour of harvesting. As sugarcane production is highly seasonal, children’s work on the plantations varies from not working at all to doing light work to being subject to hazardous labour.
Although the work they do is not always onerous, the children do experience negative effects from working in the fields. Working prevents them from attending classes regularly, from doing homework, and sometimes from going to school altogether. When at work, children risk overheating, exhaustion, cuts from the tools and from the sharp cane leaves, insect bites, and pesticide poisoning, resulting in 60 per cent of the children involved in agricultural work (including sugarcane) reporting health problems, according to the Labour Force Survey of 2001. The children work in or near recently sprayed fields, thus inhaling and touching pesticides (de Boer 2005). All these factors jeopardize children's health in the short term, as well as their possibilities later in life, reinforcing a vicious circle of poverty and exploitation.

### A.3 Fishing and aquaculture

**Example A. Child labour in fishing, El Salvador**

This concerns an investigation using the rapid assessment methodology into worst forms of child labour in fishing in El Salvador. It is based on Godoy (2002), *El Salvador Child Labour in Fishing: A Rapid Assessment*.

#### The context and study methodology

Researchers evaluated aspects concerning the times and places in which the fishing boys and girls interact, their working life, and the risks and dangers to which they are exposed. The investigation was carried out in a sample of geographic areas around six communities known for the high concentration of children engaged in fishing. Fishing is carried out on the shores of lakes, bays, estuaries and gulfs located in the areas covered. The areas covered are poor communities located in the region hardest hit by a hurricane in 1997 and earthquakes of 2001, which wrought further damage on the basic infrastructure of these communities. The economies of these communities depend upon small-scale fishing and, in some cases, farming and livestock. Their residents lack the capital needed to engage in production, and consequently they have to engage in plunder-style fishing as a strategy for survival. Given this situation, the work performed by boys and girls constitutes an important source of income.

The population of these communities is composed of large families (with an average of six members). In general, they are broken or separated families in which the father is almost always absent and often the mother leaves for long periods in order to carry out remunerated domestic work.

The main types of fishing identified include: trammel net, casting net, hook and line, molluscs and oyster extraction, small fry gathering, and fishing with explosives.

#### Profile of children and their work

The working children largely range in age from 8-16 years. They are initiated into fishing at an early age when, encouraged by their parents, older siblings, relatives or friends, they plunge into the water and learn to swim. Later they gradually become involved in fishing tasks. The participation of boys and girls in this activity is accepted, and is seen as natural, normal and convenient for both the family and the community. Of the children interviewed in the rapid assessment, 18 per cent were
female. Under orders from the adult women of the household, these girls attend to traditional domestic chores. Their participation in fishing is limited to marketing the products obtained by their relatives.

Some children are own-account workers; others perform part of the work of their parent/guardian. Still others work on the basis of a verbal contract. Forms of compensation include payment in cash, in kind or a mixed form of payment. The children usually hand over their earnings to the person incharge of the household. The majority of the boys and girls interviewed reported that if they fail to perform their jobs or if they make mistakes, they are physically or verbally mistreated, and in some cases not paid for their work.

Child labour in fishing encompasses every aspect of the activity (preparation, transportation, operation, selection, storage and marketing). Boys and girls involved in fishing sometimes work in a team with adults, or at other times work alone without adults nearby to protect them. Some types of fishing are nocturnal. The hours of work are exhausting, ranging from 5-13 hours, with night fishing requiring the most time. Fishing with the aid of explosives requires the least amount of time, but is one of the most hazardous types of fishing.

**Schooling**

Most of the children interviewed have had the opportunity to go to school, though not all of them remain there for long. Their school attendance is irregular and the majority drop out without having achieved a sufficient level of education. Some 42 per cent of the sample do not attend school. There is an inverse relationship between the children's level of education and their involvement in fishing; that is, the higher is their participation in fishing, the lower the level of education.

**Hazards**

Generally speaking, children's bodies show the effects of their early physical effort: wrinkled and burnt skin from constant exposure to water, sodium and wind. Often their skin is infected with fungi and bacteria. In addition, they present clinical cases of malnutrition and some suffer from illnesses.

The hazardous nature of the types of fishing are such that this work is considered to be one of the worst forms of child labour. Involvement in this type of fishing implies excessive amounts of physical, personal and collective danger, which permanently undermines the normal development of boys and girls.

The boys and girls who participate in activities related to fishing are confronted on a daily basis with the following dangers and risks inherent in their jobs:

- drowning, getting carried out by strong currents or lost at sea,
- attacks by sharks or other marine animals,
- bites and stings from insects or other land and sea animals,
- sunstroke,
- respiratory problems,
- blindness,
- hearing problems as a result of exposure to high water pressure when diving,
- addiction to stimulants (amphetamines),
- wounds and disfiguration of the hands and body,
- arthritic deformation of the hands and feet,
- damage to major bodily systems,
- early alcoholism,
- sexually transmitted diseases,
- physical and psychological mistreatment and sexual abuse.

### A.4 Mining and quarrying

#### Example A. Child labour in small-scale mining in the Philippines

The following account is based on Espino (2009), *Children in the Philippine Gold-Mining Industry*.

A 2001 survey conducted by the National Statistics Office of the Philippines reported that 4 million out of 24 million 5-17 year old Filipino children, or one out of six children, were working. About 2.4 million of these children were involved in hazardous work (ILO-IPEC and ASIADEV, 2003). The same survey reported that the mining and quarrying sector employed around 18,000 (just under 0.5%) of working children, who were subjected to daily health hazards such as noise, high temperature and humidity, inadequate illumination, slipping and falling, exposure to dust and chemicals, and absence of protective work gear. The average child labourer in small-scale mining is male, is 15-17 years of age but began working (whether in mining or in some other sectors) at age ten or even earlier, and is a school dropout.

Small-scale mining in the Philippines is predominantly a family business, where a family or group of families pool resources to find gold or other precious metals. They work known mineral deposits or excavated mining areas abandoned by bigger mining companies. In many cases, small-scale mining areas are found within large mining concessions granted to large multinational companies or their local affiliates, under various forms of co-sharing and ownership arrangements. Child labourers are usually found in gold-rush mining communities, where the location of activities shifts from one area to another depending on the volume and quality of gold ore present. … Small-scale gold mining employs a variety of methods to extract gold, including the compressor system, tunnelling and sinking, open-cast mining, and panning and sluicing. Each of these methods poses different levels of danger to the life and limb of child miners.

#### Example B. Child labour in mining in Tanzania

The following account is based on Mwami, Sanga and Nyoni (2002), *Tanzania Children Labour in Mining: A Rapid Assessment*.

The mining sector in Tanzania is a significant employer in the country. This is especially true when considering the small-scale mines. The mining sector in Tanzania has a high concentration of child labour: a 1996 study contends that more than 555,000
people are directly involved in mining activities around the country, some as full-time miners, and others as part-timers.

The present study covered three main districts in the country in order to find out the causes and incidences of child labour in the mining sector in Tanzania; to examine the working conditions, characteristics and consequences of child labour in the mining sector; and to propose tentative measures of intervention to alleviate the child labour phenomenon. The study also aimed to test and evaluate the ILO-UNICEF Rapid Assessment methodology.

**Methodology**

The study used both qualitative and quantitative methods. Since the Rapid Assessment methodology is more prone to collecting descriptive information, special effort was made to collect quantitative data as well to supplement the qualitative information. The data collection methods used included documentary analysis, child interviews with key informants, observation, and focus group discussions. The exercise of mapping was useful where child labour incidences were concentrated. One sampling problem was that working children kept on moving along the river searching for areas likely to have gold deposits.

The number of interviews according to the type of respondent were as follows.

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region/district mining officials</td>
<td>7</td>
</tr>
<tr>
<td>District education officials</td>
<td>3</td>
</tr>
<tr>
<td>Social welfare officials</td>
<td>4</td>
</tr>
<tr>
<td>District planning officials</td>
<td>3</td>
</tr>
<tr>
<td>Health officials</td>
<td>3</td>
</tr>
<tr>
<td>Ward/village officials</td>
<td>10</td>
</tr>
<tr>
<td>Village elders</td>
<td>18</td>
</tr>
<tr>
<td>Teachers</td>
<td>8</td>
</tr>
<tr>
<td>Mine owners</td>
<td>5</td>
</tr>
<tr>
<td>Working children</td>
<td>125</td>
</tr>
<tr>
<td>Parents of working children</td>
<td>48</td>
</tr>
<tr>
<td>Adults working in the mine</td>
<td>11</td>
</tr>
<tr>
<td>Non-working children</td>
<td>34</td>
</tr>
<tr>
<td>Parents of non-working children</td>
<td>12</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>291</strong></td>
</tr>
</tbody>
</table>

**Main findings**

Children below the age of 18 years were involved in different activities related to the mining sector, the largest group being children aged between 14–17 years, who accounted for around 60% of the total number of children interviewed. The main explanation for the age concentration of child labourers in the range 14–17 years is that for these children, faced with limited chances of joining secondary education or vocational training, working in the mines can be the only available option. The highest number of children involved in the mining sector was observed during the school vacations and the lowest during the school terms.

There were fewer girls (20%) than boys (80%) involved in child labour. Mining is an activity mostly for males. In some places the district officials had actually prohibited girls from working in the mines in an effort to keep them from engaging in prostitution related to the mine sites.
The most prominent explanation as to why children involve themselves in child labour was related to the economic situation of the family, mainly the parents’ inability to provide for their children. Children had to either drop out of school or had to work part-time to meet their educational expenses. Children also had to work to support their families, and parents were noted to highly value their contributions. A majority (54%) of the children who had either dropped out of school or had completed their primary education and were now engaged in mining activities, came from either female-headed families or were orphans. The disintegration of the extended family structure has also led to loss of welfare of orphans by family members. Orphans are frequently left to fend for themselves.

Children received payments, either in kind or in cash, the latter accounting for over 75%.

Children as young as 10 years old were involved in the drilling of rocks, washing of rock dust, and collecting and carrying pieces of crushed rocks - all hazardous tasks. Children worked in very hazardous conditions, predisposing themselves to a number of health problems. There were also health risks, which were subtle and indirect, making the adverse effects they caused not immediately noticeable. This was especially true in the case of exposure to mercury. Children were exposed to mercury in the amalgamation process, and mercury was also disposed of into rivers thereby contaminating water sources. Some (12%) children were involved in related activities other than actual mining, like working in restaurants, bars and shops, and carrying out other errands. Tools and equipment used in the mining process were rudimentary in nature and required muscle power. Such tools included axes, pick axes, iron rods, chisels, sacks and buckets. Children were forced to strain themselves to get the work done.

Children worked with little time to rest or play, usually under direct sunshine, exposing them to high temperatures and wind. On average children working in the mine pits worked for a total of six hours per day, but this varied among different age groups. Children working in the restaurants, bars and shops worked much longer hours (10 hours per day on average). Variations were also noted on the average working hours of children working while attending school and children engaged more permanently in mining activities. Working days varied with respect to the mining site and whether schools were closed or open.

A.5 Manufacturing and handicrafts

Example A. Children in India’s carpet industry

The following account is based on Ashraf (2009b), *Children in India’s Carpet Industry.*

Child labour in India’s carpet industry is a complex phenomenon that has been compounded by social and cultural factors encouraging the continuity of trade and skill in particular castes or communities. Most of the work performed by children is also performed by adults, and most children work side by side with adult labourers. Children possess no irreplaceable skills for this work and could be replaced by adults, but the situation is complicated by a fragmented and informal industrial structure that relies extensively on subcontracting. The Indian carpet industry has few large businesses and consists mostly of small family enterprises. Around 90–95 per cent of
exports are woollen Persian hand-knotted pile carpets. The exporter is the prime figure around whom the overall organizational structure of the carpet industry revolves. He is the link between foreign importers and local contractors. Local contractors, in turn, are the link between the exporter and the loom owners and weavers. The weavers are at the bottom of the hierarchy. Weaving is given out by the exporters through the contractors to the loom owners, who in turn hire additional weavers. Wage payments are made on a piece-rate. Contractors loan small amounts to the weavers during emergency or festivities, with deductions subsequently made against the weaving payments.

Carpet weaving is a male-dominated activity, where two or more weavers sit together in close physical proximity to each other. The loom sheds are usually part of the loom owners’ houses, and loom owners hire the weavers. Female weavers are rare. For a woman to go to somebody else’s house to work is not socially acceptable in rural India. Adult weavers are paid so little that they are forced to bring their (male) children, or the (male) children of relatives, to weave with them. About 80 per cent of weavers are illiterate.

“Weaving is a misnomer for Persian carpets, as the process consists of tying a knot of woollen yarn to a cotton thread, cutting it with a sharp, curved knife, repeating the process endlessly. It is sometimes said that children are hired because their ‘nimble fingers’ enable them to tie more knots per square inch, thus improving the quality of the carpets. This is not true. The process requires strong fingers, particularly for high-quality carpets, and the best weavers tend to be between eighteen and thirty-five years of age. Children are more liable to cut their fingers while cutting the woollen thread after tying the knot. For the carpet industry, the nimble-fingers argument is merely another excuse for employing children, thereby also maintaining a pool of trained weavers for the future. Children tend to be more pliable and less likely to be absent than adult weavers. Therefore, loom owners prefer a mix of both adult and child weavers to meet quality standards and delivery deadlines. ... Statements on the number and percentage of child weavers have not always been properly substantiated. Estimates of child labour in the industry range from 4 per cent to 80 per cent of the workforce. More reasonable and careful statistical methods, however, estimate child labour in the range of 8 per cent to 30 per cent of total workers in the carpet industry. …”

Example B. Child labour in ‘subcontracting sectors’ of Indonesia’s garment and footwear industries

The following account is based on Tjandraningsih (2009), Child Labour in Subcontracting Sectors of Indonesia’s Garment and Footwear Industries.

The garment and footwear industries in Indonesia consist of home-based, small-, medium-, and large-scale industry, with a wide range of product diversification and market orientation. These industries are major sectors in the employment of children under fifteen years of age. Child labour is concentrated in small and home-based workshops where the product orientation is to the domestic market, and the production system is based on subcontracting to household production units. Small and home-based workshops, where most children are employed, usually cluster in certain geographic areas, thus creating small industrial centres. (This
Sampling elusive populations: Applications to studies of child labour

pattern has important indications for sample design.) For example, Cibaduyut is a small industrial footwear centre well known in West Java. Several local villages support the industry, and most of the villagers work as shoemakers. In almost every household, the whole family works in shoe making, some working in other people’s workshops, and some bringing work home through the putting-out system. “Under such circumstances, working from an early age becomes common, being part of the socialization process. Parents rarely forbid their children to work. Instead, they often encourage it, and the children consider working as a way to earn money, and as an arena for playing and socialising with friends. … While the legal minimum age for work in manufacturing is fifteen, in the garment and foot-wear sectors one can easily find younger children working. Some children receive their own wages, working on a piece-rate basis. Others are unpaid members of a family-based workforce. Girls are more heavily involved in the garment industry, … This work is usually performed in the living room of the employer’s house or in the girl’s own house. Boys dominate in the footwear industry... This work is usually performed in a workshop dedicated to footwear production. In both sectors, children perform the same work as adults”. For school-age children, the work in both industries is conducted after school for 4-5 hours per day, with one day off per week. The children who do not attend school work full-time, eight hours per day. Overtime is common, however, usually amounting to 3-4 additional hours per day – and more during peak season.

Work in the garment industry is relatively safe in comparison to work in the footwear industry, which uses more chemicals, particularly the pungent glue that causes headaches. In addition, the use of blades, scissors, and other cutting utensils creates an occupational hazard.

Example C. Child labour in the Tibeto-Nepalese carpet industry

The following account is from O’Neill (2009), *Child Labour in the Tibeto-Nepalese Carpet Industry.*

How many child labourers? There is little doubt that child labour played an important part in the formative years of the industry as it grew from a cottage industry that employed a handful of Tibetan refugees in the 1960s into Nepal’s largest foreign-exchange earner in the early 1990s and the employer of hundreds of thousands of people. The actual number of child workers remains a mystery, however. “In 1992 the Nepalese child-protection NGO Child Workers in Nepal (CWIN) estimated that fully one-half of the 300,000 workers employed by the industry were under eighteen years of age. However, a 1994 estimate by USAID put the number of all workers at 108,000, and fieldwork conducted in 1995 resulted in an estimate between 80,000 and 90,000. “CWIN’s estimate was thus exaggerated, perhaps intentionally, to counter the government of Nepal’s equally implausible claim that only 11 per cent of the weaving force was underage prior to state intervention, which reduced the level to less than 1 per cent by 1994.”

The children’s lives, according to reports, “were being broken by persistent illness, undernourishment, harsh physical discipline, and sexual abuse; they were separated from their families and alone in a hostile environment, cheated of education and even of the earnings that should rightfully have been theirs. They were being trafficked from the less-developed mountainous and lowland regions that surround Kathmandu, lured
by the enticements of work and life in the big city, sold to labour contractors by their destitute families, and tricked or even kidnapped by contractors looking for a cheap and docile labour force to feed the growing industry”. These reports made sensational reading that outraged child protectionists around the world, earning the industry international condemnation (Sattaur 1993; NSAC 1998). As carpet exports increased greatly in the early 1990s, the demand for weaving labour grew proportionately. Numerous entrepreneurs and would-be entrepreneurs, many with little or no experience in carpet weaving, set up factories of all sizes to try to capitalize on global demand. Thousands of migrants moved to the Kathmandu Valley, some in families or groups of contracted workers, while others came alone to earn a wage that far exceeded what they could earn in their villages. “Unquestionably, abuses occurred during this exodus, and probably continue today, although on a much-reduced scale. But O’Neil (2009) qualifies the above picture of exceptionally harsh conditions of child workers in the industry. “The problem with such representations of child carpet weaving was that little effort was made to put them into context, the image of the child carpet weaver, particularly the image of the very young child forced through debt bondage or other forms of domination to weave carpets, is not consistent with the experience of most young carpet weavers in the industry, either then or now”.

Example D. Children in India’s glass industry

The following account is based on Ashraf (2009c), Children in India’s Glass Industry.

A range of glass products are produced in India, including domestic wares, scientific and laboratory wares, glass chandeliers, and glass bangles, a traditional accessory which has a symbolic significance in the cultural system of India. Child labour is concentrated mainly in the production of glass bangles, rather than the glass industry as a whole. The glass sector producing bangles has a complex production structure, with work divided into many separate activities. Much of the work is done through subcontracting to small, informal-sector enterprises. The different production steps are performed by a specified manufacturing unit under subcontract, involving piece-rate systems of payment. It is important to understand the subcontracting industrial structure because working conditions, pay rates, and use of child labour vary among the stages of production. The participation of child labour also varies a great deal by type of enterprise. Estimates of child labour vary from 10 per cent to 50 per cent of the workforce, depending on the stage of production and the type of work.

(1) At the factory level, the first stage of making bangles is done at the furnace by a team of around 25 people, mainly male adults, but also employing a few children. Molten glass on a long rod is softened and transformed into a thin wire, which is subsequently rolled into a spring-like spiral on a small rotating rod and cut into rings.

(2) These rings are sent to smaller informal enterprises for further processing; this is where child labour generally occurs. The glass rings are straightened by girls and women. The work is performed in their houses. Small kerosene burners are kept in a closed room, with no ventilation, so that the flame does not move, and as a result, the girls and women inhale the poisonous smoke.

(3) The cut in the bangles is joined together in small rooms, also equipped with kerosene burners. Again, ventilation in the room is poor so that the flames
remain steady. Male children and adults work side by side. Both cutting and joining involve long hours of sitting in one cramped position.

(4) The glass bangles are fragile after the joining stage and must be hardened. The hardening units are small furnaces housed in small thatched structures. Each furnace requires a team of 3–4 male persons, at least one of them a child. After the bangles have been hardened, workers, generally male adults, etch intricate designs onto them by holding them against a spinning wheel with abrasive ridges.

(5) Hardened bangles are sent for colouring. The chemical colours, toxic in nature, are applied either by small brushes or by hand. The colouring of bangles by hand is generally done by young girls. Colouring has a toxic impact on the girls, as chemical-based colours stick to their hands and are difficult to remove. When they are asked, the girls proudly say they are saving money for their marriage. The irony is that after some years of this work, many are afflicted with tuberculosis. Pollution levels are high, and respiratory problems, throat problems, and burns are common.

A.6 Construction

Example A. Child labour baseline survey in Uganda

The following description is based on report of The Child Labour Baseline Survey conducted by Uganda Bureau of Statistics (Uganda, 2010).

Study objectives and methodology

The survey was conducted in three districts of the country. The sample was a conventional two-stage area and household-based design. The study targeted communities and all households with children in the focus districts. Enumeration areas (EAs) from the 2002 Population and Housing Census, along with counts of household with children, were used as the sampling frame. Irrespective of its population size, each district was allocated the same number (36) of sample EAs. Independent representative samples of EAs were selected from each of the districts using population proportional to size (PPS) sampling, with the number of households in the EA with children taken as a measure of size. A fixed number (15) households containing children were selected from each sample EA. The sample distribution by district is shown below.

<table>
<thead>
<tr>
<th>Survey district</th>
<th>No. of households</th>
<th>% of households</th>
<th>No. of sample EAs</th>
<th>No. of sample households</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Mbale</td>
<td>36,024</td>
<td>19.6</td>
<td>36</td>
<td>537</td>
</tr>
<tr>
<td>2 Rakai</td>
<td>48,468</td>
<td>26.4</td>
<td>36</td>
<td>540</td>
</tr>
<tr>
<td>3 Wakiso</td>
<td>98,843</td>
<td>54.0</td>
<td>36</td>
<td>540</td>
</tr>
<tr>
<td>Total</td>
<td>183,335</td>
<td>100</td>
<td>108</td>
<td>1,617</td>
</tr>
</tbody>
</table>
The survey objective was to collect information on the main characteristics of working children and those of the households they live in, including demographic and work-related information. In addition to a questionnaire administered to each sample household, 4 focus group meetings and 2 key informant interviews were organized for each of the districts in order to obtain additional qualitative information.

**Children involvement in work and school**

The survey results show that around 35% of children aged 5-17 years are engaged in economic activity. Among these, 10% of children are involved in economic activity only, without going to school. About 25% of children work in an economic activity and attend school at the same time, the figure being very similar for boys and girls. The results also show that 16% of children are not involved in economic activities and are not in school. The fact of being orphaned does not seem to be strongly related to the proportion working or the proportion attending school. Those with both parents lost are slightly more likely to be economically active, but surprisingly also slightly more likely to be attending school, compared to children with parents alive.

When involvement in economic activity and household chores is assessed, over half of 7-17 year-olds were engaged in housekeeping activities or household chores, with almost a similar proportion of males and females being involved. Fetching water is the main non-economic activity. Children’s total participation in work includes involvement in economic and non-economic activities, and in many case also attending school. Among the age group 5-17, 18% are involved in all the three types of activity.

**Characteristics of children’s work**

Agriculture is the lead sector in which children work: agricultural and fisheries work accounting for nearly 85% of child labour. The table below shows the children’s location of work. Over 65% of children work in places such as plantations, farms and gardens, and 25% work in family dwellings. They start work at the age of 8 years. However, the results also show that, although most children are engaged in agriculture it is not very work intensive. Generally, children in the manufacturing or services sectors work for much longer hours.

<table>
<thead>
<tr>
<th>Per cent distribution of working children according to location of work</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Plantation/farm/garden</td>
<td>66.7</td>
</tr>
<tr>
<td>At family dwelling</td>
<td>25.1</td>
</tr>
<tr>
<td>Shop/market/Kiosk</td>
<td>2.1</td>
</tr>
<tr>
<td>At employer’s house</td>
<td>1.3</td>
</tr>
<tr>
<td><strong>Construction site</strong></td>
<td>1.9</td>
</tr>
<tr>
<td>On the street</td>
<td>0.5</td>
</tr>
<tr>
<td>Industry/factory</td>
<td>0.2</td>
</tr>
<tr>
<td>Quarrying site</td>
<td>1.0</td>
</tr>
<tr>
<td>Other</td>
<td>1.3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>100.0</td>
</tr>
</tbody>
</table>
Sector-specific versus general child labour survey: some implications for the results

The baseline survey illustrated above is based on a household sample of the general population, and not tailored to capture any particular sector of child labour, such as construction being discussed in the present section. It is in fact a common situation that separate sector-specific surveys are not affordable (or sometimes are not feasible or not considered necessary). The alternative is to use a single survey to obtain a broad picture covering all or most sectors of child labour, rather than to carry out separate surveys focussed on individual sectors. The quality and completeness of the sector-specific information obtained by the combined survey is of course likely to be inferior.

Nevertheless, such a combined survey can still yield useful information for different sectors of activity, including their relative size and basic characteristics. Sample design issues in such combined surveys will be elaborated in Chapter 5 below. The table below shows, as an example, the activities and conditions identified as hazardous in the construction sector in Uganda.

<table>
<thead>
<tr>
<th>Activities</th>
<th>Conditions under which the work is hazardous</th>
<th>Risks and possible consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brick making</td>
<td>Exposure to chemicals</td>
<td>Burns</td>
</tr>
<tr>
<td>Portering</td>
<td>Exposure to fumes, dust</td>
<td>Musculoskeletal injury</td>
</tr>
<tr>
<td>Carpentry work</td>
<td>Exposure to fire, excessive heat</td>
<td>Cuts and wounds</td>
</tr>
<tr>
<td>Building</td>
<td>Working long hours</td>
<td>Respiratory diseases</td>
</tr>
<tr>
<td>Road construction</td>
<td>Carrying heavy loads</td>
<td>Fatigue</td>
</tr>
<tr>
<td></td>
<td>Excessive noise/vibration</td>
<td>Loss of hearing</td>
</tr>
<tr>
<td></td>
<td>Exposure to dangerous tools</td>
<td>Stunted growth</td>
</tr>
<tr>
<td></td>
<td>Exposure to dangerous heights and depths</td>
<td>Deformity</td>
</tr>
</tbody>
</table>

A.7 Street work and the informal sector

Example A. Child Labour and the Urban Informal Sector in Uganda

The following account is based on Uganda (2004), Report of the sectoral study on child labour and the urban informal sector in Uganda.

Methodology

This is a study of labouring children aged between 5-17 in the urban informal sector in Uganda. The informal sector is defined as comprising of micro and small enterprises (MSEs). Micro enterprises employ less than 5 people while small enterprises employ between 5-20 people. Micro enterprises are largely dependent on family labour or own account employees and have low levels of productivity and limited skills. The small enterprises by comparison tend to use better technology, usually operate from a fixed location, attempt to maintain records and employ more wage labour. The study
employed a mix of quantitative and qualitative tools. Data were collected from urban centres having concentrations of the informal sector in four districts. Quantitative data were collected from 433 working children, 249 informal sector establishments (owners), and 172 households. Qualitative data were also collected.

The population of this study was divided into three strata namely; (a) the working children who are the victims of child labour; (b) the households (heads) who are employers and/or sources of child workers, and (c) establishments (owners) who are employers of child workers. From each of the strata, a simple random sample was drawn independently. Key informants providing quantitative data included local leaders, cultural leaders, religious leaders, and others participating in focus group discussions. Working children, household heads and employers were the principle sources of quantitative data.

**Main substantive findings**

(1) Considering all children aged 5-17, a majority (85%) are currently attending school, 11% left school while 4% never attended school. Gender disparities in school attendance are not significant. Around 27% of the children had a job in the last 12 months. Thus the proportion of child labourers in the urban informal sector is extremely high.

(2) As to working children among the above, a vast majority (88%) are out of school. The majority (78%) have left school, 10% have never been to school and only 12% are currently attending school. A majority of the working children have lost one parent (30%) or both (31%). A majority (63%) of working children have either shifted or migrated from one place to another, nearly 2/3 of them from other districts. More girls than boys have moved across districts. Children are involved in many informal sector activities including hawking, domestic work (cooking and serving food, house cleaning, trading assistance, working in bars and restaurants, and prostitution). A large majority (81%) work outside private households, the proportion being higher among girls than boys. More girls than boys are involved in bars and restaurants as attendants, in domestic service as servants, as hair dressers, or as prostitutes. Conversely, more boys are involved in hawking, garage or repair works, carpentry, or scavenging.

Child labourers work an average of 61 hours per week; female children work six more hours than male children. Of the working children 24% are not paid for the work they do, while 64% are paid in cash, 8% are paid in kind and 4% benefit from profits they get from the products they sell.

(3) Nearly all (97-98%) the children who are ‘working’ can be classified as ‘child labourers’. The difference arises because, in accordance with the standard definitions (see table below), some types of child work are not considered child labour. The difference between these two groups is very small in Uganda. This means that nearly all the working children aged 12-14 are involved in heavy work, and all the working children aged 15-17 are involved in hazardous work or have long working hours. In fact, a majority (82%) of all child labourers are involved in hazardous work.
Sampling elusive populations: Applications to studies of child labour

Child work
This term generally refers to activities that children carry out within or outside their households for income, family gain or profit, including unpaid family work. Such children are often described as being economically active, which is a broad concept that encompasses most productive activities undertaken by children, whether for the market or not, paid or unpaid, for a few hours or full-time, on a temporary or regular basis, legal or illegal. The term excludes schooling and chores undertaken in the child’s own household.

Child labour
This study defines child labour as comprising of:

i) all children involved in work aged 5-11 years;

ii) all working children aged 12-14 years involved in work beyond their capacity or work which is not ‘light work’ as provided for in national legislation, or they work for a total of 14 hours or more per week; and

iii) all working children aged 15-17 involved in hazardous work, or if they work an equivalent of 43 or more hours per week. This definition of child labour fits within the ILO Conventions on child labour and national labour legislation.

(4) The principle causes of child labour i.e. why labouring children are working, is related to: the high orphanage rate (62%); the high cost of primary and secondary education; and deepening poverty in households.

(5) Concerning child labour and the health and safety of children, seven out of every ten labouring children report falling sick in the last six months; 58% said they had to be off work for an average of 5 days because they were sick in the last 30 days. Besides sickness, nearly half the children report injuries in the past 6 months, with boys and girls nearly equally affected. Injuries declined with increasing age of children.

(6) As to attitude concerning child labour, the proportion of household heads who are against child labour based on an overall index is 36%; while 23% support child labour and the remaining 40% do not care. Employers felt that employing children was done out of sympathy to help them get basic needs such as food and in some cases school fees, but evidence suggests they do it largely to reduce labour costs (establishments employing children spent 8% less on the wage bill as a proportion of their total earnings, compared to those not employing children).

(7) In conclusion, the proportion of child labourers in the urban informal sector in Uganda is very high. Nearly all working children in the sector are doing work which is incompatible with their status, or work which affects them adversely.

Example B. Street children in Mexico

The following account is based on Peralta (2009), Street Children in Mexico.

In Mexico, at least 1.5 million children live and work on the streets as part of the informal economy. Mexico City has more than 650,000 children, with no legal protection, working and living on the streets. The greatest increase in the number of street children in Mexico in the last two decades has been experienced in cities along the U.S.-Mexico border. Street-working children are likely to come from impoverished families who reside in disadvantaged neighbourhoods. Economic conditions in Mexico
have greatly increased the number of street children. Budget cuts at the level of city government, inspired by neoliberal political-economic policies, have eliminated most social programmes for low-income families and street children (Magazine 2003).

Street children in Mexico share many similarities with their counterparts in other Latin American countries. They are disproportionately male, are on the street primarily to earn money, are more likely to be truant, come from large families, and report a history of physical abuse. In the last decade, there has been a significant increase in the number of girls living and working on the streets. Young females have migrated from the countryside to large urban centres looking for employment. There are already a significant number of young women who have given birth to children and started street families.

A series of comprehensive studies of street children in two important urban centres in Mexico identified three distinct types of street children: the independent street workers; the family street workers; and the children of the street. These distinctions are very important because each category involves different types of work, different relations with parents, different relations to governmental authorities, and has different needs. They are also likely to differ in the required sampling procedures.

**Independent Street Workers**

The independent street workers are the largest group, comprising more than 50 per cent of the street children in Mexico. The majority of these children attend school on a regular basis. Most of them live with their parents and work to supplement the household income. The majority of the children work selling items in the streets such as candy, seeds (semillas), or toys. Major problems in this group are the significant reports of child abuse by their parents and arrests by local officials.

**Family Street Workers**

Children in the family-street-worker category make up more than 30 per cent of the children working and living in the streets of Mexico. These children work in the streets with their parents, or they work with siblings and have daily contact with their parents. In this category, girls play a significant role. Most commonly, family workers sell food or candy to pedestrians. In most cases the parents prepare food items at home for sale on the streets by their children. There are also a number of children who are musicians, dancers, and other street performers. In this group only a handful of children attend school. One major problem with this group is the long hours worked. Children work after midnight on a regular basis.

**Children of the Street**

While this group represents less than 20 per cent of street children, children of the street are extremely vulnerable because the streets have become more than a workplace. Most of these children have very little contact with their parents and sleep on the streets on a full-time basis. These children do not attend school, and there is significant involvement in youth gangs, substance abuse, and youth institutions. Most of them have been arrested and have major psychological problems. Much of their work involves illegal activities such as stealing and prostitution. Children in such circumstances receive little or no adult supervision or protection. They are
exposed to the hazards of their physical and social environments, thereby facing the risks of being victimized by adults or by other youngsters. Few perceive public officials as their protectors. Instead, children rely on other children and street adults for protection. This has resulted in violence against children. Half of the children interviewed reported a traumatic experience while living or working on the streets. Most commonly they referred to being beaten, being mugged, and having their money and goods stolen. There has also been a significant increase in sexual attacks on children throughout the country.

**Example C. Street children in Vietnam**

The following account is based on Masina (2009), *Street Children in Vietnam*.

Vietnam is a poor country, but a country in which economic transition is substantially successful in creating broad-based improvement in living conditions in both urban and rural areas. This affects the reality of street children in Vietnam.

The term *street children* is used to cover a wide array of conditions of children, ranging from teenagers working in the urban informal sector to abandoned young children. While these different conditions cannot be easily categorized, it is important to distinguish among cases in terms of both the problems the children face and the policies that should be implemented to shelter them. Again, the different groups may differ in the sampling procedures required.

1. One very large group of street children is represented by young peasants migrating to urban areas for seasonal or temporary engagement in the informal sector. Many teenage girls move to towns as street vendors (but also as nannies, waiters or shop assistants). Teenage boys typically work as shoe shiners, newspapers sellers, peddlers, or market haulers. Their work in the urban informal sector is part of the livelihood strategy of their rural family. These young people are defined as street children because they often literally live and work on the street. Although they often have a caring family back in the village, in their life in town they are vulnerable. They tend to move back to their home villages after a time.

2. A second group of street children is represented by young children who have lost the protection of their families. Poverty is a contributing factor, but often not the main reason, in forcing these children onto the street. Many children report stories of broken families and abuse. Although life on the street in Vietnam is safer and less violent than life in many other developing countries, these children may be manipulated by criminal groups, exposed to drugs, or even sexually exploited - these risks are more apparent in Ho Chi Minh City than in Hanoi or smaller towns.

3. A third group of street children is represented by children moving to town with migrant parents, particularly migrant mothers, with informal occupations such as garbage collection for recycling. Very young children collecting garbage can be found in the markets of major towns, often wandering around in small groups. These children are normally far from parental protection and supervision for large parts of the day, even at very young ages, and are therefore exposed to severe risks.
Example D. Children in street trading in Nigeria

The following account is based on Aderinto (2009), *Children in Street Trading in Nigeria*.

There are no precise estimates of the number of children involved in street trading in Nigeria. National labour force data exclude children, and, even if they were included, informal-sector workers are undercounted. In spite of the scarcity of data, there is no contention about the soaring number of children involved in street trading in Nigeria. Anecdotal reports from most Nigerian cities suggest an increasing number of these children. Children involved in street trading can be found in both urban and rural areas, though different reasons account for their involvement.

On the streets, children are engaged in all manner of economic activities, including load carrying, bus conducting, water fetching, scavenging, hawking and vending, begging, and prostitution. Hazards associated with street trading are immense, including physical exertion, road accidents, contamination of cuts and sores because of walking barefoot, recruitment into gangs, use of psychoactive drugs (mainly among boys), assault, and vulnerability to ritual killings. A worrisome development is the increasing number of children who leave their homes to stay permanently on the streets, living under bridges, in motor parks, or in empty buildings.

Street traders are often exposed to early initiation into sex. Among girls, this may result in pregnancy followed by illegal abortion, single parenthood, or baby abandonment. Among commercial sex workers, the risk of AIDS is high, with HIV seroprevalence rates in prostitutes estimated at 34 per cent.

Two broad categories of children involved in street trading can be observed. The first, major (90 per cent), category involves street trading during school hours among children who have dropped out of school due to either poverty or persistent failure in school. The second, minority (10 per cent), category tallies with the assumed socialization function of street trading, and involves children who trade at specific times of the day, early in the morning before school and late in the afternoon after school. For these children, street trading is not disruptive of education, and therefore is acceptable. There is an important gender dimension to child street trading also. While boys comprise the vast majority of street traders, girls are less likely to be in school, are more likely to be on the streets during school hours, and tend to work longer hours than boys.

Example E. Child street vendors in Brazil

The following account is based on Kassouf and Ferro (2009), *Child Street Vendors in Brazil*.

In Brazil, the Brazilian Geographical and Statistical Institute (IBGE) undertakes an annual national household survey that collects information on children working from the age of five. According to the 2005 survey, among the 5-17 year old population, 12 per cent are working, and among those working, 5 per cent are working on the streets. Of the street child workers, half (52%) are selling candies and other goods, a third (33%) are polishing shoes, guarding cars or delivering papers, and the remainder (15%) collect recyclable materials. (However, since this was a household
survey, the data exclude children who are also living on the streets.) Street vendor work is concentrated in large metropolitan areas. The great majority of street working children are boys, about 80%, as compared to the percentage male among children working in other economic activities (63%).

Twenty per cent of the 7-17 year old children that are working do not go to school, compared to only 6 per cent of those not working. The loss of education caused by having to work long hours at an early age has the effect of limiting the child’s future employment opportunities to unskilled, low-paid jobs, thins locking the child into a repetition of the cycle of poverty that the child’s parents experienced.

Example F. Working street children in Turkey

The following account is based on Akşit, Karancı and Gündüz-Hoşgör (2001), Turkey Working Street Children in Three Metropolitan Cities: A Rapid Assessment.

Scope and methodology of the study

The research employed the ILO-UNICEF rapid assessment methodology (see Section 3.1.4 and 3.4). Four different types of data collection were carried out to provide a picture of the families, work conditions, school attendance and attitudes of children working in the streets: (1) semi-structured interviews with 188 working children, and 65 parents and household members in the three cities; (2) in-depth and focus-group interviews with experts from various related institutions and customers; (3) observations of the children’s work sites and their homes; and (4) literature review of study reports on children engaged in street work and other related material.

Rural-to-urban migration is an important aspect of the reality in Turkey. The study was carried out in three metropolitan cities. On the basis of census data and surveys, it was observed that the three cities constitute three different degrees of urban hierarchy and development in Turkey. A great majority of the working street came from recently migrated families, either directly from rural areas to large urban centres, or step-by-step from smaller to larger localities.

Children working on the streets can be classified into two groups. The first group of children works on the streets during the day, sometimes during the evening and night, but finally goes home to stay with their family. They are supposed to be under the protection and supervision of their families. The great majority of children working in the streets in Turkey belong to this first group, and are involved in selling napkins, chewing-gum, water, sunflower seeds, lottery tickets and other small items appropriate to the season. Sometimes they are involved in more dangerous activities.

The second group of children are working and living on the streets. They have left their homes and/or their families have disintegrated. They are children of the streets. The children in this group are involved in collecting and sorting garbage on the streets as well as at garbage dumps. Since they are outside of family and community protection and supervision, they are more likely to be involved in drug abuse, street gangs and violence.
Children’s characteristics

The children interviewed were 7-17 years old. The majority were males. They were from very poor, large nuclear families with an average household size of 7.8. The majority of the mothers were illiterate housewives. Fathers were generally more educated. There was a high rate of unemployment among fathers of these children. The majority of employed fathers were working in the informal sector. The majority of children in the study had attended or continued to attend school. Only 13% of the sample had never attended school; however, school dropout rates were high: 25% of all children interviewed. The children interviewed were having difficulties in school in terms of the attitudes and behaviours of their teachers and peers towards them. They felt that their educational needs were not being met in school due to the crowded classes and lack of personal attention. Economic hardship was the main reason given by children and their parents for dropping out of school. In other words, working in the streets was either the result or the cause of children dropping out of primary education.

Harsh working conditions and consequences

Shoe polishing, selling different goods and scavenging are the most common types of work activities. Children work in congested city streets and intersections, and are exposed to noxious pollutants risk of being run over from motor vehicles, and to abuse from older gangs of drug abusing street hawkers and customers. The work the children carry out impairs their health and socialisation, and puts their safety at risk. Children suffer from working long hours and fatigue. Children face physical hazards, for example, child garbage pickers face extreme risks of developing long-term diseases. Children working at night are in addition exposed to inappropriate adult models and sexual abuse. They may become voluntarily or involuntarily involved with petty crime and drugs and face abuse from gangs of older children or from adults. Children work for long hours throughout the week. Scavenging is the worst form of work that these children engage in. This exposes the children to infectious diseases.

Most children in the study did not like their work. They said that they would not like their younger siblings to work in the future and that they believed that children of their own age should not work. However, they were under pressure from their families to go out to work and to bring back money. The majority of children gave their earnings to their families, mainly to their mothers.

The majority of parents seemed to be ignorant about the risks their children were exposed to while working in the streets. Customers, however, generally had a sympathetic attitude towards these children and believed that giving them money was a form of social help. Often, child labour is legitimised by its cultural acceptability, and working on the street is considered to be an ‘apprenticeship’ for adult life, teaching children self-discipline and how to overcome hardships in life while contributing to family income. It was also seen as a form of family solidarity in families whose adult male members were unemployed. Child work was preferred to begging; however, it often was also recognised to be a form of begging.
Example G. Child ragpickers in Nepal

The following account is based on Bal Kumar, Gurung, Adhikari, and Subedi (2001), *Nepal Situation of Child Ragpickers: A Rapid Assessment*.

This study is based on ILO-UNICEF Rapid Assessment (RA) methodology. It was aimed at obtaining in-depth knowledge of a given phenomenon within only three months. Primary information consisted of both quantitative and qualitative data, with an emphasis on qualitative data that may shed light on the plight of children working as ragpickers in Nepal. Secondary information was obtained from the limited existing studies. It should be noted that most prior studies on this topic were developed in the broader context of street children, not ragpickers specifically.

Ragpicking and characteristics of ragpickers

The term ragpicker refers to people who collect rags or recyclable materials that can be sold for money. Ragpicking entails the sorting, collecting and selling of these various waste materials that can be found at dumpsites, riverbanks, street corners, or in residential areas, and consist primarily of plastics, bottles, cardboard, tin, aluminium, iron, brass, and copper. Ragpickers get paid according to the quality of the materials they sell to junkyards and garbage collection centres. As the market for recyclable materials has increased, many street children have turned to ragpicking as a means of survival. Based on this study it is estimated that there are nearly 4,000 children engaged in ragpicking in the various urban centres of Nepal.

Most children who work as ragpickers come from poor rural and landless families involved in non-agricultural, low-paying occupations, specifically from hill and mountain regions. The majority migrate from rural to urban areas. Many of them leave their families behind. Some children migrate with their families. More boys than girls engage in ragpicking, and the majority of boys live on the street, whereas almost all the girls reported that they were living in rented or family homes. The average age of a ragpicker is 12. This study interviewed a total of 300 ragpickers: 264 boys and 36 girls. The school dropout rate amongst ragpickers is high, though the children reported that they would attend school if the financial support were available. Literacy rates of ragpickers are comparable to national averages, with the majority literate, and boys more literate than girls overall.

Conditions of ragpicking

Ragpickers live from day to day, usually spending all the money they earn in a day, and are still often left without enough money to feed themselves. Though 92% of the children interviewed reported to work year-round, ragpickers are most likely to go hungry during the rainy season when materials are wet and dirty and collecting them is difficult due to the rain and mud. This draws the children to either borrow money from junkyards, which puts them in a bonded labour situation, or causes them to resort to petty. In addition to experiencing hunger, it is common for ragpickers to fall sick due to exposure to contaminants or bacteria, the cold, and wounds inflicted by sharp glass or metal objects. They are highly vulnerable to drugs as they are likely to be exposed to addicts and suppliers as well as to peer pressure. Ragpickers tend to be territorial about their collection sites and as a result they are involved in gangs
and are exposed to gang violence. Many of the children interviewed have been used in drug peddling and commercial sex as middlemen, or been sexually abused. They are thus vulnerable to HIV/AIDS.

Ragpickers interact relatively positively with other street children and NGOs, but face negative interactions with the police, civil society, junkyard owners and criminal gangs. They live in a world that teaches them violence and abuse and exposes them to unhealthy behaviours and lifestyles.

The best hours to work are in the mornings and evenings when most people dispose of their garbage, but many ragpickers work for six hours per day in their effort to survive.

Example H. Children of Kolkata slums

The following account is based on Bagchi (2009). Children of Kolkata Slums.

Kolkata Metropolitan Development Authority data (1993) show that there are 2,011 registered and about 3,500 unregistered slums in Kolkata, which house 1.5 million people, or a third of the total Kolkata population. Slum dwellers shelter in overcrowded conditions with little prospect of having a safe water supply, hygienic waste disposal facilities, or adequate ventilation, with grossly insufficient municipal services and an inadequate primary health-care system. Absence of skills to work in the formal economy compels an overwhelming number of people in these slums to work in the informal sector. Only 100,000 of nearly 1.5 million people in the slums are literate.

The children from the slums of Kolkata generally start working at seven or eight years of age, normally working for more than eight hours a day, six to seven days a week. They work in a wide variety of settings. Some work in family businesses; others are self-employed in the informal economy; some are domestic helpers; or work in roadside food stalls and restaurants. Children work as waste pickers, porters, van and rickshaw pullers, bidi (cigarette) rollers, shoe shiners, and paper-packet makers. Many slum children work as beggars. Some are engaged as apprentices in various skilled work such as garage work, cooking, or construction. Much of this work is hazardous, exposing children to noxious gases, harmful chemicals, tobacco, dangerous fumes, dog bites, and other accidents that are sometimes fatal. Very few workplaces have any facility for hazard prevention, self-protection, first aid, or even toilets. The children often have to walk a long way to and from their workplace because they cannot afford any other mode of transport. They frequently receive harsh treatment from their employers or adult colleagues in their work-places. In return for their daily ordeal, many children earn pitiable amounts, sometimes none at all. As to their perceptions, children recognize that their families’ financial inability to continue their schooling is the main reason they work (Bagchi, 2006). Parents tend to cite the family’s financial inability to pay for schooling, or they maintain that apprenticeship to a skilled and lucrative job is a better prospect than schooling for their children. Most employers say that they employ children in their establishments only to help the families. Most of these working children consider their lives to be very harsh, but they have somehow accepted it as their destiny. They also tend to be optimistic about their future, for example expecting that they might earn more money to support their families.
It is possible that the promulgation of laws against child labour has proved to be counterproductive for the working children in Kolkata slums. “Recent measures, such as the 2006 ban on children working as domestic servants and in roadside eateries (dhabas) have only worsened the child labour situation in Kolkata by increasing surreptitious employment of children in informal establishments, or by bringing about the outright firing of children”.

**Example I. Children of Delhi slums**

The following account is based on Mitra (2009), *Children of Delhi Slums*.

The conventional distinction between economic and noneconomic activities is quite unsatisfactory for both the analysis and the measurement of child activity patterns in slum environments. Children may be engaged in more than one activity at a time, combining work and school; their work may be intermittent, or they may change activities frequently; they may or may not be paid for their work; or they may share work with family members or other adults.

Because child labour is so imperfectly measured by economic activities, some authors have used indirect evidence, drawing inferences from indicators related to education and poverty. For example, Gupta and Mitra (1997) estimated child labour in Delhi slums, based on education criteria, at about 34 percent; based on poverty criteria, the estimate ranged up to 65 percent. These are extremely high rates, and suggest that good samples of working children may be obtainable through surveying household economic situation in combination with enrolment and dropout statistics from schools in the slum communities.

**Example J. ‘Idle’ children**

In a study by Biggeri, Guarcello, Lyon and Rosati (2009), data from six countries - Brazil, Cameroon, Guatemala, Nepal, Turkey, and Yemen – were analysed in an attempt to address the issue of apparently idle children. According to the results, the proportion of children left out of both school and economic activity varies significantly across the six countries studied. Exclusion from school and economic activity is most common in Yemen and Cameroon, accounting for one-third and one-fourth, respectively, of 7-14 year olds. It is least common in Brazil and Turkey, where only around one in twenty children are neither in school nor economically active.

What might explain this large group of children left out of both school and economic activity?

“A number of possibilities exist. First and most obviously, these children might not go to school or work in economic activities because they are needed at home to perform chores... [But] household chores such as water collection or caring for younger siblings, which are technically noneconomic activities, constitute a major time burden for only a small proportion of idle children. Less than one-half of idle children in Cameroon, and less than one-fifth in the other five countries, put in four or more hours each day on household chores. Moreover, many economically active children and students also perform chores at home, suggesting that responsibility for household chores may not play a central role in exclusion from school and economic activity. Unemployment appears to play an important role in idleness only in Turkey, the one middle-income country examined.” *(ibid)*.
A.8 Unconditional Worst Forms of Child Labour (UWFCL)

Example A. Trafficking in children for labour and sexual exploitation

The following example is based on Albania (2004), *Rapid Assessment of Trafficking in Children for Labour and Sexual Exploitation in Albania*.

**Background and study methodology**

As a country in transition, from a central economy typical of the communist era to the free market economy demanded by structural adjustment, Albania has experienced huge political, economic and social transformations since the mid-nineties and has faced a number of challenges never before experienced, such as extreme unemployment and lack of job opportunities, decreased access to basic social services including education and health, and massive migration flows. Among these challenges, one that deserves special attention is child trafficking, unknown until 1990. Evidence shows that, after 1990, girls especially were trafficked from Albania to European countries for exploitation; between 1992 and 2002, an estimated 4,000 children were trafficked.

The study was undertaken using the standard ILO-UNICEF Rapid Assessment methodology. Standardized information from questionnaires, qualitative information from semi-structured interviews and existing information were the main research tools used. The study focused on identifying push and pull factors influencing child trafficking; sketching profiles for different categories of children who had been trafficked or risk being trafficked; detailing the trafficking process and ways of exploiting children. This was partly achieved by drawing up a picture of the daily life of trafficked children: the work they have to do, the control exerted by traffickers, the ways in which children exit the trafficking net and the process of rehabilitation and social reintegration of children who have exited trafficking. Children who had been trafficked and released, and children who had never been trafficked, completed 83 questionnaires. In addition, 63 semi-structured interviews were done with such children, parents, and key informants. Twelve group discussions were held in the cities chosen for the rapid assessment. Girls exploited in prostitution were especially difficult to reach. Access to trafficking victims became more difficult as children were identified but then often re-trafficked or moved across borders illegally.

**Characteristics and conditions of trafficked children**

Traffickers were in almost all the cases responsible for transporting the child from the country of origin to the destination country. In the majority of instances, the parents and the traffickers were involved in recruiting the children. Parents interviewed often said they hoped that their children would earn some income for the family. Frequently parents were not aware of the difficulties and challenges that their children would face. The preferred country for traffickers was Greece; trafficking to Italy was considered more difficult.

The study shows that boys aged 11-16 are most likely to have been trafficked or to be at risk, and that most of these children will have worked before the trafficking incident, often on the streets. When trafficked, most of the children trafficked live...
with friends who accompanied them when they left home. In general, they come from families where the parents do not have even a minimal education, and are divorced and/or remarried. The families live in poor conditions; large extended families live together and the households are characterized by social problems such as alcoholism, illness, or domestic violence.

Trafficked children are engaged in a range of activities that have proven profitable for traffickers. The majority of the children interviewed in the study were engaged in begging and hawking on the streets in Greece. Some children were involved in illegal activities such as theft; a smaller number were involved with drugs. Other children were variously exploited in agriculture, as waiters, running games of luck (boys only), housework, street-based car washing (girls and boys) and other such activities. Of those interviewed, 17 per cent said they had been involved in sexual services (girls); these children were particularly difficult to reach.

Trafficked children live under arduous conditions: these children are considered merely tools for profit; they are generally given only minimal food and lodging. There are no social, entertainment, schooling or training activities in their lives. No gift or positive reinforcement is given for the work they do, no matter how challenging. The profits of child labour are under the total control of the bosses. In the study a large number of children did not receive everything their ‘contracts’ had promised: food, lodging and clothes. Some of the children said they had to sleep in tents in very rough conditions. Control is an important element in the life and work of trafficked children. The level of control largely depends on the type of work and the working conditions of the child: whether they are out in the open, on the street or in closed quarters; the way things have been forced upon them or deals made; and the child’s age or sex.

The main push factors influencing child trafficking continue to prevail in Albania: poverty and the desire to earn a living, parental unemployment, dysfunctional families, domestic violence, low education levels of parents and children, lack of hope within the country, inadequate legislation and/or difficulties in its enforcement. Among the main pull factors influencing child trafficking is the widely held belief that a better life is to be had abroad.

Most of the children released from trafficking were returned home by the police after they were caught. Others found ways to escape; a few were helped to exit trafficking. Most of children released from trafficking continued to work in the streets after they returned home. In general, the parents of released children found it difficult to fully understand the trauma their child has suffered, they could identify only physical injury or health problems. The psycho-emotional damage caused by trafficking often goes unnoticed and untreated.

Example B. Children in commercial sexual exploitation in the Philippines

The following illustration is based on ILO (2004c), *Girl Child Labour in Agriculture, Domestic Work and Sexual Exploitation: Rapid assessments on the cases of the Philippines, Ghana and Ecuador.*
Methodology

The study made use of purposive sampling that allows the selection of a sample population that fits the pre-defined group, in this case boys and girls engaged in prostitution. Initially, a mapping of available statistics on children in prostitution, from national to regional down to provincial level, was done. Independent surveys/studies on children in prostitution were also obtained to supplement existing statistics. The process, however, failed to produce reliable estimates, given the paucity of quantitative information available on the sector. In the light of this, a purposive sampling was undertaken in which the characteristics of the sample respondents were predetermined. The number of the sample respondents was arrived at on the basis of cost in conducting the interviews. Forty-four children participated in the survey. Interviews were conducted with 33 girls and 11 boys aged 17 and below who were engaged in prostitution. In addition, parents of 10 of these children and other key informants, such as local government officials, were also included in the list of interviewees.

Several methods of data collection were utilized. These included conducting one-on-one interviews with the child-respondents, parents and some key informants using a structured questionnaire. Focused group discussions with local government officials, health workers, social workers, employers/managers, and police officers were also conducted to verify/validate and compare information provided by the children and their parents.

Profile of the children subject to commercial sexual exploitation

The children in commercial sexual exploitation involved in the study were aged 11-17. Most of the children had migrated from nearby localities. However, there were also cases where children in prostitution came from as far away as Manila. A considerable number came from broken families, had single or widowed parents, were orphaned, or had escaped from family members who frequently engaged in quarrels, violence or alcoholism. While working, many of the children lived with persons other than their families or relatives and, depending on their work arrangements, they lived either with their ‘employers’ or ‘co-workers’. A majority of the children claimed to be their respective families’ breadwinners and said that being in the commercial sex sector was their first job. All of the respondents appeared to be no longer attending school. Most of the children came from families which barely had access to education.

In many instances, it was their respective pimps who explained the terms of their actual employment. Most of their clients were local to the city. Most of the children barely received information about prevention and treatment of sexually transmitted diseases. In many instances, the children resort to self-medication to address health concerns such as sexually-transmitted diseases or birth control. Most are at high risk of acquiring and passing on sexually-transmitted diseases as they are not keen on using condoms during intercourse. This practice is further aggravated by the fact that some of them, especially those belonging to gangs, practise ‘partner-swapping’. While check-ups from the Local Health Office are free, the children admitted to not submitting themselves for regular health inspection.

A majority of the children have plans for leaving the industry for good, that is, after they have saved enough money or when they get tired of the job. Although a number of children noted that they would not have any difficulty leaving their current jobs,
a handful were candid enough to note that getting integrated into the mainstream would be difficult for them for fear of being looked down on, maligned, rejected or ostracized. Almost all of the respondents said that they would prefer not to be involved in this kind of activity if they are given a choice. Some said they would like to continue their studies; others still aspire to having more decent jobs.

Example C. Commercial sexual exploitation of girls in Ghana

As the preceding illustration, the following is based on ILO (2004c), *Girl Child Labour in Agriculture, Domestic Work and Sexual Exploitation: Rapid assessments on the cases of the Philippines, Ghana and Ecuador.*

Child labour is prevalent in both rural and urban areas of Ghana. In urban areas children living and/or working on the street mainly engage in trading activities, while in rural areas they are active in both agricultural and trading activities.

Child prostitution is a growing problem in Ghana. It is common for young girls to trade sex for gifts or money. As a means of survival in cities following migration, girls increasingly engage in commercial sex as an occupation. Street girls who work as porters have been observed to practice prostitution by night to add to their earnings. Most often this occurs without an intermediary, although organized prostitution is becoming a serious concern in some areas. The research that was undertaken for this Rapid Assessment Study attempts to indicate the degree to which young girls engaged in porterage work and petty trading as their primary occupation, and the degree to which they also engaged in commercial sexual work as their secondary or ancillary occupation. It highlights the link between escaping the drudgery of agricultural work in rural areas in order to work in cities.

In any analysis of the causes of child labour in Ghana, it is important to address the gender dimensions. Girls have relatively lower enrolment in school than boys at all levels of the formal education system. Girls in some instances are also known to be less aware than boys about sexually transmitted diseases. High-risk behaviour not only places boys and girls at increased risk of contracting HIV/AIDS and other STDs but also increases the possibility of unwanted pregnancies for girls. The paternity of their children is often unknown, or if known, is rejected. This places an additional burden of childcare on the girl workers.

A rapid assessment study of the problem

A variety of RA research methods were deployed in order to obtain the required information on working children, particularly girls. Success in getting to these young girls, some as young as 7 years old, depended to a great extent on the interviewers’ ability to gain the respondent’s confidence and trust. This was even more critical, given some of the questions were of a sexually sensitive nature, and if handled unprofessionally the respondents were liable either to lie or to abandon the interview altogether for fear of repercussions or sanctions, from the police or from their madams/pimps. There were naturally some serious issues of security and safety for the interviewers. Some of the chosen locations were notorious haunts of criminals and prostitutes. Special care had to be taken in order to record and analyse the data in a manner that was a true reflection of the accounts given and that was not in any way sensationalist or distorted. Twelve locations were used for data collection.
The research team comprised one team leader and fifteen researchers, including one gender specialist and one child psychologist, who were trained extensively on the application of the rapid assessment methodology. Four Rapid Assessment research methods were used in undertaking the study and to achieve the required triangulation, namely:

- Interviews with girl children involved in commercial sexual work as their primary or secondary source of income.
- Focus group with children, pimps, child welfare workers and other key stakeholders.
- Observations of physical locations associated with children engaged in commercial sexual and related activities.
- Review of secondary sources of information.

A total of 398 people were interviewed through a mix of: interviews (363 girl children); focus group discussions (35 adult stakeholders); observations in 14 different locations; and documentary reviews.

**Example D. Children engaged in drug trafficking**

The following summarises the methodology and main findings reported in de Souza e Silva and Urani (2002), *Brazil Children in Drug Trafficking: A Rapid Assessment*. The study compiled and organized data concerning living standards of children working in drug trafficking schemes in several low-income communities in Rio de Janeiro.

Drug abuse and trafficking are among the most commonly reported crimes by children in Rio de Janeiro. A decrease in the age of entry into drug trafficking is also worth stressing. The average fell from between 15 and 16 years in the beginning of the 1990s to between 12 and 13 in the year 2000.

**Rapid assessment methodology**

The collection of quantitative data carried out through semi-structured interviews in the above mentioned study seeks to gather information on the life conditions of children involved in drug trafficking. Further to this, it seeks to reveal the views of community members and professionals on the topic and possible alternatives to confront this activity.

The population surveyed consisted of 40 children, all male; ten young adults, between 20 and 30 years old, three of whom were female and all of whom were involved in drug trafficking schemes; five family members of the children; five police officers; 10 members of the local branch of the judiciary system; five principals of public schools; five low-income community members, some of whom were community leaders; 20 children and young adults from low-income communities who were not involved in the trafficking business, ten being drug users and ten who did not use drugs. Altogether, one hundred (100) people were interviewed.

This sort of data survey is a delicate task to implement, as it is subject to unexpected situations. Thus, a team with extensive professional experience in working with and in low-income areas and in dealing with low-income groups was created for the fieldwork.
The team managed to collect data on children working in drug trafficking in 21 different communities. The most difficult task of the investigation process was to reach the people who should be interviewed. This was achieved in two ways: (i) members of the research team that lived in low-income communities directly contacting children active in trafficking, and (ii) interviews with children committed to intensive care institutions.

Only three girls active in trafficking were interviewed, as the time was limited and female participation in drug trafficking is relatively small. Children who had stopped working in trafficking or who come from the middle classes were not interviewed at this stage. None of the children interviewed in the institutions belonged to this social stratum. Moreover, it was not possible to find middle class children active in drug trafficking as they are dispersed throughout a vast area.

Generally, the children and adolescents involved in trafficking of drugs were interviewed in public spaces, such as bars and trailers, or in residences of people involved in the traffic. The other interviews took place in the subject’s residences or workplaces. There were no serious unexpected situations that arose during the interviews, thanks to the use of experienced interviewers.

It was important to recognize that drug trafficking does not fit the forms of labour organization, weekly workload and compensation practices that formal patterns of labour do. These practices in drug trafficking vary a great deal, even for the same individual, as people are involved in several different ways. Drug trafficking could not be considered “regular” work, and needed to be investigated taking into account its hidden and illegal nature.

Main findings

The main characteristics of the children involved in drug trafficking are as follows:

(1) They belong to the poorest families of the popular communities- favelas; their schooling is below the Brazilian average- today around 6.4 years; the great majority of children involved are black or pardo (Brazilian with partial African ascendance); they marry much earlier than the average Brazilian adolescent; they live with a partner or with friends.

(2) The children enter and remain in drug trafficking activities in order to acquire prestige and power, fulfil emotions, and earn money. Their main friendships are with others in trafficking and their bond with the group is an important factor for staying in this type of activity. Another important reason why they remain in this field is that after a while the children become known to rival groups and the police at which stage it is no longer possible to leave the social network of trafficking. The children’s greatest fears are imprisonment, death, and betrayal by their friends, which can leave them in a difficult situation in the group. Most of the children’s main desire is to buy a house outside the community. By leaving the area, their families will be exposed to less risk.

(3) According to the children, the most probable way of leaving drug trafficking would be through the accumulation of a large amount of money. This would allow them to move to another state and start some sort of business. Most of them are not able to gather much money, however, as they do not have a habit of saving. Extortion by the police are reported as the main obstacle to saving money.
Annex B

Efficiency of oversampling strata of concentration: a case study

Two-stage, two-phase sampling for Child Labour Survey in Cambesia

Efficiency in sampling a rare population can be improved by stratifying the sampling frame according to the degree of concentration of the rare population, and then oversampling strata with higher concentrations. These procedures were developed in Chapter 6.

The primary objective of this case study is to numerically illustrate the procedure in detail, and provide a quantitative indication of the gain in efficiency which might be obtained through oversampling strata with higher concentration of child labour activities. It supplements the various illustrations given in Chapter 6.

It can also be useful and interesting for some of those studying this book to have hands-on experience of repeatedly selecting different samples to apply and test the procedures, and examine the estimates of sampling bias and variance so produced. For this purpose, this case study is accompanied by a large data file providing a list frame of census enumeration areas, with a set of variables for each area which are useful for sample selection and estimation and for comparing sample estimates with the true population values as present in the frame. Specifically, the accompanying Excel file (Cambesia.xlsx) contains the following information. Each row corresponds to an enumeration area or primary sampling unit (PSU). Geographical information is given in columns D to I on the name and code of the district, constituency, and ward. The rural/urban nature and geographical code of the PSU are indicated in columns J and K, followed by census data on the number of households and population in columns L and M. Data concerning the child labour situation are in columns N to S. These cover the variables shown in Table B.2 below, namely: the number of children, the number of working children, the number and percentage of child labourers among children, and the number and percentage doing hazardous work among child labourers. The numerical illustrations presented here and the accompanying data file are based on the work of Mehran (2012).

In planning and designing any survey, it is always a good practice to begin from an assessment of the situation towards which the survey is meant to contribute. To provide a simple illustration, we begin in Section B.1 with such a description of a hypothetical – but nevertheless based on a very real – situation in a developing country.

B.1 Cambesia

B.1.1 The country

Cambesia is a small imaginary landlocked country in the southern hemisphere. According to the latest census, its population was about 716,000 in October 2012. Cambesia is one of the most highly urbanized countries in the region with about 37
per cent of the population concentrated in a few urban areas along the major transport corridors, while rural areas are sparsely populated.

While agriculture for domestic consumption is Cambesia’s largest employment sector, the money gained from exporting platinum is the single greatest source of economic growth in Cambesia. Platinum is used mostly to make jewellery, but also for manufacturing goods. It is used in petroleum refining and in construction of cars, computers, and many electrical appliances. Some such industries have recently been developed locally which employ an increasing number of Cambesians, including some children.

Most rural Cambesians are subsistence farmers producing maize, tobacco, cassava, sugar cane, rice, and wheat as well as cotton lint, cottonseed and livestock products. Some are cash products and sold in the market and even abroad. Children often help at harvest time and at other times in family enterprises for processing and selling some of the products. Many rural children are also working as domestic workers in households living in nearby towns. Some return to their home in the rural areas during the summer and some stay in town all year round.

B.1.2 Child labour survey

As part of a large national programme on elimination of child labour in Cambesia, the National Statistical Office is planning to conduct a specialized survey to assess the size and characteristics of the child worker population in Cambesia. The survey is meant to also estimate the number of children working in hazardous conditions in different branches of economic activity. This objective involves measurements on exposure to injuries and other risks. It requires detailed questions on working conditions of the children including long hours of work, unhealthy work environment, unsafe equipment, heavy loads, dangerous locations, as well as exposure to physical or other abuse.

The National Statistical Office intends to follow the international standards on this topic (ILO, 2008) and is well aware of the methodologies recommended for data collection on child labour including questionnaire design and sampling (Verma, 2008). Two options are under investigation.

One option (A) is to conduct a conventional household survey based on an area frame from the 2012 census using a stratified two-stage sampling methodology. In the first stage, a sample of urban and of rural areas is separately selected with probability proportional to size measured in terms of number of households according to the census. In the second stage, within each urban and rural stratum, a fixed number of sample households are selected in each sample area with equal probability by systematic sampling.

The other option (B) is to list the households in the sample areas after the first stage of area sampling of option A, and in the listing process obtain information on households with children in the target age group (5-17 years old) who are working or not attending school. The information is then used to stratify further the urban and rural strata in terms of suspected areas of concentration of child workers.

The sample selection according to the two options (A) and (B) are conducted and the corresponding estimates of child labour are derived in Sections B.3 and B.4, respectively. The sample selection and the derivation of the estimates are obtained for one hundred separate draws (using the sampling frame provided in the Excel data
file mentioned above), so that the two options can be compared with respect to the empirical values of the standard deviations of the estimates.

B.1.3 True child labour situation in Cambesia

The full data on child labour in Cambesia at the time of the survey are of course not available for the survey. The data presented here as the true child labour situation at the time of the survey, derived from the frame, will serve to assess the results obtained from sampling in Sections B.3 and B.4.

Based on this information, at the time of the survey, the exact population of Cambesia was 727,044, of which 457,656 was in rural areas and 269,388 in urban areas. In total, there were 142,599 households, of which 89,581 were rural and 53,018 urban. The average size of households was about 5.1 both in rural and urban areas. These data are shown in Table B.1.

<p>| Table B.1. Population and households in Cambesia at the time of the survey |
|-----------------------------------------------|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th>Number of persons</th>
<th>Number of households</th>
<th>Average household size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cambesia total</td>
<td>727,044</td>
<td>142,599</td>
</tr>
<tr>
<td>- Rural areas</td>
<td>457,656</td>
<td>89,581</td>
</tr>
<tr>
<td>- Urban areas</td>
<td>269,388</td>
<td>53,018</td>
</tr>
</tbody>
</table>

The child population (5-17 years old) constitutes about one-third of the total population, in urban and rural areas alike. Some 66,936 of them are working, 53,787 in rural areas and 13,149 in urban areas.

<p>| Table B.2. Child labour in Cambesia |
|-------------------------------------|-----------------|-----------------|
| Children (5-17 yrs)                 | Working children | Child labour | Hazardous work |</p>
<table>
<thead>
<tr>
<th>number</th>
<th>number</th>
<th>number</th>
<th>%</th>
<th>number</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cambesia</td>
<td>227,889</td>
<td>66,936</td>
<td>53,097</td>
<td>23%</td>
<td>30,714</td>
</tr>
<tr>
<td>- Rural</td>
<td>143,610</td>
<td>53,787</td>
<td>42,606</td>
<td>30%</td>
<td>24,648</td>
</tr>
<tr>
<td>- Urban</td>
<td>84,279</td>
<td>13,149</td>
<td>10,491</td>
<td>12%</td>
<td>6,066</td>
</tr>
</tbody>
</table>

Not all working children are in child labour as defined by the international standards as shown in Table B.2. There are 53,097 children in child labour in Cambesia, 42,606 in rural areas and 10,491 in urban areas. The percentage of children in child labour is therefore 23.3%, significantly more in rural areas (29.7 per cent) than in urban areas (12.4 per cent). The total number of children in child labour working in hazardous conditions is 30,714, indicating that about 58 per cent of children in child labour are exposed to hazardous work conditions. The percentage is virtually the same in rural and urban areas.

The geographical distribution of child labour in Cambesia is, however, uneven. The 2012 census of population divides Cambesia into 1,188 enumeration areas, 792 rural and 396 urban. Some 173 of the enumeration areas report no child workers at all, and in some 260 areas more than 50 per cent of the children in the age category, 5 to 17 years old are child labourers. The distribution of the 1,188 enumeration areas by prevalence of child labour is shown in Figure B.1.
There appear to be enumeration areas of three types with respect to the prevalence of child labour: areas with no or almost no child labour (198 enumeration areas); areas where the prevalence of child labour is 10%-35% (730 enumeration areas); and areas with a high prevalence of child labour, more than 50 per cent (260 enumeration areas).

**Figure B.1. Distribution of enumeration areas by prevalence of child labour**

B.2 Conventional two-stage sampling

As part of the two-stage sampling design, the National Statistical Office reviewed its budget and took into account its workforce and current fieldwork programme. Based on these considerations, it decided to consider a sample size of 2,400 households for the proposed Child Labour Survey. The sample will be based on 120 sample enumeration areas or Primary Sampling Units (PSUs) and 20 sample households per PSU.

**B.2.1 First-stage sampling**

To improve on the efficiency of the sample, the National Statistical Office decided to stratify the first-stage sampling of PSUs into urban and rural samples, and to allocate the total sample size between the two strata according to the square root allocation method (see Section 6.6.5). The square root allocation method is a compromise between the equal and proportional allocation methods as shown in the following table.
Table B.3. Allocation of Primary Sampling Units among strata

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Frame PSUs</th>
<th>Sample PSUs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Equal allocation</td>
</tr>
<tr>
<td>Cambesia</td>
<td>1,188</td>
<td>120</td>
</tr>
<tr>
<td>- Rural</td>
<td>792</td>
<td>60</td>
</tr>
<tr>
<td>- Urban</td>
<td>396</td>
<td>60</td>
</tr>
</tbody>
</table>

The decision to use the square root allocation was based on the fact that child labour characteristics are more uniform in rural areas than in urban areas and therefore a relatively smaller sample size would be required in rural areas relative to urban areas. Also, the cost of transportation is generally much higher in rural areas than in urban areas. Thus, minimizing the sample size in the rural areas relative to the urban areas would reduce the cost of the survey.

Within each stratum, the sample PSUs are to be drawn with probability proportional to size \(x_k\) for area \(k\), measured in terms of number of households according to the 2012 census. A sample of enumeration areas is selected using the systematic probability proportional to size procedure (see Sections 3.3.1-3.3.2). With 89,581 as the total number of rural households in Cambesia according to the 2012 Census, and 70 the allocated number of PSUs to be sampled in rural areas, the sampling interval for the rural stratum is

\[
I^{(r)} = (89,581/70) = 1,278,
\]

and the probability of selection of an area \(k\) in the rural stratum is

\[
\pi_k^{(r)} = x_k/I^{(r)} = x_k/1,278.
\]

Similarly for urban areas, with 53,018 as the total number of urban households in Cambesia according to the 2012 Census, and 50 the allocated number of PSUs to be sampled,

\[
I^{(u)} = (53,018/50) = 1,060,
\]

and the probability of selection of an area \(k\) in the urban stratum is

\[
\pi_k^{(u)} = x_k/I^{(u)} = x_k/1,060.
\]

The actual draw of the sample PSUs is by systematic sampling from the census frames of rural and urban PSUs sorted by geographical order.56

56 In place of working with measures of size \(x_k\) for systematic selection, one can also work with new variables \((x_k/I^{(r)})\) and \((x_k/I^{(u)})\), in which case the sampling interval becomes 1.0 in both urban and rural strata. See Section 3.3.4 above. Tillé (2006, p. 124) also provides a simple algorithm for implementing a draw of fixed-size by systematic sampling with unequal probabilities, which has been used in the Excel sheet mentioned earlier.
B.2.2 Second-stage sampling

First let us consider the second stage of the conventional two-stage design.

A household listing operation is undertaken before the survey, which involves updating the census lists by freshly relisting the households in the sample PSUs before drawing the second-stage sample. Normally no information is collected at this stage about household characteristics.

The objective of fresh listing is to create an up-to-date sampling frame from which households can be selected. Sometimes the listings are of dwelling units and then all households in selected dwelling units are included if a dwelling unit is sampled. The quality of the listing operation is one of the most important factors that affect the coverage of the target population. Independent listing of households in a sample PSU is generally preferred than the action of updating the census listing of that PSU.

Because of movement of households, creation and disappearance of households through marriage, death and other vital events, the number of households found in a sample PSU at the listing stage of a survey may differ from the number obtained in the previous census.

Design A

In area k, let the number of households in the new list be $z_k$, as distinct from $x_k$ in the last census which was used for the PPS selection of areas at the first stage. See Section 3.3.1 for a discussion of various design options for a two-stage design. A self-weighting design is obtained by selecting households within a sample PSU with probability inversely proportional to the area’s selection probability at the first stage, i.e. inversely proportional to size measure $x_k$:

$$\pi_{ik|k} = \frac{b}{x_k}$$

where $b$ is determined by the target sample size (number of households) per PSU. The overall selection probability of household $i$ in PSU $k$ is

$$\pi_{ik} = \pi_{ik|k} \pi_{i|k} = \frac{x_k}{l} \frac{b}{x_k} = \frac{b}{l}$$

This probability is uniform in a sampling domain, but may be chosen to be different in different domains, such as in urban-rural areas as in the present case.

Under this scheme, the actual number of households to be selected from area $k$ is

$$b_k = b \left( \frac{z_k}{x_k} \right)$$

which varies according to the difference between the census and current number of households in the area.

We will identify the above as “Design A”.
Design B

An alternative is a constant-take design, so that the probability of selection of a household within sample area $k$ is

$$\pi_{i|k} = \frac{b}{z_k}$$

where $b$ is now the fixed sample-take per sample PSU.

The overall selection probability of household $i$ in PSU $k$ is

$$\pi_{ik} = \pi_{i|k} \pi_{i|k|x} = \frac{x_k b}{l z_k} = \frac{b \left(\frac{x_k}{z_k}\right)}{l}$$

We will identify the above as “Design B”.

In Cambesia it was decided to use Design B (rather than the self-weighting Design A). At the second stage of sampling, the design envisages a fixed number, $b = 20$, sample households to be drawn with equal probability by systematic sampling in each sample PSU selected in the previous stage.

From the list of households in each PSU, sorted with respect to geography and other variables strongly correlated with child labour, a sample of 20 households is selected with equal probability by systematic sampling. Thus for a PSU in which 116 households were listed the probability of selection of households in that PSU is $20/116 = 0.1724$.

Note that with Design A, the uniform household selection probabilities with, $b = 20$ are

$$\pi_{ik}^{(r)} = \frac{x_k}{1278 x_k} = 1.56\%$$

$$\pi_{ik}^{(u)} = \frac{x_k}{1060 x_k} = 1.89\%$$

For the Cambesia survey, each sample rural household thus represents about 64 rural households in the country ($64 = 1/0.0156$) and each sample urban household represents about 53 urban households ($53 = 1/0.0189$). The survey estimates can thus be obtained by simply multiplying the sample counts by 64 for the rural sample and 53 for the urban sample. It should be mentioned, however, that in practice because of non-response and other realities of field operations, no design is in fact fully self-weighting.

Concerning Design B, the design is no longer self-weighting. The degree of departure from a self-weighting design depends on the change in the number of households in the PSU between the census and the relisting. Households may be selected within sampled PSUs at sampling rates that generate equal overall probabilities of selection for all households within strata instead of rates that generate a fixed number of sampled households in each PSU as proposed for the Cambesia survey. There are some advantages and limitations in each approach.

---

57 The calculation of sample weights generally involves adjustments for non-response and sometimes also for more reliable external information; see Verma (2008), Chapter 7.

58 For a discussion see United Nations (2005), Section B4.
Design C

In Cambesia the census was in 2012, and the planned survey in 2013 – not too long a gap. Therefore, consideration was also given to another option as follows.

It is a common practice in some countries, though not recommended, to directly use the census listings of households as the frame for selecting the sample households for interview in the selected PSUs. This practice is popular because it avoids the burden and expense of relisting households, the result of which should not be very different to the census listings, if the date of the last census was recent.

This practice can introduce bias in several ways. Firstly, households formed since the census (which maybe systematically different from older households) are not represented. Secondly, independently of the above effect, changes in the number of households are not reflected. For example, the sample will tend to under-estimate the population size if the population has expanded. Thirdly, changes in geographic distribution of the population since the census are not reflected, neglecting, for example regional and urban-rural migration. Finally, the household selection probabilities are not known exactly. Taking a fixed number of households per sample area, for example, gives the false impression of the sample being self-weighting at the same time. This is because the conditional probability of selecting a household within a sample area appears to be

\[ \pi_i|k = \frac{b}{x_k} \]

rather than its actual (but unknown without fresh listing) value \((b / z_k)\). Neglecting this source of variation biases downwards the estimates of variance of survey estimates.

From the list of households in each PSU, sorted with respect to geography and other variables, a sample of 20 households is selected with equal probability by systematic sampling, as in the previous design – the only difference being the source of the lists used.

B.2.3 Child labour estimates

The results of one sample draw are shown in Table B.4. According to this sample draw, the estimate of total population in Cambesia at the time of the survey in 2013 is 715,478 if the census listings were used (Design C), and 729,640 if fresh listings were made (Design B). The corresponding estimates for the rural population were 449,946 and 459,832, respectively. For the urban areas, the population estimates are 265,532 and 269,808. The survey estimates using fresh listing (Design B) are uniformly higher those using census listings (Design C), reflecting the growth in population between census date in 2012 and survey date in 2013.
Table B.4. Survey estimates of child labour in Cambesia, 2013: 2-stage sample

Estimates are based on the results from a single sample drawn from the frame.

<table>
<thead>
<tr>
<th>Two-stage sampling (One sample draw)</th>
<th>Number of persons</th>
<th>Number of households</th>
<th>Children (5-17 yrs)</th>
<th>Working children</th>
<th>Child labour</th>
<th>Child labour %</th>
<th>Hazardous work</th>
<th>Hazardous work %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Design B</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cambesia</td>
<td>729,640</td>
<td>142,777</td>
<td>229,726</td>
<td>63,28</td>
<td>50,87</td>
<td>22.1%</td>
<td>30,488</td>
<td>59.9%</td>
</tr>
<tr>
<td>- Rural</td>
<td>459,832</td>
<td>89,883</td>
<td>144,033</td>
<td>52,163</td>
<td>42,398</td>
<td>29.4%</td>
<td>25,725</td>
<td>60.7%</td>
</tr>
<tr>
<td>- Urban</td>
<td>269,808</td>
<td>52,895</td>
<td>85,693</td>
<td>11,118</td>
<td>8,472</td>
<td>9.9%</td>
<td>4,763</td>
<td>56.2%</td>
</tr>
<tr>
<td><strong>Design C</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cambesia</td>
<td>715,478</td>
<td>140,464</td>
<td>233,739</td>
<td>71,055</td>
<td>56,583</td>
<td>25.3%</td>
<td>32,533</td>
<td>57.5%</td>
</tr>
<tr>
<td>- Rural</td>
<td>449,946</td>
<td>88,163</td>
<td>141,313</td>
<td>55,417</td>
<td>44,396</td>
<td>31.4%</td>
<td>25,315</td>
<td>57.0%</td>
</tr>
<tr>
<td>- Urban</td>
<td>265,532</td>
<td>52,301</td>
<td>82,426</td>
<td>15,638</td>
<td>12,186</td>
<td>14.8%</td>
<td>7,218</td>
<td>59.2%</td>
</tr>
</tbody>
</table>

Two-stage sampling (areas then households), with stratified random sampling (urban-rural) at first stage.
Design B fresh listing for selection of households in sample areas.
Design C households selected using census listing for sample areas.

Similar results are obtained with the estimates of the number of households and the child population aged 5 to 17. The survey estimates of the total number of households in Cambesia based on the census listings are exactly the 2012 census counts of households (140,464 for total Cambesia, 88,163 for rural areas and 52,301 for urban areas), thus underestimating the change in the household population during the period 2012 to 2013. The estimates of the number of households and population based on fresh listings are closer to the true values in 2013 shown in Table B.1. Similar results are true for the population of 5-17 years old children, as can be verified against data in the first column of Table B.2.

The above pattern is not observed for child labour variables in this particular draw. This is likely to be due to sampling variability: as will be seen below, child labour variables are subject to larger sampling errors than basic demographic variables such as the total population of the country.

B.2.4 Sampling errors

As a reminder, we may note the following. In a sample survey, sampling errors arise due to the fact that the survey does not include all elements of the population but rather only a selected portion. A different sample may produce different sample elements and therefore different estimates of child labour.

In general, sampling errors may be decomposed into two components: (a) sampling bias; and (b) sampling variance. The sampling bias of an estimate is the difference between the mean of its sampling distribution (i.e. the distribution that would result from taking all possible samples from the population) and the true value of the variable being estimated. The sampling variance of the estimate is the variance of this sampling distribution. It reflects the uncertainty of the estimate due to the particular sample used...
for its calculation, among all possible other samples that could have been drawn from the same population under the identical conditions.

Table B.5 gives the average estimates of child labour in Cambesia obtained from one hundred sample draws from the sampling frame of the Child Labour Survey. The results are given for designs B and C; please also refer to similar results for the two-phase designs 1 and 2 given later in Table B.6. All designs shown, except Design C, are essentially unbiased, meaning that the expected values of the samples from them equal the true population values given in Tables B.1 and B.2. Empirically, the small differences between the averages over 100 independent samples and the true populations values are due to sampling variability: the average of 100 samples is like a sample estimation for a single sample 100 times the size of individual samples of the type we are considering. That would imply a standard error of an average estimate which is one-tenth the size of standard error in a single sample. The latter values will be shown in Table B.7 below. On this basis the discrepancies between the observed averages over 100 samples and true population values in Tables B.1-B.2 are quite consistent with the expected sampling variability.

The exception is Design C. Here, the average over samples estimates the 2012 census figures, not the current 2013 figures. Hence the bias in Design C is simply the difference between the census figures and the current figures at the time of the survey. Neglecting the relatively small effect of variability in the averaged results, biases in the estimates with Design C have the following pattern. The negative relative bias of the estimates of population and number of households under census listings (around -1.5 per cent) reflects the underestimation due to lack of adjustment for population growth between census date (2012) and the survey period (2013). For the estimates of working children, the maximum relative bias under census listings is -1.8 per cent for rural areas. It appears that general population figures have changed more since the census than the figures relating specifically to child labour. There seems to have been a decline in the rates of child labour, such that, given the increase in population size, the total numbers involved have not changed much since the last census.

### Table B.5. Average sample estimates of child labour in Cambesia, 2013: 2-stage sample

Estimates are averaged over 100 samples drawn from the frame.

<table>
<thead>
<tr>
<th></th>
<th>Two-stage sampling (One hundred sample draws)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of persons</td>
</tr>
<tr>
<td>Design B</td>
<td></td>
</tr>
<tr>
<td>Cambesia</td>
<td>726,927</td>
</tr>
<tr>
<td>- Rural</td>
<td>457,859</td>
</tr>
<tr>
<td>- Urban</td>
<td>269,068</td>
</tr>
<tr>
<td>Design C</td>
<td></td>
</tr>
<tr>
<td>Cambesia</td>
<td>716,143</td>
</tr>
<tr>
<td>- Rural</td>
<td>451,036</td>
</tr>
<tr>
<td>- Urban</td>
<td>265,107</td>
</tr>
</tbody>
</table>

Estimates are averaged over 100 samples drawn from the frame.

Two-stage sampling (areas then households), with stratified random sampling (urban-rural) at first stage.

Design B fresh listing for selection of households in sample areas.

Design C households selected using census listing for sample areas.
Later in Table B.7 are shown estimates of relative standard error of the estimates, also calculated using one hundred sample draws from the sampling frame of the Child Labour Survey. The estimates are for designs B and C. (The table also shows and compares corresponding standard error estimates for two-phase designs 1 and 2 discussed later.) The values for Design C are a little lower, but that is only because of the under-estimation in this case, as noted earlier. For instance, the relative error estimate for the number of households is zero under Design C, simply because by definition the estimates are all exactly equal to the census values and therefore there is no variation in their values.

The relative standard deviation of rural and urban estimates are larger than the corresponding estimates for total Cambesia. This of course is simply a function of sample size.

The relative standard deviations increase with decreasing size of the underlying population under both designs. Under Design B (fresh listings), for example, the relative standard deviation of the estimate of total population is 1.0%. It is 1.7% for population 5-17 years old, 6.2% for working children, 6.9% for child labour, and 7.3% for children in hazardous work.

**B.3 Targeted two-phase sampling**

In broad terms, two-phase sampling may be described as sampling followed by targeted sub-sampling.\(^{59}\) When the original frame contains little or no useful information about the target population, the sample selection may be divided into two phases: in the first phase, a large sample is selected, for which information on one or more auxiliary variables related to the target population is collected; in the second phase, the auxiliary information is used to sub-sample and target the population more efficiently.

In the present case, the first phase of sampling is the selection of the sample areas or PSUs identical to the conventional two-stage sampling described in the preceding section, except that the sample is now larger. For the second phase, the sample PSUs are freshly listed as in two-stage sampling, but here the listing involves one or more additional variables specifically related to child labour. The information is then used to stratify the sample PSUs and sub-sample them to reach the target population more efficiently.

**B.3.1 Listing to identify prevalence of child labour**

Figure B.2 shows the information to be collected in listing the households in the sample PSUs. In addition to identification data, information on number of household members 5 to 17 years old is to be obtained. If there are children in that age group, two other questions are to be asked as part of the listing operation: is any of them working? And: is any of them not attending school?

---

\(^{59}\) Särndal, Swensson and Wretman (1992), Chapter 9.
After listing, the number of households likely to have child labour may be counted for each sample PSU as illustrated with the data in Figure B.3 for one sample PSU. In this example, the PSU has a total of 100 households with 520 members (257 male and 263 female). There are 174 members 5-17 years old, 71 of them reported to be working, and 37 not attending school. In total, there are 24 households said to have at least one working child or one child not attending school (categorized as households with likely child labour).
B.3.2 Stratification of areas according to prevalence of child labour

The next step is to use the listing information to stratify the sample PSUs according to the level of concentration of households with likely child labour. Here as shown in Figure B.4, the sample PSUs are stratified in three strata: sample PSUs in which there are none or less than 10 per cent households with likely child labour (low concentration PSUs); sample PSUs where there are 50 per cent or more households with likely child labour (high concentration PSUs); and sample PSUs in-between, i.e. where the number of households with likely child labour is 10 per cent or more but less than 50 per cent (mid concentration PSUs).

**Figure B.4. Stratification of PSUs according to concentration of household with child labour**
The choice of the thresholds 50 per cent and 10 per cent was guided by the analysis of the true situation of child labour in Cambesia reported in Section B.1 (see Figure B.1). In practice, the thresholds may be determined as part of a pilot study or expert knowledge on the geographical distribution of child labour in the country. The resulting distribution of the PSUs according to strata of concentration of households with likely child labour is shown in Figure B.5.

![Figure B.5. Distribution of PSUs according to concentration of child labour](image)

In practice, not all PSUs are stratified by level of concentration, but only those sampled at the first phase of sampling. Also, the number of working children and the number of children not at school reported at the listing phase may be revealed to be inaccurate, and different numbers may be obtained at the survey phase. The purpose of the information at the listing phase is to stratify the sample PSUs by broad level of concentration for targeted sub-sampling.

After stratification of the sample PSUs according to the three levels of concentration, sample PSUs in the low concentration stratum will be sub-sampled at lower probability of selection (25 per cent in our example), sample PSUs in the mid concentration stratum will be sub-sampled at higher probability of selection (75 per cent), and sample PSUs in the high concentration stratum will be all kept in the sample (probability of selection 100 per cent).

**B.3.3 Comparison between two designs**

For the rest of our numerical example, two alternative designs are illustrated:

- **Design 1.** Selection of 120 sample PSUs at the first phase of sampling, followed by sub-sampling according to level of concentration at the second phase of sampling, selecting in all \( a = 80 \) PSUs and \( b = 30 \) sample households per sub-sample PSU, giving a sample of \( n = ab = 2,400 \) households.
- **Design 2.** Selection of 183 sample PSUs at the first phase of sampling, followed by sub-sampling according to level of concentration at the second phase of sampling.
sampling, selecting in all \(a = 120\) PSUs and \(b = 20\) sample households per sub-sample PSU, again giving a sample of \(n = ab = 2,400\) households.

As to the details: in Design 1, with an initial sample of 120 PSUs, about 70 will be in the mid concentration stratum and about 25 each in the low and high concentration stratum. Sub-sampling at the rate of 25%, 75% and 100%, in the low, mid and high strata, respectively, results in about 80 sub-sample PSUs. And, drawing 30 sample households per sub-sample PSU gives a final sample size of 2,400 households. Similarly in Design 2, with an initial sample of 183 PSUs, there will be about 38 low concentration, 106 mid concentration and 38 high concentration PSUs. After sub-sampling at the given rates, one obtains about 120 sub-sample PSUs. And, drawing 20 sample households per sub-sample PSU gives again a final sample size of 2,400 households.

Estimates for demographic and child labour variables for the 2-phase design (similar to those for designs B and C in Table B.5) are shown in Table B.6. The results from the two set of designs are very similar, as can be expected for the relatively large sample sizes in the illustration.

| Table B.6. Average sample estimates of child labour in Cambesia, 2013: 2-phase sample |
| Estimates are averaged over 100 samples drawn from the frame. |

<table>
<thead>
<tr>
<th>Two-phase sampling (One hundred sample draws)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of persons</td>
</tr>
<tr>
<td>------------------</td>
</tr>
<tr>
<td>Design 1. Sub-sampling 80 from 120 sample PSUs, 30 households per sub-sample PSU</td>
</tr>
<tr>
<td>Cambesia</td>
</tr>
<tr>
<td>- Rural</td>
</tr>
<tr>
<td>- Urban</td>
</tr>
<tr>
<td>Design 2. Sub-sampling 120 from 183 sample PSUs, 20 households per sub-sample PSU</td>
</tr>
<tr>
<td>Cambesia</td>
</tr>
<tr>
<td>- Rural</td>
</tr>
<tr>
<td>- Urban</td>
</tr>
</tbody>
</table>

2-phase, 2-stage design with oversampling of strata having higher concentrations of child labour.

Design C: more clustered sample \((a = 80; b = 30, n = ab = 2,400)\).

Design D: more dispersed sample \((a = 120; b = 20, n = ab = 2,400)\).

Table B.7 gives the corresponding relative standard deviations of the estimates. In general, the relative standard deviations are lower under Design 2 in comparison with Design 1, indicating that drawing a larger sample of PSUs in the first stage of sampling is more efficient than having more sample households at the second stage of sampling. This results from the fact that Design 2 is spread out into a larger number of PSUs, and hence having lower design effects. Of course, the cost under this design is also higher. Around 50 per cent more PSUs have to be listed compared to Design 1 (183, versus 120 under Design 1), and the number to be travelled to for the interview is increased by the same proportion (120, versus 80 under design 1).
Sampling elusive populations: Applications to studies of child labour

### Table B.7. Relative standard deviations of survey estimates of child labour in Cambesia, 2013

Estimates are averaged over 100 samples drawn for each design.

#### Two-stage sampling (One hundred sample draws)

<table>
<thead>
<tr>
<th>Design</th>
<th>Number of persons</th>
<th>Number of households</th>
<th>Children (5-17 yrs)</th>
<th>Working children</th>
<th>Child labour %</th>
<th>Child labour Hazardous work</th>
<th>Hazardous %</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cambesia</td>
<td>1.0%</td>
<td>0.1%</td>
<td>1.2%</td>
<td>6.2%</td>
<td>6.9%</td>
<td>6.5%</td>
<td>7.3%</td>
</tr>
<tr>
<td></td>
<td>- Rural</td>
<td>1.3%</td>
<td>0.2%</td>
<td>1.7%</td>
<td>6.5%</td>
<td>7.1%</td>
<td>6.7%</td>
</tr>
<tr>
<td></td>
<td>- Urban</td>
<td>1.9%</td>
<td>0.2%</td>
<td>1.9%</td>
<td>13.3%</td>
<td>14.2%</td>
<td>14.4%</td>
</tr>
</tbody>
</table>

#### Design C

<table>
<thead>
<tr>
<th>Design</th>
<th>Number of persons</th>
<th>Number of households</th>
<th>Children (5-17 yrs)</th>
<th>Working children</th>
<th>Child labour %</th>
<th>Child labour Hazardous work</th>
<th>Hazardous %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cambesia</td>
<td>0.9%</td>
<td>0.0%</td>
<td>1.1%</td>
<td>6.2%</td>
<td>6.7%</td>
<td>6.3%</td>
<td>7.1%</td>
</tr>
<tr>
<td></td>
<td>- Rural</td>
<td>1.1%</td>
<td>0.0%</td>
<td>1.5%</td>
<td>5.8%</td>
<td>6.2%</td>
<td>5.7%</td>
</tr>
<tr>
<td></td>
<td>- Urban</td>
<td>1.8%</td>
<td>0.0%</td>
<td>1.9%</td>
<td>13.3%</td>
<td>13.8%</td>
<td>14.0%</td>
</tr>
</tbody>
</table>

#### Two-phase sampling (One hundred sample draws)

<table>
<thead>
<tr>
<th>Design</th>
<th>Number of persons</th>
<th>Number of households</th>
<th>Children (5-17 yrs)</th>
<th>Working children</th>
<th>Child labour %</th>
<th>Child labour Hazardous work</th>
<th>Hazardous %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sub-sampling 80 from 120 sample PSUs, 30 households per sub-sample PSU</td>
<td>Cambesia</td>
<td>1.6%</td>
<td>1.4%</td>
<td>1.7%</td>
<td>3.2%</td>
<td>3.2%</td>
</tr>
<tr>
<td></td>
<td>- Rural</td>
<td>2.0%</td>
<td>1.8%</td>
<td>2.6%</td>
<td>3.4%</td>
<td>3.2%</td>
<td>3.0%</td>
</tr>
<tr>
<td></td>
<td>- Urban</td>
<td>3.8%</td>
<td>2.8%</td>
<td>3.7%</td>
<td>8.3%</td>
<td>7.7%</td>
<td>7.9%</td>
</tr>
<tr>
<td>2</td>
<td>Sub-sampling 120 from 183 sample PSUs, 20 households per sub-sample PSU</td>
<td>Cambesia</td>
<td>1.3%</td>
<td>0.8%</td>
<td>1.5%</td>
<td>1.8%</td>
<td>1.9%</td>
</tr>
<tr>
<td></td>
<td>- Rural</td>
<td>2.0%</td>
<td>1.6%</td>
<td>1.7%</td>
<td>2.4%</td>
<td>2.4%</td>
<td>2.6%</td>
</tr>
<tr>
<td></td>
<td>- Urban</td>
<td>3.1%</td>
<td>1.9%</td>
<td>3.9%</td>
<td>5.5%</td>
<td>5.8%</td>
<td>4.9%</td>
</tr>
</tbody>
</table>

As expected, the results of Table B.7 also show that irrespective of the design, the relative standard deviations of the estimates are larger when the population base is smaller. For example, under design 2, the relative standard deviation of the estimate of total population is 1.3%. It is 1.5% for children 5-17 years old, 1.8% for working children, 1.9% for child labour, and 2.5% for children in hazardous work. Similarly, the relative standard deviations, under both designs, are larger in urban areas than in rural areas, reflecting the smaller population and the higher heterogeneity in urban areas relative to rural areas.

The performance of targeted two-phase sampling is also compared with those of conventional two-stage sampling in Tables B.7. The variables divide sharply into two subsets: (1) general demographic variables (estimates of numbers of households, persons, children aged 5-17); and (2) child labour related variables (estimates of numbers involved in child work, child labour, hazardous child labour). Oversampling aimed at an improved sample for child labour variables reduces the efficiency of the sample in estimating general demographic variables, which tend to be distributed uniformly throughout and are generally served best by a self-weighting (uniform
selection probability) design. Consequently, designs B and C give substantially lower relative standard errors than designs 1 and 2 for general demographic variables.

However, our primary interest here is in studying child labour. For these variables, by contrast, the average values of the relative standard deviation are lower with two-phase sampling than with conventional two-stage sampling. Of course, the cost of the two-phase designs is also higher. More areas have to be visited for the screening interviews, and the screening interviews conducted. Such a screening operation can be much more demanding than the simple listing of households. On the other hand, with the present design the screening interviews do not have to be too detailed or precise, since the information collected is used only for the stratification at the level of sample areas, rather than for classification at the level of individual households.

**B.3.4 Concluding remark: some other options**

In the present case study, the first phase of sampling consists of targeted listing of the sample PSUs and the second phase is the final sample of households for the child labour survey. Other possibilities could be envisaged as shown below.

<table>
<thead>
<tr>
<th>First phase</th>
<th>Second phase</th>
<th>Third phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listing</td>
<td>---</td>
<td>CLS</td>
</tr>
<tr>
<td>Listing - LFS</td>
<td>---</td>
<td>CLS</td>
</tr>
<tr>
<td>Listing - LFS</td>
<td>---</td>
<td>CLS</td>
</tr>
</tbody>
</table>

For example, the first phase could have been a standard labour force survey (LFS) that would be used for screening households with working children for the second phase child labour survey (CLS). Another example would be a three-phase sampling scheme according to which the child labour survey sample itself would be further subsampled to conduct a more focused labouring children survey (LCS) with the objective of determining the conditions and consequences of child labour. See Verma (2008) for explanation of the CLS-LCS distinction and subsampling procedures between the three phases of the operation. The distinction is also summarised in Section 2.13 above.
Annex C

Applying sampling procedures in combination: A case study

Starting from a real situation as identified from a baseline survey in a community, in this case study we discuss scenarios for the development of a set of surveys and studies aimed at providing a comprehensive assessment of the child labour situation in the community. This scenario calls upon several of the sampling procedures discussed in this book, to be chosen from and used in combination. Chapter 10 on sampling mobile populations provides the main technical background for this discussion.

For this case study, we draw on the report of one of three baseline surveys on child labour conducted in Jamaica, namely a study of child labour in Rocky Point, Clarendon and Old Harbour Bay, St Catherine (Jamaica, 2001).

The baseline survey conducted in 2001 involved a small number (78) of interviews with children and their employers, and a survey of enrolment and attendance in three schools serving the study area. The broad objective of the baseline survey was to identify the magnitude and nature of the problem of child labour, and not to produce estimates in precise quantitative terms. Specifically, it provided (i) indicative estimates of the potential number of children at work based on school enrolment and attendance data, and (ii) from the small number of interviews with children, employers, guardians and other persons, it provided qualitative indicators and a picture of child labour characteristics (types of activity, composition by age and sex of children involved in work, degree of hazard involved, etc.)

First we will outline the study methodology as it was designed and implemented. Then on the basis of the picture of the child labour situation which emerges and rough estimates of its various characteristics, we will develop a possible design scenario for a full-scale survey, or rather a set of surveys and studies, aimed at producing more precise estimates of the size and characteristics of child labour in a community. Our interest in this discussion is more general than the particular community used here as an illustration.

Quite a number of baseline surveys of child labour have been conducted in developing countries (see Section 2.12.5). These surveys provide a wealth of information and guidance which can be valuable for developing more comprehensive surveys covering different sectors and types of child labour. The objective of the present case study is to illustrate the process through analysing an example in detail.

C.1 Lessons from a baseline survey of child labour in Jamaica

C.1.1 The baseline survey

Both the baseline survey sites, Old Harbour Bay and Rocky Point, are locations noted for their fishing industry and drug trade. Fishing is facilitated by the existence of
quays off the southern part of the island, enabling fishermen to remain off-shore and obtain larger quantities of fish than they could by fishing from the beaches. In general, however, for the two densely populated parishes, employment opportunities in small or large industry are restricted, and fishing, small-scale crop cultivation and animal husbandry provide a critical, if limited, means of livelihood.

The following quotation gives an impression of the survey conditions. Survey conditions constitute the most important practical factor to be taken into account in the design of any survey. Technical details, though undoubtedly important, should come second.

“Supportive factors:

The interviewers conducting the fieldwork were at an advantage in that they were from the Rural Family Support Organization … On learning this, some persons volunteered information and even accompanied interviewers to the homes of some of the children. People in the communities in which the survey was conducted had felt the impact of the Teenage Mothers Project [TMP]. Individuals in the communities had benefitted from the TMP and one shopkeeper volunteered to contact some of the parents of the children. There were moments of light humour – when one parent discovered that one of the interviewers had the same surname as one of her children's father, she became a valuable resource person and eagerly identified those who were involved in child labour. Also, some of the children who were involved in labour played a key role, as they were also instrumental in identifying other children (the ‘snowball’ or ‘cascade’ effect).

“Constraining factors:

By the second week of the survey, there was a noticeable difference in the number of children who were ‘hanging out’ on the beach. Interviewers had trouble in locating them. One interviewer reported that persons who were very helpful the week before became ‘cagey’ and were reluctant to give out any information. A fisherman told her – ‘don’t come back on the beach and take any pictures’.

“One contact person reported that a number of parents felt that they would ‘get into trouble’ and they were therefore not willing to talk to us. She reported also that it appeared that some of the children may have gone into hiding, because although observed in the first week, they were nowhere to be found in the second week”.

The focus of the baseline survey of child labour was on fishing industry, this being the major sector of employment for boys:

“Child labour in the fishing industry ranges from the apparently less hazardous type of cleaning out boats and disposing garbage, to the most hazardous of going out on the boats with adult fishermen and spearing fish, diving to set pots and pulling the nets in. The ‘hazard’ is with respect to the danger of the open sea for a child who cannot swim, the presence of sharks at times, and the possibility of drowning. With respect to spear-fishing and setting of pots, these fishermen do not tend to use aqualung equipment and the depths to
which the child may have to go under water can interfere with oxygen flow to the brain and ultimately can cause death”.

The baseline survey was a small one and included the following components.

- A school-based enrolment/attendance study conducted in three schools in the study areas.
- Community-based interviews with 72 actual child labourers and 3 potential labourers of which 65 were male and 10 female.
- Community-based interviews with three employers of children (fishermen).

The study focussed on the worst forms of child labour among children under 18 years of age.

The report’s main conclusions include the following.

- Large proportions of children are not enrolled at school (especially girls), or are drop-outs from school (especially boys). This is indicative of the high level of potential child labour in the area.
- The incidence of ‘hidden’ labour of females is a feature of the situation in the study communities. Much of the child labour of girls is in the household or family context, and is not captured by studies or observations outside of the household sector.
- Child labour in the fishing sector involves mainly boys, and some of it is very hazardous.

Some children may also be involved in drug trafficking, though specific follow-up of such activity was not been attempted in the baseline survey.

C.1.2 Use of the baseline survey to illustrate process of survey design of a mobile population of labouring children

The baseline survey described above, though essentially qualitative in nature, provides a lot of information on the basis of which a full-scale representative child labour survey and a set of supplementary studies may be developed. We will use the results of the baseline survey report in order to develop and illustrate ideas about a full-scale survey of child labour and supplementary studies in the sort of developing country setting described in the report. The baseline survey identifies several different sectors of child labour. Drawing on its specific results, we will discuss – in broader and more general terms – issues and options for designing components of a comprehensive child labour enquiry covering different sectors and types of activity.

Of course, the information available from the baseline survey is not sufficient for the purpose. Development of a full-scale survey and supplementary studies would require additional information from other administrative and statistical sources, past surveys, and above all, more detailed information on the situation on the ground, especially from local ‘experts’ and informed residents.

Such additional information is not available to us, except for what has been documented in the baseline survey report. Therefore the objective of the following is to consider a possible model for the procedure to develop a full-scale representative survey and supplementary studies starting from the results of the baseline survey of child labour.
Where sufficient information is not available, we will have to make some plausible assumptions. In the real situation, with additional information and better understanding of the local situation, the actual design may differ from the model proposed below in matters of detail. Nevertheless, we expect the exercise to be a useful tool for bringing out the various issues and options and for explaining a number of technical considerations in designing child labour surveys of a mobile population of children.

C.1.3. Assessing magnitude of the child labour problem based on school enrolment and attendance

The starting point of the survey design is to identify the numbers of children involved in labour in the study population, and to the extent possible, breakdown of the child labour force by age group, gender and the type of activity.

This is already well-done in the baseline survey report on the basis of the school enrolment and attendance survey. Below we will consolidate some figures from the report to facilitate a focussed discussion. For the same purpose, we will rework and refine some of the reported figures, with some additional assumptions where necessary.

Table C.1 shows some data on children’s school enrolment and attendance. Panel (A) shows the number of boys and girls by enrolment and actual attendance at grades 1-6, which roughly correspond to ages 7-12. As noted, the figures have been taken from the baseline survey report.

For the purpose of this exposition, it is reasonable to assume that in the population of children the numbers of boys and girls are the same, and that the numbers from one age to the next in the small age range being considered also do not vary.

It is noted in the baseline survey report that most families register their children for school at the point when the child would be entering primary school at grade 1, usually at age 6. The registration appears to be more complete for boys than for girls.

Let us assume to start with that the enrolment for boys at grade 1 is 100 per cent complete (we will revise this assumption later). The number of boys enrolled is 343. On the basis of the above assumptions, this is the number of children expected to be in every grade 1-6, separately for boys and girls. With this constant base population for each of the 12 cells (grades by gender categories) in Panel (A), Panel (B) shows the proportion enrolled, its complement the proportion not enrolled, and proportions actually attending by grade and gender. By grade 6, over a third of the boys and half the girls are not enrolled. Even more strikingly, a quarter of the girls are not enrolled even at grade 1. Thereafter, attendance falls sharply, especially among boys.

Panel (C) shows the same figures, but this time on the assumption that enrolment for boys at grade 1 is 90 per cent rather than 100 per cent complete – this probably being a more realistic basis.

Finally in Panel (D), these figures are converted to the numbers involved, given the base as 381 children in each grade by gender. This gives an estimation that almost 2,500 children at the primary school age are not attending school, which is more than half of the estimated total of around 4,500 children of those ages in the area. The figure is indicative of the number of children in the young age group 7-12 who may be engaged in labouring activities.
### Table C.1. Case study based on a child labour survey in Jamaica: Children’s school enrolment and attendance

#### (A) Observed numbers of children (in two schools in the community)

<table>
<thead>
<tr>
<th>Grade</th>
<th>Boys</th>
<th></th>
<th></th>
<th>Girls</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>not enrolled</td>
<td>enrolled</td>
<td>attending</td>
<td>not enrolled</td>
<td>enrolled</td>
<td>attending</td>
</tr>
<tr>
<td>1</td>
<td>343</td>
<td>254</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>308</td>
<td>213</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>239</td>
<td>179</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>292</td>
<td>225</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>224</td>
<td>163</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>221</td>
<td>133</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>grades 1-6</td>
<td>1,627</td>
<td>1,167</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### (B) Proportions, assuming proportion boys enrolled in Grade 1 = 100%

<table>
<thead>
<tr>
<th>Grade</th>
<th>Boys</th>
<th></th>
<th></th>
<th>Girls</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>not enrolled</td>
<td>enrolled</td>
<td>attending</td>
<td>not enrolled</td>
<td>enrolled</td>
<td>attending</td>
</tr>
<tr>
<td>1</td>
<td>0.00</td>
<td>1.00</td>
<td>0.74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.10</td>
<td>0.90</td>
<td>0.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>0.70</td>
<td>0.52</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.15</td>
<td>0.85</td>
<td>0.66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.25</td>
<td>0.65</td>
<td>0.48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.36</td>
<td>0.64</td>
<td>0.39</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>grades 1-6</td>
<td>0.21</td>
<td>0.79</td>
<td>0.57</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### (C) Proportions, assuming proportion boys enrolled in Grade 1 = 90%

<table>
<thead>
<tr>
<th>Grade</th>
<th>Boys</th>
<th></th>
<th></th>
<th>Girls</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>not enrolled</td>
<td>enrolled</td>
<td>attending</td>
<td>not enrolled</td>
<td>enrolled</td>
<td>attending</td>
</tr>
<tr>
<td>1</td>
<td>0.10</td>
<td>0.90</td>
<td>0.67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.19</td>
<td>0.81</td>
<td>0.56</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.37</td>
<td>0.63</td>
<td>0.47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.23</td>
<td>0.77</td>
<td>0.59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.41</td>
<td>0.59</td>
<td>0.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.42</td>
<td>0.58</td>
<td>0.35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>grades 1-6</td>
<td>0.29</td>
<td>0.71</td>
<td>0.51</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. The number of children in each grade by gender is assumed to be constant: 343 in panel (B), and 343/0.90 = 381 in panel (C).

In the small baseline survey of 75 children, a third were aged 7-12, and twice as many were in the older age group 13-17. It is possible that more of the older children were in the fishing sector (see below) and may have been better covered in the survey than other children. In any case, even if we assume the ratio between numbers of labouring children in the younger and the older age groups to be 40:60 (rather than 1:2 as in the small survey), the figures suggest that the size of the potential child labour force in the study area may be of the order of 6,000 children aged 7-17.

Such a figure is an important parameter for the survey design and the choice of sample size.
It is always important to estimate the magnitude of the problem as best as possible before deciding on the survey approach and methodology.

Table C.1 (cont.)

(D) Estimation of numbers of children by school enrollment and attendance

Number in each grade and gender = 381
Number of children of grades 1-6 = 381x12 = 4,572
(assuming same number of boys and girls in each grade)

Number of children according to enrollment and attendance

<table>
<thead>
<tr>
<th>Grade</th>
<th>Boys</th>
<th>Number not enrolled</th>
<th>Enrolled but not attending</th>
<th>Number attending</th>
<th>Not enrolled or not attending</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>38</td>
<td>89</td>
<td>254</td>
<td>127</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>73</td>
<td>95</td>
<td>213</td>
<td>168</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>142</td>
<td>60</td>
<td>179</td>
<td>202</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>89</td>
<td>67</td>
<td>225</td>
<td>156</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>157</td>
<td>61</td>
<td>163</td>
<td>218</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>160</td>
<td>88</td>
<td>133</td>
<td>248</td>
<td></td>
</tr>
<tr>
<td>Grades 1-6</td>
<td>660</td>
<td>460</td>
<td>1,167</td>
<td>1,120</td>
<td></td>
</tr>
<tr>
<td>% of children</td>
<td></td>
<td></td>
<td></td>
<td>49%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade</th>
<th>Girls</th>
<th>Number not enrolled</th>
<th>Enrolled but not attending</th>
<th>Number attending</th>
<th>Not enrolled or not attending</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>122</td>
<td>66</td>
<td>193</td>
<td>188</td>
<td>315</td>
</tr>
<tr>
<td>2</td>
<td>151</td>
<td>62</td>
<td>168</td>
<td>213</td>
<td>381</td>
</tr>
<tr>
<td>3</td>
<td>194</td>
<td>45</td>
<td>142</td>
<td>239</td>
<td>441</td>
</tr>
<tr>
<td>4</td>
<td>160</td>
<td>29</td>
<td>192</td>
<td>189</td>
<td>345</td>
</tr>
<tr>
<td>5</td>
<td>178</td>
<td>39</td>
<td>164</td>
<td>217</td>
<td>435</td>
</tr>
<tr>
<td>6</td>
<td>215</td>
<td>23</td>
<td>143</td>
<td>238</td>
<td>486</td>
</tr>
<tr>
<td>Grades 1-6</td>
<td>1,021</td>
<td>264</td>
<td>1,002</td>
<td>1,285</td>
<td>2,404</td>
</tr>
<tr>
<td>% of children</td>
<td></td>
<td></td>
<td></td>
<td>56%</td>
<td>53%</td>
</tr>
</tbody>
</table>
C.1.4 Sectors of activity and components of the survey

The small-scale baseline survey also provides a picture of the composition of child labour in the study area. Some basic figures are summarised in Table C.2.

First by gender. The baseline survey of working children covers largely boys (65 of the 75 survey respondents were boys; see Panel (A)). However, school attendance figures suggest that the child labour should be equally, if not more, prevalent among girls. Girls are likely to be more engaged in domestic work and in work related to the family enterprise (see below), neither adequately covered in the baseline survey, as indeed recognised in the survey report. These are major sectors to be covered in any full-scale survey.

Next by sector. A majority (52 of 75) of the baseline survey interviews are in the fishing sector. Mainly boys work in this sector (49 out of 52 child workers in fishing are boys). Even though other sectors of activity are probably underrepresented in the survey, fishing is the second major sector to be covered in a full-scale survey, other than family-based labour of girls.

In Panel (B) a picture of child labour by sex, age-group, and fishing/non-fishing sectors has been constructed. The figures are of course very tentative, and can be revised on the basis of more information on the actual situation of a particular survey. The assumptions made in constructing these figures are noted at the bottom of the table.

According to these figures:

i. around one-third of the total child labour is in the fishing sector, comprising mainly of boys;

ii. a third is accounted for by girls working in non-fishing, mostly in domestic and family related activities; and

iii. (a little over) a third is accounted for by other diverse activities, in which boys and girls may be engaged in similar numbers, though most likely the two groups concentrating in different types of activity.

It may be argued that (iii) is a diverse group and may therefore deserve larger than a third of the total effort in a full-scale survey and supplementing studies. Certainly its diversity would require more than one survey component, possibly several components with different methodologies.

However, the other two components also have their special importance and requirements. Component (i), the fishing sector, represents a major source of child employment in the area, a lot of it hazardous. Furthermore, it is probably also linked most closely to drug trafficking, the most hazardous and most difficult to enumerate activity. Concerning (ii), work of girls, in particular domestic and family-related activities, are notoriously under-covered in child labour surveys, and therefore also need special attention and effort.

The appropriate share of the effort to be given to these components is an empirical question, dependent on particular policy concerns and the degree of difficulty involved in surveying the different sectors. On the face of it, it appears that the three major sectors identified above might deserve approximately equal shares, or sector (iii) perhaps somewhat more than the other two.
Based on the above picture, we may identify different components of a potential full-scale enquiry, comprising a set of related studies. The different components may require different approaches and methodologies as will be discussed below.

### Table C.2. Composition of child labour in the baseline survey

(A) Some results from the small survey of 75 children aged 6-17

<table>
<thead>
<tr>
<th>Age group</th>
<th>Sector of work</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-17</td>
<td>Fishing</td>
</tr>
<tr>
<td>Boys</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>49</td>
</tr>
<tr>
<td>Girls</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>75</td>
</tr>
</tbody>
</table>

Breakdown by degree of hazard of work

<table>
<thead>
<tr>
<th>Boys</th>
<th>Sector of work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fishing</td>
<td>Other</td>
</tr>
<tr>
<td>most hazardous</td>
<td>21</td>
</tr>
<tr>
<td>moderately hazardous</td>
<td>11</td>
</tr>
<tr>
<td>least hazardous</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>49</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Girls</th>
<th>Sector of work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fishing</td>
<td>Other</td>
</tr>
<tr>
<td>most hazardous</td>
<td>1</td>
</tr>
<tr>
<td>moderately hazardous</td>
<td>2</td>
</tr>
<tr>
<td>least hazardous</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>3</td>
</tr>
</tbody>
</table>

(B) Tentative picture of child labour by sex, age group and main sector

<table>
<thead>
<tr>
<th>Fish</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-12</td>
<td>13-17</td>
</tr>
<tr>
<td>7-17</td>
<td>7-12</td>
</tr>
<tr>
<td>Boys</td>
<td>15</td>
</tr>
<tr>
<td>Girls</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
</tr>
</tbody>
</table>

Percentage of total child labour

<table>
<thead>
<tr>
<th>Fish</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-12</td>
<td>13-17</td>
</tr>
<tr>
<td>7-17</td>
<td>7-12</td>
</tr>
<tr>
<td>Boys</td>
<td>9</td>
</tr>
<tr>
<td>Girls</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
</tr>
</tbody>
</table>

Notes (all figures are approximate or plausible conjunctures)

(1) Figures from the baseline survey (rounded up for girls)
(2) Assumption: boys “other” underrepresented in non-fishing sectors by a factor of 0.5
(3) Age group 7-12: girl/boy ratio from school survey
(4) Assumption: girls 13-17 same number as boys
(5) Assumption: for boys age distribution in fishing same as in total (similarly for girls); figures are rounded
(6) Assumption: in non-fishing sectors, younger children subject to more undercoverage in baseline survey
  boys: distribution by age group (7-12 : 13-17) assumed as 1:2
  girls: distribution by age group (7-12 : 13-17) assumed as 1:1
C.2 Possible scenarios for a full-scale survey of child labour

Study components

It is clear from the baseline survey that children in the study area engage in many different types of labouring activity. The following are probably the main ones.

- Fishing (catching, cleaning, selling fish, etc.)
- Workshops, restaurants and similar places where mainly boys work
- Transporting goods
- Selling goods from more or less fixed places
- Mobile vendors
- Children working in bars, clubs, music and other entertainment places
- Children doing housework, working at home on family business, working in other people’s houses (e.g. as domestic servants) – activities mainly but not only by girls
- Children working in a family-run activity outside the house (e.g. hair dressing, upholstering) – again activities mainly but not only by girls
- Children doing occasional or casual work, or no work at all; apparently idle children, gathering outside homes, places of study or work, etc. (see Section 2.10)
- Children engaged in illicit activity (e.g. drug trafficking, sex work)

The full scale survey system would need to be developed for a smaller number of clearly defined and distinct components each with its own methodology, but components together aimed at capturing the full diversity of child labour activities in the study area. Of course it may not be possible to capture all the activities fully through conventional survey instruments.

In the following sections, we will consider appropriate methodology for the main components which may be covered in a full-scale survey and supplementary studies of child labour in the type of conditions implied in this illustrative example. It is important to note the hypothetical nature of this illustration – actual survey components will need to be developed taking into account what is needed, affordable and feasible in each concrete situation. Nevertheless, the hypothetical examples here have instructive and illustrative value. The survey components considered below in turn are the following.

- Survey of the fishing sector
- Household-based component
- Establishment-based survey component
- Child vendors
- Children on the street

Winning cooperation

Apart from providing, as seen above, a picture of the situation for developing more comprehensive studies of child labour, the baseline survey is invaluable in another
dimension. It helps to inform us about the field conditions under which any enquiries on child labour will be conducted.

As noted in C.1.1 above, the baseline survey began well, with cooperation and support from different community residents, especially mothers. However this turned into an hostile atmosphere by the second week: previously helpful persons become ‘cagey’; employers (fishermen) become hostile; and child workers around the area become reclusive.

It is important in all surveys, but it is crucial in surveys on sensitive topics such as child labour, to identify and try to control factors which affect how well (or ill) the enquiry is received by the community, and in particular by its target population. This deserves a considerable investment of time and resources before the actual survey can be launched.

A crucial factor is likely to be whether the survey population views the enquiry as beneficial or potentially beneficial in attracting help or better service to reduce some of the problems people have. Or conversely, whether the statistical enquiry is perceived to be threatening, for example by bringing certain activities of the people to the attention of hostile authorities.

C.3 Survey of the fishing sector

C.3.1 Scope of the survey and sampling units

The baseline survey identifies various kinds of work of children in the fishing sector such as the following:

- catching fish, including diving for fish with a spear gun, diving for pot/fish, setting fish pot, coiling/casting/pulling nets, picking lobster, rowing boats
- treating fish, including scaling fish, cutting fish, peeling conch, cooking fish and other food
- working on the boat, including unloading fish from boats, cleaning, disposing of garbage, washing, getting supplies for the boat
- selling fish and related activities, including market place fish vending, selling fish from a stall or on the street.

Some of the activities are hazardous or very hazardous.

The appropriate sampling and observation unit for catching this variety of activities is the boat. A sample of boats may be selected and the owning fisherman/woman interviewed concerning various activities and the employment of children for the jobs. Associated with each sampling unit will be a time interval when the unit is active and can be contacted.

Then after having made a contact with the fisherman, a sample of working children may be selected for in-depth interviewing.

If all boats have to be registered with some administrative authority, it is possible that a list can be obtained from the authority concerned to serve as the sampling frame. Otherwise a sample will have to be selected depending on where and when the boats
can be found. It could be the point and time of their departure to sea, or of their return from fishing.

In either case, as in any survey of a mobile population, it is necessary to begin with a seriously conducted exploratory phase. This would involve observing all locations for times when fishing boats leave (or return, depending on the choice), variation over time in the volume of the flow, weather pattern especially concerning winds and tides, and to the extent possible, the types and sizes of the boats involved. Suitable time for conducting the interview requires not only that children are present for the purpose, but also that they are free from other chores and constraints to answer questions. It is quite likely, for example, that they are too occupied or preoccupied for a period before leaving for and after returning from a fishing trip.

Normally more precise counts will be required in order to estimate the size of the population of boats and the sampling fractions to be applied later at the time of sample selection.

It is also desirable that, where possible, information for stratification is collected. Such information should not be allowed to become too complex or detailed. It should be kept simple, confined to items such as broad classification of boats by type and size (e.g. big, medium, small). Some such information may be already available in the frame when lists have been obtained from administrative sources.

The process of sample selection may be seen as involving three steps:

i. demarcation and obtaining a sample of locations;
ii. defining and selection of a sample of ‘observation points’ by ‘time segments’; and
iii. selecting a sample of boats and of children working on boats.

C.3.2 Demarcation and obtaining a sample of locations

The area to be covered by the survey may be divided into sampling locations covering the whole population. If the locations are not too many (for instance not exceeding 10), it may be possible to include them all in the survey. If there are many, a sample may have to be selected. In the latter case, it is desirable to select locations with probability proportional to some measure of size (PPS) such as the number of boats operating from the location. The size measures may be available from the list frame; otherwise they have to be constructed from observation and approximate counts during the exploratory phase mentioned above. With PPS sampling at the first stage, boats may be selected at the next stage with inverse-PPS, or a constant number taken from each selected location, so that an approximately equal probability sample of boats is obtained to the extent the size measures are reasonable. If no sampling is done at the first stage (i.e. all locations are included) then boats will have to be selected at the same rate at each location in order to obtain a uniform selection probability throughout. The same applies if locations have been selected with equal probability.

The objective of the above operations is to allocate the sample among locations, whether all the locations or a sample of them is included in the survey. Details of actual selection within each location (or each selected location) are discussed in the next two steps.
To summarise the above sampling scheme, there are essentially two scenarios. If a PPS sample of locations is selected, then the allocation to each location is approximately constant (meaning that the sampling rate within a location varies inversely to its size). If all locations are included in the study, or if they are sampled with equal probability, then the allocation to each location varies in proportion to its size (meaning the sampling rate is approximately the same in all locations).

C.3.3 Defining and selecting a sample of ‘PSUs’

It is often useful to divide locations into smaller ‘observation points’. Observation points may be area segments or even specific points where a more accurate count and sample selection of boats is done. With each observation point there will be an associated time period during which the point is active – for example when boats are leaving or arriving at the location. This time period may be divided into shorter time segments. The matrix of observation points by time segments forms the units for the next stage of selection – we will refer them as PSUs for convenience. The size of a PSU is a function of the number of boats leaving (or arriving) at the point and time defining the PSU. A more precise definition of the PSU size measure (say $S_1$) is considered in the next sampling step described below. The ideal is to create PSUs of constant size, but that is not always possible, nor is it essential in all cases. If these PSU size measures do not vary a great deal (say are within a range of 1:2 for the set of PSUs), then PSUs may be selected with a constant probability. Generally a PPS (probability proportional to $S_1$) sample is desirable.

C.3.4 Selecting a sample of boats and child workers

Let us assume that boats have been classified by size group on the basis of information available in the list frame, or approximately on the basis of field observation. Let us, for illustration, assume that there are three size groups: small boats; medium boats which are on the average 3-4 times larger than small boats; and large boats which are more than 8-10 times larger than small boats. Ideally the size should be measured in terms of the number of children working on the boat. But such information is mostly unavailable. Nevertheless, the exploratory phase may provide some information on how the more easily observable physical characteristics such as boat size and type may be correlated with the likelihood of employing child workers, and perhaps also on the likely numbers employed. The correlation is likely to be very location specific.

In the sampling scheme, larger boats should be selected at a higher rate taking account of two factors. (1) Firstly because larger boats are more important in terms of the information they provide. (2) They are likely to have a higher ‘per-boat overhead’ cost – including the cost of making contact with the fishermen, obtaining their co-operation for the interview concerning the boat and its operation, and obtaining agreement to approach and interview children working on the boats.

One compromise scheme is to select boats with probability proportional to square root of the size of the boat (number of employees if available, but more commonly on the basis of broad classification by type and physical size of the boat), and then to select working children with probability inversely proportional to square root of the boat size (see Section 3.3.3).
In our example above, boat sizes in relative terms were assumed to be roughly as ‘small’=1, ‘medium’=4, and ‘large’=9, meaning that with the square root scheme, the selection probabilities from small to large boats are in the ratio 1:4:9 i.e. 1:2:3. The number of working children selected will also be, from small to large boats, in the ratio 1:2:3 – provided that the boat size measures are chosen to reflect the number of child labourers on the boat.

The size measure \( S_i \) of PSU \( i \) mentioned earlier is the sum of size measures of the boats in that PSU, i.e. of all the boats leaving (or arriving at) that observation point during the time segment defining the PSU.

Suppose there are \( n \) boats associated with a particular PSU, with the following distribution by the size groups small-medium-large:

\[
 n = n_1 + n_2 + n_3 .
\]

Let the boat size measures be in the ratio (1:4:9) as assumed above. In accordance with the square root allocation, we may give weights to boats in proportion to square root of their size measures, namely as 1:2:3 for the three size groups. The sum of these weighted size measures of boats in a PSU is then the size measure \( S_i \) of the PSU for sample selection with PPS:

\[
 S_i = n_1 + 2n_2 + 3n_3 .
\]

Similarly, if the boat size groups differed less sharply, say the three size groups had average sizes roughly in the ratio 1:2:4, the above equation for the PSU size measure for sample selection would be:

\[
 S_i = n_1 + 1.4n_2 + 2n_3 .
\]

C.4 Household-based component

A special objective of the survey in this illustration is to enumerate more completely the work of girls, which is often within the household or within family-run business. Of course, such work by boys should also be covered as completely as possible.

The survey should aim to cover all activities of children living and working in their household: household chores, minding younger children, contributing to economic activity of the household, etc. It should also cover work by children in other households, as child domestic servants for instance.

Among economic activities performed in establishments at outside locations, there are certain sectors which tend to be family operated, with children – especially girls – of the household providing labour. Examples are hairdressing, upholstering, mending and tailoring garments, or providing laundry services. Such activities are likely to be captured more directly and more completely through a household-based survey of child labour. In order to avoid duplication with the establishment-based component described in the next section, sectors or activities covered through households should be demarcated and excluded from any establishment-based survey covering the same population and time period.
It is not necessary to discuss sampling for the household-based component here. The subject has been treated in detail in Verma (2008).

Child domestic workers can represent special situations, and hence special sampling problems, depending on the size of the group and its geographical distribution. A number of examples have been given in the preceding chapters, see in particular Sections 2.3 and 7.6, and also in Annex A.1.

C.5 Establishment-based survey component

In the present illustration, there are a number of sectors where activities are carried out in small establishments, generally at fixed locations. These cover for instance shops or stalls selling a variety of goods, workshops for making or repairing goods, construction sites, garages, electrical supplies and repairs, cafés and restaurants. Also should be included here ‘entertainment places’ – bars, clubs, dance and music halls, sports centres, etc. In the last mentioned group, the interest is not so much in children’s participation as customers but as workers. The coverage of types of activity should be as complete as possible, but excluding types of activity already covered through the household survey components (Section C.4).

Again, it is not necessary to discuss here issues in sampling of small, including informal sector, establishments since the subject has been dealt with in Chapter 5.

A major task involving fieldwork concerns the preparation of a list frame from which establishments can be selected. If the area covered by the study is not extensive, establishments may be selected in a single stage from a list covering the whole population. If the study area is extensive or if it is too difficult or expensive to list establishments in the whole population, a two-stage area-based sample would be more appropriate.

Another important issue concerns whether, when data are required for multiple sectors, it is more appropriate to conduct separate sectoral surveys, or an integrate survey covering several sectors simultaneously. This issue has also been discussed in Section 5.7.

C.6 Child vendors

The characteristic features distinguishing this component from the establishment-based survey component are that: (i) here the working children are mobile, or are stationary only at temporary locations such as stalls; and (ii) by-and-large children undertake the activity by themselves, rather than under the direct supervision of adults as in a typical enterprise of the type covered in the previous section.

The main constituent of the population covered is child vendors. However, a number of other types of activity, such as transporting goods, which are performed in similar circumstances may also be covered under this component. An activity mentioned in the baseline survey report used for this illustration is ‘giving/selling bicycle rides’ – children engaged in various forms of labour connected with carrying and transporting.

The survey should distinguish between two categories of activity in terms of the mobility involved:
i. Children performing the activity are mobile, such as street vending and transporting goods.

ii. Children working from temporarily fixed locations, such as by setting up stalls for the day on the roadside.

In order to survey (i), children performing mobile activities, it is necessary to study the pattern of movement of working children – where and when they move in the course of their work. Ideally one would try to identify whether there are points or spaces through which most child workers pass. It is also useful to identify how the size and nature of the flow changes over time. As in other surveys of mobile populations, the primary sampling units would consist of appropriately defined observation points by time segments in order to determine differences in selection probabilities resulting from different durations and frequencies with which children perform the activity. It is necessary to obtain information on those durations and frequencies, for example the number of days, mornings, afternoons, etc., per week the child performs such labour. Information should also be collected on whether during these times, or in addition to these activities at other times, the child ‘sets up a stall’, so that any overlap with (ii) can be properly taken care of. Another important question is whether the child works on his/her own, more or less as self-employed, or for someone else (including for own parent or other household members), so that overlap with other survey components discussed above can be identified and taken account of in the estimation.

In order to survey (ii), children working from temporarily fixed locations, it is necessary for the interviewers themselves to be mobile. The study area where such activities occur may be divided into smaller segments, and a sample of them selected and enumerated. Often it may be reasonable to include in the survey all working children in each selected segment.

The difference between this component and the establishment-based survey component in Section C.5 is that here the work locations are temporary, and prior listing and sampling of them is not feasible. In each segment selected the survey has to be a relatively rapid one, so as to catch the existing situation in a short time and avoid over- or under-coverage problems resulting from the transient nature of the units.

Questions similar to those for (i) concerning mobile activities have to be asked also in (ii) concerning work at temporarily fixed locations – in particular questions to identify the timing and frequency of work. It is desirable to avoid overlap with (i) and other components of the full-scale child labour survey.

### C.7 Children on the street

The objective of this component is to cover children who are not at home, nor at school or another establishment or institution, but move around or gather at outside places for prolonged periods – apparently without being engaged in any work-related activity. They may not be completely economically inactive, for instance they may be looking for work, or even doing some work but only on an occasional basis. They may be present by themselves, or in small groups, or in larger gatherings. They may be temporarily staying at one place or may move around. They may also include homeless ‘children of the street’.
How to obtain a reasonable sample of such a population of children depends very much on the particular circumstances and prevailing patterns, but is a difficult task in any case. Strictly probability sampling is likely to be out of the question. Small quota samples, as wide in coverage as possible in terms of space, time, situation and activity, may be a possible approach. Snow-balling and similar link-tracing strategies may also be useful. In these approaches, initially contacted children lead to other children in similar circumstances. (See Chapters 13 and 14 for discussion of these and similar sampling methodologies. In Section 14.3.1, for instance, a survey of street children using the respondent-driven sampling method has been reviewed in some detail.)

Of course not all children present outside are in the target population for the survey component being discussed. Children are not prisoners and may be outside for any number of other reasons. Hence a careful screening operation is required to identify the target population of interest.

It is important to begin the interviewing by asking about schooling: whether the child is enrolled, and if so whether the child is in regular attendance at school. This perhaps is the most useful set of screening questions. Then those not enrolled and not attending school may be asked about their activity – what do they normally do? If not work, then what? How much of the day do they spend outside? And so on.

Despite the difficulties of sampling, screening and enumeration, it is important to cover this group of children where it is large enough to be of concern. While it may not come strictly under the rubric of ‘child labour’, it is important as a survey component concerning the condition of children who are not at work and not at school.


Groger, L.; Mayberry, P.; Straker, J. 1999. “What we didn’t learn because of who would not talk to us”, in *Qualitative Health Research*, 9(6), 829-835.


Matsuno, A.; Blagbrough, J. (n.d.). “Child domestic labour in South-east and East Asia: emerging good practices to combat it”. Unpublished report, ILO.


__ 2012. Private communication, including a series of technical notes providing numerical illustration of diverse sampling procedures (multi-frame, multiplicity, adaptive cluster, snowball, location and balanced sampling, sampling of street children, and oversampling in strata of concentration) within the common framework of indirect sampling.


BiBliography

Sampling elusive populations: Applications to studies of child labour


Bibliography


BiBliography

Sampling elusive populations: Applications to studies of child labour


<table>
<thead>
<tr>
<th>Author</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aaron</td>
<td>486, 793</td>
</tr>
<tr>
<td>Abdul-Quader</td>
<td>677, 793</td>
</tr>
<tr>
<td>Aderinto</td>
<td>56, 745, 793</td>
</tr>
<tr>
<td>Adhikari</td>
<td>56, 748, 793</td>
</tr>
<tr>
<td>Akșit</td>
<td>56, 746, 793</td>
</tr>
<tr>
<td>Albania</td>
<td>59, 68, 751, 793</td>
</tr>
<tr>
<td>Alcantara</td>
<td>725, 797</td>
</tr>
<tr>
<td>Aldana</td>
<td>700, 793</td>
</tr>
<tr>
<td>Anderson</td>
<td>346, 502, 505, 508, 509, 512, 801, 804</td>
</tr>
<tr>
<td>Ashraf</td>
<td>43, 49, 728, 734, 737, 793</td>
</tr>
<tr>
<td>Atkinson</td>
<td>610, 614, 793</td>
</tr>
<tr>
<td>Bagchi</td>
<td>56, 749, 793</td>
</tr>
<tr>
<td>Bal Kumar</td>
<td>56, 748, 793</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>37, 70, 151, 152, 154, 155, 156, 161, 257, 258, 362, 793, 794</td>
</tr>
<tr>
<td>Betti</td>
<td>116, 119, 177, 231, 426, 807, 808</td>
</tr>
<tr>
<td>Biernacki</td>
<td>610, 623, 658, 661, 699, 794, 808</td>
</tr>
<tr>
<td>Biggeri</td>
<td>57, 750, 794</td>
</tr>
<tr>
<td>Binder</td>
<td>116, 794</td>
</tr>
<tr>
<td>Bishop</td>
<td>538, 794</td>
</tr>
<tr>
<td>Blagbrough</td>
<td>33, 36, 721, 794, 802</td>
</tr>
<tr>
<td>Blair</td>
<td>333, 336, 337, 795</td>
</tr>
<tr>
<td>Blangiardo</td>
<td>448, 794</td>
</tr>
<tr>
<td>Bohning</td>
<td>485, 794</td>
</tr>
<tr>
<td>Brewer</td>
<td>203, 206, 208, 794, 798</td>
</tr>
<tr>
<td>Broadhead</td>
<td>667, 798</td>
</tr>
<tr>
<td>Buckland</td>
<td>323, 619, 801</td>
</tr>
<tr>
<td>Burnham</td>
<td>502, 505, 508, 509, 512, 804</td>
</tr>
<tr>
<td>Cain</td>
<td>659, 804</td>
</tr>
<tr>
<td>CAPMAS</td>
<td>82, 486, 490, 492, 794</td>
</tr>
<tr>
<td>Chang</td>
<td>486, 793</td>
</tr>
<tr>
<td>Chao</td>
<td>557, 794</td>
</tr>
<tr>
<td>Chaudhri</td>
<td>726, 795</td>
</tr>
<tr>
<td>Chauvet</td>
<td>563, 605, 795</td>
</tr>
<tr>
<td>Citro</td>
<td>325, 795</td>
</tr>
<tr>
<td>Clémenceau</td>
<td>177, 808</td>
</tr>
<tr>
<td>Coleman</td>
<td>610, 795</td>
</tr>
<tr>
<td>Cormack</td>
<td>538, 795</td>
</tr>
<tr>
<td>Cowan</td>
<td>623, 806</td>
</tr>
<tr>
<td>Cruyff</td>
<td>485, 486, 807</td>
</tr>
<tr>
<td>Czaja</td>
<td>333, 336, 337, 795</td>
</tr>
<tr>
<td>Dalenius</td>
<td>197, 795</td>
</tr>
<tr>
<td>Damacena</td>
<td>685, 795</td>
</tr>
<tr>
<td>David</td>
<td>613, 623, 795</td>
</tr>
<tr>
<td>de Boer</td>
<td>43, 729, 730, 795</td>
</tr>
<tr>
<td>de Cock</td>
<td>63, 796</td>
</tr>
<tr>
<td>de Groot</td>
<td>43, 726, 795</td>
</tr>
<tr>
<td>Deming</td>
<td>104, 323, 486, 795, 805</td>
</tr>
<tr>
<td>de Souza e Silva</td>
<td>63, 755, 795</td>
</tr>
<tr>
<td>Deville</td>
<td>105, 116, 177, 181, 561, 562, 563, 592, 431, 604, 795</td>
</tr>
<tr>
<td>Diallo</td>
<td>36, 723, 796</td>
</tr>
<tr>
<td>Dixon</td>
<td>271, 796</td>
</tr>
<tr>
<td>Dumas</td>
<td>63, 796</td>
</tr>
<tr>
<td>Durand</td>
<td>630, 796</td>
</tr>
<tr>
<td>Early</td>
<td>206, 794</td>
</tr>
<tr>
<td>ECPAT</td>
<td>59, 796</td>
</tr>
<tr>
<td>Erickson</td>
<td>668, 796</td>
</tr>
<tr>
<td>Ernst</td>
<td>177, 796</td>
</tr>
<tr>
<td>Espino</td>
<td>47, 732, 796</td>
</tr>
<tr>
<td>Fashola</td>
<td>668, 691, 796</td>
</tr>
<tr>
<td>Faugier</td>
<td>610, 796</td>
</tr>
<tr>
<td>Fee</td>
<td>76, 796</td>
</tr>
<tr>
<td>Félix-Medina</td>
<td>621, 667, 796</td>
</tr>
<tr>
<td>Ferguson</td>
<td>51, 797</td>
</tr>
<tr>
<td>Ferretti</td>
<td>808</td>
</tr>
<tr>
<td>Ferro</td>
<td>56, 745, 801</td>
</tr>
<tr>
<td>Fienberg</td>
<td>486, 538, 794, 797</td>
</tr>
<tr>
<td>Fisher</td>
<td>485, 486, 797</td>
</tr>
<tr>
<td>Flint</td>
<td>610, 614, 793</td>
</tr>
</tbody>
</table>
Flores-Oebanda 725, 797
Foster 486, 808
Frank 609, 619, 625, 627, 667, 797
Fyfe 41, 797

G
Gagliardi 326, 808
Gaudet 359, 802
Genicot 61, 797
Ghellini 231, 177, 807, 808
Gile 609, 714, 717, 797, 798
Glasser 189, 797
Godoy 45, 730, 797
Goel 680, 714, 715, 797
Goodman, L. 609, 797
Goodman, R. 559, 560, 563, 592, 594, 604, 797
Groger 613, 797
Groves 224, 801
Guarcello 57, 750, 794
Gündüz-Hoşgör 56, 746, 793
Gupta, I. 750, 797
Gupta, R. 61, 798
Gurgel 485, 486, 798
Gurung 56, 748, 793

H
Hajek 396, 798
Handcock 609, 714, 717, 797, 798
Hanif 203, 206, 208, 794, 798
Hansen 170, 322, 323, 368, 397, 414, 429, 455, 461, 462, 692, 693, 798
Hartley 366, 798
Hatley 52, 53, 82, 681, 798
Hay 486, 798
Heckathorn 667, 668, 671, 675, 677, 686, 692, 694, 695, 696, 698, 793, 798, 804, 808
Hess 561, 799
Hindman 24, 799
Hodges 197, 795
Hogan 486, 799
Holland 538, 794
Hook 485, 799
Horvitz 353, 354, 367, 455, 481, 561, 562, 621, 799

Houwelingen 485, 486, 807
Hubble 177, 796
Hughes 667, 798
Human Rights Watch/Asia 61, 799
Hurwitz 173, 322, 348, 353, 397, 414, 429, 455, 461, 462, 692, 693, 798
Huser 52, 53, 82, 681, 798

I
ICLS 29, 30, 31, 32, 799
ILC 33, 799
ILO-IPEC 28, 29, 30, 31, 33, 36, 42, 43, 44, 45, 46, 47, 48, 49, 50, 56, 732, 796, 799
International Labour Organisation 1, 799, 800, 807
IWGDMF 485, 800

J
Jamaica 19, 774, 778, 800
Jeffri 667, 798
Johnston 667, 670, 675, 676, 677, 678, 695, 709, 800, 801, 804
Judkins 177, 796
Júnior 685, 795

K
Kajubi 671, 672, 676, 688, 801
Kalton 317, 325, 346, 347, 361, 466, 795, 801
Karancı 56, 746, 793
Kassouf 56, 745, 801
Kendall, C. 679, 702, 800, 801
Kendall, M. 323, 619, 801
Kish 4, 106, 118, 119, 126, 143, 224, 313, 321, 322, 324, 365, 559, 560, 563, 592, 594, 604, 797, 801
Kitada 37, 725, 801
Klovdahl 609, 801
Kraemer 659, 801
Krotky 224, 801

L
La Ferrara 61, 802
Lapierre 359, 802
LaPorte 486, 793
Lauby 674, 805
Author Index

811

Lavallée  177, 180, 181, 400, 431, 796, 802
Liebman 674, 805
Lohr 325, 802
Lopes 631, 802
Lyon 57, 750, 794

M
Madow 322, 798
Madran 60, 802
Magazine 743, 802
Main 53, 802
Markovic 486, 793
Marriage 62, 802
Masina 56, 744, 802
Massey, D. 630, 796
Massey, F. 271, 796
Matei 563, 806
Matsuno 33, 36, 721, 802
Mayberry 613, 797
McKeganey 486, 798
McKenzie 448, 472, 627, 802
MEASURE Evaluation 426, 802, 806, 808
Mecatti 369, 802
Mehran 110, 179, 284, 350, 373, 401, 430, 432, 547, 592, 757, 799, 803
MELS 222, 803
Milgram 695, 803
Mistiaen 448, 472, 627, 802
Mitra 56, 750, 797, 803
Monsour 204, 805
Morrison 485, 803
Muhib 469, 803
Mwami 47, 732, 803

N
Nadeau 359, 802
Nathan 343, 803
Neter 535, 803
NSAC 737, 803
Nyland 726, 795
Nyoni 47, 732, 803

O
Ohlsson 206, 803
O’Muircheataigh 325, 803
O’Neill 49, 736, 737, 803
Oosterlee 486, 803
O’Rourke 726, 795
Otis 502, 505, 508, 509, 512, 804

P
Pacis 725, 797
Pakistan 43, 254, 727, 804
Patak 116, 794
Patrick 610, 804
Pedlow 325, 803
Penrod 659, 804
Peralta 56, 742, 804
Platt 700, 705, 706, 804
Potter 106, 804
Pratesi 448, 804
Preston 659, 804
Pugh 485, 486, 797

Q
Quintero 700, 793

R
Rao 116, 325, 802, 804
RDSAT 691, 698, 804
Regal 485, 799
Rocco 448, 804
Rodrigues 631, 802
Rosati 57, 750, 794
Ryost 174, 805

S
Sabin 670, 675, 676, 677, 678, 793, 800, 801
Salganik 667, 668, 680, 696, 714, 715, 716, 797, 804
Sanga 47, 732, 803
Sargeant 610, 796
Särndal 105, 149, 207, 290, 767, 796, 804
Sattaur 737, 805
Seber 395, 415, 420, 421, 424, 425, 485, 488, 489, 515, 544, 549, 805, 806
Sekar 486, 805
Semaan 667, 673, 674, 675, 798, 805
Sergeyev 667, 798
Sichieri 631, 802
Sidel 538, 808
Sigm 204, 805
Sirken 165, 174, 327, 332, 333, 340, 623, 805, 806
Smit, P. 803, 808
Smit, F. 486, 798, 803
Snijders 613, 616, 619, 621, 623, 624, 625, 627, 795, 797, 805
Sommerfelt 37, 724, 805
Spreen 610, 623, 805
Srikantan 561, 799
Starks 659, 804
Stone 485, 803
Straker 613, 797
Subedi 56, 748, 793
Sudman 267, 319, 340, 345, 347, 623, 805
Swensson 290, 767, 804
Szwarcwald 685, 795

T
Tanzania 47, 426, 732, 733, 803, 806
Taylor 485, 486, 797
Thiemann 659, 801
Thompson, D. 170, 205, 208, 353, 354, 367, 396, 455, 481, 561, 562, 621, 799
Thompson, S. 6, 386, 389, 395, 415, 420, 421, 424, 425, 621, 667, 796, 806
Tillé 354, 561, 562, 563, 592, 604, 761, 795, 796, 806
Tilling 485, 806
Tjandraningsih 49, 735, 806
Tobi 677, 793
Tremblay 359, 802
Trow 610, 806
Turner 485, 486, 797

U
Uganda 51, 56, 70, 255, 256, 738, 740, 741
UNICEF 53, 733, 806
United Nations 137, 139, 140, 763, 806, 807
Urani 63, 755, 795

V
van der Heiden 485, 486, 807
van Meter 613, 623, 667, 807
Verma 1, 3, 25, 72, 75, 91, 114, 116, 119, 122, 177, 186, 190, 207, 221, 231, 247, 298, 314, 318, 323, 326, 426, 428, 758, 763, 773, 787, 799, 807, 808
Vink 486, 803
Volz 692, 694, 698, 808

W
Waldorf 610, 658, 661, 794
Wang 711, 713, 808
Watters 623, 699, 808
Watts, C. 486, 808
Watts, D. 695, 808
Weir 486, 521, 522, 808
Wells 325, 808
White 502, 505, 508, 509, 512, 804
Whyte 610, 808
Wilkinson 426, 808
Williams 485, 486, 808
Wittes 538, 808
Wretman 290, 767, 804
Wu 116, 804

Y
Yue 116, 804

Z
Zwi 486, 808
adaptive cluster sampling 6, 13-4, 49, 53, 110, 165-6
basic procedure 385-92, 421-5, 440-2
estimation, with illustrations 396-414
with imperfect detectability 420-1
with multiple adaptive criteria 419
with multistage sampling 416-8
with order statistics 420
overlapping adaptive clusters 426-9, 432-40
with stratification 415
with unequal selection probabilities 414-5
uses and limitations 386-8, 392-5
variance 387, 391-6, 404-5, 415, 419-22, 442
waves of adaptive sampling 425-6, 430-2
adaptive clusters 388-91, 403, 416
Africa/African 46, 50, 54-5, 62-3, 251, 667, 672, 688-92, 724, 756
agriculture sector 60, 722-4, 752-4, 758
see also: commercial agriculture
Albania 59, 68
survey illustrations 751-2
allocation 10, 67, 141, 161, 183-5, 190-204
compromise allocation 313-8
effectiveness of allocation 236-46
model-based allocation 193
optimal allocation 86, 189-90, 193-200, 302, 305-6, 312-4
power allocation 194
proportionate allocation 153-4, 190-7, 297, 302, 306-8, 314, 317-8
Neyman allocation 196-7
StrCon, allocation using 231-6, 313-8
X-optimal allocation 196
analysis units and sampling units 9-13, 84, 98, 102, 111-2, 119-22, 131, 135, 158, 172-8, 228, 286-7
defined 132
estimation, with illustrations of 110-4
imperfect correspondence between 142-51
linkages, types of 163-9
in multiplicity sampling 327-8, 333, 337, 340-3, 348-52, 356, 399-400, 440
see also: reporting and reported units
area frame 9, 12, 131, 758
defined 138-9
economic census, as source of 269
multiple frames 257-9, 358-65, 371-80
requirements and problems 139-41
Asia/Asia Pacific 36, 43, 49-50, 61, 667
survey illustrations 721-3, 726
auxiliary variables/information 16, 109, 135-9, 141, 147, 188, 193-7, 200, 323, 358-60, 395, 419, 422, 561-3, 593, 596, 600, 604, 767
balanced sampling 16, 46, 186, 354, 463
defined 561-2
basic procedure, illustration 592-5
step-by-step procedure, illustration 595-9
with time-location sampling, illustration 600-4
see also: controlled selection
Bangladesh 9, 11, 35, 47, 68-71, 253, 362, 726
survey illustrations 37-9, 151-8, 161-3, 257-60
baseline surveys and studies
defined (objectives and types of) 68-70
as data source 1, 8, 47, 65, 80
for developing full-scale child labour survey 19, 469, 774-89
as establishment surveys 131, 253-60
elements of sectors covered 71
illustrations from country studies 37-9, 43, 49, 51, 151-8, 161-3, 253-60, 727-9, 738-40
sampling frames used 9, 151-3, 159-63, 362
Bernoulli sampling 206
bias 14, 19, 86-7, 106, 117-9, 123, 261, 442, 447, 470-2, 476, 757, 764-6
in adaptive cluster sampling 385-6, 420-1, 425
in capture-recapture sampling 675, 679-80, 686, 692-6, 699-700, 704-5, 714-8
in chain referral procedures 17, 612-6, 622, 642-4, 659-60, 664, 667-71, 717
coverage and non-response bias 101, 109, 126, 148-51, 262
from incomplete lists 141-69
in multiplicity sampling 12, 340-1, 346-7, 350, 369, 376
technical (estimation) bias 123-5, 207-8
commercial agriculture 8, 24, 31, 40-43, 68, 721, 726-30
see also: agriculture sector
commercial sexual exploitation 8, 24-7, 32, 35, 57-60, 68, 78, 721-2
survey illustrations 751-5
community-level inquiries 65
compromise allocation 313-8
concentrations of population/subpopulations 3, 6, 11, 13, 19, 39, 74, 88, 90, 228, 249, 265-7
locating concentrations 268-71
strata with differing concentrations 290-6
see also: empty clusters; strata of concentration (StrCon)
construction sector 8, 24-31, 49-51, 55, 721, 738-40, 749, 787
controlled selection 16, 46, 186, 463, 559-61
basic procedure, illustration 560-1
step-by-step procedure, illustration 563-86
time-location sampling, illustration 587-91
see also: balanced sampling
correlation 91, 147, 196, 200, 500, 724, 785
see also: intra-cluster correlation
Costa Rica 68
Côte d’Ivoire 36, 40
survey illustrations 723-4
coverage error 5, 12, 99, 109, 133, 140, 225, 251, 341-2, 357
defined 126
non-coverage 104, 345
cumulation over time 324-6
deft see design effect
design effect 8, 114, 186, 227, 241-4, 251, 307, 771
analysis of 84, 118-22, 236-7
in chain referral sampling 612, 680, 698, 710, 715-6
effect of sample weights 119, 122, 126, 236-45

direct sampling 9, 98, 131-2, 164-5, 178-80, 186

domestic labour/work 8, 12, 24-7, 31-7, 40, 52-4,
60-2, 78, 81, 292, 541, 661
framework for identification 32
and household chores 31-2, 63
using sampling with multiplicity, illustrations 12,
40, 165, 327-8, 343, 350-6
using sequential test, illustration 271-83

drug trafficking 8, 24-9, 44, 57, 62-3, 68, 721, 755-6,
774-6, 780-2

dual approach to listing establishments 249-52

economic units 41, 47, 183-4, 221-3, 225, 229-32,
246-51, 260, 268-70
Ecuador 60, 68
survey illustrations 752-4
Egypt 15, 82, 426, 486
survey illustrations 490-3
El Salvador 45, 52, 68
survey illustrations 730-2
elusive population 98, 131, 143, 148, 163-5, 168,
358, 421, 486, 609-10, 616
defined 4-9

employers’ surveys 66

empty clusters 6, 11, 110, 225, 266-7, 271-83,
284-7, 291, 296, 320, 358, 433
equal probability sampling 5, 188-91, 201-2, 208,
236, 265, 331, 450, 453, 464-7, 478, 482, 501,
515, 527, 530
examples 153, 354, 373, 378, 406, 458, 471, 523,
564, 587, 600-1, 696, 758, 763-4, 784-5
see also: self-weighting samples

errors in survey data 122-7

establishment surveys 1, 8, 70, 226
sample allocation schemes 191-200
simulated illustrations 212-21
survey illustrations 151-63, 253-9
types of 66-7

see also: proportionate allocation; optimal allocation; sampling of establishments

establishments, sampling of
illustrations, baseline surveys 70, 151-63, 253-9,
783
sample allocation 691-7
sample design issues and strategy 185-7, 228-30
sampling frames 183-9, 196-201, 207, 210, 225-8,
232-7, 247, 253-61
sampling stages 186, 232, 247
selecting children see selecting children within establishments
single stage sampling schemes 136, 201-9
with single stage selection 201, 204-7, 257-9
stratification 187-91, 697
see also: establishment surveys; strata of concentration (StrCon); selecting establishments within sample areas

Estonia 18, 68, 701, 706, 714
fieldwork 198, 266, 279, 288, 395, 464, 531, 685,
700-1, 705-7, 736, 755, 775, 787
fishing and aquaculture sector 8, 19, 24, 31, 44-6,
68-9, 71, 78, 222, 258, 541, 691, 721, 729-31,
774-84
fixed probabilities vs. fixed sample size 203-4
fixed-take design 89-90, 211
focus groups, in rapid assessments 77-8
forced labour see bonded labour

Ghana 18, 60, 68
survey illustrations 681-4, 754-5
Guatemala 57, 68, 426, 750
Hansen-Hurwitz estimator 170, 368, 397, 414, 429,
455, 461-2, 481, 692-3
hazardous work 3, 11, 23-5, 30-6, 42-50, 70-1, 152,
257-8, 345, 362, 445, 486, 616, 634, 640
illustrations, studies in different sectors 647-57,
721-56
international standards 25-30
see also: sectors of hazardous work
heterogeneity of population 5, 53, 223, 228, 395,
556-7, 772
Horvitz-Thompson estimator 170, 205, 208, 353-4, 367, 396, 455, 481, 561-2, 621

household-based surveys 1, 8, 201, 230, 375-6, 416, 473, 476

in association with establishment surveys 65-6, 162, 221, 228, 249, 252, 270, 377-8, 383-4
examples of application 37, 40, 49, 151-2, 727-8, 738-40, 782, 786-7
listing and sampling 132-4, 225, 270
methodology, CLS, LCS 72-5
use with multiplicity rule 331, 339, 342, 345
compared with other survey types 2-3, 64-5, 75, 183
compared with rapid assessment (RA) studies 25, 69, 80-2

Hungary 17, 622-7
survey illustrations 623-7

‘idle’ children 57, 750

ILO conventions see international conventions concerning child labour
impact assessment 68-9
imperfect detectability 420-1
imperfect (list, area) frames see area frame, list frame

index of relative concentration 233-4

India 34, 43, 48-9, 54-6, 61, 268-9, 726
survey illustrations 728-9, 734-8, 749-50

indigenous fieldworkers sampling (IFS) 18, 700-1, 705-7
indirect sampling 9, 131-2, 164-5, 178-81, 431

Indonesia 35, 49, 722-3
survey illustrations 735-6

informal sector establishments 8-11, 24, 27, 47-9, 53, 56, 67-71, 81, 183-5, 260-2, 268, 372, 941
characteristic features of 221-4
illustrations 151-8, 161-3, 253-9, 721-6, 737, 740-9, 787
survey design issues 224-7
see also: sampling of establishments; selecting establishments within sample areas

initial sample 341, 473, 476, 771
for adaptive cluster sampling 13, 385-98, 401-2, 405-15, 418-39, 442

international child labour statistics
framework for statistical identification 32
ILO global estimation, conceptual framework 29
international comparisons 28
international standards 29

international conventions concerning child labour
Convention No.182 (1999) 25-8, 31, 35
Convention on Domestic Workers No.189 (2011) 25, 31, 33
ILO Recommendation No.190 27
International Conference of Labour Statisticians, Resolution (2008) 29
Minimum Age Convention No.138 (1973) 25-6
UN Convention on the Rights of the Child (1989) 25, 26
UN Protocol to Prevent, Suppress and Punish Trafficking in Persons (2000) 59
UN Supplementary Convention on the Abolition of Slavery, the Slave Trade and Institutions and Practices Similar to Slavery (1956) 61

intra-cluster correlation 121, 321-3
isolating empty clusters
gain in precision 284-90
see also: empty clusters

item non-response 125-7

Jamaica 19, 44, 68, 71
survey illustrations 774-89

jackknife repeated replication (JRR) 116-7, 121-2

Kenya 35

Lebanon 68

linearization 114-6

link-trace sampling 18, 132, 168-9, 178-80, 426, 609, 659, 698, 711, 717
see also: chain-referral sampling
list frame 9, 10, 39, 99, 136-9, 183-5, 292, 428, 464
example of use 19, 151-8, 161-3, 757, 784-9
imperfections in 131, 142-51
for single stage selection of establishments 201, 204-7, 257-9
use with multiple frames 358-65, 371-84
see also: selecting establishments within sample areas
listing units 248-50
see also: screening
location sampling
defined 14, 448-50
multistage 461-2, 482-3
single-stage 450-60, 481-2
Madagascar 68
Mali 82
survey illustrations 683-4
manufacturing sector 8, 24-6, 31, 47-50, 71, 253-8, 721, 727-8, 734-9, 758
Markov chain 18, 669, 696
mean square error (MSE) 108, 716
measurement errors 123-4
measures of size 88-95, 137, 147, 185-92, 202, 210-2, 217-20, 225, 416-8, 466, 563, 569, 595, 738, 784
scaling of size measures 92-4, 198-200, 246-7, 318, 564
Mexico 56, 622, 631, 742-4
migrants 41, 472-7, 627-30, 722
mining and quarrying sector 8, 24-31, 45-7, 63, 68, 78, 258, 541, 721-4, 728, 732-4
mobile populations 2-3, 7, 14-6, 19, 23, 46, 62, 71, 81, 90, 158-60, 169, 172, 184, 226, 256, 327
defined 445-7
examples of studies 468-77
sampling framework and weighting 446-7, 477-83
see also: location sampling; time-location sampling; capture-recapture sampling
modular surveys 1-2, 10, 65, 72, 185, 229-30, 253
Moldova 68
Mongolia 722
Morocco 37, 40
survey illustrations 724-5
MoS see measures of size
multiphase sampling 11, 179, 267, 285, 318-24, 473, 757, 766-73
multiple frame sampling 12-3, 85, 134, 165-6, 257, 271, 325-7
in area frames, examples 257-9, 358-65, 371-80
estimation from multiple frames 367-71
in list frames, simulated illustrations 358-65, 371-84
potential uses 357-9
overlapping frames, dealing with 362-3, 383
multiplicity rules 12, 173, 327-30, 333-53
multiplicity sampling 12-3, 131, 142, 165-72, 267, 284
basic procedure 327-32
estimation 348-50
illustrations 350-6
uses and limitations 343-8
variance 340-3, 352-5, 369, 378
see also: sampling with multiplicity; multiplicity estimator; multiplicity rules
multi-sectoral surveys 10-1, 70, 151, 185, 228, 231, 257
multistage sampling 13-5, 135-6, 604
with adaptive cluster sampling 416-8
with capture-recapture sampling 520-4
with location sampling 461-2, 482-3
Nepal 49, 56-7, 68, 726, 750
survey illustrations 736-7, 748-50

network sampling see multiplicity sampling

network structure 18, 329, 337, 350, 470-2, 609-13,
617, 648, 662-81, 685-9, 691-700, 704-10, 714-7,
756

networks see adaptive clusters

Neyman allocation see allocation

Nicaragua 55

Nigeria 56
survey illustrations 745

non-probability sampling 161, 169, 261, 614-9, 705

non-response 4, 85, 88, 125-7, 143, 148-9, 190, 207,
261, 344-7, 526, 531-2, 547, 613, 638, 653
weighting for non-response 8, 83, 100-9

non-sampling errors 83-4, 123-7, 140, 160, 174, 251,
341, 421, 716

optimal allocation see allocation

optimal cluster size 322

optimal stratification 190, 197

oversampling 39, 86, 320, 417, 526, 609, 630, 673,
700
of strata of concentration 11, 19, 267, 291, 296-318,
757-73

Pakistan 11, 43, 47, 50, 71, 726
survey illustrations 253-5, 728-7

panel surveys 177

Peru 34, 36

Philippines 34, 37, 40, 43, 47, 55, 60, 68
survey illustrations 725-6, 729-32, 752-4

Poisson sampling 10, 204-6, 209, 486

primary sampling units (PSUs) 37-9, 87-92, 96, 108,
115-8, 135-6, 152-3, 160, 226, 517

probability proportional to size (PPS) sampling 8, 38,
83, 137, 150, 199-204, 246, 251, 414-8, 464, 530
basic PPS design 88-96, 209-12

illustrations 37, 95-8, 212-21, 253-5, 466-7, 522-3,
563-86, 595-9, 738, 758, 761-2, 784-6
with systematic sampling 93-5
variations from basic PPS design 89-92, 189-90,
232, 318

probability sampling 8, 16-7, 64, 81-2, 105, 164,
251, 261-2, 388, 425, 559-61, 615, 631, 670, 700,
789
defined 83-5, 126, 134, 179

project design and monitoring, baseline surveys 69

proportionate allocation see allocation

raking procedure 104-6

rapid assessment (RA) studies
as supplement to baseline surveys 69
as data source 1, 8, 65
examples of RA studies 45-7, 56-63, 68
compared to household-based surveys 25, 80-2
illustrations from country studies 730-4, 746-56
methodology 67, 76-9, 88, 133
uses and limitations 79-80

rare populations
defined 5, 265-6
estimation 299-318
sampling techniques 11-3, 90, 162, 228, 266-7

rare traits 5, 265
see also: rare populations

reclusive populations 7, 16, 23, 40, 81, 158, 184,
447, 465, 610, 617, 622, 639, 659, 666, 670, 681,
700, 708, 783
defined 2, 6

recruitment chains 17, 674-7
incentives 679-80
recruitment by previous respondents 677-8
from single productive seed 712-3

recruitment quotas 679

referrals in link-trace sampling 17, 473-4, 609-12,
622, 627-56, 661-81, 701, 709, 712-3, 717

reporting and reported units 328-30

respondent-driven sampling 7, 17-8, 53, 79, 82, 132,
157, 169, 327, 472, 632
alternative procedures, comparison with 698-707
coupons and incentives 669-80, 686, 690-3, 697-703, 707-11, 717
‘degree’ of respondent 669-73, 693-8, 715-7
design effects 680, 698, 710, 715-6, 771
equilibrium 670-5, 688, 690, 694-6, 710-3
field experience review 709-13
homophily and heterophily 669-70, 680, 686-7, 695, 707, 717-8
illustrations of field studies 680-92
methodology 667-80
model assumptions 692-8, 708
sensitivity 717-8
simulations for assessment of network structure 714-6

Romania 68

Russia, Russian Federation 18, 701-2, 714
survey illustrations 705-7

sampling distribution 290-1, 411-3, 421, 426, 765

see also: variance

sampling frames
defined 83-8, 111, 131-8, 160, 179
in controlled selection 587
for establishment surveys see establishments, sampling of
illustrations from establishment/location surveys 70, 151-63, 253-9, 783-6
incomplete frames 1-9, 16, 47-8, 72, 81, 126, 168, 180, 225, 261, 525, 610-6, 623-4, 647-9, 661, 667
for mobile populations 446-7, 463, 469-71, 525, 531, 587
multiple and overlapping frames, multiplicity see multiple frames
for multistage sampling 135-6
for rare populations 265-8, 273, 282, 288, 291, 357-61
rules of association 84, 134-5
sources 65, 139, 268, 271, 699, 702, 738
supplementary sampling frames 270-1, 359-61
unit size measures see measures of size
for weighting 96, 99, 101, 178

see also: area frame; list frame

sampling units 40-1, 134, 226-9, 587, 600, 783, 788
and analysis units see analysis and sampling units
‘computational’ PSUs and strata 117-8
and listing units 135
macro and micro sampling units 135-6
PSUs see primary sampling units
time-location sampling PSUs 14, 447, 462-3
USUs see ultimate sampling units
‘visits’ vs. ‘visitors’ as survey units 465-6
see also: area frame; list frame; sampling frame

sampling with multiplicity 9-12, 99, 131, 165-72, 178-80, 261, 271, 327, 357, 367, 426

scaling of sample weights 8, 97, 106-9, 528-31, 536-7, 664

scaling of selection probabilities 453-4, 460-2, 582-6, 599

scaling of unit size measures see measures of size

see also: listing operation

sectors of child labour (examples)
automobile sector 155-6
battery recharging and recycling 47, 152-4
brick factories 50
carpet industry 734-7
commercial sexual exploitation 26-7, 60, 521-2, 681, 685-8
family street workers 743
glass bangles industry 253-4
glass industry 737-8
porters 55
rag-picking 54
road transport 157
slavery 26
waste-picking 55
welding sector 154-5

sectors of hazardous work 24

seeds for link-trace sampling see initial sample

segmentation 139, 247-8
selecting children within establishments 10, 87, 91, 146, 209-21
selecting establishments within sample areas 11, 139, 185, 247-52, 258-9, 266
self-weighting samples 88-90, 99, 223, 227, 236, 240, 252, 762-4, 773
see also: equal probability sampling
sequential-collocated sampling 206, 209
sequential sampling 10, 206-9
sequential test 11, 266, 271-83
simple random sampling (SRS) 118, 121, 205, 259, 488, 501, 588, 621, 680, 715
departures from SRS 8, 83-8, 414
illustrations using SRS 110-4, 212-7, 433-8
SRS within strata 190, 208, 299-307, 355
single-sector surveys 10, 185, 228-9, 257
small areas and domains 265
small establishments
see: informal sector establishments; sampling of establishments; selecting establishments within sample areas
snowball sampling
defined 609-14
comparison with other methods 625-30
estimation procedures and problems 614-22
illustrations from national studies 622-34
implementation issues 657-66
selective follow-up 631-4
simulated illustrations: inaccessible strata 647-57
simulated illustrations: structure of referrals 634-47
social networks see network structure
South Africa 55, 68
Sri Lanka 68, 726
stand-alone surveys 2, 10, 72, 75, 229-31, 253
strata of concentration 10-1, 19, 185, 229, 267, 296
constructing StrCon 231-6
oversampling StrCon 296-313, 757-73
sample allocation 231-46, 313-8
StrCon see strata of concentration
street children 1, 8, 13-8, 24, 35, 51-5, 64, 68-71, 78, 99, 152, 721, 739-54, 782, 787-9
using adaptive cluster sampling 385, 401-7, 411-3, 419, 430-2
using capture-recapture sampling 81-2, 157-9, 485-6, 490-3
illustrative population 401-5
using respondent-driven sampling 681-4
typology 51-2
street, children of the 51-5, 743-4
street, children on the 51-4, 788-9
street families 52
street-living children see street, children of the
street-vending children 55, 745-6, 787-8
street-working children see street, children on the
systematic sampling 93-101, 117, 143, 189, 199, 217-8, 246-7, 373-4, 378, 411-3, 430-1, 464-7, 594, 758-64
take-all sampling 145-6, 186, 199, 202, 213, 216, 250
with cut-off 90-1
take-all strata 189-90
Tanzania 34, 47, 68, 426
survey illustrations 732-4
target population 6, 12-7, 37-40, 63, 73-80, 132-4, 540-1
targeted sampling 5, 11, 18, 265, 623, 698-700
Thailand 68, 722
time-location sampling 6-7, 14, 18, 87, 172, 327
defined 447, 462
with controlled selection (balanced sampling) 587-91, 600-4
examples of application 466-72
sampling procedures 462-6
Turkey 56–7, 68, 71, 750
  survey illustrations 56, 746–7
two-phase sampling see multiphase sampling
Uganda 11, 18, 49–51, 56, 70–1, 253, 671, 676
  survey illustrations 255–8, 681, 688–91, 738–42
Ukraine 68
ultimate sampling units (USUs) 87, 135, 228, 252, 266
UN conventions see international conventions concerning child labour
unconditional worst forms of child labour (UWFCL) see WFCL other than hazardous
unequal selection probabilities 8, 13, 83–7, 203, 354, 414–5, 501, 761
unit non-response see non-response
United Kingdom, UK 467, 486, 706, 724
United States 41, 177, 466, 630–1, 715
variance 8, 14, 19, 108, 125–6, 143–5, 170–3, 191–9, 205, 288–90, 325, 757, 764–5
  in adaptive cluster sampling 387, 391–6, 404–5, 415, 419–22, 442
  in capture-recapture sampling 488–9, 517, 557
estimation 114–8, 185, 207–9, 297–308, 312
  in multiplicity sampling 340–3, 352–5, 369, 378
response/measurement variance 124–5
weights, effect on variance and deft 119, 122, 126, 236–8, 245
  see also: sampling error, design effect
Vietnam 56, 68, 722, 744
weighting 8, 83, 96–109, 116, 476, 687
  in capture-recapture sampling 15, 498, 517–9, 524–38, 546
design weights 81, 97–9, 102, 106–7, 147, 237, 244
effect on variance and deft 119, 122, 126, 236–45
estimation with weights 107–14
non-response weights 8, 83, 100–7
trimming 8, 105–7
weighting classes 100–3
worst forms of child labour (WFCL) 1, 3, 8, 19, 23–9, 32, 45, 53–5, 64–9, 72, 76, 490, 666
  defined 26–7
examples from surveys 151–8, 257–9, 721–56
WFCL other than hazardous 1, 8, 23–8, 32, 59–63, 69, 72, 79–81, 541, 721, 751–6
  defined 57–9
workload 89–90, 124, 251–2, 423, 464, 543, 756
Yemen 57, 750
Zimbabwe 486
  survey illustrations 521–2
Sampling elusive populations: Applications to studies of child labour

Department of Statistics and FPRW/IPEC
International Labour Organization
4, route des Morillons
CH-1211 Geneva 22
Switzerland
Tel.: +41 (0) 22 799 8181
Fax: +41 (0) 22 799 8771
ipec@ilo.org
www.ilo.org/ipec