Sampling for household-based surveys of child labour

Vijay VERMA

International Programme on the Elimination of Child Labour (IPEC)

International Labour Organization (ILO)

IPEC; VERMA, V.
Sampling for household-based surveys of child labour
Geneva: ILO, 2008

978-92-2-121503-5 (Print)
978-92-2-121504-2 (Web pdf)
978-92-2-121505-9 (CD-ROM)

International Labour Office

guide / child labour / survey / sample / statistical method /developing countries – 13.01.2

ILO Cataloguing in Publication Data

Note
This publication was elaborated by Vijay Verma for IPEC and coordinated by Mustafa Hakki Ozel from IPEC's Geneva Office.
Funding for this ILO publication was provided by the United States Department of Labor (Project INT/03/61/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.

Visit our website: www.ilo.org/ipec

Printed in Italy
Photocomposed by International Training Centre of the ILO, Turin, Italy

Acknowledgment

Thanks are due to a number of colleagues, in particular the following:

Mustafa Hakki Ozel for consistently pursuing and supporting this work; Farhad Mehran, Debi Mondal, Angel Rodriguez, Giulio Ghellini, Gianni Betti and the whole SIMPOC team for reviewing various parts of the document; participants at a review seminar held in May 2008 at ILO headquarters in Geneva; several other colleagues from the ILO for discussing and making suggestions for improving various drafts; and Giulia Ciampalini and Ilaria Vannini of the University of Siena for assisting in the preparation of the manual.

Vijay Verma

Siena, 30 June 2008.

Contents

Boxes

Tables

Figures

Chapter 1
Introduction

As noted in *Child labour statistics: Manual on methodology for data collection through surveys* (ILO, 2004, p.3)[1], child labour has become an important global issue. Detailed and up-to-date statistics on working children are needed to "determine the magnitude and nature of the problem, identify the factors behind child labour, reveal its consequences, and generate public awareness of the related constellation of issues".

This manual is concerned with sampling issues arising in the context of household-based child labour surveys. It does not aim to address special considerations involved in the design of surveys aimed at estimating the prevalence and nature of child labour in targeted sectors and activities, considerations often rather different from those involved in the conventional, more widely-based household surveys. For some purposes and in certain circumstances, child labour surveys may also involve non-household based data collection, and may even force some departures from the principles of probability sampling. Various statistical techniques used for sampling non-standard units need separate treatment, which may be taken up in other SIMPOC documentation. Nevertheless, many of the techniques discussed in the present manual can also be useful in the design of more specialized, targeted or sectoral surveys of child labour.

1.1 Sources of information on child labour

Data on child labour may be obtained from diverse sources, often used in combination. These may include data from national population censuses, secondary sources, existing national household surveys, and special child labour studies and surveys (*SIMPOC Manual*, p. 60).

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.

General national household surveys

Many countries collect socio-economic and demographic data through periodic household-based sample surveys. These include surveys on the labour force, living conditions, household income and expenditure, demography and health, etc. 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. Moreover, attaching child labour modules to such household-based surveys is also a potential source of information.

---

[1] Henceforth referred to as the *SIMPOC Manual*.

Secondary sources

Concerning secondary sources, the *SIMPOC Manual* notes that a wide range of institutions, while not primarily concerned with child labour, often produce useful information pertaining to it. Examples are annual school reports compiled by ministries of education, school surveys and inspection reports, statistical reports by national statistical offices, surveys and research conducted by international development organizations, and other studies and reports prepared by experts for national ministries and the donor community.

SIMPOC child labour studies and surveys

These include several different types of instruments. The *SIMPOC Manual* notes the following methods of data collection on child labour:

- household-based child labour surveys;
- rapid assessments (RAs);
- establishment surveys;
- street children surveys;
- school-based surveys;
- community-level enquiries; and
- base-line surveys and studies.

In the present Sampling Manual we are concerned with sampling issues arising in the context of household-based child labour surveys or, more generally, of surveys dealing with work-related activities of children living in private households.

It is important to note at the outset the potential and limitations of the information on child labour which can be collected through household-based surveys. However, before going into the role of such child labour surveys, we shall first summarize the features of sources of information other than household surveys. The various sources are generally complementary. These notes, paraphrased largely from the *SIMPOC Manual*, should help to clarify the potential and limitations of the household-based approaches.

1.2 Alternative sources of data on child labour

The following methods complement the standard household-based surveys.

Rapid assessments

Rapid assessments (RAs) use largely qualitative techniques that are employed to gather in-depth information on hidden or invisible worst forms of child labour (UWFCL) of different types, often concentrated in particular geographical areas[2]. The RA methodology has been developed for obtaining information on child labour and child workers in the most dangerous or unhealthy types of activities and occupations. It uses a participatory approach, conducting discussions and interviews to obtain information on working and living conditions of children involved in activities or occupations otherwise

[2] It should be clarified that the abbreviation UWFCL in this manual is used to refer to the worst forms of child labour other than hazardous work (commonly termed as "unconditional worst forms of child labour").

difficult to identify and characterize. An RA may use structured or semi-structured questionnaires, close observation, key informants or other knowledgeable persons, and background information derived from case studies and administrative reports.

The RA methodology is primarily intended to provide information relatively quickly and inexpensively, of the type needed for, for example, creating general awareness of the problems or formulating specific projects. Its output is mainly qualitative and descriptive and is usually limited to small geographic areas. Hence, this tool is generally of limited use for obtaining quantitative estimates of the prevalence of child labour. As with any kind of data collection, the value of the results depends on the quality and appropriateness of the study design. Often, the usefulness of qualitative information from an RA study can be enhanced by complementing it with a survey of households in the study area.

Establishment surveys

Most sources of information refer to the "supply side" of child labour, i.e. surveys of children supplying the labour. To collect information on the "demand side", i.e. the beneficiaries of child labour, interviews of employers and the conduct of other establishment- or workplace-based surveys are required.

Establishment survey questionnaires are normally administered in the workplace – whether at a factory or at 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 be compared with parallel data on adult workers. Information can also be sought regarding employers' perception of the reasons for using such labour, the advantages and drawbacks of using child workers, the recruitment methods used, etc.

An establishment survey may be based on a sample selected from a frame of establishments likely to be employing children, in which case it aims to produce estimates for the population of such establishments. More often, such surveys involve interviewing establishments selected on the basis of their link with working children identified through a household-based child labour survey. In this latter case the survey of establishments provides additional variables on children with whom the establishments are linked. It does not necessarily yield in itself a representative sample of the total population of establishments employing children.

Street children surveys

Children living on the street or in similar dire situations are among the most vulnerable child labour groups. They cannot be covered through normal household surveys which exclude homeless persons. Most of these children remain on the move from one place to another during daytime, and sleep outside buildings at night. This makes it difficult to survey them with normal sampling procedures. Specially designed street children surveys are needed to collect the relevant information. A *purposive* or *convenience* sampling approach is often needed, both in selecting the areas to be covered, and in conducting random interviews with children regarding their working conditions and interviews with employers in the informal sector employing children.

School-based surveys

School-based surveys are required for collecting statistics on how child work affects school attendance and a child's educational performance and attitudes towards schooling. Normally, school-based surveys are applied to respondents identified as working children through household surveys. These surveys aim primarily at determining the impact of work on school attendance and the performance of children enrolled at school, and often also at assessing attitudes of working children towards studying. In addition to working children, this approach also normally covers non-working children as a control group, preferably from the same schools as attended by the surveyed working children. Interviews are conducted with children, teachers, school management and parents. The survey may also attempt to assess school-related factors, such as the quality of available education, which may influence the likelihood of children engaging in economic activity[3].

Community-level inquiries

Community-level inquiries usually collect information from administrators and other community leaders to obtain a cultural, demographic and socio-economic profile of the community and other characteristics which may be related to child labour in the area. They can be useful for identifying the main variables directly or indirectly related to the presence of child labour in the community. Often, information is also collected in the area regarding 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, recreational opportunities). As noted in the *SIMPOC Manual*, community-level inquiries may "well be conducted as independent investigations collecting data on particular child labour situations … however, community-level inquiries are also often an integral part of the methodology applied in RAs [Rapid Assessments] and base-line surveys".

Baseline surveys and studies (BLSs)

Baseline surveys and studies are useful for identifying target populations and their characteristics, and analysing the determinants and consequences of child labour in particular socio-economic sectors. BLSs may involve one or more of the available methods of data collection, such as a combination of sample survey and qualitative (participatory) research techniques. For instance, such studies may use a combination of qualitative and quantitative techniques to gather data on conditions existing before an intervention or a project, mostly as a planning and implementation tool for projects concerned with improving targeting of interventions. They may also be used for monitoring project progress and assessing its impact in particular sectors or areas. For programme intervention purposes, base-line surveys may collect data on initial (baseline) conditions for use at each stage of the programme cycle: design, implementation, monitoring, and impact assessment. Information provided by base-line surveys helps in establishing targets that allow changes to be measured by means of follow-up studies and in developing monitoring systems.

[3] In the labour force framework, the concept of *economic activity* includes seeking work. However, in the context of child labour, seeking work is not included. It is in this sense that the term economic activity is used througout this manual in relation to children.

1.3 Complementary nature of the sources

While appreciating the limitations and strengths of the information on child labour which can be collected through household-based surveys, it is important to note that the various sources of information on child labour are generally complementary. The practical implication from the standpoint of survey and sampling design is that the household-based instrument does not have to be developed to meet all the information needs: in fact some of these needs are better (or sometimes can only be) met by other types of instruments. As noted in the *SIMPOC Manual* (p. 66), "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 the complexity of 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 with the unconditional worst forms of child labour (UWFCL). With UWFCL it is very difficult, if not impossible, to make contact with the child to collect the necessary information. As the unconditional worst forms of child labour 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 unconditional worst forms of child labour. Consequently, regular household-based surveys are largely ruled out for this purpose; special sampling and enumeration procedures must be employed.

1.4 Household-based surveys

Regular child labour surveys are household-based national sample surveys whose target is children, along with their parents/guardians living in the same household. Such surveys may be conducted as stand-alone surveys, or as separate but linked operations, simply or as modules attached to other national household-based surveys such as a labour force survey (LFS). The statistics generated by these surveys 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 *SIMPOC Manual* (p. 63-74) makes the following points in relation to household-based surveys of child labour.

- The household is generally the most appropriate unit for identifying children and their families, measuring their socio-economic and demographic characteristics and housing conditions, and obtaining information regarding the circumstances that force children to work and the conditions of work for household-based activities.
- Household surveys based on probability sampling provide an efficient approach for estimating the prevalence of particular forms of child labour, with the exception of certain special child labour categories such as children living on the street or those engaged in hidden forms of child labour.
- The main limitation is that household surveys are generally not suitable for collecting information from children engaged in the unconditional worst forms of child labour.
- In so far as household surveys are based on scientifically designed samples, they permit the drawing conlusions from the whole population that was sampled from the study results. For this purpose, a child labour survey sample should be selected according to established principles: it must be representative of the entire population, and one must be able to produce estimates of population parameters within known and acceptable margins of error.
- In practice, the household survey content may be detailed and specialized, 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 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. The questionnaires hitherto used in these surveys have generally been designed to obtain information regarding the magnitude, character and conditions of child labour.
- These questionnaires also collect information on working conditions, industrial activity, occupation, employment status, and effects on children's health, educational achievement and opportunities for normal development. Many child labour-related subjects can be incorporated into the survey questionnaires, including demographic and socio-economic characteristics, housing conditions, work-related characteristics of children and their families, the factors that lead children to work, and the attitudes of parents/guardians towards children's work and schooling.
- Household surveys may apply a variety of designs and organizational structures. The main factors determining design are, of course, substantive objectives concerning the content, complexity and periodicity of the information sought. These substantive requirements determine the timing, frequency, reference period, sampling arrangement and other features of the survey structure. For instance, 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 child labour surveys supported by SIMPOC have been occasional or one-time surveys.

1.5 Contents of the manual

Household-based child labour surveys can be of various types, involving different types of survey structures and designs to reflect different objectives and circumstances. Sample design of a household-based child labour survey begins with the choice of an appropriate survey structure. The survey structure refers to the manner in which different components of the survey have been arranged in relation to each other and, as relevant, possibly also in relation to other surveys.

Chapter 2 presents a typology of survey structures and discusses factors underlying the choice of an appropriate one. Features of a large number of child labour surveys in developing countries are tabulated, and the forms of linkage encountered between "base surveys" and surveys aimed at estimating prevalence of child labour, and between the latter and more detailed surveys on conditions of working children, are analyzed. Noting that a fundamental prerequisite for a good sample design is that "the sampler must understand the content of the survey for which the sample is to be designed, the objectives and methodology of the survey and the conditions of its implementation", the aim of this chapter is to provide a cross-section of these requirements from actual practice as they relate to surveys of child labour.

Chapter 3 seeks to clarify the basic sampling principles and procedures involved in the most common types of designs used for household surveys of the general population, to which a child labour survey may be attached as a module or linked in some other form. Of course these are also directly relevant to the design of stand-alone child labour surveys. This description of the basic sampling principles and procedures is important for gaining an understanding of how different arrangements for the measurement of child labour may be linked to or derived from other household surveys. The chapter deals with a number of important practical aspects concerning sampling frames, sample design and selection methods, and looks in some detail at the complex issue of sample size. Surveys aimed at estimating the prevalence of child labour tend to have sampling requirement similar to those of surveys of the general population, such as labour force surveys (LFS). This is because the base population for the two types of surveys (all children for the child labour survey, the adult population for the labour force survey) tend to be similarly distributed.

Chapter 4 describes the technical details of procedures for drawing the sample for a child labour survey (CLS) from the sample used for a larger survey of the general population such as the LFS. To obtain detailed information on the conditions and consequences of child labour, the base population of interest is the population of working children, rather than the population of all children. Surveys based on this population are known as labouring children surveys (LCS). The LCS sampling requirements differ from those of a survey of the general population such as the LFS, and likewise from those of a survey of the general population of children such as the CLS. This is because of differences in the distribution of the respective base populations between the LCS and these other types of surveys. The distinction between the two types of surveys – CLS and LCS in the terminology of this manual – is a fundamental one. This is not to imply that CLS and LCS must be two distinct surveys. They can, and indeed often do, appear as components of a single survey. However, even when these two types of

survey have to be integrated into a single operation, the distinction remains conceptually important as it helps in choosing the best compromise design for an integrated operation.

Chapter 5 discusses sample design issues for surveys where the base population of interest is the population of labouring children, and it provides technical details of procedures for drawing an LCS sample from a larger survey of the general population, such as an LFS or CLS. The basic requirement is that an intensive survey aimed at investigating details of circumstances and conditions of working children should have a sample structure reflecting the patterns of distribution of working children in the general population.

Chapter 6 aims at elucidating the process of developing sample design and selection procedures in a real situation. In this chapter a very realistic illustration developed in the context of an actual survey in a developing country is presented, with many details and numerical examples to clarify the various sample design and selection procedures.. The survey structure developed and presented in this chapter is somewhat different from the more common one underlying the preceding two chapters. There is no fixed design or structure for surveys of child labour. There are various data requirements which need to be arranged and accommodated in the most efficient way possible.

Finally, **Chapter 7** addresses selected issues relating to estimation from sample survey data from a practical perspective. It discusses issues and statistical procedures relating to the weighting of sample data and the production of estimates from the surveys. It then reviews practical procedures for the computation and analysis of information on sampling errors, and concludes by drawing some lessons concerning the choice of the sample size and structure from a practical perspective.

Chapter 2
Choosing an appropriate survey structure

Household-based child labour surveys can be of various types, involving different types of survey structures and designs reflecting different objectives and circumstances.

In designing the sample for a household-based child labour survey, the first step is to choose an appropriate survey structure. A child labour survey needs to collect different types of information, from estimating the prevalence of child labour to studying in depth the consequences of the labour on working children. The types of data differ in their content, mode of collection, and – especially in the case of sampling – in the base population to which they relate.

The different types of data involved may be viewed as constituting different "components" of the survey. By "survey structure" we mean the manner in which the various components of the survey have been arranged in relation to each other. For instance, the components may be combined into a single integrated whole; or they may remain distinct but linked in various ways; or they may form more or less separate "stand-alone" operations, each like a survey in itself. Similarly, the child labour survey or its components may be related in various ways to other existing surveys, such as the labour force survey.

There is, of course, a great deal in common between child labour surveys of different types; but there can also be major differences between them, and it is very useful to identify and classify these differences.

A clear typology of surveys can help in identifying the type of survey structure and design needed to meet the given objectives and the practical constraints in undertaking a survey. Many examples can be found from national experiences of child labour surveys where a lack of understanding of the type of survey required resulted in choosing an inappropriate structure, design or sample size.

A clear terminology is also a useful tool in that it provides a framework for describing and documenting the survey methodology and design clearly and in detail, and for discussing the relative merits and shortcomings of different choices.

In this chapter a large number of examples are provided to illustrate various aspects of the structure and relationship of surveys of child labour to each other and to surveys on other topics. As noted in the introduction, a fundamental prerequisite for a good sample design is that the sampler understands the content of the survey for which the sample is to be designed, the objectives and methodology of the survey and the conditions of its implementation. Illustrating these issues for surveys on child labour from actual practice is the aim of this chapter.

Mostly of the examples given here are taken from national reports on child labour surveys and are quoted without editing. Comments bringing out points of particular

interest from these illustrations appear in the main body of text. For ease of reference, the boxes included are listed below:

In addition, at the end of this chapter (Sections 2.8 and 2.9) more detailed reviews are provided with commentary on design aspects of two surveys concerning child labour: the Survey of Activities of Young People in South Africa (1999), and the Jamaica Youth Activity Survey (2002). The objective is to use these examples to illustrate and critically discuss a number of important points relating to survey structure and sampling.

In Tables 2.1 and 2.2 which appear at the end of Section 2.1, we have tried, using the terminology described in this chapter, to classify the survey structures used in around 30 child labour surveys conducted during the past few years. At various places in the chapter, we shall comment on this diversity of survey structures in order to bring out important sample design issues.

2.1 A fundamental distinction: Child labour survey (CLS) versus labouring children survey (LCS)

2.1.1 Child Labour Survey (CLS)

As a generic term we use "a (regular household-based) child labour survey" to indicate a household sample survey whose main objective 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 prefix "regular" is used, when required, to emphasize that the context is that of a broad household-based survey, as distinct from other types of studies concerning children not residing in – or at least not identified for enumeration through – private households.

While retaining the above more general, descriptive use of the term "child labour survey", it is very useful to keep in mind two different types of such surveys. These differ in their objectives, or at least in the emphasis given to different types of objectives.

The first type 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 area (urban-rural), household type and characteristics, the household's employment and income situation, the children's age and gender, and similar factors.

The target population of a survey with this type of objective 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 among the general population. The size and structure of the sample are 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.

We propose to limit our use of the term "child labour survey" (CLS) to surveys whose primary objective is thus to measure the prevalence of child labour, as distinguished from the "labouring children survey" (LCS) described below. This distinction was first introduced in the *SIMPOC Manual* (pp. 176, 188). The defining factor in this distinction is the base population for which the survey estimates are generated – essentially, all children within specified age limits for the CLS, and only those considered to be engaged in child labour for the LCS.

2.1.2 Labouring Children Survey (LCS)

We have a different type of survey when the primary objective is to investigate the circumstances, characteristics and consequences of child labour: what types 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. We refer to this type of survey as a "labouring children survey" (LCS).

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 should be determined largely by the size and distribution of the population of working children.

At the same time, it is important to clarify that the concept of a "labouring children survey" does not imply that the ultimate units enumerated in the survey are only labouring children. On the contrary, it will normally be 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 those subject to child labour. Nevertheless, the sample size and design of a LCS are determined primarily by the need to represent the population of labouring children; any sample of non-labouring children is supplementary, selected and added to the main sample as necessary for analytical purposes.

2.1.3 Sampling implications

There are some important differences in terms of sampling aspects between child labour surveys and labouring children surveys in the above sense.

The target population of the CLS, being all children in a certain age bracket, tends to be distributed in very much like the general population. Thus, the required structure and distribution of the CLS sample are likely to be quite similar to that of a survey of the general population, in particular to that of the labour force survey (LFS), which the CLS very closely resembles in concept, definitions and even survey content.

The target population of the LCS – whether it is the population of children engaged in any work-related activity or defined more narrowly as those engaged in specific forms of child labour – is, by comparison, smaller and more unevenly distributed, often being concentrated in specific areas. Consequently, the sample design required is also generally different from that of an LFS or a CLS.

The two types of survey differ in their size and complexity. The CLS is normally less intensive (that is, it involves a simpler and shorter survey interview) and requires larger samples. 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 estimates.

By contrast, for investigating the detailed conditions and consequences of child labour, the LCS entails more intensive data collection, often involving separate interviews with the guardians as well as the children concerned, the collection of attitudinal and other qualitative data, and the conduct of associated enquiries, for example at the school or place of work of the 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 are in any case 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 (see Section 3.5 on the choice of sample size).

The above does not imply that CLS and LCS must always be two separate surveys. On the contrary, as noted in Section 2.2 below, the two components (CLS and LCS) identified here are, in practice, often operationally integrated into a single survey. This does not of course obviate the need to consider the sample size and design requirements of each component in its own right: the common design of the integrated survey should be an appropriate compromise between those requirements.

2.1.4 Diverse target populations: An illustration

In practice, the target populations of interest can be more diverse than the "all children versus labouring children" distinction of CLS and LCS. The various objectives of a survey on child labour, and the various population groups to which the results from the survey apply, are well illustrated by the following description from Portugal (1998). The survey identifies seven different target groups of questions (see Box 2.1 below for details).

For three of these groups, namely:

group 1: families and the activities of children,

group 2: children and activities in general

group 7: aspects of the lives of children,

the target population is all children (and their families). The questionnaire modules pertaining to these constitute the "CLS" component of the survey. Group 1 is the basis for estimating the proportion of children engaged in child labour.

For another three groups, namely

group 3: children engaged in an economic activity,

group 5: characterization of those responsible for children engaged in an economic activity,

group 6: attitudes and perceptions towards child labour of the child, and of adults responsible for the child.

the target population is labouring children, defined here as children engaged in any economic activity. The questionnaire modules pertaining to these constitute the "LCS" component of the survey. Group 3 – children engaged in an economic activity – is the basis; the population of persons responsible for children in groups 5 and 6 is defined through association with the base population in group 3.

Group 4: children who carry out domestic chores – covers a somewhat different group of children and is often considered less important than group 3 in determining the LCS sample size requirements.[4]

[4] There is considerable debate as to the definition of what constitutes "child labour", and in particular whether "substantial" domestic chores should be included. See *SIMPOC Manual*, Chapter 2, and also Box 2.6 below.

The final sample size is normally a compromise between the requirements of the "CLS" and "LCS" components. It is of course possible, in principle, to introduce sub-sampling from group 1 to group 3 (i.e. to follow up only a sub-sample of the labouring children identified from group 1); or to introduce sub-sampling from group 3 to groups 5-6 (i.e. to follow up only a sub-sample of adults responsible for the labouring children). There is greater flexibility in this respect when the various components involved are operationally separable from each other.

Box 2.1: Survey objectives and target populations – Portugal 1998

Survey objective

Among the range of inquiries already adopted by the ILO to measure child labour which were mentioned previously, three types of operation were proposed for Portugal: a family questionnaire; a school questionnaire (teachers and pupils); the collection of ongoing information from a group of personalities to be defined (key informers). The general objective of the family questionnaire was to quantify the work of children in Portugal at a global and regional level and to characterize the factors that determine and explain the phenomenon. This general objective can be split into two specific objectives:

- to quantify the work of children in the light of ILO recommendations, Portuguese legislation, the concept of work in an economic sense and in a wider sense, and defining work with reference to time in the preceding year and also to the time when the questionnaire was carried out;
- to characterize the work of children in terms of factors that explain its existence, of determining factors in the social and family context and of the attitude of families vis-à-vis children's work .....

Survey questionnaire

Following the model of related questionnaires carried out by the ILO, the questionnaire was divided into two modules:

1. the first, to be answered by the representative of the family group, contains three types of question: about the family in general (e.g. habitation, furnishings, family income, etc.), about each of the members of the family group (sex, age, qualification, working position), and about children within the parameters of this study (relation between those responsible and the children, activities of children, etc.).

2. the second, to be answered by the children themselves; these questions covered aspects of characterization, schooling, relation to school, free-time occupation; habits and in the case of those involved in activities, a second type of questioning concerned the character of the activity, its relation to work, schedules, and the feelings of the child in relation to the activity. ...

Target population groups

The target population chosen was of children between 6 years, the age when children normally start primary school, and 15, when children under normal conditions finish obligatory schooling. Fifteen years of age is also the legal limit for entry into the world of work, except in certain exceptional circumstances and within certain limits. …

Among the objectives sought in the drafting of the questionnaire was that it should provide a detailed response to various groups of questions.

Group 1: Families and the activities of children

This group of questions is designed to establish in how many families there is child labour, what its socio-economic relevance is to the lives and activities of children, and whether there is a simple relationship between the standard of living, child labour and other activities of children or whether other aspects intervene such as regional, social and cultural factors.

Group 2: Children and activities in general

While not encompassing economic activity in the strict sense, the questionnaire tried to find out about the occupations of children in a broader sense - namely in the area of school, home, work in the economic sense, domestic work and free time - and to understand the relation between the different activities of children and sex, age, schooling and the regions where they live. While the main reference point of the questionnaire was the week preceding questioning, it also sought to collect data relating to holiday-time activities and the preceding year.

Group 3: Children engaged in an economic activity

Children engaged in an economic activity make up the group of child labour in terms of the questionnaire. Firstly, however, two important subgroups have to be distinguished to characterize child labour: children who work in family-run economic units and those who work for a third party outside the family. Within this distinction a set of characteristics was sought, such as regions, sex, age, schooling and the way they relate to work. Another objective within this group of questions was to characterize child labour in terms of activity sectors, types of economic unit where children work, reasons for carrying out these activities, schedules and pay (for those working for a third party), while also considering some related data concerning sickness and accidents caused by child labour.

Group 4: Children who carry out domestic chores

Children who carry out domestic chores are not covered by child labour. However, when this type of activity is outside certain limits it too can prejudice schooling and other aspects previously referred to. Children with domestic activities are therefore quantified and characterized in a similar way to those engaged in an economic activity. The length and type of these occupations and the way in which they relate to other aspects like schooling and free time are also a line of research of this questionnaire.

Group 5: Characterization of those responsible for children engaged in an economic activity

This group of questions seeks to relate the parents (or those responsible) to the economic activity of children: what age they started work, what level of schooling they reached, their professional situation and level of remuneration.

Group 6: Attitudes and perceptions of those responsible and children towards child labour

With regard to children who work, an attempt is made to establish how parents feel about and justify their children working and how the children themselves accept and approve of their situation.

Group 7: Aspects of the lives of children

The questionnaire was also projected to find out about other aspects of the lives of children both related and not related to their activities. Free time and holiday occupations, relations with other children and family heads, the time they normally went to bed and got up, the number of hours spent watching television, whether or not they received pocket money and how much are some of the questions aimed at finding out about certain aspects of children's lives that are dealt with in this study.

Source: *Child Labour in Portugal: Social Characterisation of School Age Children and Their Families*, 1998. Portugal Ministry of Labour and Solidarity (MTS).

2.1.5 Examples of various structures of surveys of child labour

Tables 2.1 and 2.2 provide some essential information on the samples and the structure of linkages for around 30 surveys on child labour. The tables show the diversity of the survey structure encountered. The structural concepts are explained in more detail in other sections of this chapter.

Linkage of CLS to a base survey

Table 2.1 classifies the type of linkage between the CLS and its base survey, if any. If the CLS is not linked to any base survey, then it is classified as a "stand-alone" survey.

As summarized below, the two "extreme" forms of arrangements have dominated in past surveys. Half the surveys (15 of the 29 reviewed) are *stand-alone,* meaning that the survey is exclusively or primarily concerned with child labour, so that the survey content, design and procedures can be chosen independently (Section 2.2.2).

The other common arrangement is the collection of the base survey and CLS information during the same operation (this applies to 13 of the 29 surveys reviewed). Here we distinguish between *modular surveys* of child labour, which are essentially added on to an existing base survey interview as a module (Section 2.2.1), and the more substantial *combined surveys,* in which the child labour part more clearly affects the design and operations of the base survey (Section 2.2.3). The LFS forms the base for most of the

combined surveys. This is also the case for several modular CLSs, but a variety of other types of survey have also served as the base, as shown in Table 2.1.

Another possibility, but a rare one, is to have a *linked survey*, where the interview is operationally separated from that of the base survey but the sample and some substantive information is fed-forward from the base survey (Section 2.2.3).

Forms of base-CLS linkage: Summary

Modular (8)	Azerbaijan (2006), Cambodia (2006), Honduras (2002), Nepal (1996), Nicaragua (2000), Uganda (2000-2001), United Republic of Tanzania (2000-2001), Zambia (2000) .
Combined (5)	Costa Rica (2003), Kenya (1998-99), Turkey (1994), Turkey (1999), Ukraine (1999).
Linked (1)	Zimbabwe (1999).
Stand alone (15)	Bangladesh (2002-2003), Belize (2001), Cambodia (2001), Dominican Republic (2002), Ethiopia (2001), Ghana (2001), Mongolia (2002-03), Namibia (1999), Nigeria (1999), Pakistan (1996), Panama (2000), Philippines (2001), Portugal (1998), South Africa (1999), Sri Lanka (1999).

Table 2.1 also shows that a CLS is usually a single-round, one-time survey, though in one out of six cases it has involved multiple rounds - typically four rounds corresponding to the quarters of the year. Presumably this is to capture seasonal variations. Note that unlike the LFS, which generally uses rotational samples, the CLS rounds are likely to be based on independent or non-overlapping samples. This permits more efficient cumulation of the data from one round to the next.

A multi-round CLS is not an affordable option in most circumstances and so is not discussed further in this manual.

Other information shown in the table concerns the sample size and its division into the number of clusters and sample-takes per cluster. The sample size varies greatly, from 6,000 to 48,000 in the cases reviewed. The one exception is where a brief CLS module was combined with a very large household listing operation (Pakistan 1996, Section 2.3).

The sample-take per cluster is even more variable – in the range of 5 to 50 (with the same single exception as above). To a large extent, this variation may simply be the result of lack of information on design effects and cost parameters, or the failure to analyze and use such information in the design (Sections 3.4.3, 7.7, 7.8). Information on the number of clusters and sample-takes per cluster is not even available in some survey reports.

Linkage between CLS and LCS

Table 2.2 classifies the type of linkage between CLS and LCS components. It also indicates the substantive scope of the LCS - mainly, whether it concerns primarily child labour (i.e. working children) or is a more general survey of children or child activities.

By far the predominant form has been an *integrated* CLS-LCS operation. By this is meant that the information on prevalence of child labour (the CLS component) and more detailed information on children who are found to be working are both collected in a "single interview operation" in the sense explained in Section 2.4. Generally (or perhaps in all cases so far) it has meant no sub-sampling between the two components.

Undoubtedly, it can be more convenient and faster to cover both components in a single operation. It may also be seen as more economical. But unfortunately this is not necessarily the case. The CLS and LCS have quite different sample design and sample size requirements, as discussed in Chapters 4 and 5. An integrated design with no sub-sampling in between may often mean a sample too small for the CLS and/or too inefficient yet too large for the LCS component. Too large a sample for the LCS is very likely to have a negative impact on the quality of the complex data the survey is meant to collect.

An alternative is to have *linked* surveys which permit a degree of operational separation between CLS and LCS (Section 2.5), though this arrangement has hitherto been used in only six of the 31 surveys reviewed in Table 2.2. There is a wide variety of forms of linkages possible (Section 2.5 lists 13 possibilities). In practice most cases in Table 2.2 have involved taking for the LCS all or a sub-sample of households containing a working child as identified during the CLS.

Concerning the substantive scope of the LCS (whether integrated or linked to the CLS), a majority are concerned primarily with the *economic activity* of working children; only a little less than one-half are broader in scope. There are a number of *child activity* surveys covering all types of activities of children, including non-economic activities; then there are even broader *children's surveys* covering in addition other areas such as children's health, housing and living conditions (Section 2.6). This broader scope requires the coverage of a broader population – essentially, all children. Therefore in structure the "LCS sample" in this case should resemble that for the CLS, but still a substantial difference in sample size would normally be desirable because of the different objectives and complexity of the two components. The sub-sampling procedures of the type discussed in Chapter 4 may be more appropriate for the LCS with a wider scope in this sense. Those discussed in Chapter 5 are more geared to the "conventional" LCS focused on the economic activity of children.

Table 2.1. Child labour surveys: Base-to-CLS - Examples of various structures

Survey	Year	(1) Base-to-CLS	(2) Whether multi-round	(3) Sample size and clustering n=a*b	a	b
Azerbaijan	2006	Modular (LFS)		17,000	850	20
Portugal	1998	Stand-alone		25,000	1,150	22
Turkey	1994	Combined (LFS)		13,500		?
Turkey	1999	Combined (LFS)		20,000		?
Ukraine	1999	Combined (LFS)	4 rounds	48,000		?
Bangladesh	2002-2003	Stand-alone		40,000	1,000	40
Cambodia	1996	Modular (Socio- economic Survey)	2 rounds	9,000	750	12
Cambodia	2001	Stand-alone		12,000	6,000	20
Mangolia	2002-2003	Stand-alone (same sample as as LFS)	4 rounds	12,000	1,200	10
Nepal	1996	Modular (Migration and Employment Survey)		20,000	600	33
Pakistan	1996	Stand-alone *only a listing survey to identify target households*		140,000	1,860	75
Philippines	2001	Stand-alone		27,000	2,250	12
Sri Lanka	1999	Stand-alone	4 rounds	15,000	1,000	15
Ethiopia	2001	Stand-alone		44,000	1,250	35
Ghana	2001	Stand-alone		10,000	500	20
Kenya	1998-1999	Combined: LFS, Informal sector survey, child labour survey		13,000	1,100	12
Namibia	1999	Stand-alone *only a listing survey to identify target households*		8,000	270	30
Nigeria	1999	Stand-alone		22,000	2,200	10
South Africa	1999	Stand-alone		26,000	900	30
Tanzania	2000-2001	Modular (LFS)	4 rounds	11,000	220	50
Uganda	2000-2001	Modular (Demographic and Health Survey)		8,000	300	27
Zambia	2000	Modular (Multiple Indicator Survey)		8,000	360	22
Zimbabwe	1999	Linked (Indicator Monitoring (IM)-LFS) *IM-LFS provids lists of children in all sample households*		14,000	400	35
Belize	2001	Stand-alone		6,000	200	30
Costa Rica	2003	Combined (Multipurpose Household Survey)		11,000	?	?
Dominican Rep.	2000	Stand-alone		8,000	800	10
Honduras	2002	Modular (Permanent Multipurpose Survey)		9,000	1,800	5
Nicaragua	2000	Modular (ad-hoc LFS)		8,500	1,700	5
Panama	2000	Stand-alone		15,000	1,500	10
Georgia, Romania: similar to Ukraine						

Source: Compiled from national reports on surveys of child labour.

Table 2.2. Child labour surveys: CLS-to-LCS - Examples of various structures

Survey	Year	1. Base-to-CLS	2. CLS-to-LCS	3. Sub-samling	4. LCS scope
Azerbaijan	2006	Modular	Linked	Sub-sample of areas and hhs with working children n=4.000 a=400 b=10 + special design for refugee children	Child labour (CL)
Portugal	1998	Stand-alone	Integrated		Chidren survey
Turkey	1994	Combined	Integrated		CL
Turkey	1999	Combined	Integrated		CL
Ukraine	1999	Combined	Integrated		Chidren survey
Bangladesh	2002-2003	Stand-alone	Integrated		CL'+employers' questionnaire
Cambodia	1996	Modular	Integrated		Children survey' + employers' questionnaire
Cambodia	2001	Stand-alone	Integrated		Chidren survey
Mangolia	2002-2003	Stand-alone	Integrated		Child activity survey
Nepal	1996	Modular	Integrated		CL
Pakistan	1996	Stand-alone	Linked	All households with labouring children (no subsampling) n=10.500 a=1.400 b=8	CL
Philippines	2001	Stand-alone	Linked	All households with labouring children (no subsampling)	CL
Sri Lanka	1999	Stand-alone	Integrated		Child activity survey
Ethiopia	2001	Stand-alone	Integrated		CL + schooling
Ghana	2001	Stand-alone	Integrated		Child activity survey
Kenya	1998-1999	Modular	Integrated		CL
Namibia	1999	Stand-alone	Linked	All households with labouring children (no subsampling)	CL
Nigeria	1999	Stand-alone	Integrated		Child activity survey'+ street children survey
South Africa	1999	Stand-alone	Linked	Sub-sample of hhs with a labouring child (all areas taken)	CL
Tanzania	2000-2001	Modular	Integrated		CL
Uganda	2000-2001	Modular	Integrated		CL
Zambia	2000	Modular	Integrated		CL
Zimbabwe	1999	Linked (IM-LFS)	Integrated		CL
Belize	2001	Stand-alone	Integrated		Chidren survey
Costa Rica	2003	Combined	integrated		Child activity survey
Dominican Rep.	2000	Stand-alone	Integrated		Child activity survey
Honduras	2002	Modular	Linked	2 in 5 subsample of household from all sample areas n=3.600 a=1800 b=2	CL
Nicaragua	2000	Modular	Integrated		CL
Panama	2000	Stand-alone	Integrated		Chidren survey
Georgia, Romania: similar to Ukraine					

Source: Compiled from national reports on surveys of child labour.

2.2 Relationship of a CLS to a base survey

Child labour survey structures can vary according to their context. The surveys may differ in how and to what extent they are linked to surveys on other topics.

In most countries regular surveys of child labour are not established, and they have for the most part been undertaken as one-time or occasional operations. There are two main forms in which these surveys have been organized:

1. modular surveys; and
2. stand-alone child labour surveys.

2.2.1 Modular surveys

The collection of information on child labour in conjunction with a broad-based survey of the general population such as a labour force survey (LFS) very often takes the form of child labour questions attached to the LFS as a module.

The essential CLS information, namely estimates of the proportion of children in various categories who are engaged in child labour, may be obtained by extending downwards the lower age limit for the standard LFS questions on economic activity. This is a possibility when child labour is defined in terms of the standard LFS concept of economic activity. However, different and generally more elaborate questioning will be required, with a different interpretation of what is meant by "child labour".

The major attraction of modular child labour surveys is that they provide an economical and convenient arrangement for obtaining essential information on child labour. Modular child labour surveys also allow the use of items enumerated in the base survey as explanatory and classification variables in the analysis of child labour data.

But, as noted in the *SIMPOC Manual*, modular surveys present two potential problems:

- the number and detail of child labour items that may reasonably be inserted into operations primarily concerned with other topics are quite limited;
- to ensure high-quality data, the various survey topics must be compatible in terms of concepts, definitions, survey methods, reference periods, coverage, and design requirements. This compatibility requirement may demand compromises that subsequently limit the usefulness of the resulting data on child labour.

2.2.2 Stand-alone child labour surveys

Stand-alone surveys are exclusively or primarily concerned with child labour topics. Hence a stand-alone survey is characterized by its single-subject focus, and by a considerable degree of separation from other surveys in its design and execution. Such separation often provides better control and greater flexibility in the design and operations of the child labour survey. The survey content, design and procedures can be chosen to meet the survey's substantive requirements more precisely.

While clearer focus on the topic of child labour and flexibility in the design and operations are the strong points of stand-alone surveys, their major drawback is their

high cost. It is difficult to sustain stand-alone surveys on a regular basis to serve as a source for obtaining regular information on child labour.

Normally, a stand-alone survey follows a listing operation from which the sample of households or addresses for the survey is selected. The single-subject focus of stand-alone surveys, however, does not preclude operational co-ordination and the sharing of facilities and arrangements with other surveys, or the use of common coverage, concepts, definitions, and classifications. In practice, the notion of a stand-alone survey is actually a matter of degree, and a sample survey is hardly ever designed and implemented in complete isolation from other contemporary surveys.

2.2.3 Other arrangements: Combined and linked surveys

In Table 2.1 most of the child labour surveys have been classified as being stand-alone or modular. In a few cases, we have identified sub-types which specify the particular survey structure more precisely. These are "combined" surveys and "linked" surveys.

Combined surveys

These are a more comprehensive version of the modular survey. When a set of child labour questions is said to have been attached as a "module" to an existing base survey, it is generally understood that the number of additional questions involved is small enough not to affect the base survey data collection significantly. In any case, the sample design is already established, or is determined, to reflect almost exclusively the requirements of the base survey.

We use the term "combined survey" to indicate a situation where the child labour questions constitute a substantial addition to the existing survey that influences its data collection operations, and/or where the survey design and sample size are influenced by the requirements of the child labour component.

Such was the case, for instance, with the two child labour surveys in Turkey, and the surveys in Ukraine, Kenya and Costa Rica. Boxes 2.2 and 2.3 give details of two of these surveys, Costa Rica 2003 and Turkey 1994.

In the "combined" survey in Costa Rica, the child labour component was allowed to influence the overall sample design: "Given the country's interest in obtaining data on children's work through special modules, the national sample was distributed by area (urban and rural) according to the variability of the economic participation rate of the population aged 5 to 17. This rate was estimated, based on the 1995 Child Labour Module". (Survey report, Costa Rica 2003).

In the case of the combined (LFS+CLS) survey in Turkey, the sample design was entirely pre-established for the household labour force survey (HLFS). The child labour survey involved "sub-sampling", which was achieved simply by applying the child labour component during only one of the two LFS rounds in the year. Nevertheless, the substantial number of additional child labour questions involved meant that the resulting survey is better described as a "combined (HLFS+CLS) operation". As noted in Box 2.3: "The simultaneous application of the HLFS and CLS meant that the respondents were required to answer a large number of questions. These lengthy

questionnaires often annoyed the respondents who became rather reluctant to answer the questions toward the end of the survey". (Survey report, Turkey 1994.)

Linked surveys

A linked CLS means a survey that is dependent on the base LFS survey for its sample, and possibly other information fed forward, but is otherwise operationally separated from the latter for the purpose of data collection.

Such a variant is represented, for example, by the survey in Zimbabwe. The base survey is the indicator monitoring labour force survey (IM-LFS). Exceptionally, this survey provides not only lists of the addresses and names of household heads in the sample households, but also *the names of the children age 5–17 years in each household* for the subsequent child labour survey. The CLS is therefore not a "stand-alone" survey. However, it is operationally separated from the IM-LFS, and is therefore not a module attached to that survey in the usual sense. We have termed this arrangement a "linked survey" to distinguish it from the three arrangements described above. (Box 2.4, taken from the survey report for Zimbabwe 1999, illustrates the wide variety of arrangements which are possible for a child labour survey.)

A linked survey permits more detailed and accurate measurement of child labour than is normally possible in a modular survey, but of course there is the extra cost of the separate operation involved.

Further comments

A few further remarks on the above illustrations will be useful. The example from Costa Rica is quite exceptional in that usually child labour modules have to accept the sample size and design of the existing base survey to which they are attached. The sampling requirements of the child labour questions have to be met primarily by varying the sub-sampling procedures and rates applied to the existing sample of the base survey.

In the illustration from Turkey, the CLS is applied during only one of the two annual rounds. This automatically provides a 50 per cent sub-sampling of the LFS interviews to be followed up in the CLS.

The important point already noted is that, generally speaking, national labour force surveys are relatively large, established and continuing, while child labour surveys tend to be smaller, occasional or one-time operations. Increasingly, labour force surveys are being carried out more than once a year – twice a year in the above example from Turkey, or even quarterly or monthly in some more developed countries. This increased frequency is motivated by the need both to have more up-to-date data and to monitor short-term and seasonal variations in the labour force.

Of course, capturing seasonal variations is important in relation to both adult economic activity (LFS) and the labour of children (CLS). Yet data in these two topics are not important to the same degree, and have their own cost and other practical constraints. Even basic data on child labour are scarce and much needed and, relatively speaking, fewer resources are available to improve the situation. Therefore, it is often wiser to give more weight to obtaining a reliable "fix" on the basic child labour situation in the country (including its geographical and sectoral variations) than, at the present stage, to devote increased resources to monitoring its seasonal variation.

There can be an enormous difference in the amount of care and work required, and therefore in the cost and time involved, in liking a CLS to a base survey, depending on the type of information to be transferred between the two surveys at the micro level. Having a common sample of addresses or households (i.e. the CLS using the same sample or a sub-sample of the LFS) requires simply the identification, sub-sampling and re-enumeration of LFS households. The permissible time lag between the two is determined by the extent to which addresses or households are stable units over time (addresses are usually more stable than households). The next option is to use the LFS to identify households containing children to serve as the frame or the sample for the CLS. Such units are somewhat more susceptible to change over time, being affected also by changes in household composition, and require more information to be transferred from the LFS to the CLS. A procedure going considerably further in the same direction is to use names and other identifiers of the actual children to be followed up – but including all children, whether working or not, in the follow-up. This was the option taken in the 1999 CLS in Zimbabwe. Some surveys go even further, following up only the particular children identified as working in the earlier, larger survey. (See Section 2.5 for a more detailed discussion of the options in linking the various survey components.)

Box 2.2: Example of "combined" LFS-CLS – Costa Rica 2003

The sample design corresponds to a probability, area-based, stratified and two-stage approach. …. The sample design requires self-weighting in each stratum. Since the enumeration areas have the same probabilities, self-weighting was maintained by selecting a fixed fraction of dwellings during the second stage – one fourth (1/4) in urban enumeration areas and one third (1/3) in rural areas. … The size and distribution of the sample were calculated as follows:

- The sample size necessary to obtain unemployment rate estimates of 5%, with a 1% margin of error, a 95% confidence level and a design effect of 2.45, was calculated.
- Since four study domains were established (Urban Central Region, Rural Central Region, Other Urban and Other Rural), the sample size for the country as a whole was obtained by multiplying the result of the operation mentioned above by four.
- Given the country's interest in obtaining data on children's work through special modules, the national sample was distributed by area (urban and rural), according to the variability of the economic participation rate of the population aged 5 to 17. *This rate was estimated based on the 1995 Child Labour Module* [emphasis added].
- The resulting sample for each area was distributed by region, according to the relative variability of the number of unemployed persons.
- Sample size was calculated using the same procedure followed for simple random samples. Since the sample in question is complex, an adjustment was made for sample design effects. …

Source: National Report on the Results of the Child and Adolescent Labour Survey in Costa Rica, 2003. International Labour Organization, 2004.

Box 2.3: Example of "combined" LFS-CLS: Turkey 1994

The 1994 October round of the Household Labour Force Survey (HLFS), which has been conducted by the State Institute of Statistics (SIS) since 1988 on a semi-annual basis, included two additional questionnaires aimed at measuring for the first time the incidence of child labour in Turkey. In Turkey, the potential labour force is defined to be those above the age of 12 who are not mentally or physically disabled. … This necessitated a special survey to be conducted on child labour covering working children between the ages of 6-14. … In the October round of the 1994 HLFS, households where at least one child between the ages of 6-14 is found are applied both the HLFS and CLS.

The HLFS consists of two questionnaires: Form A and Form B. The first part of Form A covered demographic characteristics of the household members, and the second part covered the employment status of the household head. The survey questions in the second part of Form A are repeated in Form B to investigate this time the employment status of all household members above the age of 6.

The CLS, applied as a part of the 1994 October HLFS, consists of two questionnaires as well: Form C and Form D.

Form C is used to determine the various characteristics of the household such as the characteristics of the household dwelling (type of dwelling, useable area, etc.), migration history of the household members, monthly household income, number of employed household members, and employment and educational status of children, by interviewing the head of the household or another member of the household in the absence of the household head.

Form D is used to get more detailed information regarding the educational and employment status of children by directly interviewing the children themselves. Besides the usual questions like the hours worked per week, type of economic activity engaged in, job status and monthly earnings, attitudinal questions like children's assessment of the working conditions, their relationship with the employer and their aspirations for the future are asked. In addition to questions that investigate the economic activities of children, questions aimed at measuring their domestic activities (chores) and playing time are also inquired.

Problems encountered during the pilot survey included the following:

(1) *The simultaneous application of the HLFS and CLS meant that the respondents were required to answer a large number of questions. These lengthy questionnaires often annoyed the respondents who became rather reluctant to answer the questions toward the end of the survey* [emphasis added].

(2) Some respondents seemed reluctant to provide information in regards to their young working children.

(3) Questions on migration were not answered clearly …

(4) Problems were encountered in communicating with young children.

(5) Households were quite reluctant in providing information regarding their household incomes.

(6) Since the CLS was applied as a part of the HLFS, in order to assure a logical sequence, some questions were repeated several times. Having to answer the same question over again often bored the respondents.

Source: Child Labour 1994. State Institute of Statistics Prime Ministry, Republic of Turkey, 1997.

Box 2.4: Example of "linked" LFS-CLS – Zimbabwe 1999

The child labour survey (CLS) was a module to the Indicator Monitoring Labour Force Survey (IM-LFS) conducted by Central Statistical Office (CSO) in June 1999. The IM-LFS is a component of the on-going integrated household-based Indicators Monitoring Surveys started by CSO in 1993. These surveys have a fixed and limited data content designed for monitoring living conditions. In this respect the child labour survey was intended to be an in-depth inquiry to provide a national picture on the nature, magnitude and causes of child labour in Zimbabwe. ….

The area-sampling frame used for the CLS was the 1992 Zimbabwe Master Sample (ZMS 92) developed by CSO following the 1992 Population Census. The ZMS 92 included 395-enumeration areas (EAs) stratified by province and land use sector. The ZMS 92 was revised in 1997 for the Inter-Censal Demographic Survey (ICDS).

….. A two stage stratified sampling design was applied with enumeration areas (EAs) as the first stage, and households as the second stage sampling units. In total 395 EAs were selected with Probability Proportional to Size (PPS), the size being the members of the households enumerated in the 1992 Population Census. The selection of the EAs was a systematic, one-stage operation, carried out independently for each of the 34 strata. Within each of these EAs, a complete household listing and mapping exercise was conducted, forming the basis for the second stage sampling. The resultant households were selected by random systematic sampling. The self-weighting sampling technique ensured that all households in each province had an equal probability of being selected. It also simplified the analysis of data collected through the direct computation of percentages, means and ratios of population parameters from the sample, without necessarily weighting or raising factors.

A total of 13.591 households was selected from the household lists of 55.176 households.

During the CLS, *for each EA selected, the name of head of household and names of the children age 5–17 years in each household were provided to each enumerator to enable them to easily identify eligible interviewees*. …. [emphasis added].

A list of selected EAs was provided to each Provincial office. Besides the list of EAs, a list containing the name of head of household, names of children age 5–17 years in each household and physical address of the household to be interviewed were also provided to each enumerator.

For responding households, a questionnaire similar to IM-LFS questionnaire was used to collect data on child activity. Since the subject matter under this IM-LFS and the CLS was similar the relevant questions that were already contained in the IM-LFS were not repeated. The questionnaire for the CLS therefore only concentrated on variables dealing with children activities and their working conditions.

… The approximate interview time to complete the CLS questionnaire was between 20 to 30 minutes per respondent. … Data collection started on 20 September 1999 and ended on 30 September 1999. Each enumerator covered an average of about 70 households during data collection period.

Source: 1999 National Child Labour Survey, Country Report. Ministry of Public Service, Labour and Social Welfare, Zimbabwe.

2.3 Implication for sampling of the definition of "child labour"

The CLS component normally collects a range of information on household characteristics, on basic characteristics of all household members, and on characteristics and activities of children. The "defining" information in this set is that collected for the identification of *work-related activity of children*, on the basis of which estimates can be made of the prevalence of child labour. This being so, during the implementation of the LCS component the children's parents or guardians - and usually the children themselves - are subject to further, often quite detailed questioning on the nature and consequence of their work-related activity. Hence the CLS component provides "screening" for the identification of sample cases for the LCS component. This relationship holds irrespective of the form in which the CLS and LCS components are operationally linked to each other. The relationship may vary from complete integration to only a weak linkage.

Survey practice has varied as to what is included under "work-related activity" for this purpose. It may include only paid economic activity in the strict sense, or all economic activity whether paid or unpaid; it may also include appropriately defined "substantial" household chores, or indeed any other activities, such as looking after other children or other persons in the household needing care, which may potentially have an impact on a child's development and well-being. Similarly, the reference period for the activity may vary, from the current or past week to an extended period such as the past year. [See Box 2.5 for an illustration of the meaning given to terms such as "activity" and "work" of children in a particular survey (Portugal 1998), and Box 2.6 for international recommendations on the subject (ILO, 2004).]

Of course, the choice of the basic concepts of the survey is primarily based on substantive considerations, determined by the survey objectives, and not on considerations of sample design. In other words, what is included under "child labour" is a substantive issue and, as such, not really a sampling issue. In practice, however, *the definition of child labour adopted in the survey can have important implications for the choice of survey structure and sample design*. This is for at least two reasons.

Complexity of the definition

First, the amount and complexity of the information which has to be collected to identify "labouring children" from among all children influences the type and size of the survey required for this purpose. For instance, if child labour is defined as standard "economic activity" according to the ILO concept of the labour force, it can be captured by simply extending the lower age limit for questioning in a large-scale survey such as the LFS. Sometimes, information on the economic activity of children has even been obtained through a very large household listing operation (e.g. Pakistan, 1996). We do not recommend this practice, however, because of the high cost and low quality of such a procedure, unless some specific provision is made to extend the listing operation to include a "mini questionnaire" for the identification of child labour (Chapter 6). On the other hand, a very complex definition of child labour, in terms of particular characteristics involving all non-school and non-leisure activities of children, would require much more elaborate questions as to how children spend their time. Such information cannot be incorporated easily into an existing large-scale survey such as the LFS. In any case, the sample size resulting from such incorporation is likely to be unreasonably large for the complex information to be collected, taking into account cost, data quality and practical constraints.

Restrictiveness

The second sampling implication of the particular definition of child labour adopted is that this choice determines the proportion of children classified as being engaged in labour and thus the resulting sample size of the *part of the enquiry concerned with investigating the conditions and consequences of child labour* – that is, the part for which the relevant population is that of labouring children rather than all children.

For instance, a very broad definition of child labour could mean that a large proportion of the children are recorded as being engaged in labour. This may result in too large a sample of labouring children for studying the conditions and consequences of child labour in detail, and may necessitate sub-sampling of the total number of children identified as "working" in the original sample before their inclusion in the detailed study of labouring children.

By contrast, too restrictive or narrow a definition of child labour can result in only a small proportion of the children being regarded as engaged in labour. A sufficiently large sample of labouring children for the study of conditions and consequences of child labour may require increasing the total number of children covered in the original (CLS) sample by increasing its sample size. This may not always be possible, for example if the original sample is from an existing survey with a pre-fixed sample size. Alternative solutions may have to be found. For example, in a child labour survey in Azerbaijan in

2006, the existing LFS sample of nearly 9,000 households was not able to provide the required sample of over 5,000 households containing labouring children according to the standard definition of "economic activity". The child labour survey objectives were met more adequately by broadening the definition of child labour to include engagement in "substantial household chores", thus increasing in the CLS sample the number of "labouring children" so defined, all of whom were included in the LCS which followed.[5]

To summarize, the substantive choices made regarding what is included under child labour can have important consequences for the resulting sample for a detailed survey on labouring children. This is because the sample size which can be achieved for the LCS component depends on how restrictive or broad is the adopted definition of "work-related activity". The relative sample sizes in terms of the number of children of the LCS and CLS components are constrained by the proportion of children engaged in work-related activities; and in a given situation that proportion depends on the decision as to what is included under "work-related activities".

In this connection, it is instructive to note the concept of child labour as discussed in the *SIMPOC Manual* (see Box 2.6).

The screening function of CLS

The basic idea of the CLS-LCS system is that the CLS provides information on the basis of which a more efficient and targeted sample of labouring children for the LCS can be identified.

Essentially, the CLS identifies units (areas, households, or even individual children) containing labouring children. These identified units may contain "false positives" (child labour reported when none existed), as well as "false negatives" (failures to report existing child labour). From the substantive and also the practical point of view, false negatives present a much more serious problem than false positives. The LCS is meant to be a more sensitive instrument for identifying the labouring status of children precisely. False positives from the CLS can be identified more readily during the LCS phase. However, false negatives (missed cases of child labour) are a small part of the large population of children, and tend to be much more difficult and expensive to identify. They are missed altogether if the LCS is confined only to cases identified as positive in the CLS.

A practical way of reducing this problem is to use a more inclusive, i.e. broader, definition of child labour in the CLS, so as to permit a narrower definition for the LCS.

---

[5] The following methodological information appears in the survey report:
"The CLS collected information on 16,482 children aged 5–17 from 8,785 households across the country. … The LCS covered 10,543 children aged 5–17 from 5,355 households. While the main aim of the LCS was to gather additional information on working children, all children in the households covered by the survey were interviewed, thus producing another data set (in addition to the CLS data set) that included both working and non-working children. .. Because the non-working children in the LCS sample share the same household characteristics of working children – and can thus be considered to be at higher risk of employment than non-working children in general – it is likely that the non-working children in the CLS sample are more representative of the non-working child population as a whole. For this reason, this study relied primarily on the results of the CLS in analysing the main characteristics of working children and the determinants of child labour.., as well as the determinants of schooling .., whereas the LCS results were used to evaluate the possible effects of work on health and schooling. ..".

This will reduce the number of false negatives in the CLS *in terms of the LCS definition* (which is meant to be the definition actually of interest). The proportion identified as false positives in the LCS may be used to adjust appropriately the estimates of prevalence of child labour from the CLS.

Box 2.5: Implications for sampling of terminology and key definitions – Portugal 1998

…The term "activity" is used instead of "work" more than once in the present text, the latter term being employed in a precise sense. Thus, the word "activity" is used in a very general sense, which could be in the economic sense of the word (work) or even household and school activities. The expression "economic activity" is therefore used to designate situations of third party employment or unpaid family work as it is used in the field of employment and unemployment statistics. Given the improbability of the situation in the case of children, self-employment is not considered. Therefore, in the terms of the questionnaire "children engaged in an economic activity" corresponds to child labour. And with this definition the restricting legal concept of child labour is widened to cover anything which could be considered as infringing present legislation, including work which could (by its characteristics) be the object of a contractual relationship.

In its turn, domestic activity is also a focus of the questionnaire, given the negative effect it could have on the development of a child – similar, even, to the negative effect of child labour. However, it is not an activity in the economic sense, and even less in the area of working rights. Therefore, domestic activities should be considered in relation to their difficulty and duration or as a factor which impedes schooling and reduces its benefits.

In the questionnaire, domestic activities are considered to be those which are carried out with a certain regularity and as a help to the family, according to the statements of the children and heads of households. They are not scheduled activities but light duties performed within a process of family upbringing, such as tidying up clothes, making beds or clearing the table.

The need to quantify the different situations covered by the term "activity" means that economic activity is considered a determinant in the classification of children. Therefore, whenever a child has an economic activity, whatever its importance, it is included in this category even if it is part of a range of domestic chores.

Children without an activity are those who have no activity in the economic sense. However, in the questionnaire analysis the term "school activities" is also used. This involves participation in lessons, study and homework, but no economic value. School attendance is thus quantified. School activities are those carried out by children whether they have any other activity or not.

Source: Child Labour in Portugal: Social Characterisation of School Age Children and Their Families, 1998. Portugal Ministry of Labour and Solidarity (MTS).

Box 2.6: On the concept of child labour – *SIMPOC Manual* (ILO, 2004)

What is child labour? There is no universally accepted definition of child labour. Widely differing positions prevail among researchers regarding what kind of activities may be classified as child labour…

Herein lies the crux of the debate: which is the better approach to defining "work" with regard to children – to adopt the general production boundary criteria, or adhere to the SNA production boundary? …

[At one extreme, some] observers maintain that child labour should include only those economic activities that deny a child the possibility of normal development into a responsible adult. This perspective reserves the term "child labour" for strenuous or hazardous employment in economic activities by younger children, as well as children in worst form of child labour (WFCL) sectors. …

Some researchers, however, adopt a wider concept of "work by children", one that, following the general production boundary, includes common non-economic activities such as domestic chores. …. Supporters of a wider definition – one that includes non-economic activities such as domestic chores – argue that adhering to the narrower definition of child labour as a subset of only economic activities carries the risk of gender-biased data. …

[At the other extreme, some observers] take the view that all non-school and non-leisure activities of children constitute child labour. According to this view, child labour would include light work in household enterprises after school, or even help with domestic chores such as home cleaning or looking after younger siblings.

The ILO/SIMPOC approach … defines child labour as a subset of economic activities (i.e. work) performed by children. This is based on the standard ILO definition of work as applied to persons of working age, which is linked to the production boundary as determined by the UN system of national accounts (SNA). …. Starting with a broad framework and moving to a narrower concept – the range of statistical definitions of "child labour" adopted by researchers measuring children engaged in various non-schooling activities could include:

- "children in non-economic activities (including household chores)" plus "working children";
- "working children" only (i.e. children engaged in only economic activities); and
- "child labour" as adopted in the ILO/SIMPOC approach, understood to represent a subset of "working children". …

Including non-economic activities in child labour estimates, even if this determined using an internationally standard definition for non-economic activities, could itself prove problematic. [A number of steps are required in its implementation including:] refining the criteria for including household chores, defining a "light work" threshold [only above which can an activity be considered as "labour"], refining definitions of "hazardous work", establishing more universally accepted lower age limits for child respondents, and improving questionnaire design.

ILO global child labour estimates

… In the absence of a universal definition of child labour that could fit all countries and all circumstances, the 2002 ILO global child labour estimates were guided by benchmark ages reflected in the provisions of ILO resolutions and the relevant international instruments to which several countries are signatories.

Assuming a minimum age for light work of 12 years and a minimum age of 15 years for admission into employment, the 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 the unconditional WFCL [worst forms of child labour].

Light work was defined as non-hazardous work performed by girls and boys 12 years of age and older performed for fewer than 14 hours per week (on average two hours per day). Hazardous work, on the other hand, 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…

Note that these guidelines for classifying economically active children are specific to that particular global estimate exercise, and do not represent an ILO recommendation for national statistical surveys.

Source: SIMPOC Manual (extracts from pp. 17, 21, 27)

2.4 Integrated LCS

In Table 2.2 (column 2), two forms of the relationship between the CLS and LCS have been identified: an integrated survey; and linked surveys.

The distinction between the two types of surveys (CLS and LCS), discussed in Section 2.1, by no means implies that they must (or even can) always be organized as two separate operations. In fact, in a majority of the national surveys conducted so far they have been completely integrated into a single survey operation – the same survey covering the two different types of objectives.

2.4.1 Characteristics of the integrated design

By an "integrated" LCS we mean that the information for the CLS and LCS components is collected as a single interview operation. Note that a "single interview operation" does not necessarily imply that only a single integrated questionnaire is used, or that all interviewing takes place during a single visit to the household, or even that all the information is obtained from a single respondent in the household. Multiple questionnaires, repeated interviewer visits and different respondents (head of household, parents or guardians, children themselves, and sometimes even employers and teachers, etc.) may be involved. A "single interview operation" is meant to indicate that the information on the CLS and LCS components is collected at the same time or at least within a short space of time , and that all the information collected during the CLS part of the operation is directly available to (and is generally not repeated in) the LCS component.

Normally, the integrated survey entails the LCS questions being attached to the survey questionnaire as additional sections, which become applicable if the child concerned is found to be engaged in work-related activity during a specified reference period.

Simultaneous implementation of (or at least the absence of any significant time lag between) the CLS and LCS components implies that normally no sub-sampling from the one to the other can be introduced (i.e. all children identified as working during the CLS part are subject to the LCS part of the interview). In any case, it is desirable in practice that any sub-sampling involved is straightforward - the LCS part of the interview being applied only to a pre-selected subset of CLS survey rounds or sample areas, for example.

In so far as sub-sampling cannot be (or is not) introduced between the LCS and CLS components of an integrated survey, the relative sample sizes in terms of the number of children of the two components are determined entirely by the proportion of children engaged in work-related activities; and in a given situation, that proportion depends on what is chosen for inclusion under "work-related activities", as discussed above.

In an integrated survey, the sample size requirements of both the CLS and LCS components have to be met. Hence, *even in a single "integrated" survey the distinction between the two types of survey components still remains a conceptually useful one; it reminds us that an integrated design has to be a compromise between different types of objective*. In practical terms, therefore, this entails the CLS being large enough to provide estimates of prevalence of child labour with the requisite precision, and also to yield sufficient numbers of working children for the LCS.

Unfortunately, in practice the survey design has often been determined more or less unilaterally – with an over-emphasis on one type of objective at the expense of another. For instance, some surveys have been too small to yield useful estimates of the extent and distribution of prevailing child labour (the CLS component was too small in sample size), while others have been too large in size to permit sufficiently in-depth investigation of the characteristics and consequences of child labour (the LCS component was too large in sample size from the practical point of view). By contrast, there are examples where the sample size, while adequate for the CLS, has turned out to be inadequate in providing enough cases for the LCS component for the purpose of investigating child labour activities in detail.

It is worth re-emphasizing that, with an integrated design, the sample for the LCS component is entirely determined by the results of the screening provided by the CLS component. Not only the sample size but also – and more importantly – the quality of coverage of the population of working children in the LCS component is determined by the quality of the screening questions in the CLS part for identifying those children.

In general the consideration mentioned above also applies to separate but "linked" CLS and LCS surveys (discussed in the next section), though in that case the operational separation between the two surveys can permit some control of LCS coverage and reduce its dependence on the quality of screening during the CLS. This separation also makes it easier to introduce sub-sampling between the two operations.

2.4.2 Advantages and disadvantages of an integrated design

There can be practical and cost advantages to integrating the CLS and LCS components into a single operation. Clearly, it is cheaper and more convenient to collect all the required information in one go. This is a major advantage, and it may explain why a large majority of child labour surveys to date have chosen the integrated arrangement.

However, there are also some disadvantages to employing an integrated (CLS+LCS) design, rather than two operations conducted separately, albeit linked in varying degree.

- An integrated implementation tends to limit the information that can be collected during the LCS component without jeopardizing the quality of the CLS component because of the increased respondent burden.
- Apart from the obvious upper limit imposed by the number of eligible cases (working children) identified in the CLS component, there is no independent control over the size and distribution of the sample for which the LCS information is collected. Introducing sub-sampling from CLS to LCS in an integrated design is not easy in practice. As noted above, this is more feasible when the two components are separated in time.
- The level of child labour may be very unevenly distributed over the sample area, and the interview workload may thus vary greatly. This becomes all the more troublesome when the LCS part involves lengthy interviews with adults as well as with children individually.
- The increased burden associated with the LCS part can have serious consequences for the quality of coverage of the sample enumerated in the CLS

part. This is because, in order to reduce the (often substantial) workload involved in the LCS part of the interview, the interviews may sometimes have a tendency to under-report working children. Such tendencies have in fact been widely reported in other surveys involving heavy modules, modules which apply only to a part of the population identified as being eligible during the preceding stage of the interview.[6]

2.5 Linked LCS

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 separated) surveys but must be linked to each other in some way. This is because the LCS depends on the CLS for the identification of its sample of working children – in terms of identification of CLS sample areas containing high concentrations of working children, and of households containing such children or of individual working children themselves.

The LCS sample can thus be identified on the basis of the CLS results in different ways, and this introduces different forms of linkage between the two surveys.

2.5.1 Forms of the CLS-LCS linkage

There are a number of ways in which the LCS may be related to a previously conducted CLS. The best solution depends on the particular situation and objectives, but primarily on two factors: (1) *the time gap between the two surveys*; (2) the *quality of screening* provided by the CLS in identifying the presence of working children.

The diversity of options in CLS-LCS linkage includes the following.

A. In relation to the selection of the sample within the CLS sample areas, we may take all or select a sample of:

   1. working children previously identified individually;
   2. households previously identified as containing a working child;
   3. households previously identified as containing any child in the age bracket of interest;
   4. all households interviewed in the previous sample;
   5. or possibly, all households selected in the previous sample (including non-respondents);
   6. all households listed in the CLS sample areas (i.e. obtaining a new LCS sample from existing CLS household lists); or
   7. households from re-listing of the CLS sample areas (i.e. updating the area lists before selecting a new sample).

[6] Demographic and Health Surveys (DHS), for instance, have often found an under-reporting of women of child-bearing ages, and also of infants, when the interview has involved collecting additional information for these sub-populations.

B. In relation to the "eligibility for inclusion" of the CLS areas, we may take as "eligible":

8. CLS sample areas containing at least one working child (or, equivalently, at least one household containing a working child); the lower limit may be raised to "at least x", where x>1;

9. CLS sample areas containing at least one child (or at least x children, x>1) in the age bracket of interest; or

10. all CLS sample areas.

C. In relation to the selection of areas, we may:

11. take a sub-sample of the CLS sample areas;

12. take all CLS sample areas; or

13. select additional areas for LCS, linked to the CLS sample areas

Some of the options in the three sets A-C can be combined, for instance:

- households identified previously as containing any child in the age bracket of interest (option 3),
- but only from a sub-sample of areas (option 11),
- selected from CLS sample areas containing at least one working child (option 8).

Sampling within CLS areas

Option 1 provides the tightest link between the two surveys: the LCS sample is confined to the particular children identified as being engaged in child labour during the CLS. This makes the option almost the same as an integrated CLS-LCS design, except for the possibility of sub-sampling between the two operations because of their operational separation in time.

Note that, of options 1 to 7, all except option 1 require fresh identification of labouring children in the households included in the LCS sample. With option 1, the LCS sample is confined to the particular set of labouring children identified during the CLS interview. Any working children unidentified during CLS remain so during the subsequent LCS operation. The opposite – but usually much less important error of children not working being identified as working can of course be identified during the subsequent operation.

With options 2 to 7, the LCS has a greater potential to refine the CLS estimates of child labour.

The options taken in reverse order, from 13 to 1, reflect increasingly close linkages between the two surveys. From 13 to 1, and especially from 7 to 1, the options generally become:

- more restrictive (the LCS sample is increasingly restricted to units identified in the CLS sample);
- more focused (on the particular population of interest, i.e. labouring children); and

- more efficient exploiters of information already collected or of the design already implemented in the CLS (hence more cost effective, and less burdensome for the respondent).

But the options increasingly require:

- more information to be collected during the CLS, and then fed forward to the LCS (with the operation becoming more costly and time consuming); and also
- more sensitive to changes over time (and thus less suitable in the presence of long time gaps between the two surveys).

Option 1 fails to cover children in the population (not just in the CLS sample) who started working subsequently to the time of CLS enumeration. This introduces a bias, unless the time interval between the two surveys is very short, i.e. a few weeks at the most. By contrast, option 7 requires stability of the area units only, and hence can be used even when the two surveys are widely separated in time.

Options 1 to 3 require information on some characteristics of households or persons; options 4 to 7 require information only on household addresses.

Note that options 1 and 2 are quite different in terms of the information needed for their application. Specified households or addresses, as in option 2, are much easier to identify (and identify correctly) than specified children, as in option 1.

Households are likely to be much more stable in terms of whether or not they contain children, than in terms of whether or not they contain *working* children. This means that 3 is more suitable, when the time lag between the surveys is not very short, than either of options 1 or 2. Also, option 3 avoids visits to households that are unlikely to contain a child, hence it may be significantly more efficient than option 4.

Exclusion of certain CLS areas

The difference between options 9 and 10 may be minor when the sample areas are relatively large (in so far as most of the areas contain at least one child). Option 9 is commonly used to avoid wasting effort on areas known to be highly unlikely to contain a child in the specified age group.

One should be careful of potential bias with option 8 – including only areas containing a household with a working child – if the time lag between the two surveys is large (several months, for example).

However, there can be strong practical reasons for choosing this option. Child labour is usually very unevenly distributed geographically. Often large concentrations of it are found in a limited proportion of areas, and low prevalence in a large proportion of areas. The proportion of areas reporting no child labour at all can also be high. This pattern of distribution – and to a large extent even the position of individual areas in it – is determined by prevailing socio-economic conditions, which tend to be persistent over time.

Therefore – despite the cautioning note above concerning the choice of option 8 – information classifying the CLS areas according to the level of child labour can be fairly

stable and thus very valuable in determining the sample design and implementation strategy for the LCS.[7]

Apart from using it to determine the general structure of the sub-sampling of LCS from the CLS areas, the information on the reported level of child labour in the latter is often useful in determining the cut-off level below which the areas may be altogether excluded from selection into the LCS. This is often necessary for practical reasons; it may simply not be cost effective to attempt the LCS in areas containing no reported working children or only very few. The cost considerations have to be balanced against the bias which such exclusion introduces.

Sub-sampling of CLS areas

Sub-sampling of CLS sample areas may be introduced in order to:

- reduce the LCS sample size; and specifically
- make the LCS sample more concentrated, i.e. confined to fewer sample areas with higher levels of child labour.

Special techniques of sub-sampling to achieve the latter objective are described in Chapter 5.

Expanding the original CLS areas

A brief comment will be useful on option 13. This option may be used when the LCS sample needs to include additional areas, beyond the CLS areas. This may, for instance, be because the CLS sample areas do not yield a sufficient number of sample cases (labouring children) for the LCS. This can happen if, for instance, the original CLS design was based on an unrealistically high expectation of the number of working children which would be found. Another motivation can be to enhance the LCS sample by selecting more cases from and around CLS areas that are found to contain concentrations of labouring children. These requirements are likely to arise more often in special surveys, such as surveys of places where children congregate to work (e.g. porters, garbage collectors, etc.), but they can also arise in household-based surveys of child labour.

An obvious way to expand the LCS sample is to include in it additional areas in the "neighbourhood" of CLS sample areas. Various techniques can be used for expanding the sample in this way, while retaining its probability nature. A couple of simple procedures are noted in Section 5.4. More sophisticated procedures, such as "adaptive cluster sampling" have also been developed.

In all the other options listed above (1 to 12), the LCS sample is confined to CLS sample areas (though not necessarily confined to CLS sample households). Hence we may consider option 13 to be the loosest link between the two surveys.

[7] It is very important to note, however, that many forms of child labour are subject to significant seasonal variations (school vacation periods, crop seasons, tourist seasons, etc.). In using information from one survey to another, the two should be conducted during the same or similar seasons.

2.5.2 Quality of screening provided by the CLS

Two basic factors determining the appropriate linkage between the CLS and LCS were noted above as (1) the time gap between the two surveys, and (2) the quality of the screening provided by the CLS in identifying the presence of working children in areas, households or individually.

The issue of time lag has already been mentioned above in connection with various options. Basically, a longer time lag requires linkages through more stable units (such as areas, and not individual children) and less detailed information that needs to be fed forward from one survey to the other.

Depending on the detail of its questionnaire, the CLS can vary greatly in the efficiency and specificity of its screening function for the identification of child labour. When it is a stand-alone survey or forms a fairly detailed module of, say, the LFS, it has the potential to identify the presence of working children more precisely, using the proper definition of what constitutes "child labour" according to the survey. In this case the link with the LCS component can be close, or the two may even be integrated into a single operation, unless otherwise required by considerations relating to sub-sampling or the complexity of the total interview.

When the CLS has to be confined to a limited set of more straightforward questions, it is likely to be less complete and precise in identifying cases of child labour. Usually, a simplified version of the concept of child work has to be used in any case. The information it provides on individual cases (children) is naturally less reliable. The link between the two components therefore needs to be less close at the micro level compared to the previous case. Nevertheless, the information from the CLS can still be extremely useful in determining parameters for an efficient design of the LCS – for example, in the selection of areas taking into account the patterns of concentration of child labour, and the stratification of households according to the likely presence or otherwise of working children. A combination of, for instance, options 11, 8 and 3 of Section 2.5.1 may indeed be quite appropriate in many situations.

2.6 Surveys of children, and surveys of child activities

2.6.1 Scope of surveys

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. In Table 2.2, "LCS scope" (column 4), we have distinguished the following three types of situations.

1. A majority of the LCS surveys have as their main focus the study of conditions and consequences of child labour. In the table, the LCS scope for this type of survey situation has been indicated as "CL (child labour)".

2. In a number of countries, the scope is broader, and covers all types of activities of children, including economic and non-economic activities, education, leisure and even non-activity. Many such national surveys are actually named a "child activity survey" (CAS). In the table, therefore, the LCS scope for this type of survey situation has been indicated as "child activity survey".

3 In a few countries, the scope is even broader and includes more general information about children beyond their economic and non-economic activity, such as information on children's health, housing conditions, etc. These are termed a "children's survey".

2.6.2 Sampling considerations

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 (Chapter 5). With their broader and more diffuse scope, children's surveys (option 3), would require a design similar to that of the CLS (Chapter 4), or may even incorporate the latter. However, for this type of survey, and especially for child activity surveys, option 2 - the measurement of the incidence and conditions of child labour as such - is likely to remain a special objective. A compromise design is therefore desirable which covers both working and non-working children but which gives greater weight to the former. A design on the lines of Chapter 6 – with its compromise allocation of area selection probability – may indeed be an appropriate choice in many situations. 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 survey areas.

We shall now consider some more specific sampling aspects for children's and child activity surveys.

Firstly, we should note that, irrespective of any additional information collected on children, the *core* in these various types of survey remains the study of conditions and consequences of child labour. Thus in all cases this requirement should influence the survey sample size and design. The base population for this aspect is, of course, the *population of working children*.

In the overall sample design, we in fact have to consider three components:

(a) the "CLS component", which has the primary objective of estimating the proportions (or numbers) of children engaged in child labour;

(b) a "supplementary component", with the objective of obtaining a wider range of information on all children – on child activities of all types, and possibly also on other aspects of children's life;

(c) the "LCS component", with the primary objective of obtaining detailed information on conditions and consequences of child labour.

Generally, we can expect that the more general *children's surveys* will give somewhat more emphasis to objective (b) than to the other objectives, while (a) and (c) are likely to be somewhat more emphasized in *child activity surveys*. Leaving that aside, particular surveys of either type may differ in the relative emphasis each gives to objective (a) versus objective (c).

Giving more emphasis to (a) – i.e. the measurement of prevalence of child labour – would make the design requirements more similar to that of a CLS (Chapter 4). In any case, the base population for (a) and (b) is the same, namely the population of all children. However, putting these two objectives together still requires compromises. These mainly concern the related aspects of interview complexity and sample size. Surveys of all activities of children, and especially more general children's surveys, would usually involve a much heavier interview than a pure CLS. This would *limit the feasible sample size*.

At the same time, in so far as objective (b) influences the choice, the sample size *required* also tends to be smaller. Estimates of prevalence of child labour are frequently required for various sub-populations and domains in the country, and generally require large sample sizes. By contrast, the objective of (b) is normally to provide more descriptive and detailed information on children's activities and other aspects of their life. Often this sort of information does not, in itself, require a sample size as large as is the case for (a).

Furthermore objective (c), even if less emphasized, cannot be disregarded. It argues for smaller and more targeted samples. This consideration is likely to be even more important in a survey giving emphasis to objective (c) vis-à-vis (a), i.e. to the assessment of conditions and consequences of child labour as distinct from the estimation of its prevalence. This is more likely to be the situation in what have been called child activity surveys as opposed to children's surveys. In this case the sample should be smaller and more targeted at areas of higher concentration of child labour.

One possibility is to determine the selection of areas primarily (though by no means exclusively) on the basis of the specific requirements of objective (c), but at the next stage to chose samples representing all children (possibly with substantial over-sampling of working children) within each sample area.

2.6.3 Examples of children's and child activity surveys

The objective of the following examples, extracted verbatim from national survey reports, is to provide a more concrete picture of the objectives, content and diversity of such surveys which have been conducted in developing and transition countries, generally under the rubric of "child labour statistics".

Box 2.7: Example of a children's survey – Ukraine 1999

Survey units consisted of children 5–17 years of age and one parent or guardian. …

The reference period covered the three months prior to the month of interview. Interviews were conducted from the last two weeks of the final month of the quarter to the first week of the following quarter.

The survey instruments comprised the following:

Child labour survey questionnaire for parents. This was completed according to answers provided by parents/guardians of children 5–17 years of age and included 33 questions.

Child labour survey questionnaire for children. This was completed according to the answers furnished by children 5–17 years of age and included 43 questions about education and leisure time, economic activity, working conditions, health care and domestic work.

The above questionnaires for children and for parents/guardians were supplementary, with certain questions - i.e. questions on children's activities, reasons for working, job availability, health status and household commitments - appearing on both questionnaires in order to allow for the comparison of the children's perception of their work with that of their parents. To ensure the non-interference parents in children's answers, children were interviewed in the absence of their parents. During data analysis, answers furnished by parents and children to the same questions were compared to ensure their impartiality and reliability.

Source: Child Labour In Ukraine 1999, Statistical Bulletin. International Labour Organization and State Statistics Committee of Ukraine, 2001.

Box 2.8: Example of a children's survey – Cambodia 1996

Objectives of the survey: to gather information on the demographic and economic characteristics and the health status of mothers and children, on labour force characteristics, on child labour, etc.

Items of information collected in the survey:

Demographic and economic characteristics

For all household members: relationship to household head; age; sex; disability; migration.

For household members aged 5 years and over: migration in relation to employment; education; literacy; usual activity in the last 12 months; current activity the preceding week; primary occupation; secondary occupation.

For household members aged 10 years and over: marital status.

For females aged 15 to 45 years: pregnancy.

Children aged 5 to 17 years (child labour)

School attendance in the past week. Reasons for dropping out or not attending school. Main reason for working or having a job. Age the child started to work. Place of work. Proportion of the child's earnings given to the household. Illnesses, injuries and other health problems of the working child. Recruitment of children to work elsewhere.

Health for children aged less than 5 years

Illnesses; treatment received; information for children aged less than 2 years; information for children aged 1 to 2 years; information for children aged from 6 months to less than 5 years.

Housing and other household particulars

Type of building and year the building was constructed; tenure status and use of housing unit; household expenditures; use of water, lighting, cooking fuel and toilet facilities in the household; means of transport, appliances and other amenities owned by the household; access to basic services; landholdings; economic activities; credit behaviour; accidents in the household; health practices of household members.

Source: Report on Child labour in Cambodia 1996 (draft). National Institute of Statistics, Phnom Penh, Cambodia, July 1997.

Box 2.9: Example of a children's survey – Panama 2000

The general objective of the child labour survey was to measure child labour or, more specifically, to measure the socio-economic characteristics of dwellings with population between 5 and 17 years of age, to ascertain the housing conditions of the child population entering the labour force, to measure the conditions under which child labour occurs and to obtain information on the incidence of occupational hazards and lesions to the population of boy, girl and adolescent workers. The survey provides crucial information for preparing specific policies for the population between the ages of 5 and 17 years, as well as for monitoring and evaluating programmes being carried out by different social agencies in order to eradicate the worst forms of child labour. It is worth noting that the data presented here allow a detailed study to be made of the population between 5 and 17 years of age among the non-indigenous population, taking into consideration provincial boundaries and the country's rural and urban areas. In Panama province, it is disaggregated into the Panama and San Miguelito districts.

Sampling aspects. The child labour survey was carried out nationwide and included indigenous and inaccessible areas in order to interview customary residents (de jure survey). The segments interviewed were those selected from a sample framework of segments that were previously known to contain a population of children between 5 and 17 years of age. Fieldwork was carried out with personal interviews throughout the country, in those dwellings where a population aged 5 to 17 years had been detected, regardless of whether or not this population was working.

Source: National report on the results of the child labour survey in Panama 2000. International Labour Organization, 2003.

Box 2.10: Example of a child activity survey – Belize 2001

Belize's 2001 Child Activity Survey (CAS) was envisaged for the statistical count of the number of economically active children along with more disaggregated data. It was also intended that it provide needed information on children engaged in economic and non-economic activities and the comprehensive demographic and socio-economic characteristics of all school-age children and of working children: working conditions, safety and health aspects (focusing on the type, frequency and gravity of injuries and illnesses) and reasons for working. The survey also sought to identify the demographic and socio-economic characteristics of the parents of any child aged 5 to 17 years.

The working children. The survey distinguished between "economically active" and "non-economically active" children (according to ILO definitions), the former being children engaged in any form of economic activity for at least one hour per week and the latter being children engaged in essentially unpaid domestic labour. A further distinction was made in that – in accordance with, in particular, the ILO's Minimum Age Convention, 1973 (No. 138) - "child labour" was defined as applying to all children aged 5–17 years who were economically active, except for:

- children aged 12-14 years engaged in light work; and

- children aged 15–17 years engaged in work of a non-hazardous nature (including light work).

"Working children" were those who were active, whether economically or non-economically.

Conclusions and recommendations. A number of conclusions were drawn from the CAS and associated data and information, and a number of recommendations were presented according to the following areas of the study:

- international Conventions and national laws and policies;
- non-economically active children;
- economically active (working) children;
- working children and household chores, school and health;
- children's savings and contribution to the household;
- child labour; and
- worst forms of child labour.

The CAS covered a wide cross-section of topics, including housing characteristics, migration status of households, characteristics of children living away from the household, respondent characteristics, demographic characteristics of the children, migration status of the children, economic activity of the children (current economic activity, place of work, employers, earnings, hours of work during the past week and usual economic activity), children in non-economic activity, idle children, health and safety aspects of children who had worked at any time in the past, perception of parents or guardians of the children, and related questions directed at the children. Background characteristics relating to the demographic and socio-economic status of the population surveyed were also included. Results were presented by sex, age group, urban and rural areas of residence, district, ethnic group and level of education, as well as by other demographic and socio-economic characteristics.

Source: Child Labour in Belize: a Statistical Report 2001. International Labour Organization and Central Statistical Office, Government of Belize, 2003.

Box 2.11: Example of a child activity survey – Ghana 2001

The overall strategic objective is to provide quantitative data on children's activities, including schooling and economic or non-economic activities as well as household chores. It is also intended to establish a database that will serve as the benchmark for measuring progress with regard to the elimination of the problem.

Concepts and definitions

Child labour. The name of the survey might suggest interest in children engaged in economic activities only, but the scope goes beyond this; it seeks information on *the general use of children's time and the effect on their health, education and normal growth*. In general, therefore, any activity, economic or non-economic performed by a child which has the potential to affect negatively his/her health, education, moral and normal development would constitute child labour. In determining what constitutes child labour, factors such as the number of hours worked, the type of work, the working environment and others, have been taken into account.

Children's activity. Economic activity refers to any work or activity performed during a specified reference period for pay (in cash or in kind), profit or family gain. In the survey, two reference periods – the preceding seven days and the preceding 12 months - were used. All other activities were considered non-economic (i.e. household chores or work of a domestic nature performed within the household, voluntary and charitable activities, etc.). Since children do carry out housekeeping activities in their parents'/guardians' households, the survey also investigated children's activities of this nature.

The household questionnaire collected information on housing/household characteristics, socio-demographic characteristics of all household members and the economic activity, health and other conditions of children. The street children questionnaire collected information on socio-demographic characteristics, living arrangements, parental background, economic activities, health, safety and other related street issues and on the assistance that street children expected of society and government.

Source: The 2001 Ghana Child Labour Survey. Ghana Statistical Service, March 2003.

Box 2.12: Example of a child activity survey – Mongolia 2002-03

The Mongolia NCLS 2002-03 was conducted, jointly with the LFS, to provide reliable estimates of child labour at the national, urban and rural level, as well as by region. The NCLS covered the child population aged 5 to 17 years living in the households, while children living in the streets or institutions such as prisons, orphanages or welfare centres were excluded. This was a stand-alone survey and the sample size and the coverage of the survey were such that it could furnish fairly reliable key estimates by region of the country. The survey had been designed to obtain estimates on as many variables or parameters as possible, particularly in relation to the economic and non-economic activities of the children in the age group 5–17 years under the usual circumstances. Since it was a household-based national level survey, it was not able to capture children in the WFCL (worst forms of child labour) sectors, or children who were on their own and living in public places. …

The NCLS (that is, LFS + CAM, the child activity module) was designed as a household-based survey, investigating activities of children, defined for NCLS purposes as those aged between 5 and 17 years. The strategic objectives of the NCLS were to generate quantitative data on child activities (including schooling, economic and non-economic activities) in Mongolia, and to begin the process of establishing a database containing both quantitative and qualitative information on the activities of children. …

The required information was generated in a two-pronged approach. First, a large part of the data was collected through personal interviews with the heads of the household (or a responsible and knowledgeable adult member of the household). These persons were asked questions regarding the general demographic and economic characteristics of each of the household members, including the activities of children. The second part of the interview was directed at the children themselves, about their activities (including schooling), working conditions, reasons for them to be at work, their perception of working, and future plans.

The survey report provides comprehensive information on all activities of children in the 5–17 year age group who are living in households in Mongolia. The children are broadly classified as:

- attending school only (no other activity);
- attending school and also engaged in an economic activity;
- attending school and also engaged in a non-economic activity;
- attending school and also engaged in an economic and a non-economic activity;
- engaged in an economic activity only;
- engaged in a non-economic activity only;
- engaged both in economic and non-economic activities;
- not attending school; and
- not attending school and not engaged in any economic and/or non-economic activities.

Source: Report on National Child Labour Survey 2002-2003. National Statistical Office of Mongolia, Ulaanbaatar 2004.

Box 2.13: Example of a child activity survey – Sri Lanka 1999

Survey field work was done in four rounds during the period from November 1998 to June 1999. Children were mostly not available during the time of the interviews and required information had to be obtained from a proxy respondent (very frequently from a parent of the child).

Survey teams received very good co-operation from the respondents (children and their parents) as regards the information collected. This was so because the questionnaire was not just confined to economic activities of the children, but was designed to collect information on educational, leisure and housekeeping activities of the children before seeking the information on economic activities; this sequence of questioning helped the survey enumerators greatly in obtaining the co-operation of the respondents. However, the reliability of the information provided by respondents as to the duration of time spent on each activity is questionable, especially when the involvement of a child in a particular activity is minimal.

Source: Child Activity Survey Sri Lanka 1999. Department of Census and Statistics, Ministry of Finance and Planning, 1999.

Box 2.14: Example of a child activity survey – Dominican Republic 2000

A children's survey. A national child labour survey was carried out in order to generate quantitative and qualitative data on the economic, domestic, school and recreational activities of children between the ages of 5 and 17 years throughout the country. …

The main population of interest in the survey consisted of the children with ages running from 5 to 17 years residing on Dominican soil. To collect the data, two different questionnaires were employed: one was aimed at the households, to collect data on the dwelling, the household and the usual and visiting residents; the other questionnaire was designed for each individual child between 5 and 17 years of age found in the households surveyed.

The topics covered by the survey included the following.

General characteristics of the population between 5 and 17 years of age and their households (the household population; breakdown by sex, age and area of residence; literacy and education; characteristics of households and dwellings; composition of the households with children; dwellings inhabited by children; and availability of basic services

Children's activities. Domestic chores in the own home; school activities; working children; magnitude and characteristics of working children; industry; category in employment and type of work; shifts and hours worked; children's income; schooling among working children

Factors related to working children. Reasons for work and job satisfaction; work-related injuries and illnesses and safety among working children; parents' or guardians' opinion of child and adolescent labour; opinion of the working children; the household context of the children.

Source: Report on the results of the National Child Labour Survey in the Dominican Republic, 2000. International Labour Office and State Department of Labour of the Dominican Republic, 2004.

Box 2.15: Example of a child activity survey – Costa Rica 2003

Costa Rica's multiple purpose household survey (MPHS) is a valuable source of information on various issues relating to the country's households and their members. The survey's employment module, focusing on the work force and its characteristics (employment, unemployment, underemployment and income), is particularly important. The MPHS also provides supplementary information on the demographic, socio-economic and educational characteristics of individuals and households. The MPHS has been conducted in July of each year since 1987. The addition of a child labour module in 2002 was part of the development of the country's *child labour survey* and data base development project.

The objective of the module was to collect information on the magnitude and characteristics of the participation of girls and boys aged 5 to 17 in economic, recreational, educational and domestic activities, as well as the main demographic and socio-economic characteristics of working minors.

Source: National Report on the Results of the Child and Adolescent Labour Survey in Costa Rica, 2003. International Labour Organization, 2004.

2.7 Surveying non-household units

The following may be noted with respect to the relationship of establishment and school surveys to household-based surveys.

Two types of studies involving establishments may be distinguished. The first consists of surveys that are representative of the population of establishments employing children. Here, the survey's target establishments can be selected from available directories or lists, including those of producers' associations and cooperatives, or from lists drawn up during a community-level inquiry or household-based child labour survey. Alternatively, lists can be based on local enquiries in the area to be investigated, using informants such as unions, government agencies, NGOs, community leaders, religious groups or charitable associations.

The second type involves linked non-household units, meaning that the units are brought into the survey on the basis of their linkage with a household or person in the sample.

This concerns special enquiries involving different types of units - such as schools, establishments or other places where children work, and possibly other types of non-residential institutions affecting working children – which are linked to household-based surveys of child labour. For instance, in Bangladesh (2002-2003) and Cambodia (1996) special employers' questionnaires were used. In enquiries linked to household-based surveys the non-household units included in the study are normally identified through working children enumerated in the main household-based survey. Generally, no new sampling operations are involved in the selection of such units, except perhaps for the selection of a sub-sample of children in the main survey for whom the linked non-household type units are to be enumerated. Furthermore, the scope and coverage of the attached survey of non-household units are usually quite restricted or selective. For these reasons it is generally not appropriate to view the information from

such enquiries as representative of the population of the non-household units involved (i.e. as representative of the population of all schools, the population of all work-places employing children, etc.). An adequate representation of such populations would generally require samples of a very different size and design than the set of units obtained simply on the basis of association with a sample of children. It is more appropriate to regard the data resulting from such restricted samples of non-household units as providing additional variables which can be attached to other information on the child concerned. To the extent that the sample of children - through which the non-household units are linked with the survey - is representative of the population of children, the results may be generalized for that population.

An example is surveying employers and establishments identified through association with working children in a sample of a household-based child labour survey. Similarly, school-based surveys are normally of the above mentioned type, involving surveying teachers and schools identified through association with children in the sample of a household-based child labour survey. As noted, in this type of study the sample is not designed to be representative of the total population of teachers or schools; rather, interviews in such units are conducted to provide additional variables on the sample of children associated with these units[8].

2.8 Survey of activities of young people in South Africa 1999: Description and comments

The 1999 *S*urvey of *A*ctivities of *Y*oung *P*eople (SAYP) was the first survey of its kind in South Africa.

The following information is derived from *Country report on children's work-related activities*, Statistics South Africa (2001).

2.8.1 Objectives and structure

Objectives of the survey

The objectives of the survey were:

1. to produce comprehensive statistical data on the work activities of young persons at the national level;

2. to create a special database on work activities of young persons in South Africa, to be updated as fresh statistical information becomes available through new surveys;

[8] It has been emphasized above that, generally speaking, one cannot expect in practice to obtain a representative sample of institutions such as schools or establishments employing children through association with children in a household-based sample. In principle it is possible to do so, of course, provided the sample of children is "sufficiently large and broad-based with good coverage". What is involved here is "indirect sampling" of institutions. This is the sampling in a situation where there is no direct access to the target population (institutions) for sample selection, but only to populations related to it (working children in private households). Procedures are available for producing estimates for the population of units selected indirectly in this way. For new and recent developments, see Lavallé, P. (2007). *Indirect Sampling*. Springer.

3. to provide a comprehensive analysis of the state of the nation's working children, identifying major parameters, priority groups and patterns, the extent and determinants of child work, conditions and effects of work, etc.;

4. to look closely at factors such as economic activity, excessive household chores and maintenance activities at school which may be affecting young people's abilities to attend school or engage in other childhood activities;

5. to disseminate as widely as possible the results of the nationwide survey on the activities of young persons, particularly in the areas where such activities are most intensive;

6. to formulate a module on the activities of young persons to be attached to one of the rounds of the newly introduced half-yearly labour force survey, once the latter is fully developed and becomes operational;

7. to enhance the capacity of Stats SA to conduct national surveys on such activities more regularly in the future.

Structure of the survey

The survey gathered detailed information in two phases.

In Phase 1 interviews were conducted to determine the extent of work activities among children and to gather general demographic information about the households in order to allow analysis of the link between these factors and working children. The main aim of this phase was *to identify households where at least one child has been engaged in child labour activity*. It also included questions regarding factors which may influence whether or not children are engaged in child labour, to enable comparison between those households with child labour and those without. The questions in the first questionnaire were directed at a responsible adult, preferably female, who took responsibility for children in the household.

In the Phase 2 follow-up, interviews were conducted in Phase 1 households that had at least one child engaged in a "work-related activity" as defined below. Hence a central design feature of the survey was that the first phase questionnaire was administered in all selected households and that answers to certain questions were used to screen or select households for the second phase. Subject to further sub-sampling in some cases, all households with certain characteristics were selected for the second phase.

The screening questions were directed at the main respondent to the first phase questionnaire. A household was considered as having a child engaged in "work-related activity" if any child in the household:

(a) had been engaged, at any time in the preceding 12 months, in any of the following economic activities *for pay, profit and/or economic family gain:*

- running any kind of business, big or small for the child him/herself;
- helping unpaid in a family business;
- helping in farming activities on the family plot, food garden, cattle post or kraal;

- catching or gathering any fish, prawns, shellfish, wild animals or any other food, for sale or for family consumption;
- doing any work for a wage, salary or any payment in kind; or
- begging for money or food in public;

and/or

(b) had been engaged regularly for one hour per day or more on any or all of the following activities:

- housekeeping activities within their households;
- fetching wood and/or water or in unpaid domestic work; or
- helping in cleaning and improvements at school.

From the above it is clear that two distinct criteria were used for different types of economic activities:

1. In the case of economic activities for pay, profit and/or economic family gain, a *very wide* screening criterion was used. Every household where any child between 5 and 17 years of age had been engaged in any such activities, at any time in the preceding 12 months (even if only once), was selected for the second phase. No time-based filter was applied here. Consequently, a household was selected even when a child had been involved in such activity for a very short time (e.g. for half an hour) over the preceding 12-month period.

2. In the case of fetching wood and/or water or unpaid domestic work, a *more restricted filter* was applied. A household was selected because of a child's involvement in such activities only if he or she had been engaged in them regularly for at least one hour a day.

The main aim of Phase 2 was to explore child labour in more detail. More extensive questions about the nature of work the children were doing were put to an adult in the household, and to the child or children involved in these activities.

Comments

1. In the terminology of this manual, Phase 1 is a stand-alone child labour survey (CLS), while Phase 2 is a linked labouring children's survey (LCS). In the former, the base population is all children in a specified age bracket, and the main objective is to measure the proportion of them that are engaged in child labour. In the LCS the base population is labouring children, and the objective is to study its characteristics and circumstances. This does not preclude obtaining similar information from a sample of non-labouring children for comparison.

   Different forms of the link between the two surveys (CLS and LCS) are possible: from completely independent sampling (stand-alone LCS), at the one end, to the LCS conducted over a set of individual children with specified characteristics identified from CLS enumeration, at the other. In the present case, the linkage is quite close: the LCS is conducted among a set of households containing one or more children with specified characteristics, as identified from CLS enumeration.

2. Information on households which have no children aged between 5 and 17, and more importantly on households which have only non-labouring children in this age bracket, is obtained only in Phase 1. Detailed information in Phase 2 is confined to (1) the group of children engaged in "work-related activity", and (2) children who, while not themselves engaged in work-related activity, live in a household where some other child is so engaged. Comparisons between these two groups provide the basis for an analysis of the consequences of child labour. However, it can be argued that the information is incomplete, and may indeed lead to biased results in the analysis of determinants of child labour. This is because both the groups covered are confined to "potential child-labour households" – identified in the survey as households currently in the presence of child labour. Children living in "non-potential child-labour households" in this sense (i.e. households with children but none engaged in work-related activity) are not covered. This can greatly limit the analysis of household-level determinants of child labour, such as the income, consumption and employment status of the household and other household characteristics.

   The information required for such an analysis can come from either of the following two sources:

   1. collection of the required information in sufficient detail on all types of households containing children and on all categories of children in those households during Phase 1, and

   2. inclusion in Phase 2 of a sample of households containing children but with no child engaged in work-related activity.

The present design does not provide sufficient information of this type.

On the other hand, the primary concern of the survey is with working children and a major part of the resources should be devoted to surveying them. Most of the survey objectives listed in the survey report are concerned with the condition and circumstances of working children. Of course, non-working children also need to be covered to provide a "control group", but it is not economical – in a child labour survey as distinct from a general survey of children – to devote too much of the survey resources to that aspect.

Target population and coverage of the survey

South Africa's SAYP was a household-based survey, and data were collected in face-to-face interviews with respondents. The target population of Phase 1 was all children aged under 18 who normally resided in a private household within the country. Consequently, it excluded children who did not live in a household, for example street children and children living permanently in institutions. Coverage rules for the survey were that all children who were *usual residents* should be included even if they were not present at the time of the survey. This meant that most boarding school pupils were included in their parents' household. Activities of children under the age of five were not examined in this survey, because they were thought to be too young to answer the relevant questions.

A *household* was defined as consisting of "a single person or a group of people related or unrelated who usually live together for at least four nights a week, who eat together and who share resources". If a usual household member had been absent for more than 30 days he or she was not considered to be part of the household. Guests and visitors who had stayed for 30 days or longer were counted as household members. A household might occupy more than one structure. People who occupied the same dwelling unit, but who did not share food or other essentials, were regarded as constituting separate households. Domestic workers living in separate quarters, or who were paid a cash wage by the main household (even if they had most of their meals with the household) were regarded as constituting a separate household.

The survey was conducted throughout the country, in both urban and rural areas, in all nine provinces. The sample population excluded all inmates of prisons, patients in hospitals, boarders in boarding schools and individuals residing in boarding houses, hotels and workers' hostels. Families living in workers' hostels were, however, included. Single person households were screened out in all areas before the sample was drawn, as they contained no members of the target population (children aged 5–17) and were therefore of no interest to the survey.

Comments

1. Note that the target population is "all children" rather than the total population residing in private households. This means that households not containing any children are, as such, of no interest to the survey. Such households can be excluded from the survey if information has been collected during the household listing operation to identify them before sample selection. Otherwise, they have to be subject to the sampling process, and identified and eliminated during the field interviewing.

   The SAYP excluded a majority of households without children by identifying and rejecting single person households at the listing stage. This is a cost-effective strategy.

2. In child labour surveys it is generally preferable, as in this survey, to use the de jure (rather than the de facto) coverage definition. This is to facilitate coverage of household members who may be temporarily staying away from the household for reasons to do with work.

3. If one of the objectives had also been to provide some estimates of the economic activity of the entire population, then it would have been necessary to include in the sample households without children as well.

2.8.2 Content

Questionnaires

Two questionnaires were developed for the survey, one for each phase.

Phase 1

The Phase 1 questionnaire was used for screening purposes and covered basic information on household characteristics and all members of the household. An

adult responsible for children in the household (usually female) was asked to provide the information requested in this phase. This included the basic demographic information about the household, such as ages of household members, family relationships between household members, highest level of education, household income and economic and non-economic activities of children. More specifically, the questionnaire covered the following topics:

- living conditions of the household, including the type of dwelling, fuels used for cooking, lighting and heating, water source for domestic use, land ownership, tenure and cultivation;
- demographic information on members of the household, i.e. both adults and children; questions covered the age, gender, marital status, level of education and relationship of each household member;
- migration of the household in the two years prior to the survey;
- household income;
- school attendance of children aged 5–17 years;
- among the children aged 5–17 years, information on economic and non-economic activities in the 12 months prior to the survey.

Phase 2

The analysis of economic child labour activities in the main survey report was drawn from questionnaires administered in households selected after the Phase 1 screening process. The Phase 2 questionnaire was administered to the sampled subset of households in which at least one child was involved in some form of work in the year prior to the interview (with sub-sampling of such households if there were too many among the primary sampling units (PSUs) in Phase 1). An adult responsible for children in the household (usually female) was again asked to provide information on the economic status and income of adults in the households, as well as certain basic information on children (with more detail about younger children aged five to nine years). Next, children themselves were asked to provide most of the information about their activities, with limited overlap with questions answered by the adults. The analysis in the main survey report is based, when available, on the information supplied by the children themselves. It is believed that in a vast majority of cases these questions could be answered by the children.

One or more adults in the household answered questions on the following points:

- the employment status of all adults in the household aged 18 years or more;
- details of the type of work in which the employed adults were engaged;
- income earned by each adult in the past 12 months;
- type of work-related activity (if any) engaged in by each child aged 5–17 years in the household;
- reasons for the child/children to engage in these activities;
- school attendance and problems at school in respect of children aged 5-9 years;

- safety and health, illness and injury connected with work-related activities, in respect of children aged 5-9 years.

Each child in the household aged between 5 and 17 years was asked to answer questions on the following points:

- whether the child was engaged in work-related activities during the preceding year and during the preceding seven days and the type of activity (if any);
- details of the type of work, sector and occupation, in respect of children engaged in economic activity for pay, profit or economic family gain;
- times of the year and times during the day when those involved in these activities worked;
- reasons for engaging in these activities;
- conditions under which the work took place, including whether the child (if a paid employee) experienced sexual harassment or abuse at work;
- income earned and proportion of earnings paid to adults in the household; safety and health, illness and injury related to economic activity (asked only of children aged 10 to 17 years);
- whether the child was looking for work;
- main activity of the child;
- school attendance and (if attending school) difficulties experienced at school;
- reasons for missing school or not attending school.

Comments

1. The information that was collected in Phase 1 on the characteristics of households and individual children in households where no one was engaged in child labour is quite limited (for instance, the labour force status does not seem to have been included). More such information would have been useful for the analysis of determinants of child labour.

2. In the present design questions relating to the economic activity of children are critical as they identify cases for enumeration in Phase 2. There are a number of reasons for the trend towards under-enumeration of working children in general household surveys.

3. Thus, the sample for the LCS (i.e. Phase 2) is conditional on successful screening to identify households with labouring children, identified primarily in terms of usual economic activity (including significant household chores) rather than of a more comprehensive definition of child labour. This is quite a common practice.

   The total exclusion of households not reporting a working child prevents evaluation of the quality of the screening operation. The positive points are (1) that the performance of significant household chores was also included as a criterion, and (2) that in the households identified all children – and not only those identified previously as economically active – were covered in Phase 2 (the LCS component).

Timing and reference period

The survey was conducted in June and early July 1999. The following reference periods were used:

- For most of the questions concerning the children's economic activities for pay, profit or economic family gain (defined below), the reference period was the 12 months preceding the interview.
- A few questions were asked about the children's current economic activities for pay, profit or economic family gain. The reference period for this section was the seven days prior to the interview.
- For most questions regarding non-economic activities, collecting of wood and/or water and unpaid domestic work (defined below) the reference period was the seven days prior to the interview.
- Some questions regarding economic and non-economic activities referred to "usual" activities, without giving a reference period.
- A few questions regarding health and safety risks (such as incidence of injury) had an open-ended reference period. The following are two examples: "Have you ever been injured while doing any of these economic activities for pay, profit or economic family gain?" and "Do or did you have to do heavy physical work?"

Comment

The survey fieldwork was conducted during a relatively short period of a few weeks. Such an arrangement is common, and can have advantages in terms of survey cost and the quality of the data collected. However, the fieldwork needs to be stretched out if there are important seasonal effects to be controlled or captured.

Testing the methodology and the questionnaire: Screening test

It was necessary to obtain an overall idea of the proportion of households that contain children engaged in work-related activities, i.e. the anticipated "hit rate". This information was necessary for the design and planning of the pilot test as well as the main survey. This was achieved through a screening test, which was held in January 1999. The test was conducted in 28 enumerator areas (EAs), as follows: six urban formal areas; six urban informal areas; eight commercial farming areas; and eight other rural areas (primarily ex-homeland areas). EAs for the screening test were deliberately selected to reflect a variety of conditions. The following hit rates were found:

- 11 per cent with respect to economic activities for pay, profit or economic family gain; that is to say that, in 11 per cent of households, at least one child had been engaged in such activities in the 12 months prior to the interview; and
- 23 per cent with respect to unpaid domestic work, fetching wood and/or water, household chores and/or school labour (excluding households affected by the first hit rate). This means that, in 23 per cent of households at least one child had been involved in these activities regularly, for at least one hour a day, and no children were involved in economic activities for pay, profit or economic

family gain. These results indicated that the planned sampling strategy (based on a minimum hit rate of 20 per cent for both sets of activities together) was broadly on target. It was agreed to increase the sample size in urban informal areas and in commercial farming areas. The screening test also indicated that, because the hit rate was often higher than 20 per cent, sub-sampling would be necessary in some areas during the second phase of the survey to limit the achieved sample size.

Comment

In many situations, insufficient information is available on the proportion of children who are engaged in child labour. Such information is crucial in planning and controlling the sample size of the labouring children's survey. The South African survey provides an excellent example of good practice. "Screening tests" of the above type can be useful for improving the information available for survey design. Even so, adjustments have sometimes to be made to the sample selection rates at a later stage in an attempt to achieve the planned LCS sample size.

2.8.3 Sampling aspects

Sampling frame and stratification

The sampling frame used for the selection of areas was based on Enumeration Areas (EAs) of the 1996 Population Census. For that census the whole country was divided into about 86,000 EAs, grouped into sixteen EA types.

The enumerator areas were explicitly stratified by province. Within a province they were further stratified by the four area types constructed by consolidating the 16 EA types used in the 1996 population census. The four area types were formal urban, informal urban, other rural, and commercial farming areas, as defined below. EAs consisting solely of institutions were excluded.

Urban areas are areas that have been legally proclaimed as urban. These include towns, cities and metropolitan areas. Urban areas are divided into:

- *urban formal areas,* consisting mainly of dwellings made of formal building materials such as brick; and
- *urban informal areas,* consisting mainly of shack dwellings made of informal materials such as cardboard or corrugated iron.

Non-urban areas include commercial farms, small settlements, villages, traditional lands and other rural areas which are further away from towns and cities. Non-urban areas are divided into:

- *commercial farming areas,* consisting of areas with farms which sell most of their produce for a profit;
- *other rural areas,* consisting of most non-urban areas other than commercial farming areas. These are found mainly, but not exclusively, in the former "homelands".

Most EAs had fewer than 100 households. Small pockets of land consisting of at least 100 households each, called primary sampling units (PSUs), were constructed from the EAs. A PSU was taken to be: either a single enumerator area from the 1996 population census, if it contained at least 100 households; or otherwise, a combination of adjacent EAs, if they contained fewer than 100 households. This was to meet the requirement of a minimum of 100 households in any PSU.

Comment

Such exhaustive and uniform grouping of areas in the whole frame should be introduced only if it can be done cheaply and without too much difficulty as an office operation. Otherwise cheaper alternatives should be explored. Several such procedures are available in text books on sampling.

Sampling scheme

Two-phase sampling

Given the two-phase data collection, the sample for the survey was also drawn in two phases.

In the first phase, 900 PSUs each consisting of at least 100 households were selected using probability-sampling techniques. Of these PSUs, 579 were situated in urban areas (372 "urban formal"; 207 "urban informal"), and 321 were situated in non-urban areas (180 "commercial farming"; 141 "other rural").

The sample size was disproportionately allocated to the explicit strata by using the square-root method (i.e. the sample allocated in proportion to the square-root of the stratum population size). The allocated number of PSUs was systematically selected with a probability proportional to size in each explicit stratum (the measure of size being the number of households in a PSU). Within the strata the EAs were classified by magisterial district, and within that by EA-type and then area type – thus providing "implicit stratification".

Listing

Before selecting households in Phase 1, all dwellings and households within the sampled PSUs were listed prior to the fieldwork. The ultimate sampling unit was the dwelling. Formally, this unit was defined as follows.

> "A *dwelling unit* is any structure in which people can live. A household can occupy one or more than one dwelling unit. Conversely, more than one household can occupy one dwelling unit. Moreover, any structure or part of a structure which is vacant but can be lived in is also a dwelling unit. Any structure under construction, which may be lived in, is also listed as a separate dwelling unit. A dwelling unit may be a house, flat, hut, houseboat, etc. where a household lives or can live."

Special dwellings, not privately occupied by a household, were not considered as dwelling units, and were therefore not taken into consideration in the selection of households. Special dwellings included areas for patients in hospitals, inmates in prisons and reformatories, individuals in boarding houses and hotels, inmates in

homes for special-care citizens (e.g. disabled, aged, etc.) and boarders in boarding schools, provided that meals were served from a common kitchen. Workers' hostels with a communal kitchen were also treated as special dwellings. However, if parts of such hostels housed self-catering families, households in these parts were listed and included for purposes of selection. Dwellings in South Africa, particularly in rural areas and informal settlements, do not necessarily have addresses. It was therefore important to make a complete listing of households in each selected PSU, to ensure that the same household could be identified on a map and on the ground and re-visited.

Selection of ultimate units

The ultimate sampling units were occupied dwellings. Where multiple households were found in a selected dwelling, all of them were interviewed.

For Phase 1, within each PSU in urban areas 25 households were interviewed, while within each PSU in non-urban areas 50 households were interviewed for screening purposes. These households were selected by means of systematic sampling. The Phase 1 interviews were conducted in 26,081 households throughout the country, where information was gathered on all children between the ages of 5 and 17 years inclusive.

For Phase 2, a systematic sub-sample of at most five households in urban strata and at most ten households in rural strata was drawn from among those households where there was evidence of at least one child per household being engaged in child labour. This automatically confined the sample to PSUs with at least one child who worked. The Phase 2 interviews were conducted in 4,494 of the eligible households. During this phase information was gathered on about 10,000 children between the ages of 5 and 17 years.

Screening of households for the Phase 2 questionnaire

As indicated above, only households with at least one child engaged in work-related activities qualified for selection for the follow-up interviews. During Phase 1 households were identified with children aged 5–17 years, any of whom were engaged in any kind of work. After the interview the enumerator had to allocate activity codes to each household:

Activity code 1: Households with any child engaged in any activities for *pay, profit or economic family gain.*

Activity code 2: Households with any child only engaged in fetching water and/or collecting firewood, and/or in housekeeping and/or in helping at school for more than one hour a day.

Activity code 3: Households with no child engaged in any of the activities related to work, or households with no children.

The survey's major objective was to examine the characteristics of households with children engaged in activities for pay, profit and/or economic family gain. The second major objective was to examine the characteristics of those households with children

who were engaged only in housework, work at school, fetching water and/or collecting firewood. Thus the major interest in Phase 2 was to interview households with activity code 1 and the second interest was those with activity code 2. If a household had a child or children engaged in activities under both code 1 and code 2, code 1 was given preference.

Comments

1. In this survey, sampling from Phase 1 to Phase 2 is at the household level. In each sample area, the Phase 2 sample size is up to a maximum of 20 per cent of the Phase 1 sample size, except that the Phase 2 sample includes only households with at least one working child while Phase 1 covers all multi-person households. It follows that all Phase 1 sample areas are retained in the Phase 2 sample, except for areas with not a single working child.

2. An alternative sampling strategy would be to reduce the number of areas in the Phase 2 sample by introducing sub-sampling from Phase 1 to Phase 2 at the *area level*. (If required, the rate of sub-sampling households within sample areas can be correspondingly increased to retain the desired Phase 2 household sample size.) This alternative design can have the advantage of lower cost and improved quality control, since the resulting sample would be less scattered. Also, with an appropriate sampling scheme, the number of areas with very few households of interest can be reduced in the Phase 2 sample. Concentrating the sample into fewer areas of course increases sampling variance and design effects, but the alternatives mentioned above are likely to improve the cost efficiency of the design.

Selection of households for the second phase questionnaire

The quotas for Phase 2 interviews were five interviews in urban areas, and ten interviews in rural areas.

The five households in urban areas were selected as follows:

- If there were five or fewer households with activity code 1 or 2, then all these were in the Phase 2 sample and there was no need to sample.
- If there were more than five households with activity code 1 or 2 but three or fewer with activity code 1, then all activity code 1 households were included and the households with activity code 2 were sampled to bring the total to five households.
- If there were more than five households with activity code 1 or 2 but two or fewer with activity code 2, then all activity code 2 households were included and the households with activity code 1 were sampled to bring the total to five households.
- In all other cases households with activity code 1 were sampled to obtain a sample of three households and those with activity code 2 were sampled to obtain a sample of two households, making a total of five households.

The ten households in rural areas for Phase 2 were selected in exactly the same way as above, except that the numbers of households involved were doubled everywhere.

Comments

1. The objective of the above sampling scheme is to control the composition of the resulting sample in every area in terms of the household activity code.

2. The household sampling procedure is rather complex, relying on a lot of information and detailed classification of the lists of households. Such rigidity is often not necessary or useful, and is always cumbersome. With some flexibility in the size and composition of the sample from individual areas, alternative schemes can be devised to obtain probability samples which still yield very good control over size and composition of the total sample. This becomes easier to achieve when there is sub-sampling from Phase 1 to Phase 2 also at the area level.

3. Sample weights must be computed for households in each category in each area. This requires the denominator for the sampling rate – the number of actual households (excluding blanks) in that category in the list. This is demanding from the standpoint of the information required, error prone, and in any case time consuming. Simpler and better methods are possible.

2.9 Jamaica Youth Activity Survey 2002: Description and comments

The following information has been extracted from the *Report of Youth Activity Survey 2002*, Statistical Institute of Jamaica. This is an example of a modular survey whose size, design and timing are entirely determined by those of the parent survey to which it is attached.

2.9.1 Objectives and content

Survey objectives

The Jamaica Youth Activity Survey (YAS) was conducted nationally as a *module* of the April 2002 labour force survey. In all households interviewed for the labour force survey, when children in the 5–17 age group were found living in the household, these eligible children were interviewed for the Youth Activity Survey.

The overall objective of the YAS was to determine the character, nature, size and reasons for child labour in Jamaica, and the conditions of work and their effects on the health, education and normal development of the working child. Specifically, the survey obtained information regarding:

1. demographic and socio-economic characteristics, such as level of education and training (enrolment and attendance), occupation and skill levels, hours of work, earnings and other working and living conditions;

2. characteristics of the sectors where children were working: public/private sector;

3. where and how long the children had been working and factors causing children to work and/or families to put children to work;

4. the perceptions of the parents/guardians, children and employers about child labour, regulations, laws and legislation, etc.;
5. participation in programmes with a positive impact on the elimination of child labour;
6. status of working children's health, welfare and education situation.

Target population

The target population for the survey was the population of children in the 5–17 age group living in private households. Children living in institutions such as hostels, hospitals and "places of safety" were not covered in the survey. This target age group was selected in the light of the following considerations:

- Upper limit: the United Nations Convention on the Rights of the Child (1989) and the ILO's Worst Forms of Child Labour Convention, No. 182 (1999), defined children as those below 18 years of age.
- Lower limit: information on children prior to the compulsory age of schooling was needed, but the cut-off at five years of age was chosen because there would be very few working children below that age.

Comment

Despite the wide scope of the information sought, the focus of the survey is clearly on child labour according to the stated survey objectives.

2.9.2 Survey interview

Information for the Youth Activity Survey was obtained from interviews with both adults and the relevant child/children. In the case of the adult, the respondent of preference was the parent/guardian of the child. Adults were asked questions about the economic and non-economic activities of each child, as well as general questions about the household. The children were asked questions on their economic and non-economic activities. Interviewers were required first to complete the labour force survey questionnaires for the household, and then to complete the youth activity survey questionnaire for the eligible children.

The questionnaire was designed to gather detailed information specifically on children 5 to 17 years of age inclusive. Information was collected on their education, present/past economic and non-economic activities, occupation, industry, income, hours of work, etc. The questionnaire was divided into 9 sections as follows:

Section 1: Household roster of children 5–17 years old

Section 2: Education and vocational training

Section 3A: Current economic activity and Section 3B: Usual economic activity

Section 4: Current non-economic activity

Section 5: Work-related health and safety issues

Section 6: Parent's perception of child

Section 7: Migration status of child

Section 8: Household characteristics

Section 9: Questions to be addressed to children 5–17 years of age.

Responses for sections 1-8 were to be provided by the parent/guardian of the target child; if that person was unavailable, then another responsible adult over 17 years of age was selected. Questions in section 9 were to be addressed to each child in the 5–17 age group.

The first section of the questionnaire that the interviewer completed was Section 1 – the household roster. This listed the names and ages of all eligible children in the household, which copied by the interviewer from the labour force survey. All children were listed on the roster before the start of the interviews, thus ensuring that no eligible child in the household was excluded from the survey.

For each eligible child, interviews with the adult were repeated for sections 1–7. When all these interviews were complete, the housing information – Section 8 – was collected from the adult.

The eligible children were then interviewed, with each child being asked the questions in Section 9. Where possible, the children were interviewed in order of age, following the same order as the listing on the household roster. Hence the questionnaire provided responses from both parents/guardians and the children themselves. However, there were large discrepancies between the two when similar questions were asked, which may be in part be due to the age of the children. Therefore, the survey report used mainly the questions addressed to the parent/guardian, except when a question was asked only of the child.

Comment

Note that, apart from the questions related specifically to child labour, the module is applied to all children in the specified age group. Thus, the module may be described more appropriately as a "child module" rather than as a "child labour module". This arrangement has the advantage that it collects comparable information on all children, so that those working can be compared with those not working in order better to identify the conditions and consequences of child labour. Equally importantly, it allows exploration of the whole range of activities of children, not simply what is termed economic activity. Particularly important is the substantial engagement of children in household chores.

The major drawback of this approach is that a disproportionately large share of effort is spent in surveying non-working children, at the expense of working children who are the primary target group. This is particularly so when only a small proportion of the children are engaged in child labour.

Such information may be extremely valuable in its own right, and the above comments are meant to apply only to the relevance of the data to the study of child labour.

2.9.3 Sampling aspects

The Jamaica labour force surveys' sample design

Jamaica's labour force surveys are an integral part of the continuous household survey programme of the Statistical Institute of Jamaica and have been conducted on a regular basis since the 1960s. The surveys are conducted each year on a quarterly basis - in January, April, July and October, with the usual reference week being the last full working week of the preceding quarter. They are conducted in a 1 per cent sample of private households, which are selected using a stratified random sample design. Excluded from the sample are non-private households including group dwellings, e.g. military camps, boarding schools, mental institutions, hospitals and prisons.

In Jamaica's labour force surveys the labour force consists of persons 14 years old and above who are working, looking for work or would like to work and are not at school full time. Interviews of children under 14 years of age are limited to collecting data on age, sex and relationship to the head of the household.

The sample design for the 2002 Youth Activity Survey was based on the design for the 2002 labour force survey (LFS). The basic design for the survey was a two stage stratified systematic sample design with the first stage being a selection of areas called Enumeration Districts (EDs) and the second stage a selection of dwellings within the selected EDs.

First stage of selection

The main focus of the LFS design was to draw a representative sample that was distributed proportionately between the urban and rural areas on the basis of the number of enumeration districts (EDs). The ideal situation is to select the sample based on the proportion to the size of the population but this information was not available. Population data from the 1991 population census had become obsolete owing to the number of demographic changes in the country and data from the 2001 population census was not yet available.

The sample frame used for the first-stage selection was a list of enumeration districts (EDs) stratified into urban and rural strata. This list was quite up-to-date as it was modified prior to the 2001 population census. For the second stage a listing of the selected EDs was established before the survey. For this exercise, every dwelling unit in the selected EDs was visited by an interviewer and relevant information pertaining to the composition and description of the dwelling unit was collected. The EDs are stratified into 2,543 urban and 2,693 rural EDs. Each selection was made on a parish basis without any overlapping and was delineated in such a way that they encompassed the entire island. EDs were then placed in primary sampling units (PSUs), which were defined as having a minimum of 80 dwellings. The second stage was a selection of dwellings also using a systematic sampling method from a current list of dwellings within the PSUs.

The first-stage selection process adopted a proportionate allocation method in which a 10 per cent sample of EDs was selected from each parish and allocated proportionately to urban/rural strata. The selection was based on the following procedure.

1. The enumeration districts (EDs) were grouped by urban and rural strata within each parish.

2. The numbers of EDs that were to be selected in the urban and rural strata within each parish were computed as:

$$m_{hu} = m_h . \left(\frac{M_{hu}}{M_h} \right) = m. \left(\frac{M_{hu}}{M} \right),$$

$$m_{hr} = m_h . \left(\frac{M_{hr}}{M_h} \right) = m. \left(\frac{M_{hr}}{M} \right),$$

where

m_{hu} = the number of EDs to be selected, out of a total of M_{hu} from the urban (u) stratum of parish h

m_{hr} = the number of EDs to be selected, out of a total of M_{hr} from the rural (r) stratum of parish h

$M_h = M_{hu} + M_{hr}$, $m_h = m_{hu} + m_{hr}$, total EDs and the total number selected from parish h

$M = \Sigma_h M_h$, $m = \Sigma_h m_h$, total EDs and the total number selected from the whole country.

3. The EDs in the urban and rural strata of each parish were selected using a systematic procedure.

Second stage of selection

In the second stage of sampling a listing was made of all the dwellings in the selected EDs from the first stage. The list was made circular and a sample of 32 dwellings was selected systematically. After selection they were then placed in panels of equal size from which 16 were used for the survey.

Comments

1. Note that this procedure does not result in a self-weighting sample of households or persons. While the EDs are selected with a constant probability throughout the country, households within any ED are selected with a probability inversely proportional to the total number of households in the ED. Hence households in larger EDs receive a lower chance of selection than households in smaller EDs. The difference will be small only if the EDs are of uniform size.

2. Note that the differences in average ED size between domains, e.g. between urban and rural sectors, will be carried over to similar differences in sampling allocation and rates.

3. Sample weights have to be computed from the listing. The number of households listed, excluding blanks, must be known and recorded for the purpose.

Modular child labour survey

The sample for the Jamaica Youth Activity Survey (YAS) was determined entirely by that of the LFS, to which it was attached as a module. Apart from the addition of this module to the LFS questionnaire, only the following modifications were made to the given LFS procedures.

1. The 2002 YAS was attached to the April round of the 2002 LFS. Due to the fact that the YAS was a module of the LFS, the reference period was changed to accommodate the YAS. The reference week used for both surveys was Sunday, March 17, to Saturday, March 23, 2002 and not what would have been usual – March 24–30, 2002. This was changed because the children would have been on Easter vacation from school during the usual reference week and this would have affected responses to the questions on school attendance during reference week.
2. Special provision was made for controlling the distribution of the population (5–17 age group) in the post-stratification of the sample weights, as noted below (Section 2.9.5).

Comment

A very important issue is whether, and if so the extent to which, the size, design and operations of the parent survey were modified to make them better suited to the requirements of the child labour module. The scope for making modifications for this purpose is more limited when the parent survey, in this case the April round of the 2002 LFS, forms an element of a regular time-series.

Even when other aspects of the design remain unchanged, the addition of the child labour module increases the length of the interview involved, which may affect the quality of the data collected in the parent survey. This limits the scope of the additional information which can be collected in the child labour module.

Sample weighting

The major purpose of weighting is to adjust for differential probabilities of selection used in the sampling process, to reduce the sampling variability and to compensate for any under-enumeration or non-response in the survey.

Sampled dwellings that were in the scope of the survey, but where no collection of information was possible because they were vacant, closed or the occupants refused to cooperate during the time of the survey, were classified as non-response. The purpose of this adjustment was to reduce the bias arising from the fact that non-respondents are different from those who respond to the survey.

Post-stratification is a method used to adjust the weighted survey estimates so that they agree with the population estimates. It is normally used to compensate for different response rates by demographic subgroups, increase the precision of survey estimates and reduce the bias present in the estimates. The method that was used was a simple ratio adjustment that was applied to the sampling weight using the projected 2002 population total for each strata defined by age and gender for each parish.

The method of weighting for the YAS ensured that the estimates conformed to the distribution of the population (5–17 age group) by age, sex and parish.

Comments

It is an error to consider vacant and unoccupied dwellings as non-responses. These are blanks to be disregarded. They are quite different from refusals and other cases of genuine non-response, for which compensation by weighting is appropriate.

Sample size

The sample size of the YAS was determined by that of the LFS.

The survey report notes that the sample was too small to facilitate inferential statistical analysis. Though in regular labour force surveys there is some information about child labour among children 14 years of age and older, there is little information regarding the younger age groups. In 1994 a preliminary study estimated that about 23,000 Jamaican children between 6 and 16 years of age were engaged in child labour activities. For a 1 per cent sample, this would amount to around 230 sample cases. However, in the final YAS sample of 6,189 children, only 143 reported that they worked during the reference week. Thus the sample of working children was small and limited the range of analysis.

Furthermore, the results regarding children's involvement in economic activity must be interpreted with caution. As noted above, while the questionnaire provided responses from both parents/guardians and the children themselves, there were large discrepancies between the two when similar questions were asked, which may be due in part to the age of the children.

Comments

1. The proportion of working children in the main survey turned out in practice to be much smaller than that reported in the pre-test. Compare the above figures for the main survey with those of the pre-test as noted in the survey report:

   > "In the pre-test there were a total of 340 children from 162 households interviewed in the four parishes of Kingston, St. Andrew, St. Elizabeth and St. Catherine. …. Approximately 10.5 per cent of the children in the pre-test sample worked during the past 12 months. Of the 36 children who worked, 22 were employed in the wholesale and agriculture industries. The number of children who worked during the past week was 17, or 5 per cent of the sample. Of these, 11 were employed in the wholesale and agriculture industries. The majority of children – 75.7 per cent – helped with household chores in the pre-test households, although a significant percentage – 23.1 per cent – reportedly did not."

   Apart from the sampling variability in these figures, this difference between the pre-test and main survey results may have resulted in part from the pre-test sample not being representative of the whole country. However, it may also have been caused by the greater risk of omission in an operation of much larger size.

2. While the available sample size of the child labour module is determined by the sample size of the parent survey, the situation can be ameliorated – when, as in the present case, the latter is a regular survey – by repeating the module in more than one round of that survey.

3. In the analysis it may also be useful to consider a wider definition of child labour: to include not only children engaged in economic activity in the strict sense, but also children providing appropriately defined "substantial" help in household chores.

Chapter 3
Sampling for a typical population-based survey

3.1 Introduction

The *collection of information on child labour in conjunction with a broad-based survey of the general population* such as the labour force survey (LFS), may very often take the form of child labour questions attached to the base survey as a module. Alternatively, the collection of the two types of data may involve separate enumeration, but based on a common design – usually the *child labour survey* (CLS) – forming a sub-sample of the base survey.

While it is not the objective here to discuss general principles of sample design applicable to a LFS or a similar CLS, this chapter seeks to *clarify the sampling procedures involved in the most common type of designs used for such surveys*. This is important for gaining an understanding of how different arrangements for the measurement of child labour may be linked to or derived from such a design.

Sections 3.2 to 3.4 deal with some important practical aspects of sample design and the sampling frame, including stratification and multi-stage sampling. The important but complex issue of the choice of sample size is discussed at some length in Section 3.5. Sections 3.6 to 3.10 discuss practical procedures for sample selection, including systematic sampling with probability proportional to size (PPS) in the presence of imperfect size measures, and dealing with very large and very small primary sampling units (PSUs). Finally, Section 3.11 provides a numerical illustration of the procedures, based on some simulated data.

3.2 Sample design: Practical orientation

The objective of this chapter is to point out some good sampling practices in the design and implementation of surveys of child labour. We must begin from some basic principles. The important requirement is that, at least as concerns the standard household-based child labour surveys, they must be based on "probability and measurable samples" as described below.

3.2.1 Reasons for sampling

Sample surveys are one of the main sources of statistical information. Their importance is further increased as a result of the lack of alternative sources of information, such as administrative records and registers in developing countries. A sample survey means that information is obtained only on a subset of units in the population, from which inferences are drawn about the whole population. Clearly, the validity of these inferences depends on the manner in which the sample is drawn and the size of the sample.

Compared to complete or census coverage, the sampling method can have many advantages. These advantages arise from the fact that restricting the collection of information to a sample reduces the scale of the operation. Consequently, the cost is reduced and the results can be produced more quickly; more elaborate information can be sought; and more intensive methods which can provide more accurate data can be employed. Another major advantage of sample surveys is their vastly greater flexibility in comparison with complete censuses: flexibility in terms of timing, coverage, size, content, and other aspects of design and methodology.

The principal limitation of the sampling method is the sampling variability to which the survey estimates are subject. Above all, this limits the extent to which the survey results can be disaggregated geographically, over population subgroups, or over time because of the sampling variability resulting from limited size. This is seen particularly clearly in the inability of most sample surveys to yield direct estimates for local areas and other small domains.

It is also important to emphasize that the well-known advantages and limitations of the sampling method compared to the census apply also when we consider samples of different sizes. Smaller surveys can be cheaper, more timely, more intensive in procedures, richer in content, able to provide more accurate measurement, and repeatable over time. On the other hand they are subject to larger sampling errors, which limit the extent to which the estimates produced can be relied upon, and especially the extent to which the survey results can be disaggregated geographically, over population subgroups, or over time. We will discuss the fundamental question concerning the choice of sample size later in this chapter.

The design of samples is a highly technical task. In outline, it involves determining:

1. the sample size, i.e. the number of units of analysis to be selected for the survey;
2. sample structure, i.e. how those units are to be selected;
3. estimation procedures, i.e. how the results from the sample are to be used to draw inferences about the entire population of interest 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. *The size and structure of the sample determine the magnitude not only of the sampling error, but of most other components of the error as well; and of course it also determines the cost of the survey.* For instance, if one tried to minimize the sampling error by conducting a face-to-face interview survey on a random sample of persons throughout the country, travel costs may turn out to be prohibitive and supervision of the interviewers' work difficult. Also, depending on the situation, good and complete lists of persons for sample selection may not be available, and failures to locate the individuals selected in the field may be common. All these factors often result in greatly increased costs and errors.

It is true that non-sampling components of the survey error depend also on many factors other than sample size and structure. However, it is a mistake – unfortunately a

common one – to assume that the primary impact of sample size and structure is only on the magnitude of sampling error. Clearly, at least beyond a certain limit, *increasing the sample size can adversely affect all aspects of data quality*, including the relevance of the survey (by restricting the amount of information which can be manageably collected per case in a big sample) timeliness (delay in collecting and processing the increased volume of information), and of course its response quality (inadequate interviewer training, supervision and control due to the increased volume of fieldwork). An inappropriate sample structure (for instance, excessive geographical scatter of the sample, repeated interviewing and respondent follow-up, etc.) can have similar effects on data quality.

These considerations dictate the need for a thoroughly practical orientation in the design of samples. First of all, it is desirable to depend on methods and procedures that are reliable and practical in the particular circumstances, in order to ensure that the actual performance in the field and office do not depart too far from the requirements of the design. There is a long chain of operations from design to sample selection, field data collection, then back to coding, computing and statistical analysis. To complete that circuit in reasonable safety, *sampling procedures must be robust, i.e. insensitive to at least modest departures from the ideal*. It is preferable to adopt a design and procedures which, even if not the most precise and refined under optimum conditions, can nevertheless withstand the unexpected and unknown situations which are invariably encountered in practical survey work. Practical sampling must be rooted in the nature and inter-relationships of the field and office operations, not only in the implementation of the particular sample design but also in the data-collection stage of the survey as a whole. The sampler must remain constantly aware of what can and cannot be expected of the field and office workers actually involved in implementing the sample.

The need for practical orientation does not mean that sampling theory is unimportant. On the contrary, it is not possible to be a good practical sampler without having a good understanding of the underlying theory which guides practical work.

3.2.2 Probability and measurable samples

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 randomized 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 design of a random sample specifies the type of randomized procedure applied in sample selection. It also specifies how the population parameters are to be estimated from the sample results. The selection procedure and the estimation procedure are two aspects of the sample design.

As to the selection procedure, many types of 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 greater chance than others. We may select the elements individually, or we may first group them into

larger *clusters* and apply the selection procedure to those clusters. We may partition the population into *strata* and apply any of the above procedures separately within each stratum. Each randomized procedure in fact determines a different set of samples which can in principle be selected using that procedure, and the chance of selecting a particular sample from among them. But, to be random, any selection procedure must ensure that every unit in the population receives a specified non-zero chance of appearing in the sample to be selected.

The estimation procedure involves the statistical or mathematical formulae in terms of sample values, and possibly also of information from other sources external to the sample; it provides *estimators* which are used to produce sample estimates of population parameters of interest. The procedure also includes the estimation of measures of uncertainty ("sampling error", "confidence intervals" etc.) to which the sample results are subject.

3.2.3 Obtaining a probability sample

As noted, any survey aimed at applying observations drawn from a sample to the whole population of interest has to be based on probability sampling. 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 must be represented explicitly or implicitly in the frame from which the sample is selected. For this we need a good *sampling frame* (Section 3.3).

2. The sample must be selected from the frame by a process involving one or more steps of automatic randomization, which gives each unit a specified probability of selection. This means specifying clear selection procedures prior to the actual selection of units for the sample, and not altering the sample after it has been selected.

3. 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.

4. 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. These are called *design weights*. Actually, it is often necessary to modify the design weights to reduce the impact of other shortcomings of the sample. *Sample weights* are the set of weights to be applied to each unit enumerated in the sample to obtain the corresponding estimates for the population. Procedures for computing sample weights are described in Chapter 7 below.

However, in practice, some approximations in the implementation of these ideal requirements are often necessary owing to, for example:

- the failure to include some units in the frame (under-coverage);
- distortions in selection probabilities due to other coverage and sample selection errors;

- failure to enumerate or obtain full information on all the units selected (non-response);
- the use of approximate procedures at the estimation stage, in particular failure to take fully into account the selection method actually used (estimation bias).

The level of error up to which a sample may still be considered a probability sample is a matter of practical judgement.

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 of arbitrary assumptions or subjective judgement.

3.2.4 Measurable samples

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. In other words, a sample is measurable if, from the variability observed between units within the sample, usable estimates of the sampling variance (i.e. of the variability between different possible samples) can be obtained. To be measurable, it is normally required that the sample be a probability sample; it should also meet certain other requirements in order to ensure that sampling variability can be estimated from the observed variability between units in the one sample that is available. Again, assumptions and approximations may be involved in the variance estimation procedures without necessarily losing the measurability of the sample in the practical sense.

The essential requirement is that the sample should be designed, selected and implemented, and information on its structure recorded, so that the sample can be divided into partitions or "replications" in such a way that (some function of) the observed variability between those replications yields a practically useful ("valid") estimate of the sampling variance. Some practical variance estimation procedures are discussed in a later chapter.

3.2.5 Sampling for a "typical" population-based survey

The following sections describe technical features of the most commonly used sampling design for the labour force survey (LFS) and similar population-based surveys. Surveys on child labour may be based on the same sample, or on a sample linked to or derived from such population-based surveys.

Most samples of households and persons are selected in a number of sampling stages. For instance, in a national household survey, the whole country may be divided into area units such as localities or census enumeration areas (EAs), and a sample of these areas selected at the first stage. The types of units selected at the first stage are called primary sampling units (PSUs). For the first stage of selection, a frame of PSUs is needed which lists the units covering the entire population exhaustively and without overlaps, and which also provides information for the selection of units efficiently. Such a frame is called the primary sampling frame (PSF). The second 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 second stage units (SSUs) from each selected PSU. This process may continue until a sample of sufficiently small ultimate area units (UAUs) is obtained. It is very common to use designs with only a one area stage – as is the case for instance with most of the child labour surveys carried out with the support of the ILO.

At the last stage, in each selected sample area (or UAU), individual households may be listed and a sample selected with households or persons as the ultimate sampling units (USUs). In the survey, information may be collected and analyzed for the USUs themselves; or for other types of units ("elements") associated with the selected USUs, such as individual persons or children within sample households.

The sample design involves the choice of the number of stages to use and, at each stage, the type of units, method of selection, and the sampling rate or the number of units to be selected. This requires a sampling frame that represents the target population.

3.3 The survey population: The sampling frame

3.3.1 The survey population

The definition of the population to which the sample results are to be applied is a fundamental aspect of survey planning and design. While basic decisions about the nature and scope of the population to be covered are taken early in the survey planning process, the content and extent of the population have to be specified more precisely at the stage of technical design. This specification is in terms of:

- population content, i.e. definition of the type and characteristics of the elementary units comprising it;
- population extent in space, i.e. the boundaries of its geographical coverage;
- and its extent in time, i.e. the time period to which it refers.

Some examples may be considered. In most labour force surveys, the population of interest is private households and all persons residing in them. A de facto or a de jure definition may be used for the coverage of members. In child labour surveys (CLS) aimed at measuring the proportion of children engaged in labouring activities, the target population is all children within a specified age bracket. In "regular" CLSs the coverage may be confined to children living in private households; correspondingly, households with no children are not in the target population. In "special" CLSs the coverage may be extended – or even confined, as in the case of street children's surveys – to children living outside private households. In labouring children surveys (LCS), aimed at studying characteristics and conditions of children engaged in labouring activities, the target population is children so engaged; however, this is normally supplemented by including in the sample some non-labouring children for the purpose of comparison.

Some general, but important, points in relation to the choice of the target 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 when drawing inferences from the survey results, and when 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 usually required to control these errors and assess their effect on the survey results. Their magnitude depends on the quality of the sampling frame and sample implementation.

3.3.2 The 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 the frame may be partly 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 more than one frame in combination to represent the same population more adequately. The use of multiple frames raises special issues in the accounting of the units' selection probabilities.

Area-based frames

In practice, the required frame is defined in relation to the required structure of the sample and the procedure for selecting it. In some more advanced countries, it is possible to select a sample of persons directly from population registers. More generally, however, especially in developing countries, the frame for household-based surveys consists of *one or more stages of area units, followed by lists of households or dwellings within selected ultimate area units:*

- The primary sampling frame (PSF) is a frame of the primary sampling units (PSUs) and must cover the entire population exhaustively and without overlaps. Following the first stage of selection, the list of units at any lower stage is required only within the larger units selected at the preceding stage.
- Possibly, a hierarchy may be established of secondary area-based frames - consisting of all the units at each stage which exist in each unit selected at the preceding stage - till a frame of the lowest or ultimate area units (UAUs) is obtained. Below the UAUs, the sampling process moves from areas to the listing and selection of individual dwellings, households or persons. It is common practice for household survey samples to involve only a single area stage, in which case PSUs are the same as UAUs.

- There may be explicit lists of the ultimate sampling units (USUs), such as dwellings or households within the selected sample areas.
- The elements for data collection and analysis in the survey may be the USUs themselves, or may be other units uniquely identifiable from the USUs through definite rules of association. For example, persons (elements) may be associated with selected households (UAUs) on the basis of a de facto or de jure coverage definition.

In most labour force and stand-alone child labour surveys to date, the primary sampling frame has been based on census enumeration areas, and the design has involved only a single area stage. Normally, fresh lists of households or dwellings are prepared within each selected area (PSU).

The durability of the frame declines as we move down the hierarchy of the units from PSUs to USUs. The primary sampling frame (and to a lesser extent, frames of intermediate level units) usually constitutes a major investment for long-term use. By contrast, in most surveys it is necessary to prepare fresh lists of USUs shortly before the survey enumeration. It is a major advantage in a survey if it can utilize lists prepared for the purpose of some other recent survey or census, such as LFS lists in the case of a CLS linked to a LFS. 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 different "durability" of different types of units is particularly important in the context of linked surveys where the second is based on a sub-sample of the first, as in the case of a child labour survey based on a larger labour force survey. Lists of different levels of units may be provided by the LFS and a sub-sample of that level of units selected for the CLS, with any further sampling below that level in the CLS proceeding independently of the LFS. For instance the CLS sample may be drawn from (1) all areas in the LFS sample, (2) only the areas containing households with children, (3) only areas containing households with working children, (4) actual lists of all households selected in the LFS, (5) only the households successfully enumerated in the LFS, (6) only households which contain children, (7) only households containing labouring children, (8) actual lists of children listed in LFS households, (9) lists only of children identified as workers, or, ultimately, (x) even information on the characteristics of such working children that has been carried forward.

Clearly the durability of the units involved in the LFS-CLS sample links, and consequently also the acceptable time lag between the two operations, declines as we move down the above list. (See Section 4.4 below for further comments.)

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 a failure to define and identify the physical boundaries of the area units correctly, or from the poor quality of the lists of the ultimate units such as dwellings or households within sample areas. Common imperfections of area frames include the following:

- *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, and scattered populations outside the listed localities may not be covered. Under-coverage also increases as the frame becomes outdated with time.
- *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 the boundaries of units on which other relevant information is available in the frame (for example, information on size and density of the population) or from the boundaries of the actual sampling units.
- *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.
- *Lack of auxiliary information.* Information on the size and other characteristics of the units, which is required for efficient sample selection, may be inaccurate or simply unavailable.
- *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 is to be used repeatedly for many surveys and survey rounds.

Problems with list frames

Problems can arise in the absence of one-to-one correspondence between *listings* (which are the units actually subject to the selection process) and the *elementary units* (of which obtaining a sample with specified probabilities is the actual objective). The lack of correspondence can arise in several forms.

Blanks: meaning 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 represented in the list is not known.

Blanks in the list *do not* indicate non-response. A common error is to treat blanks as non-response, and substitute other units for them or adjust the sample weights to compensate for them.

Duplications: meaning that the same unit is represented by more than one listing. Sometimes the problem arises from the nature of the frame – 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. 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. Steps have to be taken to compensate for or avoid such variations in selection probabilities; *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. In such cases painstaking work to eliminate all duplications in the list may be the only solution.

Clustering of elements: meaning that more than one unit may be represented by the same listing. As such, this does not distort the selection probabilities, since each unit receives the selection probability of the listing representing it. *Selecting one unit at random from the clustering is often unnecessary, but in certain situations it is unavoidable or even desirable.* A common example is the selection of one adult from each sample household for inclusion in the survey. *If done, the results have to be weighted to reflect the changed selection probabilities.* With such a design, a unit (e.g. persons) receives a probability in proportion to the number of units in its listing (household).

Under-coverage: meaning units 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 at random, but is *selective* in terms of the characteristics of the units. There is no simple or cheap solutions to the problem of under-coverage. The only advice that can be given is to spend a greater effort in the preparation or compilation of lists.

Failure to locate units: meaning the failure to identify which unit(s) a selected listing represents. 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 names of children). It can also be caused by insufficient effort by the field workers. *The failure to locate selected units which actually exist constitutes a 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.

3.4 Departures from simple random sampling: Stratification, clustering and unequal probabilities

3.4.1 Departures from simple random sampling

The simplest design is one in which every possible set of, say, S=1 to n units from a population of N units receives the same chance of selection. This is called a simple random sample (SRS). There are

$$\frac{N!}{(N-n)!n!}$$

such samples, and each receives a probability of selection equal to inverse of to the above number. In fact, as noted, in an SRS any set of s, $1 \leq s \leq n$, units receives the same chance of being in the sample as any other set of the same size. Different units

(which corresponds to s=1) all receive the same chance, the chance being n/N. Other designs depart from simple random sampling by:

1. suppressing some of the possible samples noted above, i.e. not allowing certain combinations of units to appear in the same sample, and/or
2. by giving different units (and hence different samples) different probabilities of selection.

The sample design may depart from simple random sampling in a number of ways, the three common and important ones being the following.

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 clustering or multi-stage sampling, and may differ from one stratum to another. The main objectives of stratification are to gain flexibility in sample design and allocation for different parts of the population and to increase the statistical efficiency of the design. (Among all the samples possible under SRS, only those containing a fixed or expected number of units from each stratum can result from this design.)

Clustering or multi-stage sampling

Just as stratification refers to partitioning the population before sample selection, clustering refers to the grouping of units.

Often it is economical and convenient to group the population elements into larger units ("clusters") and to apply the selection procedures to such groups rather than directly to individual elementary units. In practice, such clustering is in fact often the only option available because the individual elements are too numerous and widely scattered to be sampled directly. (Among all the samples possible under SRS, only those with the ultimate units confined to the units selected at the preceding stages can result from this design.)

The selection procedure may be more elaborate than simply selecting a sample of clusters. For example, some large units may be selected first; then each selected unit may be divided into smaller units and a sample of the latter selected; and finally, in each of the smaller units selected, a sample of individual elements may be selected. In this way we get a *multi-stage design*. The objective of such a design is to confine the elements appearing in the sample to larger units selected at the previous stage(s). This is normally done to reduce survey costs and improve control over the data collection operation in the survey.

Unequal selection probabilities

Sometimes there are reasons to select some classes of elements with higher (or lower) probabilities than others. For instance, certain domains, i.e. parts of the population such as urban areas or smaller regions of a country, may be over-sampled so as to improve the precision of their results. This can be particularly necessary in child labour surveys when the population of interest is unevenly distributed in the country. Normally,

unequal selection probabilities require weighting of the sample data at the estimation stage. Such weighting may also be introduced for other reasons, such as to compensate for non-response, or to calibrate (some) characteristics of the sample to match external standards (see Chapter 7).

Effect on variance

The effect of these departures from simple random sampling on variance may be seen as follows. With the SRS design of a given size n, all possible combination of any given number s,$1 \leq s \leq n$, elements are equally likely to appear in the sample. What clustering or stratification does is to suppress some of these possible combinations (and hence suppress the subset of samples which contain any of them). Unequal selection probabilities of units make some sample outcomes more likely than others.

Clustering tends to suppress relatively more of the "good" samples (that is, samples giving statistics close to the true population parameters being estimated) and to retain more of the "bad" samples which give results further from the true population values. Consequently, the variability between the retained samples tends to be larger, i.e. the magnitude of the sampling error is increased because of clustering.

The opposite is likely to be the case with stratification, which tends to suppress relatively more of the "bad" samples and retain more of the "good" samples. Consequently, the variability between the retained samples tends to be smaller, i.e. the magnitude of the sampling error is reduced because of stratification.

The effect of unequal selection probabilities is more complex. When these variations are essentially random, or only weakly related to unit characteristics, they normally introduce additional variability into the sample results. However, techniques such as "optimal allocation" and "calibration", which introduce unequal selection rates or weights, can sometimes be effective in reducing variance.

3.4.2 Stratification: some practical aspects

The purpose of stratification

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 segments of the population (strata) can be sampled differently, through the use of different sampling rates and designs. Although not essential to the idea of stratification, the separation may also be retained at the stage of sample implementation, estimation and analysis. It is common, for instance, 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 derive from the control it allows over sample design and selection within each stratum:

- 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.

- When data of specified precision are required separately for subdivisions of the population, it is desirable to treat each subdivision as a "population" in its own right and to select a sample of the required size and design from each independently. Stratification makes this possible.
- Sampling requirements and problems - as regards 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.
- 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.
- Stratification may also be introduced for administrative convenience; for instance, sample selection and implementation may be entrusted to different field offices, each looking after its own "stratum".

Stratification in practice

In practice, considerable care and effort are normally warranted in stratifying the list or frame before sample selection, for the following reasons.

Stratification often reduces sampling variance at little additional cost. Furthermore, the costs tend to be lower and the advantages greater in the stratification of higher-stage units in a multi-stage design, compared to the advantages of stratification of lower-stage units or in an element sample (see below). It is often desirable to pursue stratification to the limit, where only one or two PSUs are selected per stratum. Indeed, special techniques known as "controlled selection" can be employed to create even more strata than the number of units to be selected, linking the selections in different strata so as to achieve the required distribution of the sample.

- In so far as the samples are selected independently, and where they are of sufficient size, the results from the individual strata can be analyzed and presented separately. More commonly, the results are aggregated over several strata to produce estimates for major domains of the population. Efficiency is improved by defining strata as lying within (i.e. not cutting across) the reporting domains.
- A very important use of stratification is to provide flexibility in the choice of sample allocation, design and procedures in different parts of the population.
- Strata can provide natural partitions for organizing, controlling and phasing the survey work. Generally, stratification in no way complicates field operation at the data collection stage. Instead, any added complexity is confined to the operation of sample selection, which is usually more centralized and hence more easily controlled.
- Systematic sampling from ordered lists is a cheap and efficient means of achieving the effect of stratification. This procedure tends to be much simpler to implement than selection with the use of random numbers.

Stratification in multi-stage sampling

The argument for careful and elaborate stratification becomes much stronger when we consider multi-stage designs.

- The essential point is that the gain in precision by means of stratification is usually much more important in multi-stage sampling than it is in element sampling.
- Usually, much more information is available for the stratification of larger units, such as census enumeration areas or localities serving as PSUs and other higher-stage units in a multi-stage design.
- It is easier to stratify the larger, higher-stage units, which tend to be far fewer than the number of elements in the population.
- In so far as the number of higher-stage units selected is small, it becomes important to ensure that distribution of the sample is controlled. This is achieved by sampling separately within strata.
- In multi-stage sampling, it is more necessary, and also more feasible, to vary the sampling procedure in different parts of the population.

Stratification criteria

Below are presented several useful hints for the choice of criteria for stratification.

- Since stratification is carried out prior to sample selection, subjective choices can be made in determining the defining criteria, number and boundaries of the strata. Uniformity (i.e. using the same procedure in all strata) and objectivity (using pre-determined criteria and procedures not involving judgement) are not required at this stage. This stands in contrast to the need for objective procedures at the stage of actual selection so as to achieve a probability sample.
- Generally, 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.
- Stratification by unit size is useful when the units vary greatly in size. One objective of such stratification is to control the sample sizes, though special procedures such as "probability proportional to size" (PPS) sampling can also achieve this control to some extent. Another objective of stratification by size is to control the distribution of characteristics that are related to the size of the unit. This is useful even when PPS sampling has been employed to control sample takes within clusters. (See Section 3.6 for an explanation of the PPS sampling method.)
- In many situations, geographical, administrative and urban-rural classification is the most effective form of stratification. Such stratification is simple and requires little auxiliary information. It also tends to be suitable for surveys covering different topics.
- The above applies more clearly to the stratification of higher-stage units in multi-stage designs. Individual characteristics (such as household size, or gender and age of persons) can of course be additional effective stratification criteria when direct sampling of individual units is involved.

The most common type of stratification used for the selection of PSUs and other areas in household-based surveys of the general population is geographic: stratification according to the type of place (urban-rural, or several categories by degree of urbanization or size of locality); location (province, region or some other administrative division) and climatic or ecological zone. More complex systems of stratification are usually necessary for the sampling of units with special characteristics, particularly when the units of interest are unevenly distributed in the population.

Within sample areas, households may be stratified according to size, socio-economic status, employment of the head, etc., to the extent that such information is available. But in practice such stratification at the household (or personal) level can be very demanding in terms of the detailed data required, and hence expensive. Furthermore, its effectiveness in reducing variance can be very minor indeed. In typical multi-stage samples, the main gains tend to come from stratification at higher stages.

3.4.3 Clustering

Reasons for multi-stage sampling

In certain situations and for certain purposes, direct selection of elements in a single stage can be simpler and more efficient than selecting the sample in multiple stages. However, in most circumstances, especially for large-scale household surveys in developing countries, the direct selection of elements (households or persons) is not a feasible option. Multi-stage sampling is introduced for several reasons.

- By concentrating the units to be enumerated into clusters, it 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 the quality of the data collected.
- Administrative convenience in implementation of the survey can be another important reason.
- Selecting the sample in several stages reduces the work and cost involved in the preparation and maintenance of the sampling frame. Firstly, with multi-stage sampling a frame covering the entire population is required only for selecting the PSUs at the first stage; at any lower stage, a frame is required only within the units selected at the preceding stage. By contrast, for the direct sampling of elements such as households or persons, a complete list covering all elements in the population will be required. Secondly, frames where the units involved are large in size tend to be more durable and therefore usable over longer periods of time; lists of small units such as households, and especially of persons, tend to become outdated within a short period of time.
- The work involved in sample selection can also be reduced by multi-stage sampling. Dealing with a few hundred or few thousand area units in a country is, for example, easier than selecting a sample from lists with millions of households. It is easier to classify and stratify larger units, and usually much more information is available for this purpose.

Costs of clustering

The above advantages have to be balanced against various costs of introducing multi-stage sampling.

- The major cost of clustered or multi-stage sampling is the increase in sampling error compared with that in a simple random sample of the same size (i.e. with the same number of elements enumerated). The increase in variance varies according to the relative homogeneity of elements within the higher stage units, and the manner and number of units selected at each stage. If elements clustered together within a higher-stage unit are rather similar to each other, each of the units gives, in a sense, less new information than would be obtained if all elements were selected at random from the entire population. This tends to make the sample less efficient. The loss in efficiency will be higher if the number of elements selected per cluster is increased, or if the elements are more closely clustered together in compact units, or when the elements within the same cluster tend to be more homogeneous with respect to the variable of interest.
- There can also be some loss in flexibility in the sample design and in targeting the sample at populations with particular characteristics. This is because elements of different types are generally mixed up within higher-stage units, so that the selection of units of any given type cannot be controlled separately.
- Complexity of the design also increases the complexity of certain types of analysis of the survey data. This applies in particular to the estimation of sampling errors, which must take into account the structure of the sample.

Increasingly, in statistically more developed countries the balance is shifting in favour of direct element sampling, or at least towards reduced clustering of samples in household surveys. Contributing factors include:

- the increasing cost of the time needed for actually obtaining and analysing interview data compared to travel costs;
- the new modes of data collection such as telephone interviewing;
- the need to make samples more efficient so as to permit reduced sample sizes; and
- the availability of better frames permitting direct selection of elements.

Choice of unit type to serve as sample areas

With multi-stage sampling, the choice of type of area units to be used in the survey and the number of such units to be selected for the sample are important.

The choice of unit type has major implications for the sample, since the type of unit chosen to serve as the PSUs and other higher-stage units can greatly affect the survey quality, cost and operations.

First some general advice:

- It is neither necessary nor always efficient to insist on using units of the same type or same size as PSUs in all the population domains to be sampled.

- It is quite common for very different types of units to have the same administrative label. It is important not to confuse formal administrative labels with the actual type of units involved.
- The appropriate type and size of units depends upon survey circumstances and objectives.
- Nevertheless, the choice is constrained severely by what is available in the sampling frame.
- Since the appropriate choice depends on circumstances, no standard or single practice can be recommended.

Clearly, the type of units chosen to serve as the PSUs can greatly influence survey quality and cost. For an area sample, the relevant units need to be well defined, with clear boundaries. Good maps and descriptions for identification and demarcation are needed, together with up-to-date information on size and characteristics. The areas should cover the survey population exhaustively and without overlaps. Stability over time is another important requirement, especially if their use is to extend over a long period. The PSUs should be of an "appropriate" size, in line with the organization and cost structure of the survey data collection operation. If the units are too large, it may not be possible to include a sufficient number of them to obtain a good spread of the sample. Furthermore, the cost of listing, sub-sampling and data collection within big units may become excessive. On the other hand, if the units are too small and compact, it may be difficult to ensure sufficient spread within the units to obtain an efficient sample. Small area units also tend to lack clear boundaries and stability over time. Also, the smaller the units, the more of them will be needed to achieve the final sample size.

What constitutes an appropriate size for units serving as the PSUs depends upon survey circumstances and objectives. Various practical considerations need simultaneous attention; among them cost, quality control, administration, availability of the frame for sample selection, and the efficiency of the resulting design. A thorough understanding of these various considerations presupposes a sound knowledge of sampling theory and plenty of practice.

Indeed, several patterns can be identified from a variety of national surveys conducted in both developed and developing countries.

- In many developing countries, major road networks are sufficiently developed to facilitate travel between areas, but local travel can be more difficult and time consuming, even if the physical distances involved are small. This precludes the use in such situations of a mode often used in surveys in developed countries, which involves highly mobile enumerators each covering a very large and extensive PSU. On the other hand, in the absence of good maps and other materials to define suitable small area units, the use of numerous small PSUs is also precluded. The number of PSUs likewise has to be kept limited in order to control travel and supervision costs. Indeed, a common requirement in choice of design is to ensure that each PSU yields a sample large enough to keep the enumerators occupied for a sufficient length of time (e.g. a few days) in each area. This requirement becomes even more constraining when the enumerators are deployed as teams.

Thus many surveys in developing countries use census enumeration areas (EAs) as the PSUs, an EA typically consisting of 100-300 households.

- By contrast, in urban areas in many developing countries (especially in Latin America), as well as in household surveys in developed countries generally, many samples are based on numerous very small PSUs. Each PSU may be a small cluster of households (say 5-10 households), some or all of which are taken into the sample. Such a system may be suitable in densely populated urban areas where lists of housing units and/or very detailed maps of small area segments are available, and where, because of well developed transport facilities and the short distances involved, travel between units within the same locality presents no particular problem.
- Such activities as intensive surveys of particular sectors or types of child labour may encounter practical constraints that necessitate confining the survey to a very small number of sites. While studies based on limited samples may provide useful information for specific purposes, such an arrangement is generally inadequate for surveys aimed at producing statistically precise estimates.

The number of sample areas (or sample-take per area)

An examination of the types of samples hitherto used in child labour surveys shows a surprisingly wide variation in the cluster sizes (sample-takes per area) used. The variation is greater than 10-50 households per area. While this may partly reflect differing national circumstances and differences in the type of units involved, it is likely that much of the variation does not derive from real statistical or cost differences.

Based on information on sampling errors and design effects and on the distribution of field costs between and within areas, our general assessment of the quality of designs hitherto used in national child labour surveys is that *quite often, the sample designs used for child labour surveys have been inefficient and need to be (and can be) made more efficient*. This can be done only on the basis of a more careful evaluation of the implications of design choices on the survey costs, variances and other aspects of the data quality. This is essential even if the evaluation of these aspects can only be approximate and partial.

3.4.4 Unequal selection probabilities and weights

There can be various reasons for introducing unequal selection probabilities (or their inverse, sample weights) for the ultimate units in a 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. Examples are over-sampling of small domains to ensure adequate sample size for separate reporting by domain. Similarly, there can be over-sampling of small or special sub-populations or rare events. These may be seen as “necessary” departures from uniform sampling rates, dictated by the substantive objectives of the survey.

2. Then there are the "unnecessary" departures, unnecessary in the sense that they are not dictated by the survey objectives but arise from the particular sampling procedures adopted, shortcomings in the sampling frame or other circumstances of the survey. For instance, there may be multiple occurrences of units in lists which impart them different selection probabilities, but which can be discovered only after sample selection. It is desirable to avoid such variations, though this is not always possible in practice.

3. Thirdly, some types of variation in sampling probabilities may be introduced to make the sample design more efficient (e.g. reduce variances and/or costs). An example is "optimal allocation" involving over-sampling 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 total and distribution, may be introduced to reduce sampling variance and bias.

Variations in unit selection probabilities of 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 the sampling error in the total sample. This effect tends to be rather uniform across different types of statistics, variables and population subgroups. The inflation in sampling error is determined by the coefficient of variation of the sampling probabilities or the associated weights of units in the population.

By the contrast, variations of type 3 *can* result in reduced variance or bias, but not necessarily always or significantly.

3.5 Choice of sample size

3.5.1 The question of sample size

In any survey the choice of the sample size is clearly the most basic and important question, yet it is a difficult question, one that eludes any purely quantitative answer. The choice of sample size must balance between:

- what is *required* from the point of view of sampling precision, and
- what is *feasible* from the point of view of practical application (e.g. budget, field and office staff, technical resources, quality control, time constraints, manageability, sustainability).

The choice of sample size is determined by a balance between various statistical, practical and cost considerations. This balance is a complex issue. This section presents a discussion in broad terms of the considerations involved, with the aim of providing useful guidelines for the choice of sample size in specific situations.

"Scale" of a survey operation

It should first be appreciated that the scale of a survey is more than just a matter of its sample size. By "scale" we mean a measure of the total burden the survey entails. It is more than, and includes, the size of the sample. It also increases according to the complexity and volume of the information collected per case, how sensitive and

burdensome the information collected is for the respondent (and for the interviewer), how complex and demanding the survey operations are, etc.

Determining the scale of the survey is fundamental. It depends upon, and in turn affects, almost everything, including: the mode of data collection, the planning and organization of operations, the relevance, timeliness and accuracy of the resulting data and, above all, the cost of the survey. The following observations on sample size apply in fact to this whole broad concept of the scale of the survey operation.

Need for compromise

It is useful to begin by discounting extreme or one-sided approaches to the question of sample size, though each of these can have some merit in certain circumstances.

- A common practice is to begin by identifying a single estimate or at most a few "critical" estimates and to choose some more or less arbitrary level of desired precision to compute the required sample size. Such computations are usually straightforward in themselves, but they are hardly ever realistic. This is because most surveys have complex, multiple objectives that cannot be reduced to the estimation of a single "critical" figure.
- Nor can the precision objectives be regarded as entirely predetermined; they are often modified, more or less radically, by what is feasible under the cost, timing, administrative and technical constraints. Precision requirements can hardly ever be expressed precisely or with sufficient objectivity. Even if they are assumed, the simple calculations often yield unrealistically large sample size requirements, resulting in the arbitrary revision of the (often equally arbitrary) initial precision requirements.
- Another position argues that the choice of sample size is primarily a non-technical issue, determined almost exclusively by cost and other practical considerations, previous practice (if not prejudice), desire for parity with other surveys and countries and, generally, the desire to secure the maximum sample size allowable under the given circumstances.
- It is also common to argue that non-sampling errors generally predominate over sampling errors in large-scale surveys. This sounds likely enough in practice, but it is not a meaningful or useful statement when expressed in such absolute terms. It simply amounts to making the rather extraordinary assertion that the sample sizes used in practice are too large *in general*. It ignores two basic and obvious points. Firstly, in practice there is always a trade-off between the magnitude of sampling and non-sampling errors. Secondly, sampling error always becomes the predominant component as estimates are produced for increasingly small domains.

The appropriate approach is a balanced combination of these perspectives, avoiding simplistic views of the complex issue of sample size.

Minimum requirements

Situations do indeed arise in which, at least as a starting point, sample size is determined primarily on the basis of the precision requirements of a few "critical" or

basic statistics. Usually, however, precision requirements are much more complex, even if only a few critical variables are considered. Tabulation and analysis of survey results involve numerous types of estimate for diverse sub-populations. *In most situations, the primary consideration is the degree of detail with which the sample data can be classified meaningfully.* However, beyond a certain minimum below which the survey may not be useful, this degree of detail is not entirely predetermined. It is itself conditioned, in fact, by what is possible with various choices of sample size.

Despite the complexity of substantive requirements, it is possible to identify situations where a particular sample size is too small to yield useful results. This knowledge derives mainly from the experience of previous surveys (one's own and those of others) on similar topics subject to similar analyses. Practically no survey is entirely new or conducted in isolation. In planning a survey one begins with questions like the following.

- What sort of sample sizes have been used for this sort of survey in similar contexts in the past?
- What sort of information (in how much detail? with what precision?) did they yield? Negative experiences - e.g. where sample sizes proved inadequate - can be particularly instructive.
- Is there evidence that, despite being justified in terms of precision requirements, sample sizes have been so large as to affect survey implementation and the quality of the resulting data adversely?
- How do the present requirements and conditions differ from such previous surveys? For instance, are the results required for more (or for less) detailed classifications?
- What is already known about the subject to be investigated? Upon what level of existing knowledge must the new survey improve?

It is particularly important to try and identify *the minimum requirements for the survey to be sufficiently useful*, given its opportunity cost. As noted, no single estimate can as a rule be considered "critical" in determining these minimum requirements; but different types of estimates differ in their relative importance, and it is often possible to identify a subset of objectives that can be considered basic in determining the minimum sample size requirements.

Controlling sampling error

Survey estimates are normally required not just for the whole population but also for many subgroups in the population. The relative magnitude of sampling error vis-à-vis other types of error increases as we move from estimates for the total population to estimates for individual subgroups and differences between subgroups. *Beyond a certain level of detailed classification, sampling error would predominate over other sources of error*.

Number of major reporting domains

As noted above, the minimum sample size required will increase with the number of population subgroups or domains for which the results have to be reported separately. For several reasons, however, this increase is usually considerably less than

proportionate. Consider, for example, the effect of the need to produce separate estimates for individual regions of a country.

- Usually more precise information is already available at the national level than at the level of individual sub-national domains; to add something new to what is already known, the survey has to meet higher precision requirements at the total national level.
- Precision requirements are also less stringent for individual domains than for the country as a whole, because the results at the national level are usually of greater policy and substantive interest.
- More varied and detailed analyses are required at the total level, compared to those for individual domains.
- Individual domains are often less heterogeneous than the national population as a whole, thus requiring a smaller sample for the same precision.
- For certain types of domain, such as subclasses well-distributed over the population, the sample design is usually more efficient (because of smaller clustering effects), so the same precision can be achieved with a smaller sample size.

Upper limit in practice

At the other end of the scale, practical considerations of cost, timeliness and quality control determine – although perhaps not sharply or rigidly - the practicable upper limit of the sample size. Increasing the sample size yields diminishing returns as regards improved sampling precision of the results. (Sampling error is reduced approximately only in proportion to the square root of the increased sample size.) The negative effect of an increase in sample size on quality control, however, tends to grow - in many cases explosively - beyond a certain sample size as the survey becomes unmanageable.

Increasing the sample size, at least beyond a certain limit, can adversely affect all aspects of data quality – relevance, timeliness, response quality, etc.

Cost considerations, of course, can impose a much more inflexible limit on the maximum sample size allowable.

Compromise between minimum and maximum limits

In practice, the process of determining the sample size may be along the following lines. Given the major domains for which separate results are to be reported, and the type of estimates required, the minimum sample size may be determined so as to meet the most critical precision requirements. It is important, however, that the sample size so determined does not exceed the practicable maximum, given prevailing constraints. If the initial choice exceeds this maximum, then it is best to reconsider and adjust the objectives and reporting requirements of the survey, rather than try and impose an unrealistically large sample size. When this contradiction cannot be resolved, cancelling the whole survey may be the only option.

Where scope exists for a compromise between the minimum and maximum limits, the choice of the sample size depends on several considerations. The value of the survey

increases as sample size is increased, since more precise and detailed analysis of its results becomes possible. On the other hand, there may be a detrimental effect on survey quality, timeliness and cost. Balancing these contending considerations is a matter of judgment based on theoretical considerations, empirical information, past experience and knowledge of practical conditions of survey-taking - even though none of these factors can in itself provide precise quantitative results.

Different requirements for surveys of different types

Experience has established useful norms for many types of surveys from which we can derive reasonable guidelines. An important point is that, though theory and practice cannot provide precise answers regarding sample size in absolute terms, they can be much more helpful if the questions are put in *relative terms*. For instance:

- For a survey on a given topic, should the sample size be larger, smaller or the same in comparison with past surveys? In comparison with similar surveys in other countries? In comparison with established norms where they exist? If so, by how much (even if roughly), given the differing objectives and conditions of the surveys?
- What should be the relationship regarding sample size between surveys of different types on different topics, but conducted under similar circumstances?

Concerning the second point, we know, for instance, that surveys involving physical measurement or intensive follow-up are best conducted with relatively small samples, while labour force surveys for instance, which have to produce relatively simple but disaggregated estimates, need much larger samples. A great deal of information is available on this from the experience of different countries and from different types of surveys.

In given circumstances, different types of surveys tend to be less diverse in terms of "scale" than in terms simply of the sample size. As noted above, by scale is meant the total burden a survey constitutes, depending on the size of the sample, the sensitivity and burdensomeness of the information, the complexity and volume of the information collected per case, and of course the cost per unit. We can expect that the *appropriate sample* size for more complex, demanding and expensive surveys (such as those on income and household budgets) would normally be smaller than simpler and cheaper surveys (such as those on the labour force), the differences in appropriate sample sizes increasing with differences in the complexity and per-unit cost.

Surveys for the measurement of prevalence of child labour (CLSs) would normally be of smaller, frequently much smaller, size than national labour force surveys, but much larger than surveys that are designed to assess certain aspects of child labour (LCSs)

3.5.2 Specification of precision requirements

Sampling precision in terms of effective sample size

Sampling precision is determined by size of the sample, as well as by its design, i.e. its efficiency or "design effect". Both of these factors are specific to the statistic being considered.

Let us assume that we have identified the most critical statistic or set of statistics for the stated purpose. Even so, it is helpful to separate the issue of the design effect, which depends on the sample structure, from that of the sample size. The precision requirements are more clearly expressed and understood in terms of the "effective" rather than the actual sample size. By the effective size of a sample with complex design, we mean the size of a simple random sample of analysis units which has the same precision as the complex design. The effective size of a complex sample of size n with design effect $deft^2$ is

$$n_{eff} = \frac{n}{deft^2}.$$

The design effect, $deft^2$, (or its square-root, deft, which is sometimes called the design factor) is a comprehensive summary measure of the effect on sampling error of various complexities in the design. The deft is the ratio of standard error (of a given statistic) under the actual sampling design to what that error would have been under a simple random sample of the same size. (This concept is discussed further in a later chapter.)

Relative and absolute levels of error

For estimates of mean values or general ratios, the precision requirements are conveniently expressed in terms of relative standard error (r), i.e. standard error as a percentage of the mean value

$$n_{eff} = \left(\frac{cv}{r}\right)^2,$$

where cv is the coefficient of variation of the variable of interest among individual elements in the population.

For a proportion (p), cv is estimated by a very simple expression, $cv^2=(1-p)/p$. However, for mean values and more complex statistics, empirical information is required to estimate the value of the parameter cv in the population. It is convenient that this parameter tends to be highly "portable" across different situations and can often be estimated fairly easily and reliably from past surveys, other sources, or similar populations. For instance, for household income, cv is mostly found to be in the range 0.7-1.0 across diverse populations. The above equation implies that in this case, with cv=0.7-1.0, a 1 per cent relative error (r) would require an effective sample size of 5,000–10,000. Similarly, a relative error of 2 per cent would require an effective sample size of 1,250–2,500. The actual sample size has to be larger to the extent $deft^2$ exceeds 1.0 for the particular design chosen.

A note on terminology

It is very common in reporting survey results to use the term "error" or "error level", which combines the concept of relative error as defined above (meaning standard error as a percentage of the estimate to which it refers) with that of an implied confidence

level[9]. This practice is based on the expectation that, for the general reader, it is more useful and convenient to know with a certain level of confidence the range within which the "true" value of the statistic under consideration lies. While different levels of confidence may be chosen for different purposes, a common criterion asserts that the population value estimated from the sample lies within a range equal to twice the standard error on either side. This can be asserted with a high (95 per cent) level of confidence; in other words, one can say that the chances are only one in twenty that the true value is outside this range (this is a theoretical result based on the assumed normality of the sampling distribution).

In this formulation, "error" or "error level" (say R_z) is error (r) multiplied by a factor (z): $R_z = z \cdot r$. The value of z depends on the desired level of confidence – in the above example, $z=2$ for 95 per cent confidence level. Larger values of z give higher levels of confidence. For example with $z=3$, one can say that the chances are approximately only one in a hundred that the true value is outside the range equal to around 2.6 times the standard error on either side of the estimate from the survey; chances are only three per thousand for the true value to be outside 3 times the standard error on either side of the estimate.

In terms of Rz, the above expression for n_{eff} is

$$n_{eff} = z^2 \cdot (cv/R_z)^2.$$

The appropriate choice of z is specific to the objectives and implications of the conclusions to be drawn from the survey, which depend on the nature and context of the information being considered. It also involves a degree of arbitrariness, or simply convention. For these reasons, we consider it preferable to specify and present information on sampling errors in terms of the standard error, whether in absolute terms (e) or relative terms (r), with the reader being left free to convert this to intervals with the desired confidence level.

Specification for proportions

In child labour surveys (just as for many other types of survey), most statistics of interest are likely to be in the form of proportions rather than mean values or general ratios. For proportions or percentages, it is important to keep a clear distinction between the error expressed in relative terms (as a percentage of the proportion p) and that expressed in terms of absolute percentage points.[10] Both forms are relevant. For large proportions, the error is often better expressed in relative terms, while for very small proportions its expression in terms of absolute percentage points is often more meaningful.[11]

---

[9] Actually the above-mentioned terms are often used to refer to multiples of *relative error* for mean values, but for proportions, by contrast, these are used to refer to multiples of error in *absolute percentage points*.

[10] For example, a child labour rate of 22 per cent differs from a rate of 20 per cent by 10 per cent in relative terms, but by 2 percentage points in absolute terms.

[11] This is true for the purpose of statistical design. However, it is common in presentations for the general public to quote errors in absolute terms in absolute percentage points (usually as +/- 2 standard errors). For mean values, it is more meaningful to report errors in relative terms because the scale of measurement is essentially arbitrary for such measures.

The effective sample size, in terms of the precision required for a proportion (p), can be written as:

$$n_{eff} = \left(\frac{p*(1-p)}{e_p^2} \right) = \left(\frac{(1-p)}{p*r_p^2} \right),$$

where e_p is the standard error in absolute percentage points and $r_p=(e/p)$ is the same error expressed in relative terms. These quantities have been written with the subscript "p" to emphasize that, in practical situations, realistic precision requirements depend on the value of the proportion p under consideration. In fact as p gets smaller, the meaningful value of e_p is also reduced, while that of r_p tends to be correspondingly increased. For instance, if e=5% is a reasonable choice for estimating a proportion close to p=0.50%, then the same choice of error level e is not a reasonable one, at least for very small values of p. Similarly, while a relative error of, say, r=25% may be acceptable in the case of a very small proportion, it may be useless for a large proportion such as p=0.50%. This important point is often forgotten in discussions of precision requirements for different p values encountered in a survey.

Indeed, it is not unreasonable - unless there are other reasons to consider different values of this term for different estimates - to take, at least as a starting point, n_{eff} as a constant for different proportions p to be estimated in a given survey. This would imply taking e_p to be proportional to the square-root of [(1-p)*p] for the range of values of the proportion p encountered. For instance, if an absolute error e=5% is considered necessary for a proportion p=0.5, the corresponding appropriate value would be the smaller error e= 4% for a proportion p=0.2 (or 0.8); for p=0.1 (or 0.9), we will have e=3%, reducing further to e=1% for p=0.01 (or 0.99). Surely, this is a more meaningful variation than assuming a constant value for e (e.g. e=5% even when p is as small as 0.1 or even 0.01). While the precise assumption in the above argument may be disputed, the implied variation in the precision requirements is at least in the right direction.

A similar but opposite pattern holds in terms of the required precision expressed in relative terms. The assumption implies taking r_p as proportional to the square-root of [(1-p)/p] for the range of values of the proportion p encountered. In the example, we have r=10% for p=0.5, 20% for p=0.2, 30% for p=0.1, and r increasing to 100% for an extremely small p=0.01. Again, these appear more meaningful figures than assuming the requirement for a constant level of relative error irrespective of the value of p involved.

It should be noted that the particular assumption in the above illustration implies that the precision requirements are modified by p in such a way that the resulting effective sample size is independent of the p value being considered. This form of variation assumed above may indeed be simplistic, though it is intended to be merely an illustration that we have adopted for its extreme simplicity. Nevertheless, it provides a useful starting point for the specification of precision requirements e_p and r_p (in absolute and relative terms, respectively) as functions of the proportion p of interest.

In practice, some increase in n_{eff} with decreasing p is normally warranted.

In short, it is recommended that, in considering sample size and precision requirements for proportions, the sampling error in both absolute and relative terms should receive attention. These precision requirements should be varied as a function of the different proportions p to be estimated from the survey, but in such a way as to limit very much the implied variation with p of the required sample size.

Table 3.1 provides illustrations of how the sample size may be increased, as a function of p, with a decreasing proportion p to be estimated. We assume the sample size to have been varied as

$$n \propto 1/p^a$$

where a is a parameter, shown in the range 0 to 0.25 in the table. The base has been taken as n=100 for p=50%. Consider a=0.10, which in fact may be a reasonable choice in practice. This implies increasing the sample size from, say, 100 at p=50% to 120 for estimating a smaller proportion p=20%, to around 160 for p=5%, and to 220 for a very small p=1%. Note the implied variation in the required relative and absolute levels of error with a varying p. In practice, a nearly constant sample size may be acceptable for a wide range of p values – for example n=120 for p=10-30% in this illustration.

Table 3.1. Some illustrations of varying sample size according the proportion being estimated

	a=0			0.05			0.10			0.20			0.25		
p(%)	n	%e(p)	%r(p)	n	%e(p)	%r(p)	n	%e(p)	%r(p)	n	%e(p)	%r(p)	n	%e(p)	%r(p)
50.0	100	5.0	10	100	5.0	10	100	5.0	10	100	5.0	10	100	5.0	10
40.0	100	4.9	12	102	4.8	12	105	4.8	12	109	4.7	12	112	4.6	12
30.0	100	4.6	15	105	4.5	15	111	4.4	15	123	4.1	14	129	4.0	13
20.0	100	4.0	20	110	3.8	19	120	3.6	18	144	3.3	17	158	3.2	16
10.0	100	3.0	30	117	2.8	28	138	2.6	26	190	2.2	22	224	2.0	20
5.0	100	2.2	44	126	1.9	39	158	1.7	35	251	1.4	28	316	1.2	25
2.0	100	1.4	70	138	1.2	60	190	1.0	51	362	0.7	37	500	0.6	31
1.0	100	1.0	99	148	0.8	82	219	0.7	67	478	0.5	46	707	0.4	37

p%= percentage being estimated
a= parameter determining n
n increased with decreasing p: n=k/(p ^ 2a)
n= (effective) sample size
%e(p)= error in (absolute) percentage points
%r(p)= relative error (%)

3.5.3 Factors determining the actual sample size

Variation of design effect with cluster size

In a multi-stage design the required sample size to achieve a given level of precision depends on the design effect. The effect of clustering can be summarized in terms of two parameters: the intra-cluster correlation (roh), and the cluster size (b)[12].

[12] Sampling errors are discussed in more detail in chapter 7.

The first parameter (roh) is a measure of relative homogeneity of elements coming from the same sample clusters. It depends primarily on the nature of the variables of interest (values encountered for different variables can differ greatly). The nature and the structure of sampling within sample clusters also have an effect.

The second parameter (b) is the average number of elements (the ultimate units of interest in the analysis) selected per PSU in the sample. The relevant expression, for a proportion p for instance, can be written as

$$n = n_{eff} \cdot deft^2 = \left[\frac{p*(1-p)}{e_p^2}\right]\left[1+(b-1)roh\right].$$

On the basis of sampling error computations from similar past surveys, a reasonable idea may be formed of the pattern of roh values encountered for different types of variable and for various designs.

For a required level of precision (expressed in terms of e_p or n_{eff}), the choice of cluster size b (and hence of the number of clusters to be selected, a =n/b) affects the required sample size n. Table 3.2 shows the factor $(n/n_{eff})=(1+(b-1).roh$ by which the required sample size has to be increased, according to the roh and b values for the variable and the sample design concerned.

Table 3.2. Increase required in the sample size with increasing cluster size b and intra-cluster correlation roh

		roh				
		0.00	**0.02**	**0.05**	**0.10**	**0.20**
b	1	1.00	1.00	1.00	1.00	1.00
	5	1.00	1.08	1.20	1.40	1.80
	10	1.00	1.18	1.45	1.90	2.80
	20	1.00	1.38	1.95	2.90	4.80
	50	1.00	1.98	3.45	5.90	10.80

The intra-cluster correlation (roh) is a more "portable" measure than deft in the sense that it also removes the effect of average cluster size b. It is defined in terms of deft and b as

$$roh = \frac{deft^2 - 1}{(b-1)}$$

The practical point to be emphasized is that the sample size n required to achieve a specified degree of precision of the estimates (or specified n_{eff}) cannot be determined independently of the efficiency or $deft^2$ of the design. In a clustered sample the two parameters, namely *the sample size*, *and the number of sample areas* (or alternatively, the average sample-take per area or cluster) *have to be determined simultaneously*. The larger the intra-cluster correlation (roh), the stronger the relationship.

Effect of variation in sample weights

Often, the survey is designed to produce different types of estimates and at different levels of aggregation. The different objectives result in conflicts, which require compromises in sample allocation. A common example is the need to over-sample smaller domains so as to obtain a more adequate sample size, and therefore correspondingly under-sample larger domains. For any given objective, such as producing estimates for the total population, the compromise allocation constitutes weighting which is essentially random or arbitrary.

The effect of such weighting is to inflate variances concerning that objective. This inflation tends to be fairly uniform for different types of variables and population groups in the domain concerned. The increase in variance is approximately given by $[1+cv^2(w)]$ in terms of the coefficient of variation of the individual sample weights.

Units of analysis versus units of sampling

Precision and sample size requirements are, of course, specified in terms of the units of interest in the analysis. These may, for instance, be particular categories of individuals, such as any adults, old persons, women or children, and they may differ from the ultimate units used in sampling. The latter may, for example, be addresses or households. Operationally, the sample size has to be expressed in terms of these latter types of unit.

Let us assume that the household is the ultimate unit of sampling, and that on average a household contains h analysis units of interest (e.g. h adults, or h working children). The required sample size in terms of number of households (n_h) may be expressed as

$$n_h = \frac{n}{h} = \left(\frac{n_{eff}}{h}\right) \cdot \left[1 + (h * b_h - 1).roh\right]$$

Here b_h is the average number of households per sample cluster, and as before, neff is the effective sample size required in terms of the analysis units. Here are some examples.

1. The simplest situation is when the two units are of the same type: h=1. We can take this as the standard for comparison.

2. Suppose that in a sample of households the units of analysis are school-age children 6-14 years old. The average number of such units per household is likely to be substantially less than 1.0, say h=1/3. We need to take proportionally more households in the sample in order to obtain a sample size of n children. At the same time, the effective cluster size is proportionally reduced to $(h*b_h)=b_h/3$. Therefore, we can increase the number of households per cluster as h goes down without necessarily increasing the design effect for a given value of roh.

3. Suppose that we are interested in a sample of adults and take all adult in each selected households into the sample. The number of analysis units per sampling unit would be larger than 1.0, say h=2. To achieve a sample of n adults, we would need proportionately fewer households. In order to retain the same efficiency, we

need to reduce the number of sample households per cluster to retain the same design effect for a given roh value.

4. An alternative to design 3 is to select exactly one adult from each sample household. In so far as each household contains at least one adult by definition, then h=1 and the situation appears to be the same as design 1. However, there are some important differences between the two situations. On the one hand, design 4 avoids the added homogeneity among persons coming from the same household, thus reducing design effect compare to design 1. On the other hand, however, assuming the households were selected with uniform probabilities, the relative sampling probabilities of the individuals are altered, and become inversely proportional to the number of persons through which the household could have been selected. This requires weighting at the estimation stage, which in turn affects the efficiency of the sample and hence the required actual sample size for a given n_{eff}.

Often this second, negative effect predominates, making design 4 less efficient than direct sampling of persons as in 1.

Loss due to non-response

In addition, a part of the sample is lost due to non-response. The size of the sample selected has to be larger than the achieved sample size by the factor (1/R), where R is the response rate.

3.5.4 Experience of child labour surveys

Some information on sample sizes and structures of past child labour surveys was given in Table 2.1.

In the child labour surveys that we have reviewed, we have found the following general picture. (See Section 2.1.5 for some further information.)

A vast majority of the sample sizes for the CLS component are in the range 8.000-20.000, though in a few cases much bigger sample sizes have been used. Little information is available in the literature and national reports on how appropriate these choices have been in relation to the specific objectives in particular cases. Unfortunately, we have found the same situation to hold for many other types of survey. It is necessary to evaluate how appropriate these choices of sample size have been, so as to provide guidance for improving future practice.

3.5.5 Practical advice: Moderation in the choice of sample size

The sample size has to be large enough, of course, to meet the substantive and specific requirements of the survey. Nevertheless, in many cases *the quality of surveys has suffered because of the choice of inappropriately large sample sizes*. Inappropriate choices can result from several factors:

- over-emphasis on sampling precision, and neglect of the need to control non-sampling errors and to ensure the relevance and timeliness of the results;
- adoption of unreasonable and unnecessarily high standards of precision;

- the desire to produce too many breakdowns in too much detail;
- failure to explore alternative procedures to enhance the precision of the results, such as the cumulating of results over time in periodic surveys, the use of suitable models and relationships, the adjustment of the survey data on the basis of control totals from more reliable external sources, etc.;
- use of inefficient designs, which increase the sample size required to achieve a given level of precision; and, above all,
- underestimation of the per-unit cost and effort required in survey data collection, processing and analysis.

3.6 PPS sampling of area units

This is by far the most common technique for selecting population-based samples, especially in developing countries. The design involves the selection of area units in one or more stages (often in only one stage) *with probability proportional to a measure of population size* of the area (p_i) and within each selected area, the selection of ultimate units *with probability inversely proportional to the size measure*:

Selection of areas

$$f1_i = \left(\frac{a}{\Sigma p_i}\right).p_i = \frac{p_i}{I}, \text{ say.} \qquad [1]$$

Selection of ultimate units within selected area

$$f2_i = \left(\frac{b}{p_i}\right) \qquad [2]$$

Overall selection probability of an ultimate unit

$$f_i = f1_i * f2_i = \left(\frac{b}{I}\right) = f, \text{ a constant.} \qquad [3]$$

The summation Σ is over all areas in the population. Parameter *a* in equation [1] refers to the number of areas (strictly "ultimate area units", UAUs) selected. If p_i is strictly the current number of individuals of interest in each area, then b=constant is the number of ultimate units selected from any sample area, and n=a*b is the resulting sample size.

Hence, in the ideal case, such a design yields the dual advantage of: (i) control over sample size and fixed workload *b* per sample area; and (ii) a uniform overall sampling probability f for each ultimate unit (e.g. each household, child, another person).

In practice, it is unlikely that both or even one of these conditions is satisfied exactly. For one thing, the unit in terms of which the size measures are available is often different from the type of unit used at the ultimate stage of selection, which again may be different from the analysis units; examples are persons versus households, or dwellings

versus children. But the real problem is that in most circumstances, the size measures available in the frame are outdated and may not correspond well to the actual (current) sizes of the area units.

Some common (and often close) variants of this basic design are described below (Section 3.8). It is useful first to clarify a method of sample selection that is very commonly used in practice, namely the procedure of systematic selection.

3.7 Systematic sampling

A common method of sample selection is to select the units systematically from an ordered list. Systematic sampling with equal probability is commonly used for the selection of ultimate units, such as households or persons, in sample areas. More important is the use of systematic sampling with probability proportional to size (PPS) sampling, commonly used for the selection of area units.

Below is a brief, formal description of the procedure (for a numerical illustration see Section 6.7).

3.7.1 Systematic sampling with equal probability

The procedure for selecting an equal probability sample systematically from a list is basically as follows. Suppose that an equal probability sample of n units (listings) is required from a population of N units. From the list of units - preferably arranged in some useful way and the units numbered sequentially from 1 to N - one unit is selected from every $I=N/n$ units in the list.

First, assume for simplicity that the selection interval I happens to be an integer. A random number r between 1 and I identifies the sequence number of the first unit selected. Then, starting with r, every I^{th} unit is selected. That is, the sequence of numbers selected is

$$r, r+I, r+2*I, ..., r+(n-1)*I.$$

The general case when $I=N/n$ is not an integer is easily dealt with. One procedure is as follows. Starting with a real random number r (not necessarily an integer) in the range $0<r\leq I$, the sequence as defined above is constructed. Each term of this sequence, rounded up to the nearest integer, identifies a selected unit.

To the extent that the units in the original list appear in a random order, the resulting sample is equivalent to a simple random sample. Existing lists, however, are practically never randomly ordered. In any case, systematic sampling aims to make use of the order available to achieve a better spread of the sample according to some meaningful criterion, such as geographical location of the units. In this manner, systematic sampling provides implicit stratification; it can be regarded as stratification of the population into zones or strata of size I, and the selection of one unit per zone or "implicit stratum".

Widespread use of systematic sampling is also explained by its great convenience in many situations.

3.7.2 Systematic sampling with probability proportional to size (PPS)

To select units with probability proportional to some measure of unit size, such as the size of the population or the number of children in the area, the main difference from the above is that the simple count of units is replaced by the *cumulation of their measures of size*. Let

p_i be the size measure of unit i

P_i the cumulation of these measures for all units 1 to i, ordered in some meaningful way,

P this sum over all units in the population, and

a the number of units to be selected with PPS.

The sampling interval to be applied to the cumulative size measures is

$$I=P/a.$$

A random number r in the range $0<r\leq I$ identifies the first unit selected: it is the first unit whose cumulative size equals or exceeds r, i.e. the unit sequence number i satisfies the relationship

$$P_{i-1} < r \leq P_i.$$

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

$$r'=r+I,\ r+2*I,\ \ldots..,\ r+(n-1)*I,$$

and each unit with cumulative size measure satisfying the relationship

$$P_{i-1} < r' \leq P_i$$

for any i in the list of cumulative size measures P_i is selected. The chance that a selection point falls on a particular unit is proportional to

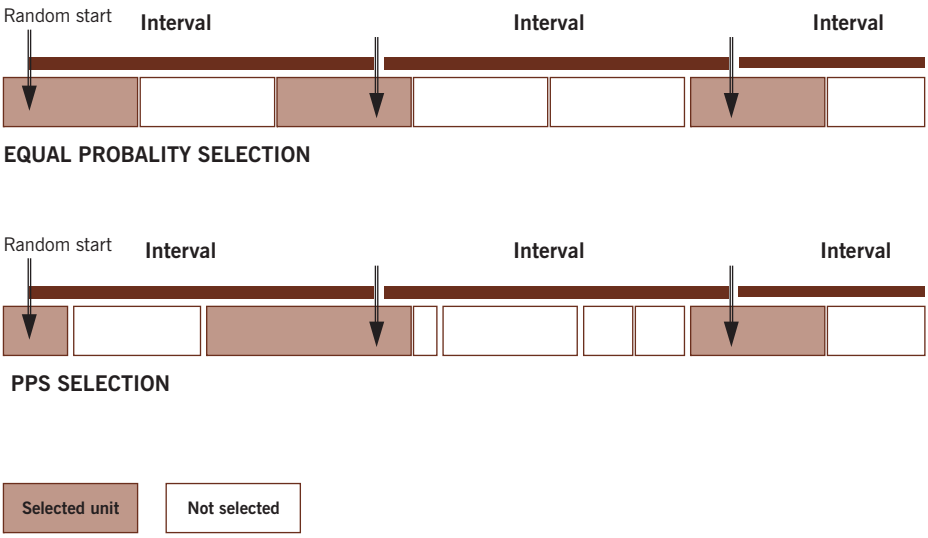
$$P_i - P_{i-1} = p_i$$

i.e. to the unit's measure of size. The absolute value of the selection probability is:

$$f1_i = \left(\frac{a}{P}\right).p_i = \frac{p_i}{I}$$

The following diagram provides a physical illustration of the PPS systematic sampling procedure.

Figure 1. Illustration of systematic sampling procedure



3.8 Imperfect size measures

The size measures (p_i) determining selection probabilities of area units have to be based on past information, generally at the time the area frame is compiled. This is because these measures are required for all area units in the population, and cannot be updated in the sample alone. These may therefore differ from the actual or current sizes (p'_i).

Inaccuracy of the size measures is a common problem and therefore important. With the higher stage units selected with probabilities proportional to estimated sizes that depart significantly from the actual unit sizes, it is important to keep clear the distinction between two types of designs for the next stage of selection:

(a) a strictly self-weighting design of ultimate units, on the one hand, and

(b) a fixed-take design, on the other.

(c) In the case of child labour surveys, a third option can also be appropriate, i.e. take-all sampling, possibly with a cut-off on the *maximum sample take from any area*.

3.8.1 Self-weighting and fixed-take designs

In the *self-weighting design* described above, the ultimate units all receive the same overall probability of selection, and the number of these units selected is allowed to vary to the extent the measures of size used in the selection of the PSUs differ from their actual size. From equation [2] above (see Section 3.6), the number of units selected from a sample area is:

$$b_i = f2_i * p'_i = b * \left(\frac{p'_i}{p_i} \right), \qquad [4]$$

with (albeit smaller) variation also in the overall sample size $n' = \Sigma_s b_i$. (Here the sum is over all areas in the sample, as indicated by subscript s.)[13]

In the fixed-take design, the number of ultimate units selected per sample area is fixed. With

$$f2_i = \left(\frac{b}{p'_i} \right); \text{ we have } b_i = f2_i * p'_i = b, \text{ a constant.} \qquad [5]$$

As a consequence, however, the overall sampling probabilities vary in proportion to the inaccuracies in size measures:

$$f_i = f1_i * f2_i = f * \left(\frac{p_i}{p'_i} \right). \qquad [6]$$

Taking samples of a fixed predetermined size is common in situations where complete lists of units are unavailable before selection – but then the result is not, strictly speaking, a probability sample. 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. But these are exceptions.

Nevertheless, using fixed-take rather than self-weighting designs seems to be very common in practice and is the case for a majority of the child labour surveys conducted so far. This may be because possible lack of control over the overall sample size is undoubtedly the most serious drawback of the self-weighting design. Excessive variation in work-load between sample areas can also be a serious inconvenience. (See Section 3.8.3 for compromise or alternative procedures to reduce the problem.)

Despite this, however, we recommend that, as a general rule, one should opt for a self-weighting rather than a fixed-take design, because fixed-take designs have a number of practical disadvantages. These arise primarily as a result of fixing the sample size to be obtained, rather than of the selection probabilities to be applied.

- Firstly, arbitrary variations in the ultimate selection of probabilities, which are involved in a fixed-take design, are generally undesirable.
- Since the overall selection probabilities are not fixed in advance, they can be computed only if an account is kept of the actual number of eligible units listed in each sample area. This may appear straightforward, but many surveys fail to keep such records in practice. Computation of sample weights requires knowing not simply the number of total units in the list but the numbers that are

[13] Self-weighting here refers to the (lack of) variation in ultimate selection probabilities across clusters in a given stratum. It does not concern possible variations in sampling rates across strata or domains. The same context applies to the fixed-take designs being discussed.

eligible for the survey, e.g. not simply the number of dwellings but of dwellings occupied by private households. This is not always a trivial requirement and may require additional information to be collected at the listing stage. It can further increase the burden of the listing work, and possibly also adversely affect the completeness of its coverage.

- It is most common in practice to select the ultimate units systematically from lists. With a self-weighting design, the selection interval can be pre-specified for each sample area. The selection interval to be applied with a self-weighting design is based on the area size measures as already available in the frame, and thus is not dependent on the re-listing of sample areas. Furthermore, if necessary, the pre-specified selection intervals can be applied to all units listed, without distinguishing between eligible and non-eligible units, and without affecting the selection probabilities of eligible units.
- This selection procedure is more complex with a fixed-take design, however, since the selection interval has to be computed on the basis of that actual number of units found in the area. This can cause problems, especially when the selection has to be performed by lower-level staff in the field.
- Under-coverage in sampling frames is a common and often a serious problem, irrespective of the sample selection method chosen. However, the following added problem can appear with a fixed-take design.
- Comparison of the sample size actually selected and the target total sample size can bring to light problems of under-coverage in the frame, such as incompleteness of the list of areas included in the frame, or major mistakes in sample selection. Examples of the latter include omitting a part of the frame, systematic preferences among field-workers for selecting smaller segments and, perhaps most commonly, systematic omission due to outdated or poor listing of ultimate units in the sample areas. Many surveys have encountered the problem of the actual selected sample size falling seriously short of the target or design size. A major disadvantage of a fixed-take design is that, by automatically pre-fixing the sample size, it conceals problems of this type.
- Worse still, a fixed-take design may actually encourage incomplete listing, especially when the same field-workers are responsible for both listing and interviewing. Of course, closer supervision can help to control this problem, but that is more difficult if the listing and enumeration are carried out as a single operation, with little or no chance of supervisory intervention between the two. There have been surveys – even at the national level - where only a fixed number – at worst only as many units as required to get the fixed number of sample cases – were listed in each area, thereby debasing the design and leaving an enormous potential for bias against less accessible units.
- Fixing the sample size can encourage uncontrolled substitution for non-responding cases, and discourage persistent effort with the more difficult cases. This happens when the interviewers' performance is judged primarily on the basis of success in obtaining the required pre-fixed number of completed interviews in each sample area, rather than obtaining the maximum response rates.

- Even with a fixed number of interviews in each cluster, the actual workload remains variable because of inevitable variations in circumstances and the amount of effort required per interview in different areas of the sample.
- In any case, fixing the sample sizes rigidly is often unattainable because of non-response and other problems at the implementation stage. The same applies when the units to be enumerated are not of the same type as the units used for listing and sampling, or when the survey is concerned with a sub-population not controlled in the design.

These problems can be even more serious in surveys of labouring children, compared with more broad-based surveys of the general population. Hence, the general advice is to *adopt, wherever possible, a self-weighting design* (or, more generally and appropriately, a design with pre-fixed probabilities rather than pre-fixed sizes) *and to tolerate variations in sample-takes per cluster and in the overall sample size*. We often underestimate the flexibility that is possible in practice in dealing with variations in the sample-takes. Extra staff can often be assigned to areas with exceptionally large workloads. There is also more flexibility when interviewers work in teams rather than alone within sample areas.

There are often good reasons to vary sampling rates across different parts of the survey population, such as over-sampling of urban areas, small regions or other small domains of special interest. The points made above are not related to this type of necessary departure from a self-weighting design, but to the desirability of avoiding unnecessary variations in selection probabilities of the ultimate units in a multi-stage sampling design.

3.8.2 Control of overall sample size

It should be noted that even in a self-weighting design in the above sense, when the individual area workload varies considerably, the overall sample size is much better controlled than the sample-takes from individual areas, in so far as much of the variation in the latter tends to cancel out when aggregated over the whole set of sample areas.

In addition, some simple steps can usually be taken to reduce variation in the *overall sample size*. For instance, it may be possible to adjust the last-stage selection rates $f2_i$ in the light of more accurate external information on the total population size (P, replacing the total size measure $\sum p_i$ in the frame), even though the size measures pi for individual areas cannot be updated:

$$f2_i = \left(\frac{b}{p_i} \right) \cdot \left(\frac{\Sigma p_i}{P} \right).$$

This adjustment reduces the impact of changes in the total population size on the size of the sample obtained. For instance, if the population has expanded (larger P), the final stage sampling rate $f2_i$ is correspondingly reduced to maintain the required total sample size.

The following adjustment may be possible when the population size P is not known. If the listing and the final sample selection are separated in time and intervention is possible between the two operations, the overall expected sample size can be adjusted, taking into account the outcome of listing (providing new size measures p'_i) in the sample areas, by changing the last stage sampling rate to:

$$f2_i = \left(\frac{b}{p_i}\right).\left(\frac{\Sigma_s p_i}{\Sigma_s p'_i}\right)$$

(Note that, in the above, the summations are over the sample areas only, as indicated by subscript , while in the previous equation the sum is over areas in the whole population.)

Adjustments of the above type can be useful, although their value is limited in so far they affect only the overall sample size but do not reduce the variation in workloads for individual areas. The selection probabilities, moreover, are now known only in relative rather than absolute terms, though that suffices in the vast majority of sample surveys aimed at estimating ratios or similar statistics rather than population aggregates, or where ratio estimators using external control totals are employed for estimating such aggregates.

3.8.3 Control of area sample sizes: compromise solutions

More difficult, of course, is the problem of large variations in sample size among the sample areas. When the discrepancies are large, it can help considerably to use some approach that provides a compromise between the self-weighting and constant-take options. For instance, the use of a modified size measure of the type

$$p''_i = (p_i * p'_i)^{1/2}$$

in $f2_i$ can avoid extreme variations both in the overall selection probabilities and in the sample-takes per cluster. Here pi is the measure of size of area i in the frame, as used for selecting areas, and p'_i is the actual size of the area found after listing in the areas selected into the sample. The measure p''_i, a compromise between the above two, is used to determine the selection rate as follows.

With the sampling rate for the selection of households in the area as

$$f2_i = \left(\frac{b}{p''_i}\right),$$

we have, for the sample take in the area,

$$b_i = f2_i * p'_i = b.\left(\frac{p'_i}{p_i}\right)^{1/2}, \text{ and } f_i = f1 * f2_i = f.\left(\frac{p_i}{p'_i}\right)^{1/2}.$$

For units which have quadrupled in size, for instance, the sample-takes as well as the sampling rates would change only by a factor of 2, and by a factor of around 3 in the

extreme case of a ten-fold change in the size measures. The statistical and practical effects of variations in sample-takes or sampling rates tend to be negligible for small departures, small for moderate departures, but increasingly large for large departures. Hence such compromise solutions can enhance efficiency.

3.8.4 Take-all sampling possibly with a cut-off

There are situations where it is appropriate to take all the ultimate units (households) in the area selected at the last area-stage. The Such a scheme is called "take-all" or "compact cluster" sampling. The obvious example is a household listing operation, which must cover each sample area exhaustively.

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 areas.

With take-all sampling the selection probability of the ultimate units is the same as that of the clusters they come from. To obtain a self-weighting sample of the former, for instance, the areas have to be selected with equal probabilities. When the area sizes vary considerably, stratification by size for their selection can become important.

It is also common to select clusters with probabilities proportional to some measure of size (PPS). For the listing operation this would normally be the case for controlling sample sizes for the *next* operation such as an LFS or CLS, for which the sample may be selected with inverse PPS as described above.

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.

A commonly used option in such designs is to impose a limit on the maximum number of ultimate units which will be selected from any area. For a given situation, design and target sample size, the upper limit determines the number of areas to be selected for the sample – and, conversely, the number of areas determines the appropriate limit for meeting the desired sample size.

3.9 Dealing with very large units

"Very large" in the context of PPS sampling means a unit whose size measure in equation [1] exceeds the sampling interval, i.e. $p_i > I$, so that the implied probability of selection f1i exceeds 1.0, which is not possible. There are several methods of dealing with this problem:

1. Large units in the sampling frame may be segmented (divided into smaller areas) such that no segment exceeds I in size. The segments then form the appropriate sampling units.

2. The division of large units may only be conceptual. A large unit is considered as if it were made of t units as defined below:

   $t = \mathrm{Int}(p_i/I)$ with probability $\quad x = \mathrm{Int}(p_i/I) + 1 - (p_i/I)$, and

   $t = \mathrm{Int}(p_i/I) + 1$ with probability $\quad (1-x) = (p_i/I) - \mathrm{Int}(p_i/I)$.

   The unit is assumed to have been selected t times, and t separate sub-samples of the ultimate units are selected from it in the usual way. This type of procedure is often most conveniently used with systematic sampling. Random variation in the value of t makes the sample size variable, but the selection probabilities are correctly maintained.

   Note that with systematic sampling, it is not necessary to actually determine t, which is a random variable, for each large area. When the units are put in a list and the systematic sampling interval I is applied to their cumulative size measures, t is simply the number of selection points which fall in the area. The important point is that once an area is selected t times in this way, not one but t separate sub-samples have to be selected from it. Alternatively, a single sub-sample but t times the size of the normal sub-sample may be selected from the entire area.

3. When treating large units as automatically selected, the area unit is taken out of the sampling frame and is assumed to have been selected automatically ("self-representing" unit) and is then sub-sampled with the required overall selection probability. This gives $f1_i = 1; \quad f2_i = f.$

   This procedure amounts to assigning to every large unit with size $p_i > I$ a new size measure $= I$ and selecting the sample of areas systematically with interval I. In a self-weighting design, the sample size from the area is then proportional to its current size, $b_i = f * p'_i$.

4. In a design with a fixed number of units b=f*I selected per area, the following procedure may be used. The large unit is taken into the sample automatically and the b ultimate units are selected at the second stage. This means

   $f1_i = 1,\ f2_i = b/p''_i = f*(I/p''_i).$

   Therefore the data must be weighted up by (p'_i/I); clearly, such a design is unsuitable if the area unit's size is significantly larger than I.

Method 1 is in principle the best but it can be expensive. It is required when a high proportion of the units in the frame are "very large" in the above sense. Method 2 has sometimes been used for convenience as it avoids the need to remove the large units from the frame used for sample selection; however, it increases variability in the final sample size. Method 4 is reasonable only if no sizes p_i greatly exceed the limit I.

Method 3 is normally the recommended procedure.

3.10 Dealing with very small units

"Very small" in the context of PPS sampling means a unit whose size measure in equation [2] is smaller than the required sample-take, i.e. $p_i < b$, so that the implied probability of sub-sampling $f2_i$ exceeds 1.0, which is not possible. Again, there are several methods of dealing with this problem:

1. Small units in the sampling frame may be grouped together (merged to form larger areas) such that no group is smaller than the required sample-take b. Each group then forms an appropriate primary sampling unit.

   In order to retain geographical contiguity, small units are often merged with neighbouring units, whether small or large. This requires knowing the physical location and boundaries of units in the frame, which is not always available in list frames of area units without accompanying maps. However, available lists are often geographically ordered, and this can help in identifying approximately contiguous units.

   An alternative is to make a separate stratum of small units and then to simply select an equal probability sample from that. This, however, does not control geographical spread of the small units in the sample.

2. It is often more convenient first to select a sample of the original area units in the usual way, and only then group the selected units as necessary to achieve the required minimum size before the next stage of sampling.

   This is permissible when the rules used for the formation of groups are objective, in the sense that how units are grouped is independent of what particular units happen to be selected in the first sample. Such rules can be easily devised.

3. Another method is simply to assign to all units a minimum size measure $=b$. In other words, all units in the frame are given a minimum size b at the area selection stage, irrespective of their actual size. With units for which the original size was equal to or smaller than b, all ultimate units are taken into the sample at the last stage ("take-all" or "compact cluster" sampling). The original ultimate selection probabilities are retained unchanged but, for a given sample size, the number of area units in the sample is increased. This can be inconvenient and expensive if there are too many small units in the frame.

   Note that this procedure simply means that units classified as "very small" get selected with constant probability rather than with PPS. This is similar to forming a separate stratum of such units and selecting them at some appropriate constant rate.

4. One solution to the above problem is to retain the area size measures unchanged. This will not increase the number of area units selected, but it does require weighting up of the final samples from units with $p_i<b$ by the factor (b/p_i).

5. Excluding the smallest of small areas, with appropriate compensation has proved a useful method in a number of child labour surveys. In the presence of many small

units, this more practical solution involves a combination of methods 3 and 4. Even though this is an approximate method, it avoids the inclusion of too many very small units in the sample. It is particularly useful when the population includes many units, each of them with only a few (or even no) ultimate units of interest – as may well be the case in child labour surveys. This procedure is outlined below, and a simulated numerical example is presented in the next section.

(i) Let Σp_i be the total size measure of the set of "very small units" to be sampled, with design parameters I as the sampling interval to be applied for systematic selection of areas, f the overall sampling rate, and b=f*I the expected sample size per sample area. Under this design, the expected number of ultimate units which should be selected from this stratum is obviously $f \cdot \Sigma p_i$.

(ii) This set of small units is ordered by unit size and divided into two parts. The first part consists of $A = \Sigma p_i / b$ largest units in this small-unit set. These units are selected as in method 3 above. The size measure of each unit is increased to b, so that selection with interval I gives a sample of $a = A.b/I = \Sigma p_i / I$ area units. All ultimate units in each selected area are retained in the sample.

(iii) The second part consists of the smallest of the small-unit set. For reasons of cost and practicality, these units are altogether excluded from the sample, even though this is not properly a probability sampling procedure.

(iv) By way of compensation for this exclusion, the weight given to each selected unit in the first part is increased: all ultimate units from a selected area of size p_i are weighted up by the factor (b/p_i), just as in method 4 above. This simply makes the weighted sample-take from each selected area b=f*I, giving an equivalent of $a.b = f.\Sigma p_i$ for the selected areas.

(v) The weight arrived at in (iv) is exactly the correct expected number noted in (i).[14] In fact, it is to achieve this balance that the number A of areas included in the first part was determined in (ii) above as $A = \Sigma p_i / b$.

Fixed-take design

Methods I and 2 can be applied as above. Other options are, of course, not possible for small areas, $p_i < b$, since such areas cannot yield the pre-fixed sample-take b, except in cases where the actual size has grown to equal or exceed b.

[14] The actual number of units appearing in the sample would vary according to differences between the size measures p_i in the frame and the actual size of the units found at the time of selecting the final units. But the selection procedure is specified entirely in terms of the size measures in the frame.

3.11 A numerical illustration

3.11.1 Simulation of a population of area units

To illustrate some of the points made above, we include here a numerical example. A small data set has been generated statistically to provide a reasonably realistic example of a set of areas with their associated measures of size. The original population consisted of some P=50,000 ultimate units (households), in nearly 800 areas of an average size of around 65 households. The areas varied considerably in size, from the smallest, consisting of fewer than 10 units, to the largest with over 1,000 units. The distribution is fairly typical of real surveys, strongly skewed to the left with many small areas and a few very large ones.

A systematic PPS sample of a=100 areas was selected, the sampling interval for the selection of areas being I=(P/a)=500. The data tabulated in Table 3.3 show only the selected 100 areas, rather than the whole population of nearly 800 areas.

The columns of Table 3.3 are as follows. The 100 units selected into the sample with probability proportional to size measures p_i are listed in order of size. Units so listed are numbered sequentially 1-100 in column [2], and the corresponding size measures p_i are shown in column [3].

A minimum limit of 20 and a maximum limit of 500 are imposed on the measures used for the purpose of sample selections. The lower limit corresponds to the target number, b=20, of households to be selected per area, and the upper limit to the selection interval, I=500, used for systematic sampling. These modified measures for the selected units are shown in column [4]. The PPS selection equation used for selecting the sample shown is

$$f0_i = p_i/500$$

where pi refers here to the adjusted pi values in column [4]. These vary from 4 per cent for the smallest areas to 100 per cent for the biggest.

Columns [6]-[7] show the number of working children, c_i, in the area, and working children per household (c_i/p_i). These figures will be used for the selection of a sub-sample of areas for the CLS/LCS surveys and will be discussed in Chapters 4 and 5.

Column [1] is based on the list of areas in the whole population ordered by p_i value. It shows the cumulative number of areas in the population at points where an area was selected. Column [2] can be seen as such a cumulation, but confined to the sample areas. Thus, for instance, the smallest 30 per cent of the areas in the population account for 10 per cent of the sample. Similarly the smallest 53 per cent of the areas in the population account for 20 per cent of the smallest areas in the sample. This results from the PPS selection method.

Table 3.3. Illustration: A PPS sample of area units

"First sample": for instance for a LFS or CLS.
Parameters:
a= 100 [number of areas in sample]
b= 20 [expected number of units selected/sample area]
I= 500 [interval for systematic selection of areas]
n= 2,000 [=a*b, expected sample size]

[1]	[2]	[3]	[4]	[5]	[6]	[7]
Cumulative frequency of number of areas in		Area measure of size		Area selection probability	No. of labouring children in area	
Population	Sample	pi	Modified pi	f1 (%)	ci	ci/pi
3.0	1	8	20	4.00	9	0.44
6.0	2	11	20	4.00	5	0.26
9.0	3	12	20	4.00	9	0.44
12.1	4	16	20	4.00	0	0.00
15.1	5	17	20	4.00	3	0.14
18.1	6	18	20	4.00	1	0.03
21.1	7	18	20	4.00	12	0.62
24.1	8	20	20	4.00	5	0.24
27.1	9	21	21	4.21	0	0.00
30.0	10	22	22	4.37	18	0.81
32.8	11	24	24	4.72	0	0.01
35.3	12	25	25	5.07	11	0.45
37.7	13	28	28	5.56	1	0.03
39.9	14	28	28	5.57	0	0.02
42.0	15	30	30	6.02	6	0.21
44.1	16	33	33	6.55	5	0.15
45.9	17	34	34	6.75	0	0.01
47.7	18	35	35	6.97	11	0.31
49.4	19	36	36	7.16	4	0.11
51.1	20	38	38	7.52	17	0.45
52.7	21	39	39	7.81	8	0.21
54.2	22	40	40	7.94	1	0.01
55.8	23	40	40	8.04	0	0.00
57.3	24	43	43	8.57	20	0.46
58.7	25	43	43	8.63	27	0.62
60.1	26	44	44	8.80	23	0.51
61.4	27	46	46	9.23	18	0.39
62.8	28	48	48	9.58	23	0.48
64.0	29	50	50	10.03	7	0.15
65.2	30	52	52	10.39	1	0.01
66.4	31	52	52	10.40	0	0.01
67.5	32	53	53	10.58	11	0.20
68.7	33	57	57	11.33	9	0.16
69.7	34	59	59	11.85	0	0.00
70.8	35	60	60	11.97	53	0.89
71.8	36	64	64	12.71	11	0.17
72.7	37	65	65	13.02	12	0.18

[1]	[2]	[3]	[4]	[5]	[6]	[7]
Cumulative frequency of number of areas in		Area measure of size		Area selection probability	No. of labouring children in area	
Population	Sample	pi	Modified pi	f1 (%)	ci	ci/pi
73.6	38	67	67	13.35	11	0.17
74.5	39	68	68	13.66	2	0.03
75.4	40	69	69	13.81	3	0.04
76.3	41	69	69	13.88	1	0.02
77.2	42	70	70	13.91	61	0.87
78.0	43	72	72	14.42	3	0.04
78.9	44	72	72	14.48	35	0.48
79.7	45	73	73	14.65	51	0.69
80.5	46	74	74	14.81	20	0.27
81.3	47	79	79	15.72	41	0.52
82.1	48	79	79	15.82	7	0.09
82.9	49	83	83	16.60	74	0.90
83.6	50	92	92	18.37	1	0.01
84.3	51	92	92	18.38	20	0.21
84.9	52	97	97	19.36	22	0.23
85.5	53	97	97	19.50	15	0.16
86.2	54	100	100	20.04	93	0.92
86.8	55	103	103	20.57	73	0.71
87.4	56	107	107	21.50	16	0.15
87.9	57	112	112	22.39	21	0.19
88.5	58	114	114	22.81	25	0.22
89.0	59	114	114	22.89	20	0.18
89.5	60	115	115	22.98	68	0.60
90.0	61	124	124	24.76	27	0.21
90.5	62	128	128	25.53	0	0.00
91.0	63	138	138	27.62	3	0.03
91.4	64	156	156	31.19	16	0.11
91.8	65	157	157	31.48	117	0.74
92.2	66	159	159	31.89	1	0.01
92.6	67	163	163	32.55	51	0.31
92.9	68	171	171	34.28	22	0.13
93.3	69	175	175	34.94	11	0.06
93.6	70	181	181	36.29	40	0.22
94.0	71	186	186	37.10	23	0.12
94.3	72	189	189	37.87	20	0.11
94.6	73	196	196	39.29	190	0.97
94.9	74	200	200	40.03	1	0.00
95.2	75	204	204	40.73	8	0.04
95.5	76	213	213	42.66	55	0.26
95.8	77	213	213	42.67	1	0.01
96.1	78	229	229	45.79	107	0.47
96.4	79	229	229	45.86	42	0.18
96.6	80	237	237	47.36	9	0.04

[1]	[2]	[3]	[4]	[5]	[6]	[7]
Cumulative frequency of number of areas in		Area measure of size		Area selection probability	No. of labouring children in area	
Population	Sample	pi	Modified pi	f1 (%)	ci	ci/pi
96.9	81	238	238	47.53	2	0.01
97.1	82	240	240	47.96	7	0.03
97.4	83	246	246	49.30	1	0.00
97.6	84	299	299	59.79	268	0.90
97.8	85	304	304	60.76	36	0.12
98.0	86	331	331	66.29	220	0.66
98.2	87	345	345	68.99	1	0.00
98.4	88	358	358	71.68	26	0.07
98.5	89	368	368	73.67	162	0.44
98.7	90	383	383	76.58	2	0.01
98.9	91	424	424	84.73	0	0.00
99.0	92	425	425	85.06	382	0.90
99.2	93	527	500	100.00	292	0.58
99.3	94	566	500	100.00	92	0.18
99.4	95	659	500	100.00	0	0.00
99.5	96	681	500	100.00	60	0.12
99.6	97	719	500	100.00	383	0.77
99.8	98	898	500	100.00	0	0.00
99.9	99	908	500	100.00	9	0.02
100.0	100	1076	500	100.00	39	0.08

With PPS sampling, the average size of the area in the sample can be much larger than the average size in the population – in this example around 170 versus 65. This difference depends on how variable the unit sizes are in the population. In fact, the ratio of average area sizes between the sample and the population is given by

$$\frac{\overline{B}_{sample}}{\overline{B}_{population}} = 1 + cv^2(B_i)$$

where cv (B_i) is the coefficient of variation of sizes of individual areas in the population[15]. The ratio of sample to population average area size in the illustrative population corresponds to a cv of 1.3, indicating a high degree of variability in the area sizes.

The practical implication of this point is that PPS sampling increases the work required for listing, sample selection and enumeration in the survey, compared with what might be expected from the average size of area units in the population. The increase depends on the variability of the unit sizes; the PPS sampling scheme brings into the sample a disproportionately large number of bigger units, which is of course compensated by a correspondingly lower proportion of ultimate units selected into the sample from each bigger area at the next stage of sampling.

[15] Note that the above equality is in terms of "expected values", i.e. it holds for quantities averaged over all possible samples with a given design. For any particular sample, the values are of course subject to sampling variability.

Let us suppose that this is the first stage of a two-stage self-weighting sample with overall probability f=f1·f2=4%. The selection equation at the second stage is $f2_i=20/p_i$, giving $f=p'_i/500\cdot 20/p_i=4/100$.

For simplicity, we will generally assume in the illustrations in this and the following chapters that the actual sizes of the areas (p'_i) at the time of enumeration are the same as the original size measures pi used in the PPS selection of areas. In reality, these quantities will differ primarily in accordance with the age of sampling frame.

The sample size of the ultimate units from a cluster is $p'_i\cdot f2_i=20(p'_i/p_i)$, which is a constant (=20) with the above assumption.

3.11.2 Dealing with very small and very large units

For the illustration, it was assumed that the target sample-take per area is b=20. This determines the minimum size measure which can be assigned to any unit. (There are 7 units out of 100 in the sample, unit numbers 1-7, with size measures below this level.) Units with sizes up to this limit have been selected with probability $f1_i=(b/I)=(20/500)=4\%$, and all ultimate units in each area selected are taken into the sample ($f2_i=1.0$). This gives the required overall sampling rate f_i also =4%. This is the constant overall sampling rate, which applies to all ultimate units in the population.

The sampling interval I=500 determines the maximum size measure that can be assigned to any unit. (There are 8 sample units in the example, unit numbers 93-100, with size measure exceeding I.). Units with sizes at or above this limit are automatically taken into the sample ($f1_i=1.0$), and are sampled at the rate $f2_i=(b/I)=(20/500)=4\%$ at the last stage. This selection also gives the overall sampling rate f_i =4%. The number of ultimate units from such an area which are included into the sample will exceed the target sample size b=20 to the extent that the area's original size measure b_i exceeds the sampling interval I (500). For instance, the largest area (size=1,076) will give 20*1,076/500=43 ultimate units for the sample.

3.12 Non-probability selections

In the context of official statistics, almost all surveys need to be based on probability and measurable samples. Nevertheless, various forms of departure from probability samples are commonly encountered in other contexts, such as in market and opinion research, mainly for reasons of cost and convenience. This may also be the case for particularly intensive and/or restricted surveys of child labour, such as surveys of special populations of labouring children in particular sectors or types of activity.

The objective of this section is to describe some of the commonly encountered non-probability sampling methods, noting their limitations but also their potential uses.

It is important to emphasize at the outset that, as a matter of good practice, every effort should be made to obtain probability samples. The discussion below should be interpreted in this context.

Limitations and possible uses of non-probability sampling methods

For probability sampling, randomization, rather than assumptions about the structure of the population, is the characteristic feature of the selection process. By contrast, non-probability sampling does not involve *random* selection. In non-probability sampling there is no way to estimate the probability of any particular unit being included in the sample, or to ensure that each unit has a chance of being included, making it impossible to estimate sampling variability or bias in the survey results. Non-probability samples cannot depend upon the rationale of probability theory: there is no "sampling distribution" to use in deriving results from the sample. Instead, we have to use some assumed probability model. This may be a model of the underlying population characteristics or of the selection process.

With such samples, we may or may not represent the population well, and it is difficult to evaluate the outcome. Reliability cannot be measured in non-probability sampling; the only way to address data quality is to compare the survey results with information about the population available from other sources.

Does that mean that non-probability samples can provide no useful information? No, not necessarily. Despite these drawbacks, non-probability sampling methods can be useful. There may be circumstances where it is not feasible, practical or theoretically sensible to do random sampling. Also non-probability methods are often quick, inexpensive and convenient. Their possible advantages include:

- They are at least feasible when probability sampling is not possible – for example when a sampling frame is not available.
- They are normally cheaper and quicker. These advantages can be particularly useful when the population is large or widely dispersed.
- They are often useful in exploratory studies such as studies for hypothesis generation or questionnaire design. For instance, in questionnaire testing we may be more interested in obtaining an idea of the range of responses than in working out what proportion of population gives a particular response.

When non-probability methods may be useful

There can be three types of reasons for using non-probability methods for sample selection.

1. A probability sample is considered desirable but is not feasible

This refers to situations *where* the choice is only between taking a non-probability sample or not doing the survey at all. For example, it is not possible to use a strict probability method of sampling when a suitable sampling frame is not available for certain populations or particular groups of the population, such as a list of children not attending school or of households containing such children; or when the expected non-response rate is so high that a probability sample cannot be achieved even if it was selected to be a probability sample; or when the cost and time required for obtaining a probability sample are not affordable.

2. A probability sample is considered unnecessary

This applies to situations *where* the population is relatively homogeneous (units have similar characteristics) or is well mixed (the arrangement of units in the population is randomized), so that it does not matter much exactly how the sample is selected. The so-called haphazard methods of sampling described below are based on this assumption. In such situations, non-probability selection methods can suffice, and they are chosen to the extent that they are quicker, cheaper or more convenient.

3. A non-probability sample is preferable

This must be considered a rare situation, but it is definitely a possibility. Essentially, it arises in situations *where* certain types of objectives are served better by using judgement to select the units than on the basis of a totally randomized procedure. The assumption is that the bias resulting from the expert's judgement is smaller than the variability in random samples of small size. The so-called judgement methods of sampling described below are based on this assumption.

Non-probability samples in official statistical agencies

Most official statistical organizations use probability sampling for their surveys, but for questionnaire testing and preliminary studies during the development stage of a survey they generally use non-probability sampling. In a number of countries major official sample surveys of businesses use purposive selection because of severe problems in obtaining respondent cooperation. Often parts of the consumer price index are based on non-probability sampling (for example, in selecting the outlets for price), largely on cost grounds. Another example: many official surveys of expenditure by tourists depend on quota samples (based on country of residence, length of trip, demographic profile, airport or station, etc.).

Departures from probability sampling can be more acceptable in situations where the information is used mainly by a single or a very restricted set of clients. In the public sector, however, there is no single well-defined user and it is highly desirable that the sampling method has wide acceptability. Random sampling methods provide that wide acceptability. Official statisticians are responsible for generating data that can be used by all of society, and therefore they should not tolerate any controllable bias in their products and should carry out sample surveys using probability sampling methods.

Various non-probability sampling methods in use

A wide range of non-probability methods of sampling is encountered in practice, and the labels used to identify and distinguish them are not quite standardized. Below we try to categorize the methods a little more systematically, and to provide a brief description of the most common.

3.12.1 Haphazard sampling

Haphazard sampling means *taking units into the sample without any structure, rules or fixed procedures*. As noted, the basic assumption of this approach is that the population is relatively homogeneous or is well mixed, so that it does not matter much exactly how the sample is selected.

Convenience sampling

Particular units are taken into the sample simply because it is more convenient to do so than to take some other, different units.

Such a sample is not normally representative of the target population because sample units are only selected if they can be accessed easily and conveniently.

There are many times when the average person uses convenience sampling in normal day-to-day life. Another example: television reporters often seek so-called "people-on-the-street interviews" to find out how people view an issue.

The obvious advantage is that the method is easy to use, but that advantage is offset by the great danger of bias. Although useful applications of the technique are limited, it can deliver accurate results when the population is homogeneous or well mixed.

Volunteer sampling

As the term implies, this type of sampling occurs when people volunteer their services for the study. In psychological experiments or pharmaceutical trials, for example, it would be difficult, even unethical, to enlist random participants from the general public. In these instances, the sample is taken from a group of volunteers. In clinical practice, we might use clients who are available to us as our sample. Television and radio media often use "call-in" polls to query an audience on their views. In many research contexts, we sample simply by asking for volunteers. Sometimes, the researcher offers payment to entice respondents.

Clearly, the problem with all of these types of samples is that we have no evidence that the respondents are representative of the target population, and in many cases it is clear that they are not.

Sampling voluntary participants as opposed to the general population may introduce strong biases. For example in opinion polling, often only the people who care strongly about the subject one way or another tend to respond. It can be assumed that generally people who participate in these surveys have different views from those who do not.

Snowball sampling

Snowball sampling begins with the identification of someone who meets the criteria for inclusion in the study. That person is then asked to recommend others who they know who also meet the study criteria. The basic idea is to expand the sample originally selected by allowing units selected to bring into the sample other units related to them in some way, The most common application, however, is to apply the process until a sufficient number of cases of the population have been generated for the purpose of the survey, though the ultimate goal may also be to construct a sampling frame of the

population of interest. *Often the emphasis is on getting enough cases rather than on the units constituting a representative sample.*

Snowball sampling is especially useful when trying to reach populations that are otherwise inaccessible or hard to find. For instance, if you are studying the homeless, you are not likely to be able to find a good list of homeless people within a specific geographical area. However, if you go to that area and identify just a few homeless persons, they may very well know who the other homeless people in their vicinity are and how they can be found.

A necessary condition for the application of this procedure is that members of the population are connected with or know each other. By the very nature of "snowballing", members of a rare population who have many contacts with other members of that population tend to be over-represented in the survey, while those who are isolated from others tend to be under-represented. The procedure therefore is unlikely to yield a representative sample, though it may still succeed in providing useful information and in certain situations may be the only option available for studying the elusive population.

3.12.2 Judgement sampling

The term *judgement sampling* is used to refer to the selection of units on the basis of the judgement that the selected units are somehow "representative" of the population (or of some specific aspects of the population) of interest.

Often this method is used in exploratory studies like the pre-testing of questionnaires, in focus groups, in laboratory settings where the choice of experimental subjects reflects the investigator's pre-existing beliefs about the population. The underlying assumption is that the investigator will select units that are characteristics of the population for the purpose at hand. The critical issue here is objectivity: how much can judgment be relied upon to arrive at a typical sample?

Purposive sampling

Such a procedure may be used when the number of units to be selected is so small that variability with random selection will be excessively large, and potentially more damaging than the bias inherent in selection by judgement. The judgement is made about whether or not to include particular units into the study, rather than about some mechanism of selecting units for the sample.

Studies based on a very small number of areas or sites are typical examples, and the standard illustration in sampling textbooks is where we want to get information about the urban population but can afford to sample in only one city. In that case, the textbooks are clear that it would be better not to use probability-based sampling to choose a "representative" city and they suggest using judgement instead. However, though the areas included may be determined on the basis of judgement, the selection of the ultimate units (households, persons, children, etc.) within each area may be randomized.

The assumption of the procedure is that the phenomenon of interest in the general population is represented, and can be captured, in a restricted number of carefully

selected units. It is assumed that the main variability lies within, rather than across, units from which only a very limited number has been included for study.

Restricted sampling locations

Restricted sampling locations may be used when a significant part of the sub-population of interest is believed to be confined to a limited number of known areas. On the basis of such judgement or assessment, the sample is often restricted to those locations. This can greatly reduce survey costs, but again bias is introduced to the extent that the uncovered part of the population lying outside the areas of concentration has different characteristics than the part covered.

Despite the need to restrict the number of locations to be included, it is desirable to retain the probability nature of the sample to the maximum extent possible, for instance by using random sampling of individual units within each location used for the restricted sample.

Modal instance sampling

The mode refers to the most frequently occurring value in a distribution. When we take a modal instance sample, we are trying to include the type of cases which are most frequent or typical in the population. Many public opinion polls, for instance, claim to interview "typical" voters. There are number of problems with this sampling approach. Above all, how do we know what the "typical" or "modal" case is? Clearly, modal instance sampling is only worth considering for informal sampling.

Sampling for extremes

This is judgement sampling in which we look for extreme cases, for instance to understand or bring out underlying factors, causes, consequences, etc. This sort of approach can, for example, be useful for surveying the worst forms of child labour.

3.12.3 Quota and other 'structured' sampling

This refers to forms of non-probability sampling in which the selection of units from the population is mediated through a structure or set of constraints. However, within that structure the selections are non-random. The tightness of the structure determines how close the resulting sample is to a probability sample. An example is the allocation of sample quotas to different parts of the population, as described below.

Quota sampling

This is one of the most common forms of non-probability sampling and refers to selection with controls ensuring that specified numbers (quotas) are obtained from each specified subgroup in the population (such as households or persons classified by relevant characteristics), but with essentially no randomization of the selection of units within the subgroups. The basic idea in quota sampling is to produce a sample matching the target population with respect to certain characteristics (e.g. age, sex) by filling quotas for each of these characteristics. It is presumed that, if the sample matches the population in these characteristics, it may also match it in the quantities we are trying to measure. Note that the method requires that good data on the whole

population be available to set the quotas. For example, if we are setting age and sex quotas, we need to know the age and sex distribution of the population.

The assumption of the procedure is that the main variability lies across, rather than within, the subgroups chosen, so that once sufficiently small and homogeneous groups have been defined and properly represented, it does not matter very much which particular individual units within a group are enumerated.

Sampling is done until a specific number of units (quotas) for various sub-populations has been selected. Since there are no rules as to how these quotas are to be filled, quota sampling is really a means for satisfying sample size objectives for certain sub-populations. Quota sampling can be considered preferable to other forms of non-probability sampling (e.g. judgement sampling) because it forces the inclusion of members of different sub-populations.

Quota sampling is often used by market researchers instead of stratified probability sampling, and is usually justified in terms of its convenience, speed and economy. It is easier to administer and has the desirable property of satisfying population proportions. This is so because it avoids the task of listing the whole population, randomly selecting a sample and following-up on non-respondents. Quota sampling is an effective sampling method when information is required quickly. It may be the only appropriate method when there is no suitable list of the population to be surveyed.

However, quota sampling disguises a potentially significant bias. As with all other non-probability sampling methods, in order to make inferences about the population it is necessary to assume that the persons selected are similar to those not selected. Such strong assumptions are rarely valid.

Although interviewers are constrained by the quotas, they are still using some elements of judgement in the choice of the sample. *The amount of flexibility interviewers have varies from survey to survey, and it is these rules and guidelines that determine how far the quota sample departs from probability-based stratified sampling.*

Just as there are many probability-based sample designs, quota sampling is not a single method. A quota sample may be drawn in stages. It is common, but not necessary, for quota samples to use random selection procedures at the beginning stages, much in the same way as probability sampling does. For instance, the first step in multi-stage sampling would be to select the geographic areas randomly. The difference is in the selection of the units in the final stages of the process.

Comparing probability-based and quota sampling

The main differences between probability-based and quota sampling are the following:

1. If probability-based sampling is properly carried out, there will be none of the bias which can arise from subjective judgements in sample selection. There is the possibility of such bias, however, in quota samples. For example, interviewers may consciously or unconsciously choose non-threatening or easy-to-approach respondents, or those who are easy to contact.

2. The quota method demands the formulation of a hypothetical model to fit the data. On the other hand, a probability-based survey does not, in principle, depend upon any model. The validity of the model underlying quota sampling may be open to question, and difficult to verify.

3. With probability sampling, we use statistical procedures to draw conclusions from the sample and sampling errors. In a quota sample, we cannot obtain comparable estimates of precision.

4. As a rule, non-response in a quota sample is handled simply by selecting another respondent who fits the quota. Non-response in a probability-based sample can usually be handled more effectively.

5. In general, the cost of a quota sample will be lower than that of a probability-based sample of the same size. But the question of cost is more complex and cannot be looked at in isolation from data quality. "There is no way to compare the cost of a probability sample with the cost of a judgement sample, because the two types of sample are used for different purposes. Cost has no meaning without a measure of quality, and there is no way to appraise objectively the quality of a judgement sample as there is with a probability sample." (Edward Deming).

Heterogeneity sampling

We sample for heterogeneity when the objective is to capture the full complexity of the phenomenon (e.g. opinions or views), but representing these views or the individuals proportionately is not of concern. Another term for this is sampling for diversity. In many brainstorming or similar group processes, for instance, we would use some form of heterogeneity sampling when our primary interest is in getting a broad spectrum of ideas, and not in identifying the "average" or "modal" ones; in effect, what we would like to be sampling is not people, but ideas. We imagine that there is a universe of all possible ideas relevant to some topic and that we want to sample this population, not the population of people who have the ideas. Clearly, in order to get all of the ideas, and especially the outlying or unusual ones, we have to include a broad and diverse range of participants. Heterogeneity sampling is, in this sense, almost the opposite of modal instance sampling described above.

This type of sampling can take the form of a somewhat less restrictive non-proportional quota sampling. In this method, the minimum numbers of units to be sampled in each category are specified. There is no concern with having numbers that match the proportions in the population.

Sampling at location

This technique is used for sampling populations which are defined on the basis of some activity/state in which they engage/exist at certain specific locations. Examples are passenger surveys at airports or bus or rail stations, surveys of visitors to museums, shopping centres, etc.

In general, this type of sampling scheme involves the selection of locations, of observation times, of individual units, and possibly also of a subset of activities or states to be observed.

An important distinction is whether the survey is designed to estimate "units" or "activities" as the elements for analysis. An example is visitors as opposed to visits to a health facility. A unit's selection probability depends on the number of "activities" through which the unit may be enumerated. Special procedures are required when a sample of the former (units) is constructed from the latter (activities, etc.). By contrast, for many purposes sampling at locations is done primarily to study activities or visits etc. rather than individual persons. This is more straightforward in terms of sampling.

Common difficulties in such surveys include the adverse conditions of, and the limited time available for, sample selection and data collection for enumerating flows at fixed locations. This can result in large selection and measurement biases as well.

Chapter 4

Child labour survey (CLS): Linked sample of reduced size

As defined in Chapter 2 (Section 2.1.1), we use the term *child labour survey* (CLS) to refer specifically to a survey or survey component designed with the primary objective of measuring the prevalence of child labour, possibly including, if relevant, information on the variation of the prevalence by geographical/administrative division, type of place, and various household and personal characteristics.

The complexity of the data collection and questionnaires required for this purpose is affected by the scope and detail of the information required for identifying child labour according to the definition adopted in the survey in question.

In defining the structure of a CLS, we have to consider its links with the operations, if any, which precede it, and those which follow it. This chapter is concerned with the linkage of a CLS with the operation preceding it. There are three dimensions: (1) the preceding operation may be a household listing operation, or it may be a large-scale survey such as the LFS; (2) the CLS may be combined with that operation, or be conducted subsequently as a separate operation; and (3) the CLS may be conducted on the same sample, or on a sub-sample of the preceding operation. The combinations of these are shown below. Of course some of these combinations are more likely (more meaningful, practical) than others.

Linkage of a CLS with the operation preceding it				
Whether the two operations are integrated or separate				
Preceding operation	**Integrated**		**Separate**	
Household listing	CLS involving brief screening questions	Same (full) sample	Stand-alone CLS	Use of full listing sample unlikely
		Sub-sampling unlikely		Sub-sampling
LFS or similar	Modular or "combined" CLS	Same (full) sample mostly	Linked CLS	Same (full) sample sometimes
		Sub-sampling sometimes		Sub-sampling mostly

Integration of the CLS with a household listing operation requires that it involve no more than a few brief questions which can be incorporated into the listing form. All households within each sample area have to be listed, and generally no sub-sampling is possible for the CLS questions as well. The resulting data may have the advantage of being based on a large sample, but the measurement of child labour is likely to be approximate. It is more appropriate to view this type of CLS as a *screening operation* (see Section 2.3).

The child labour survey may be a "stand-alone" survey, normally based on a household listing operation, and not linked to another survey such as an LFS. Sub-sampling within

the areas listed will be normally required. In this situation, the general sampling principles for different types of population surveys discussed in the previous chapter apply to the CLS as well. However, a stand-alone child labour survey is not always a feasible or even a desirable option.

An alternative is to base the CLS, at the least in terms of sampling, on another larger household survey, such as the LFS. It is common in this situation to incorporate the CLS in the base survey (LFS) as a *module* (or in the form of a "combined" survey in the sense described in Section 2.2.3). Normally this arrangement involving operational integration implies that the CLS module is applied on the same sample of household as the base survey. Sub-sampling at the area level (i.e. introducing the CLS as a module only in a sub-sample of LFS areas) is, however, possible. A particularly convenient form of this practice is to include the CLS as a module only during some of the rounds of a continuing LFS.

An alternative option is to have a *linked CLS* in the sense explained in Section 2.2.3 – as a survey dependent on the base survey for its sample and possibly other information fed forward, but otherwise operationally separated from it. Here more elaborate sub-sampling from the base survey, both at the area and the household level, is possible.

For reasons further elaborated in Section 4.1.3, we feel that *the last-mentioned "linked CLS with sub-sampling" option deserves more favourable consideration than appears to have been the case in past child labour surveys.*

This chapter describes technical details of procedures for drawing a sample for a child labour survey (CLS) on the basis of the sample used for a larger survey of the general population, in particular the LFS. The two surveys tend to have similar requirements as regards the basic structure of the sample, but are expected to differ in terms of various design parameters and the forms of linkage between the two. Such differences have to be kept in mind when deciding on the sub-sampling procedures described below.

4.1 CLS survey structure and linkages

4.1.1 Common structure

In measuring the incidence of child labour, the base population of interest in a CLS is the *population of children exposed to the risk of child labour*. This base population is defined essentially in terms of age limits, which tend to be well distributed in the general population. Even within small areas, the size of this population is approximately proportional to the total population of the area in most cases. Hence, while the size measure in the LFS is the general population (or the population of working age), for a CLS it is an appropriately defined population of children exposed to the risk of child labour. It is generally the case that these two populations are closely related in size – the average difference between them being primarily a scaling factor.

These similarities in the basic structure of the samples have important implications for the choice of appropriate survey structure for the CLS. Two basic aspects are the

allocation of the sample among population domains or strata; and the selection of sample areas.

Sample allocation

Leaving aside any disproportionate allocation of the sample to meet special reporting requirements (such as over-sampling of small domains in order to achieve minimum precision requirements), it is generally appropriate and common in labour force surveys to allocate the sample proportionately, i.e. in proportion to the domain population size. In theory, the allocation may be "optimized" by taking into account differences in variance and cost among the domains. Usually, however, the gains are too small to justify a departure from the simpler and more practical proportionate allocation. This is particularly true for estimates of proportions (such as the unemployment rate).

The situation is very similar in the case of a CLS aimed at estimating the prevalence of child labour. Only in the presence of very pronounced differences among the domains in child labour prevalence rates can disproportionate (optimal) allocation be justified. For estimating a proportion p, the theoretically optimum sampling rate is in proportion to $[p.(1-p)]^{1/2}$. This variation in allocation is very insensitive to the value of p. lying between 0.3-0.5 for p in the range 0.1-0.9. Consequently, proportionate allocation remains generally appropriate for the CLS, just as for the LFS.

The required allocation between the two type of survey may differ, but primarily because of differences in reporting requirements for sub-national domains. (See Section 4.1.2.)

Selection of sample areas

The procedure for selecting sample areas can also be the same in so far as the relevant base population sizes for the two surveys are nearly proportional to each other. For instance, if selection with probability proportional to size p_i is suitable for the LFS, it is reasonable to assume that, for practical purposes, the same size measures p_i used in the LFS for PPS selection of areas are appropriate for the same purpose in the CLS. Consequently a CLS aimed at measuring the incidence of child labour among the population of children is likely to require a *sample structure similar to that for a survey of the general population* such as the LFS.

Apart from a similarity in the structure of the base population of interest, the two types of survey also tend to be similar in the mode of data collection, and in substantive aspects (concepts, definitions, classifications, questions, reference period etc.).

4.1.2 Design parameters

Despite the similarity in the structure and distribution of the population to be sampled, the CLS and LFS will as a rule differ in a number of design requirements and parameters. In many countries the LFS is a well established, regular or even continuous. The CLS is more likely to be a new or recently instituted survey, conducted at best periodically and often only on an ad hoc basis. The resources available for the CLS are likely to be more limited, and often less certain. Increasingly, the LFS is required to produce separate estimates for different regions and sub-populations in the country and produce these more frequently, such as annually or even quarterly, while in most cases

the primary objective of the CLS still has to be, first and foremost, the production of national-level estimates from time to time. In short, *the LFS may be seen as a large, extensive and regular survey* and, by comparison, *the CLS as a smaller, more intensive and less frequent survey*. Consequently, the two types of surveys differ in relation to the choice of design parameters, despite the similarity in the basic structure – design parameters such as sample size, the number of areas selected for the sample and the related sample size per area, the allocation of the sample across different domains in view of different reporting requirements in the two surveys, the details of the stratification, and the ultimate units for which data are collected in the survey.

4.1.3 Linkages

As noted above, it is possible in principle to conduct a CLS with exactly the same sample as an LFS (or some similar large-scale survey). Indeed, in a number of countries child labour questions have been simply added as a module to an ongoing LFS. Such an arrangement has a number of advantages in terms of cost saving, convenience and greater sustainability. The information from the two surveys (or rather the two parts of the same survey in this case) can be linked and analyzed together. On the other hand, there can be some serious disadvantages. The content of the supplementing module (the CLS) has to be kept limited, which may not adequately meet the data requirements for a comprehensive assessment of child labour. The response burden deriving from the combined interview is increased, with a possible negative effect on the response rates and response quality, especially for the main (LFS) component. The completely integrated system is also likely to be too rigid in the face of differing requirements for the LFS and CLS components.

At the other extreme, as already mentioned, there have been a number of cases where a national child labour survey has been conducted as an independent stand-alone activity. While this may be able to produce more reliable data on child labour, a stand-alone survey is clearly an expensive option, often difficult to sustain or repeat.

Comprehensive but costly stand-alone child labour surveys on the hand, and much cheaper and convenient but restricted supplementary modules on child labour, on the other, are two possible and quite commonly used options.

However, in many situations, neither of these extreme solutions – complete integration or complete separation – is the appropriate solution. Often a more practical solution is to seek to link the CLS to some large-scale survey of the general population, preferably to the LFS. This linkage can take many forms and can operate in different degrees.

As to the sampling aspects of such linkages, a convenient and practical option is to draw the child labour survey sample as a sub-sample from a larger survey of the general population, most appropriately the labour force survey.

Of course, there are a number of other possibilities. At one extreme, the two surveys may be based on independent samples, but even here it is desirable and efficient to draw them from a common area frame, or from a “master sample” of areas from which samples can be drawn for different surveys. This permits the sharing of the cost of preparation and maintenance of the area frame or the master sample of areas.

More appropriately, where possible, it is desirable to base the two surveys on the same sample of areas. In principle, all the areas in the LFS may be included in the CLS sample. However, sub-sampling of the LFS areas is often desirable and appropriate, from the point of view both of survey objectives and the availability of resources. Generally speaking, labour force surveys are required to produce estimates for detailed domains – different regions and other administrative divisions, urban and rural areas, different demographic and other subgroups in the population, and so on. By contrast, the information available on child labour is often very limited or non-existent, and initially results have to be produced at the national level, or at most only for a few major domains. The sample size for the CLS can therefore be much smaller. Similarly, a time-series of data on the general labour force, based on regular and frequent surveys, is normally required, while on child labour the first priority is usually to produce reliable information of a more "structural" nature. Child labour surveys therefore tend to be one-time surveys, at the most repeated occasionally but generally without linkages between samples over time. Furthermore, as already pointed out, the resources available for child labour surveys tend to be more limited.

Within the common sample areas, various possibilities exist in terms of the relationship between the ultimate units (e.g. households) in the two samples – from entirely independent samples drawn from household lists in the common areas to confining the CLS to a sample of children actually identified during the LFS interview. (This is described in Section 4.6.)

4.2 Selection of areas

We shall now consider the procedure for selecting a smaller number of clusters for a CLS sample from a larger number in, say, an LFS sample.

The principles of such sub-sampling are quite straightforward, so the objective of this and the following sections is to provide details of the actual procedures, along with some numerical illustrations where helpful.

We begin here with the simplest situations: for a given domain, a reduced number of areas are to be selected for the CLS from a given sample of LFS areas. Section 4.3 deals with a situation involving several domains. There are some added complications when we have to deal with areas for which special procedures had been used in the existing LFS sample from which the sub-sampling has to be done. There are, for example, areas that were treated, in the sense described in Sections 3.9-10, as being "too large" (with size measures larger than the sampling interval I for systematic PPS sampling), or "too small" (with size measures smaller than the target sample take b) during their previous selection. These special problems are considered in later sections. We are concerned here with sub-sampling at the level of areas; sampling of households within areas will also be considered later.

One extremely important practical point should be noted at the outset when a sample is obtained by sub-sampling from an existing sample, namely, that full details must not only be recorded of the sub-sampling procedures, but must also be available for the existing sample used for sub-sampling. Unfortunately one finds examples of surveys in

which details of the design of the existing sample were either not properly documented or had not been preserved. The most critical piece of information is the probabilities of selection applied in the original sample. If such details are not available for an existing sample, it is desirable to look to alternative sources for sub-sampling.

4.2.1 Areas selected with constant probability in a domain

Suppose that, in a domain of the existing sample, a_1 areas have been selected from A_1 areas in the population with constant probability $f_1=a_1/A_1$.

The objective is to obtain a sub-sample with fewer areas $a<a_1$ while retaining the same structure, i.e. uniform selection probabilities for the areas $f=a/A_1$.

Obviously, the sub-sampling rate required is

$$g_1=a/a_1.$$

We can obtain the required sub-sample of areas, for instance, by systematic sampling from the existing sample with constant probability by using the sampling interval

$$k_1=1/g_1=a_1/a.$$

It is encouraging to see g_1 as a ratio of sampling rates or probabilities in both samples, as that helps to generalize exactly the same procedure for other more complex situations:

$$g_1=(a/A)/(a_1/A_1)=f/f_1,$$

with $A_1=A$ in this case, of course.

The sampling rate to be applied is the *ratio of the new to the existing sampling probabilities or rates*.

Similarly, the required sub-sampling interval is $k_1=f_1/f$.

4.2.2 Areas selected with probabilities proportional to size in a domain

Let us assume that the LFS is based on the commonly used PPS design described in Chapter 3, and that this sample contain a_1 areas selected with probability proportional to a measure of population size (p_i). The objective is to select for the CLS a reduced number of areas, $a=a_1/k_1$, with $k_1>1$, also with probability proportional to the same measure of size p_i. In other words, it is assumed here that for each area i, the measure of size p_i used in the original LFS selection is the same as the size measure to be used for the CLS. This is normally the case when the base population for the two surveys are the same or nearly proportionate and similarly distributed – for example, all adults for the LFS and children in the same population for the CLS. (The situation *where* different types of size measures are used in the two samples will be discussed in Chapter 5.)

Given the common size measures, the basic procedure for sampling of areas from the LFS to the CLS is straightforward, as follows:

- Simply select a sub-sample of LFS areas with a *constant* probability $g_1 = a/a_1$.
- Selection with a constant probability (g_1) can be achieved simply by applying to the LFS the equal probability systematic sampling procedure with interval $k_1 = 1/g_1 = a_1/a$.

The result is a PPS sample of *a* areas with probability proportional to the population size measure p_i. The selection equations for the LFS and the CLS areas are as follows.

LFS: $f_1 = \left(\frac{a_1}{\Sigma p_i}\right).p_i = \frac{p_i}{I_1}$, for example, and

CLS: $f = \left(\frac{a}{\Sigma p_i}\right).p_i = \left(\frac{g_1 * a_1}{\Sigma p_i}\right).p_i = g_1 \cdot \left(\frac{p_i}{I_1}\right)$

where Σp_i is the sum of size measures for all areas in the population from which the LFS sample was selected. This parameter remains unchanged for the CLS with the same size measures p_i.

Again, g_1 may be seen as the ratio of sampling probabilities (f/f_1), which is a constant for all units even in an existing PPS sample. This is because both f_1 and f for any area are defined as proportional to the size measure p_i, which cancels out in the above ratio.

Thus, as required, the area samples for the two surveys have the same structure, differing only in the number of areas selected.[16]

4.3 Sample allocation and reporting domains

4.3.1 Reducing the number of reporting domains

Apart from differences in the required sample size and clustering, the CLS may differ from the "parent" LFS also in the requirements with respect to sample allocation and stratification.

For instance, the LFS may be allocated disproportionately for the purpose of producing sub-national estimates with over-sampling of small regions or other reporting domains, while this may not be required for a CLS when it is based on a smaller sample aimed primarily at producing national-level estimates, or producing breakdowns for only a few major domains.

In any case, we can generally expect the CLS to have a smaller number of reporting domains than a bigger survey like the LFS; furthermore, the sampling rates in the CLS

[16] It should be pointed out that while sub-sampling from an existing PPS sample is very simple (the PPS nature is retained unchanged with sub-sampling at a constant rate), the process of *adding* additional units to an existing PPS sample is much more complex, and the resulting probabilities of the units difficult if not impossible to calculate. We assume throughout that the required number of areas for the CLS do not exceed the number already available in the LFS sample in any sampling stratum.

sample are often more uniform. Normally, the reporting domains for the CLS would be groupings of the LFS reporting domains – e.g. major regions rather than individual provinces or districts[17].

It is to illustrate this situation that the sub-sampling procedures will be described below.

Of course it is possible that the CLS may need a disproportionately large sample for sub-populations of special interest, and that this over-sampling would cut across the LFS domains. It may also be argued that, from the point of view of optimal allocation, domains with a higher level of child labour should be sampled at a higher rate. However, such disproportionate allocation for the purpose of optimization is not justified in most situations, given the very modest gains in efficiency it is likely to provide, as noted in Section 4.1.1.

Significant differences in the required stratification are unlikely. Certain common stratification criteria such as geographic location and type of place (urban-rural, degree of urbanization, etc.) are commonly used in almost every household survey. In many cases they are practically the only criteria available for stratification. Where available, additional stratification criteria which are likely to be useful (such as ethnicity, predominating occupation, literacy rate, mean level of income, etc, for the sample area) tend also to be similar for different social surveys. The CLS and LFS are likely to be even more similar in terms of stratification because of their shared or similar subject matter. Any differences in stratification requirements are likely to arise only in relation to differing requirements in terms of distribution (allocation) of the sample between the two surveys.

In any case, different requirements in terms of distribution of the sample do not *in themselves* require the two surveys to be stratified differently. It is possible to keep the sample allocation requirements (which determine the required probabilities of selection) separate from the stratification and sample selection aspects (which determine how the actual selection procedures are implemented). The selection method described in Section 4.4 can be useful for this purpose.

4.3.2 Sub-sampling procedure when grouping of LFS domains is involved

Units selected with constant probabilities

Again it is convenient for illustration purposes to begin from this simple case.

Let us suppose that a number of domains (j) of the LFS are grouped to form a single reporting domain for the CLS. The LFS sample has been allocated disproportionately among these domains. Let a_j be the number of areas selected from A_j areas in domain j with uniform probability $f_j = a_j/A_j$. The objective is to obtain a CLS sample of *a* areas, selected at a uniform rate $f = a/A$, with $A = \Sigma A_j$, the sum of A_j values being over all the LFS domains in the group put together.

[17] It is unlikely that the more detailed LFS domains would cut across the boundaries of the more aggregated CLS domains.

As before, the required sub-sampling rate is given by the ratio of the new to the existing sampling rates in the domain concerned:

$$g_j = f/f_j = (a/A)/(a_j/A_j) = (a'_j/a_j)$$

where $a'_j = a \cdot (A_j/A)$ is the expected number of CLS sample areas in LFS domain j.

We assume that in every domain the CLS sample is the same or a sub-sample of the LFS sample:

$$a_j' \leq a_j, \text{ i.e. } g_j \leq 1.$$

This is a realistic assumption.[18]

The required sub-sampling interval in going from the LFS to CLS sample areas is

$$k_j = 1/g_j = (a_j/A_j)/(a/A) = (a_j/\ a'_j).$$

The required sub-sampling may be done separately for each domain – fixing the required number a'_j to be selected and applying the selection interval k_j, which generally differs from one domain to another.

A more convenient procedure is to put one domain after another in a single list, and select a single systematic sample of *a* units. The problem caused by different selection intervals k_j to be applied to different parts of the list can be avoided by using the trick of "rescaling the size measures" of units as described in Section 4.4. This rescaling is done in such a way that the application of a common sampling interval to all the domains yields the required CLS sample.

Sub-sampling a PPS sample

Now let us suppose that the LFS areas have been selected with probabilities proportional to some measure of size, p_i.

For a given domain j, let P_j be the sum of measures of size of all units in the *population* of the domain (and not just over LFS sample areas), and let a_j be the number of areas selected in the LFS.

We assume that, for the CLS, a PPS sample of size a is required of areas selected with the *same* measure of size p_i. The selection equations for area i are

LFS domain j: $f^{(i)}_j = (a_j/P_j) \cdot p_i$

CLS, all domains: $f^{(i)} = (a/P) \cdot p_i$

where $P = \Sigma P_j$ the sum of P_j values is over all domains, and superscript i indicates that the area selection probabilities in the PPS sample vary according to unit size measures.

[18] If the above g_j turns out to exceed 1.0 for a particular domain, one option is simply to retain all the existing LFS areas for the CLS in that domain, and thereafter simply remove the domain from the computations being described here.

The required sub-sampling rate from LFS to CLS for domain j is

$g_j = f^{(i)}/f^{(i)}_j = (a/P)/(a_j/P_j)\ (a_j'/a_j)$ where $a_j' = a \cdot (P_j/P)$,

with the required interval for systematic selection

$k_j = (a_j/P_j)/(a/P) = (a_j/\ a_j')$.

This rate does not depend on the unit or its size, since p_i cancels out in the above. Hence the sub-sampling procedure is exactly the same as for the constant probability case discussed earlier: within each domain, units are selected from the existing sample at a *constant* rate defined above, irrespective of the size measures of the units, provided that the unit size measures for the two samples are the same.

The device of re-scaled size measure described below can be used to perform the selection as a single operation for all the LFS domains in the group put together.

4.4 Rescaling the size measures to facilitate sample selection

The technique is based on *rescaling measures of size* used for sampling to accommodate required variations in sampling rates and sample sizes across different parts of the list. In fact, the usefulness of this technique is much wider than the specific issue of LFS-to-CLS sub-sampling being discussed here.

Suppose that a systematic sample has to be selected from a list of units. The list is divided into a number of parts or domains, and the required selection intervals differ from one part of the list to another.

The design covers a variety of situations. For instance, the objective may be to select a constant probability or PPS sample of the units. Or it may entail sub-sampling units from an existing sample, itself with constant or variable probabilities. Such details do not affect the procedure described below inasmuch as they are incorporated in the definition of the sampling interval I_j.

One option is to select a systematic sample for each part j separately in the usual way, using the selection interval I_j applicable to that part.

A simpler alternative procedure involves the following[19].

Let $a = \Sigma a_j$ be the total number of units to be selected from the combined list, and $P' = \Sigma P'_j$ the sum over all domains combined of the modified size measures as defined above.

[19] The two procedures are not identical. The simpler alternative described here allows some random variation in the exact number of selections which fall in any original domain. But it avoids the need to allocate pre-fixed rounded numbers of selections to each of the original domains separately – something which can be unnecessary and also inconvenient where a large number of small domains are involved, especially when the expected number of selections per domain is small. The procedure avoids the need to round the number of units to be selected.

Applying a uniform sampling interval I=P'/a to the list with modified unit size measures is equivalent to applying different selection intervals I_j to units with original size measures. With the size measures scaled as above, the selection interval to be applied becomes uniform throughout, across the domains. The domains can therefore be placed one after another, in a single list, for sample selection with a uniform interval.

If, in a set of domains, P_j is the sum of measures of size of units (PSUs) in the *population* of domain j, and if, from the domain, a_j units are to be selected with PPS, the selection interval is $I_j=P_j/a_j$.

And if all size measures of units in domain j are multiplied by (I/I_j) where I is an arbitrary constant, the new sum of size measures for the domains becomes $P'_j=P_j\cdot(I/I_j)$.

In order to select the same number of unit a_j as before from this domain, but using the rescaled size measures, the selection interval required is

$$[P_j\cdot(I/I_j)/a_j]=I\cdot[(P_j/I_j)/(P_j/I_j)]=I,$$

equal to the arbitrarily chosen constant I for every domain.

A constant probability sample is just a special case of the above, with the original unit size measures p_i=constant=1, for instance. In this case the sampling to be done is treated as PPS, with each unit in j having the same measure of size (I/I_j).

This technique can be very convenient when a large number of parts are involved, each requiring a different sampling interval, but when those part are not required to form explicit strata by the sampling design.

This convenient technique can also be applied separately at different stages in a multi-stage sample, simply by appropriately re-scaling the size measures independently at each stage of selection. For instance, consider a two-stage self-weighting sample in which units are selected in the usual way, using PPS at stage 1 and inverse-PPS at stage 2:

$$f1_i=\left(\frac{a}{\Sigma p_i}\right).p_i=\frac{p_i}{I};\quad f2_i=\frac{b}{p_i};\quad f=f1_i*f2_i=\frac{b}{I}.$$

Now suppose that in one part of the population different sampling rates are required, by some factor k_1 at the area stage and by k_2 at the final stage. For the part to be sampled differently, the selection equations become:

$$f1_i=\left(\frac{k_1*a}{\Sigma p_i}\right).p_i;\quad f2_i=\frac{(k_2*b)}{p_i};\quad f=(k_1.k_2)*f.$$

Of course, this can be achieved by putting the part to be sampled at a different rate into a separate stratum and carrying out the selection process separately in the resulting two parts. This entails using different sampling intervals for the systematic selection of units in the domain concerned.

However, the same result is obtained by keeping the sampling intervals for systematic selection unchanged, but inflating the size measure of the areas concerned by the factor k_1 for the selection at the first stage (giving the new size measure as $P'_i = k_i \cdot p_i$), and by $1/k_2$ for the second stage selection within sample areas in the particular domain concerned (giving the new size measure $P''_i = P_i/k_2$). The same selection equations are seen as:

$$f1_i = \frac{(k_1 * p_i)}{I} = \left(\frac{P'_i}{I}\right); \; f2_i = \frac{b}{(p_i/k_2)} = \left(\frac{b}{P''_i}\right); \; f = f1_i * f2_i = (k_1.k_2) * f.$$

After this adjustment, there is no need to select the sample separately from the domain concerned. With systematic sampling for instance, the same common selection interval I can be applied to the whole list for the selection of areas. The measures of size multiplied by the factor k_1 automatically allocate the sample of areas as required to the domain concerned. Similarly, the sample-takes per area are adjusted automatically when original "target sample-take" selection interval *b* is used, with the size measures of the areas in the domain concerned divided by the factor k_2 to adjust the number of ultimate units to be selected from each sample area. Note that for the same area(s), the size measures have been scaled differently at the two stages (by k_1 and $1/k_2$ respectively).

An illustration

The procedure is based simply on the observation that in PPS sampling, the number of units selected

$$a = \Sigma p_i / I$$

remains unchanged if both the size measures and the interval of selection are multiplied the same arbitrary constant. By choosing the appropriate constant for each domain to be sampled at a different rate, we can make the required selection interval the same in all domains.

Let us suppose that there are two strata. The first consists of $A_1 = 100$ areas and $a_1 = 10$ are to be selected with constant probability, i.e. $f_1 = a_1/A_1 = 1/10$ and selection interval $I_1 = 10$. In the second domain the respective figures are $A_2 = 40$, $a_2 = 8$, hence $f_2 = 1/5$, $I_2 = 5$.

We wish to select the sample systematically, but for convenience using a single systematic sampling operation. Let us replace the constant probability selections by equivalent systematic PPS sampling as follows: each unit in stratum 1 is given a size measure of 1 (i.e. there is no change); in stratum 2 we assign a size measure of 2 to each area, and correspondingly also double the selection interval to be applied to 5*2=10, which now is identical to that for stratum 1.

	Stratum 1	Stratum 2	Total
No. of units in population	100	40	140
Assigned size measures to each unit	1	2	
Total size measure	100	80	180
Expected number selected (with I=10)	10	8	18

The "price" of putting the two strata together into a single list is that the actual number selected from a particular stratum may vary slightly at random (though the selection probabilities remain strictly unchanged). But the procedure can be convenient when many separate domains have to be handled.

4.5 Dealing with "very large" areas in sub-sampling

Additional steps are involved in the sub-sampling procedure when dealing with areas in the original sample that were selected using special procedures because they were considered to be too "small" or too "large" for the normal PPS procedure. This has been described in Sections 3.9-3.10.

For the sub-sampling being discussed in this chapter, dealing with areas classified as "small" in the above sense is somewhat more complicated since, in a multi-stage design, it introduces issues that have to do with the selection of households within sample areas. This will be taken up in Sections 4.7.

We first consider very large units. As defined in Section 3.9, "very large" in the context of PPS sampling means a unit whose size measure exceeds the sampling interval, i.e. $p_i > I$. Let us assume that in the LFS such areas have been dealt with as in option 3 of Section 3.9, which is the recommended option. This involves taking the large unit as automatically selected (such units are usually referred to as "self-representing units"). If ultimate units in it are then selected with the required overall selection probability, say f, this gives a self-weighting sample with

$f1_i = 1; \quad f2_i = f$ for the LFS.

For the purpose of selecting a sub-sample (1/k) of LFS area units for the CLS, two groups of these "self-representing" LFS areas need to be distinguished. Below, p_i refers to the unit size measure and I to the PPS selection interval.

Group 1: $p_i > k * I$

These are the largest units. All of these units must be retained in the CLS, with probability $=1$, as in the LFS. At the final stage, ultimate units can be selected with the overall selection rate required for the CLS.

Group 2: $I < p_i \leq k * I$

All these large units do not get selected into the CLS automatically, although that was the case in the LFS. For these units, a proper PPS sample of areas can be selected for the CLS with

$$f1'_i = \frac{p_i}{(k * I)} \leq 1.$$

For the final self-weighting sample, for instance, the final sampling within selected areas is at a rate inversely proportional to p_i, with appropriate constants to obtain the overall selection rate required for the CLS.

An illustration

Table 4.1 shows the list of "large" areas as defined in Table 3.3. Area numbers 93-100 are "large" for the LFS because their size measures exceed 500, the sampling interval for PPS selection of areas assumed for the LFS.

These areas were treated as if the size measure of each was 500, so that the area was selected with probability 1.0 for the LFS. (If at the next stage the selection of households is done with inverse-PPS, the area will give more households to the sample in proportion to its actual size.)

Let us assume that the selection interval for the CLS is 700. This divides the "large" areas in the LFS into two groups. Area numbers 93-96 are no longer "large" as regards the CLS – their size measures are smaller than the CLS sampling interval of 700. From the LFS, these areas need to be selected with a probability proportional to size, i.e. with a probability equal to the unit size measure divided by 700 (col. [7] of Table 4.1).

The second group, area numbers 97-100 remain "large" for CLS sampling interval. They are retained in the CLS sample with probability 1.0. (As in the case of the LFS, if the selection of households at the next stage is done with inverse-PPS, a "large" area will give more households to the sample in proportion to its actual size).

Table 4.1. Illustration: subsampling of "large"areas from LFS to CLS

"First sample": for instance for a LFS.

Parameters:

a = 100 [number of areas in sample]
b = 20 [expected number of units selected/sample area]
I = 500 [interval for systematic selection of areas]
n = 2,000 [=a*b, expected sample size]

[1]	[2]	[3]	[4]	[5]	[6]	[7]
Cumulative frequency of number of areas in		Area measure of size (MoS)	MoS modified for LFS	Area selection probability LFS	MoS modified for CLS	Area selection probability CLS
Population	Sample	pi	Modified pi	f1 (%)	Modified pi	f1' (%)
Very small areas						
3.0	1	8	20	4.00	20	1.09
6.0	2	11	20	4.00	20	1.61
9.0	3	12	20	4.00	20	1.72
12.1	4	16	20	4.00	20	2.25
15.1	5	17	20	4.00	20	2.42
18.1	6	18	20	4.00	20	2.50
21.1	7	18	20	4.00	20	2.63
24.1	8	20	20	4.00	20	2.85
Normal areas						
…..						
99.0	92	425	425	85.06	425	60.76
…..						
Very large areas						
99.2	93	527	500	100.00	527	75.24
99.3	94	566	500	100.00	566	80.83
99.4	95	659	500	100.00	659	94.08
99.5	96	681	500	100.00	681	97.34
99.6	97	719	500	100.00	700	100.00
99.8	98	898	500	100.00	700	100.00
99.9	99	908	500	100.00	700	100.00
100.0	100	1076	500	100.00	700	100.00

4.6 LFS to CLS: Sub-sampling of ultimate units within sample areas

It is by no means automatically necessary or useful to confine the CLS to the same or a sub-sample of ultimate units (household, persons, children) included in the LFS, even when the two surveys share a common set of sample areas. There is a range of choices depending on the circumstance and requirements.

1. At the one extreme, the common LFS-CLS sample areas may be re-listed to obtain a more up-to-date frame for the CLS, and an entirely new sample of the ultimate units selected. However, listing tends to be an expensive operation, and the cost of re-listing can be justified only if the two surveys are separated by a long time interval, such as one year or more in many circumstances. On the other hand, using old lists can introduce coverage errors in so far as newly created units (such as new dwellings or households) are not represented. Old lists also make it more difficult to identify the selected units in the field. The presence in the list of units that no longer exist can make the identification of true non-respondents more difficult; it also tends to reduce the efficiency and control over the sampling operation. Durability of lists depends on the type of units listed. Addresses or dwelling units tend to be more durable than households, and households more durable than persons.

2. When the same lists are used, the two samples (LFS and CLS) may still be selected independently or at least the overlap between them minimized. This is desirable when respondent fatigue is a concern, or when the first sample is subject to high rates of non-response. The latter problem may be serious if the CLS sample is to be based on a "heavy" survey rather than a typical LFS. Independent or at least additional sampling is also required when the first survey is not able to yield a sufficiently large sample for the CLS.

3. The next option is to base the CLS on all or a sub-sample of ultimate units included in the first sample. The outcome depends on the type and characteristics of the units involved in the sub-sampling. Selecting a sub-sample of addresses is often the simplest. The use of households as units for sub-sampling comes next. It is simpler if households from the first survey are subject to sub-sampling without reference to any particular characteristics of the households involved.

4. However, sometimes information on various household characteristics is evoked for the purpose of stratification or for applying different sampling rates. This involves the collection of such information in the LFS and its preservation and transfer to the CLS. This can be expensive and cumbersome, and in many cases not very effective in improving the efficiency of the resulting sample.

5. It is also common to exclude certain types of household from the selection, such as households not found to contain any child relevant to the CLS. This can improve control over the CLS sample size, and also the efficiency of its fieldwork. It assumes that the situation of the households with respect to the exclusion criteria has not changed during the interval between the two surveys. Another problem is that not

only the address list but also some additional information on the households has to be transferred between the two surveys.

6. Using children identified during the first survey as the units for sampling for the CLS is also an option. This is a demanding choice however, in that lists of children identified have to be prepared, transferred to the CLS operations for sampling, and then the selected children have to be identified during fieldwork. Mis-identification of individual children can easily occur. As a rule, such an option should be followed only if the interval between the two surveys is very short, no longer than a few weeks.

7. At the extreme is the procedure where information on various characteristics of the children is used for the purpose of stratification or for applying different sampling rates, such as their educational and/or activity status. Such a procedure may appear attractive when the CLS sample size is very small and its structure needs to be tightly controlled. However, generally *this is a demanding and expensive procedure, prone to implementation errors*. It should be used only when the CLS approximates the conditions of being a "module" of the first survey – nearly simultaneous or close in timing, and drawing on, to a significant extent, the substantive information collected in the first survey.

4.7 Dealing with "very small" areas in sub-sampling

The sub-sampling procedure becomes more complex when dealing with units of very small size. Care is required to ensure that correct selection probabilities are achieved in the CLS.

In a two-stage design, the procedure for handling very small areas in LFS-to-CLS sub-sampling depends on the details of the LFS sampling at both stages. For describing the sub-sampling procedure, we assume the following self-weighting design. Within each LFS sample area selected with probability proportional to P_i, households are assumed to have been selected with probability inversely proportional to p_i, with an average of b=n/a household per sample area.[20]

For the CLS, sub-sampling from the LFS involves two steps: the selection of 1 in k sample areas, a"=a/k, and then the inclusion in the CLS of an average or target selection of an expected number b' of ultimate units per area, thus giving the CLS target sample size as n'=a'*b'.[21]

The full selection equations (for self-weighting design) for the two surveys are as shown in Table 4.2.

[20] Other options are also possible, such as a fixed take of b household per area, or some compromise between the self-weighting and fixed-take version of the design, as discussed in Chapter 3. We have chosen the self-weighting design here for simplicity of exposition.

[21] Note that the two samples may not necessarily overlap in the case of the ultimate units, though often they do overlap, and the condition b'≤b holds. Also, both for the LFS and the CLS, these parameters may differ from one "sampling domain" or stratum to another. It is sufficient here to describe the procedure for one such domain.

Table 4.2. Selection equations for LFS and CLS, as assumed for the description of the LFS-to-CLS sub-sampling procedure

	(Last) Area stage	Ultimate stage	Overall selection probability
LFS	$f1_i = \frac{p_i}{I}$	$f2_i = \frac{b}{p_i}$	$f = f1_i * f2_i = \frac{b}{I}$
CLS[22]	$f1'_i = \frac{p_i}{I'}, I' = k \cdot I$	$f2'_i = \frac{b'}{p_i}$	$f' = f1'_i * f2'_i = \frac{b'}{I'} = \left(\frac{1}{k} \cdot \frac{b'}{b}\right) * f$

As defined in Section 3.10, "very small" in the context of PPS sampling means a unit whose size measure is smaller than the required sample take, i.e. $p_i < b$ in the LFS. As to what needs to be done for the sub-sampling from the LFS to the CLS, this depends on how these areas were treated while selecting the LFS sample itself. Several methods of dealing with this problem were discussed in Section 3.10. In addition, we have also to consider the relationship of the size measure p_i to the required sample-take b' in the CLS.

It is important to point out that, while we are considering a CLS sample which is smaller in overall size compared to the LFS (n'<n) and has a smaller number of sample areas (a'<a), it does not follow that the sample-take per area (b') in the CLS must necessarily be smaller than the take (b) in the LFS. For instance, within the common sample areas, the CLS may use the same samples of the ultimate units as the LFS (b'=b), as would be the case when the former forms a module of the latter; or the CLS may use a sub-sample of the units in the common clusters (b'<b); or the two surveys may use entirely different samples of ultimate units even in the common sample areas. In the last case, any of the three possibilities exists: b'<b, b'=b, or b'>b. This tends to make the proper treatment of small areas in the sub-sampling from LFS to the CLS a little complicated. Nevertheless, the issue is of practical importance and needs to be handled correctly.

Option 3 of Section 3.10

Let us suppose that in the LFS very small areas were selected according to the above-mentioned option in the LFS. This option implied that all very small areas, with p_i<b, were given a size measure =b, and hence were selected with probability $f1_i$=(b/I) at the area stage; all ultimate units in a selected area were taken into the LFS sample ($f2_i$=1), giving the required overall probability f=(b/I) for the LFS.

Table 4.3 shows all the details of the selection of areas from LFS to CLS for very small units. The procedures for "normal", and for very large areas as discussed in Section 3.9 and 4.5, are also shown for comparison. The column "Area sub-sampling rate" gives the proportion of the LFS sample areas, depending on their measure of size p_i, which are to be retained in the CLS to obtain the required overall selection rate f'=f1'*f2'=b'/(k*I).

[22] Note that in the above equations, b' is to be defined in the same scale as p_i or b, such as the total population or number of households. If b' is to some other scale, such as the number of children, then p_i and b in the equations for the CLS are assumed to have been rescaled to the same unit by a constant factor.

A simple way to make this selection is to assign to the LFS sample areas modified measures of size (different from the size measures p_i used originally for the selection of LFS areas), as defined in the last column of Table 4.3. We can then select a systematic sub-sample of LFS sample areas with interval k to obtain the CLS sample of areas. This procedure has been explained in Section 4.4.

Option 5 of Section 3.10

It is also useful to give details for this option, since it is the recommended one. In this option, the very small areas in LFS are divided into two parts: (i) a certain number of the smallest areas, say with $p_i<b_0$, are given a zero measure of size, and hence are effectively excluded from the sampling frame, as they are considered too small to be included in any LFS sample; (ii) the remaining areas, the largest among the "very small" areas, are each given a measure of size b and are treated as in option (3).

It is assumed that, to compensate for (i), the weights of any of these units selected into the LFS have been increased by the factor (b/p_i). As noted in Section 3.10, for this compensation, part (ii) should include $A = \Sigma p_i / b$ of the largest units in the "very small" set, where Σp_i is the total size measure of all units in the set.

With the LFS areas selected with option 5 as described above, the only difference from option 3 concerns cases 5 and 6 in Table 4.3.

Areas with $p_i<b_0$ of course do not appear in the CLS sample, as they were excluded even from the LFS.

For areas with $p_i \geq b_0$, the above-mentioned inflation of the weight in the LFS sample $(=b/p_i)$ is retained for any area selected for the CLS. A simpler alternative is to change the last two columns for cases 5 and 6 in Table 4.3 by this factor (b/p_i), and hence avoid the need for explicitly including such inflation of the weights in the CLS. This simply makes the treatment of case 6 the same as that of case 4, and the treatment of case 5 exactly the same as case 3. This has been done in the last two column of Table 4.3.

Table 4.3. Sub-sampling of LFS areas for the CLS

	Condition	LFS		CLS		Area sub-sampling rate f'/f1	MoS[23]
		f1	f2	f1'	f2'		
Very large areas							
1	$k * I \leq p_i$	1	$\frac{b}{I}$	1	$\frac{b'}{k * I}$	1	k
2	$I \leq p_i < k * I$	1	$\frac{b}{I}$	$\frac{p_i}{k * I}$	$\frac{b'}{p_i}$	$\frac{1}{k} * \frac{p_i}{I}$	$\frac{p_i}{I}$
Normal areas (majority of the areas)							
3	$b, b' \leq p_i < I$	$\frac{p_i}{I}$	$\frac{b}{p_i}$	$\frac{p_i}{k * I}$	$\frac{b'}{p_i}$	$\frac{1}{k}$	1
Very small areas (selected in LFS using option 3 of Section 3.10)							
4[24]	$b \leq p_i < b'$	$\frac{p_i}{I}$	$\frac{b}{p_i}$	$\frac{b'}{k * I}$	1	$\frac{1}{k} * \frac{b'}{p_i}$	$\frac{b'}{p_i}$
5	$b' \leq p_i < b$	$\frac{b}{I}$	1	$\frac{p_i}{k * I}$	$\frac{b'}{p_i}$	$\frac{1}{k} * \frac{p_i}{b}$	$\frac{p_i}{b}$
6	$p_i \leq b, b'$	$\frac{b}{I}$	1	$\frac{b'}{k * I}$	1	$\frac{1}{k} * \frac{b'}{b}$	$\frac{b'}{b}$
Very small areas (difference from above if area selected in LFS using option 5 of Section 3.10)							
5	$b_0, b' \leq p_i < b$	As in case 5 above				$\frac{1}{k}$	1
6	$b_0 \leq p_i \leq b, b'$	As in case 6 above				$\frac{1}{k} * \frac{b'}{p_i}$	$\frac{b'}{p_i}$
7	$p_i < b_0$	Areas not included in LFS or CLS sample				-	0

[23] Modified measure of size assigned to LFS areas for sub-sampling with interval k to obtain the CLS sample areas in a simple way. Thus, for example, every LFS area in row 1 is assigned a size measure k, so that selecting with interval k simply retains all these areas in the CLS sample – as required. Similarly, in row 2, each LFS area is given a size measure p_i/I, so that selecting with the same interval k gives the required sub-sampling rate (f1'/f1) specified in the preceding column.

[24] Note that only one of the two options 4 and 5, can apply in any particular situation.

Chapter 5
Labouring children survey

5.1 Approach to the survey

As defined in Chapter 2 (Section 2.1.2), we use the term labouring children survey (LCS) to refer specifically to a survey or a survey component designed with the objective of determining the conditions and consequences of child labour, as distinct from its prevalence among all children (which is the objective of the survey or survey component termed CLS).

The primary objective of the LCS is to investigate circumstances, characteristics and consequences of child labour: what type of children are engaged in work-related activities, what type 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. Therefore, the relevant base population in the LCS is the *population of working children*.

A critical issue of practical importance is whether the distinct CLS and LCS objectives can be satisfactorily met through a single integrated survey, or it is better to organize them as two separate – but nevertheless linked – operations. A related, but distinct, question is whether the two components can be based on the same sample of households, or whether the LCS should be a sub-sample of the CLS, of smaller size and possibly also with a different structure.

Undoubtedly, it can be more convenient and economical to cover both components in a single operation, and by far the predominant form in countries has been an integrated CLS-LCS operation. Such an arrangement may also be seen as more economical. But unfortunately this is not necessarily the case. From our study of past national surveys of child labour, it is our considered view that issues regarding the differing substantive and statistical requirements, practical aspects such as respondent burden and even survey costs, and especially data quality, have not always been thoroughly considered in deciding on the appropriate structure of the survey.

An integrated CLS-LCS operation may well be the most suitable option under certain conditions such as the following.

1. The CLS does not require a very large sample, which would be the case when it is not required to produce estimates of the prevalence of child labour for many different regions, population groups, sectors of activity, or other types of domains.

2. The CLS is a stand-alone survey, so that its sample can be designed as a compromise for meeting both types of information needs – of estimating the prevalence of child labour with necessary precision on the one hand, and of investigating the conditions and consequences of child labour with necessary detail on the other.

3. Child labour is not too heterogeneous or extremely unevenly distributed for it to be "captured" in a reasonable way by a general purpose sample of the population of children.

4. The resulting compromise sample size is not too large for the in-depth investigation which the LCS component usually requires.

5. But in any case, the resulting integrated interview is not too heavy to have an adverse effect on the quality (particularly completeness) of measuring the prevalence of child labour which is the concern of the CLS component. This is a common problem which has been encountered in many other types of survey with similarly dual objectives.

When one or more of the above conditions are violated, it is necessary at least to consider the possibility of operationally separating the CLS and LCS components.

In the case of such a separation, it would be generally appropriate to consider basing the LCS component on a sub-sample of the CLS. The objectives of the sub-sampling would be both to reduce the sample size for the LCS, and also to make it more concentrated and targeted to reflect the uneven geographical distribution of child labour.

This chapter sets out specifically to describe technical details of such CLS-to-LCS sub-sampling. Our primary concern is with issues of sample design for a survey where the base population of interest is the population of labouring children. It is taken as given that, as discussed in other chapters of this manual, the population of interest is clearly defined on the basis of substantive and policy considerations, including those pertaining to the need for internationally comparable data.

It is important to clarify that the concept of a "labouring children survey" is not meant to imply in any way that the ultimate units enumerated in the survey should be only labouring children. On the contrary, it will normally be necessary in such a survey to enumerate also children not engaged in labour, so as to provide a control group for comparison with the characteristics and circumstances of those subject to child labour. What is meant by the LCS concept is that, when the objective is to determine the circumstances and consequences of child labour rather than merely its prevalence, then the structure and size of the sample should be determined mainly by the size and distribution of the *population of labouring children*, rather than by the size and distribution of the general population of all children. Furthermore, for these and other substantive and practical reasons, it is often desirable to *link* the LCS appropriately with the normal CLS described in the preceding chapter, where opportunities exist for such linkage.

Regular household-based versus targeted surveys

The limitation of the scope of the sampling issues discussed in this chapter should be noted. We shall be dealing here with what may be called "regular" household-based labouring children surveys.

The estimation of the incidence and nature of child labour in targeted sectors and activities, and also of some aspects of the worst forms of child labour, involve special design features, some of them quite different from the more widely-based LFS/CLS/LCS type of operations. For some purposes and in certain circumstances, they may involve

non-household based data collection, and even a departure from the principles of probability sampling. As noted by the SIMPOC External Advisory Committee, various statistical techniques used for sampling non-standard units need to be developed and documented. This is a major issue needing a separate treatment, and it is not considered in this manual. Nevertheless, many of the techniques discussed here can also be useful in the design of more specialized, targeted or sectoral surveys of labouring children.

Sampling frame for the LCS

The procedures described below for the selection of LCS sample areas require information on the *number of labouring children* in each area in the "frame" from which the survey areas are to be selected. Such information is not normally available in general-purpose population-based frames, and it is this that makes it necessary to select the LCS as a sub-sample of areas for which such information has been collected, such as sample areas from a larger CLS.

The sample design considerations here can be expected to be similar in many respects to those discussed in the previous chapter for the c*hild labour survey* (CLS). The main difference is that in the CLS the primary focus was on the measurement of the prevalence of child labour among an appropriately defined *population of all children*, and consequently that that population formed the basis for the design and selection of the CLS sample.

For the development and exposition of the LCS sampling procedure, we assume that the basic sampling scheme is as follows.

We begin with the CLS. A most commonly used sampling scheme for the CLS is the "self-weighting PPS" design described earlier: a design involving the selection of area units with probability proportional to a measure of population size of the area (p_i) and then, within each selected area, the selection of individuals with probability inversely proportional to the size measure. For this we shall use the basic selection equations of this design given earlier:

$$f1_i = \left(\frac{a}{\Sigma p_i}\right).p_i = \frac{p_i}{I}; \quad f2_i = \frac{b}{p_i}; \quad f = f1_i * f2_i = \frac{b}{I}. \qquad [1]$$

Here *a* is the number of areas selected and, if p_i is strictly the current number of individuals of interest in each area, *b* is the constant number of ultimate units finally selected from each sample area and n=a*b the resulting sample size. $I = a/\Sigma p_i$ is the interval which would be used in systematic-PPS selection of areas, the sum of p_i values being over all units in the population. Each ultimate unit in the population has the same probability of selection. Some common variants of this basic design were described in Chapter 3[25]. The design may refer to a survey of the general population such as the LFS, or to a c*hild labour survey* with a similar structure to that described in Chapter 4.

[25] The procedures described in this chapter are easily adapted to other designs such as "fixed-take" per cluster, or constant probability rather than PPS sampling of areas. There are no differences in the principles involved.

Design features

Similarly, in a survey where the base population of interest is *labouring children*, an appropriate design will involve the selection of area units with probability proportional to the number of labouring children (c_i) in the area, and then the selection, within each selected area, of such children with probability inversely proportional to c_i:[26]

$$f1'_i = \left(\frac{a'}{\Sigma c_i}\right).c_i = \frac{c_i}{I'}; \quad f2'_i = \frac{b'}{c_i}; \quad f' = f1'_i * f2'_i = \frac{b'}{I'}. \qquad [2]$$

Here *a*' is the number of areas selected. If c_i is strictly the current number of individuals of interest (labouring children) in each area, then *b*' is the constant number of ultimate units finally selected from each sample area and $n'=a'\cdot b'$ the resulting sample size. The actual sample size from a cluster will depart from *b*' to the extent that c_i, the measure of size for the cluster, departs from the actual size. The parameter $I' = \sum c_i / a'$ is the interval which would be used in systematic-PPS selection of areas, the sum of c_i values being over all units in the population. In this basic design, each labouring child in the population has the same probability of selection. As before, some variations on this basic design are possible. Some of the considerations involved in the choice of these design parameters have been discussed in Chapters 2 and 3, but for the present let us assume that they have been appropriately determined.

The LCS design refers to a survey of the total population of labouring children, but it may also be confined to children engaged in specific types of child labour activities.

This design differs from that of the CLS discussed earlier in a number of respects.

Firstly, as noted, the base population and measures of size are c_i, the number of children engaged in child labour. These values cannot be assumed to be known for all areas in the population. It is taken that these are obtained or estimated in the "first survey" (LFS or CLS for instance), but only for the areas enumerated in that survey. Hence the labouring children survey must be confined to a sub-sample of areas in the first survey. Within common sample areas, the samples of ultimate units may of course be different or overlapping.

Secondly, the base population of labouring children is likely to be geographicaly much more unevenly distributed than the general population of children. A few areas may contain high concentrations, and many areas only very low numbers of working children. In particular, there may be many "zeros", i.e. areas containing no labouring children of interest in the LCS. Problems such as the presence of extreme ("very large" or "very small") areas, defined in terms of the number of working children the area contains, are likely to be much more widespread than those discussed earlier for the CLS.

Thirdly, the sample size of the LCS is likely to be (or at least should be in a good quality survey) much smaller because of its intensive nature.

[26] This is a particular instance of the more general structure illustrated in Chapter 6 (see Section 6.4.2), where the area selection probability is taken to be a function of both c_i and p_i. We have taken the simpler model here to bring out the basic idea more clearly. Note that this model amounts to excluding areas not containing any labouring children according to the base survey ($c_i=0$).

5.2 Selection of areas

Given a sample of areas of the type described by equation [1] in the first survey, how can we sub-sample it so as to obtain a sample of areas of the type described by equation [2] for the LCS?

This can be achieved by selecting a sub-sample of areas from the first sample with PPS, the area measures of size for this sub-sampling being the ratio (c_i/p_i). This can be expressed as:

$$g_i = a'^* \frac{(c_i/p_i)}{\Sigma_s (c_i/p_i)}, \quad f1'_i = g_i \cdot f1_i \qquad [3]$$

where *a'* is the number of areas to be selected for LCS, and the sum is taken over *all areas in the first or base (CLS or LFS) sample* (as indicated by the subscript s). Here g_i the probability of an area (i) from the base sample being selected into the LCS; $f1'_i$ is the total probability of selection of area i into the LCS, and $f1_i$ the same for the area in the base sample defined in equation [1].

This, with equation [1], gives

$$f1'_i = \left[a * \frac{p_i}{\Sigma p_i} \right] * \left[a'^* \frac{(c_i/p_i)}{\Sigma_s (c_i/p_i)} \right] = \left(\frac{a'}{k_s} \right) . \frac{c_i}{\Sigma c_i}, \text{ where } k_s = \frac{\Sigma_s (c_i/p_i)/a}{\Sigma c_i / \Sigma p_i} \qquad [4]$$

Thus, the sub-sampling procedure [3] results in a sample of areas selected with probabilities proportional to size measure c_i, as required.

The quantities in the above equation are as follows.

$f1'_i$ the final probability of area i in the LCS sample. It is made up of two factors shown in square brackets. The first factor is the selection probability of the area into the base sample. The second factor is the selection probability of the area from the base into the LCS;

Σ the sum over areas in the population;

Σ_s the sum over areas in the base sample;

p_i the measure of size of the area, as used for the selection of the base sample. This may refer to the number of households, persons or children;

c_i the measure of size of the area, as used for the selection of the LCS sample. This may refer to the number of working children as estimated in the base survey, the number of households containing working children, or some other measure related to the extent of child labour in the area. Quantities c_i and p_i need not be measured in the same units or to the same scale (since the equations are dimensionally independent of them);

k_s a constant determined by the population and base sample characteristics, independent of the LCS sample or particular area i.

Note that if the LCS sample of a' cluster were selected directly from the population, with area selection probability proportional to the size measure c_i (a function of the number of labouring children in the area), the selection equation would have been:

$$f1_i'' = a' \cdot (c_i / \sum c_i).$$

Thus k_s is a factor surmising the effect of selecting this sample "indirectly", via the base survey. If c_i is proportionate to p_i in all areas, it can be seen that $k_s = 1$.

Equation [4] is in fact identical to equation [2], except for the presence of factor k_s as noted. This factor is not known since c_i values are not known for all areas in the population. It also depends on the particular sample which happens to be selected in the first survey – hence equation [4] does not provide the "true" selection probabilities in the sense of expected values over all possible base samples. Nevertheless, the value of this factor is expected to be close to 1.0, since its numerator and denominator both estimate the ratio c/p: the numerator is the average of separate ratios c_i/p_i while the denominator is the combined ratio of the same quantities. In any case, this factor does not affect the *relative* probabilities of selection of the area units in the final sample, since the factor is the same for all these units. The units are therefore selected with relative probabilities proportional to their size measures c_i.

To summarize, the sub-sampling procedure from the existing larger LFS/CLS to the smaller LCS is as follows.

Area units from the existing sample are sub-sampled with probability

$$\frac{(c_i/p_i)}{I_s}, \text{ where } I_s = \frac{\Sigma_s (c_i/p_i)}{a'}, \qquad [5]$$

so that the actual selection probability of an area in the second sample is

$$f1_i' = f1_i \cdot \frac{c_i/p_i}{I_s} = \left(\frac{p_i}{I}\right) \cdot \frac{c_i/p_i}{I_s} = \left(\frac{1}{I \cdot I_S}\right) \cdot c_i.$$

This follows from the fact that the selection probability of the area concerned in the first sample is

$$f1_i = p_i / I.$$

5.3 Dealing with "very large" and "very small" areas

Units with extreme characteristics are likely to occur in the LCS more often than in surveys like the LFS or CLS. Care is required to ensure that correct selection probabilities are achieved for such units.

Such problems concerning unit size can of course also occur in selecting the base sample. It is assumed throughout that these have been dealt with at that stage, as explained in the previous chapters: for instance by redefining p_i as $=I$ to deal with large

units, as =b to deal with very small units, as =0 for extremely small units (respectively, method 3 of Section 3.9, and method 3 of Section 3.10).

The above-noted aspects for the base sample generally do not make the treatment of such cases in the LCS more complicated. This is because, generally, different sets of areas are involved as "extreme" cases in the base and the LCS samples, since the two use different types of size measures, so that the two sets can be dealt with separately.

5.3.1 "Very large" areas

In the context of sub-sampling from the base sample to obtain a sample of areas for the LCS, "very large" refers not to the population size but to *the degree of concentration of child labour*, i.e. to very high values of the ratio c_i/p_i. This is because the ratio c_i/p_i is used as the size measure in method [3] of Section 3.9. "Very large" refers to units for which this measure exceeds the sampling interval used in the selection of LCS areas from the base sample: $(c_i/p_i) \geq I_s$.

These areas can be treated in the same way as in other cases described in previous chapters. For instance, the measure of size $(c_i/p_i)I$ is redefined as $=I_s$, so that any such area in the first sample is taken into the LCS sample with certainty.

These areas thus retain the original probability of selection into the base sample unchanged for the LCS. The set categorized as "large" areas in the LCS is independent of the set so categorized in the base survey, since the two are defined on the basis of different size criteria.

As to the ultimate stage of selecting households or persons in the sample areas, the sampling rate $f2'_i$ at the ultimate stage may, for instance, be correspondingly adjusted so as to keep the required overall selection probability f' unchanged.

5.3.2 "Very small" areas

For large areas the PPS sampling procedure needed adjustment only because the size measure c_i/p_i exceeded the sampling interval Is. Therefore areas were identified as being large or not large on the bases of the ratio c_i/p_i.

By contrast, areas are defined as being "very small" in terms of the number of ultimate units they possess or contribute to the sample, i.e. the number (or expected number) of working children c_i.

The presence of small c_i values has important practical consequences.

First of all, it should be emphasized that in the design described above of selecting areas from the base survey with probability proportional to c_i/p_i, areas for which no working children have been reported in the base survey ($c_i=0$) are automatically excluded from the LCS sample. Formally, this is of course also true of areas with no population ($p_i=0$) for the selection of the sample areas in the base sample with probability proportional to p_i. But in practice the two situations are quite different. Areas with no population ($p_i=0$) tend to be rare and of no interest to the survey in any case, but areas with no working children ($c_i=0$) may be very common.

Secondly, the situation with respect to the later (working children) is likely to be much more changing compared to the situation with respect to the former (population). Hence the information on the presence or otherwise of working children in an area needs to be quite fresh.

Thirdly, even when not exactly zero, very small values of c_i are much more likely to occur than small p_i values. This is because of the uneven distribution of child labour across sample areas. The practical question arises as to whether a lower limit should be set for automatic exclusion from the sample of areas with c_i values below that limit.

Different procedures may be used for the selection of households within sample areas. These procedures are probably more varied in labouring children surveys than in other, more general surveys of the population.

If households within LCS sample areas are selected with inverse-PPS, the procedures for dealing with very small areas are the same as those described in the previous chapters. With "very small" areas defined in terms of the size measure c_i, small are the areas whose size measure is smaller than the required sample-take at the ultimate stage, i.e. $c_i < b'$ in the equation for $f2'_i$ (equation [2]). These units can be treated in the same way as for instance method 3 or method 5 in Section 3.10.

An obvious concern in an LCS is to ensure that the required sample size of working children can be achieved in practice. This can be a problem if the prevalence of child labour is lower than was assumed at the time of sample design, or if because of poor quality the previous survey failed to identify a high proportion of units subject to child labour. When such problems exist, a "compact cluster" or "take-all" design can be an interesting option. In this design all relevant units (e.g. households with working children) in a selected area are taken into the sample, possibly with an upper limit on the maximum number to be selected. Selection probabilities of ultimate units, as well as sample takes per area, will generally vary in such a design.

5.4 Expanding the size of first sample areas

It can happen that the type of areas originally selected in the base sample are too small to yield the required number of cases for the LCS. In such situations it may be necessary to consider whether some of the areas - perhaps those with high concentrations of labouring children, which are also likely to have such high concentrations in the neighbourhood - can be expanded in physical size to include additional neighbouring areas. The following paragraphs describe a simple procedure to replace existing sample areas by larger areas.

Let us suppose that the type of area units used for selecting the first sample (A) are subdivisions of some higher level units (B), and that the latter are considered to be suitable units for the expanded sample. An area (a level-A unit) in the existing sample can be replaced by the larger level-B unit to which it belongs. The resulting sample would be equivalent, statistically speaking, to the selection of whole level-B units, *with the probability of selection of a level-B unit taken to equal the sum of the selection probabilities of all the level-A units contained within it*. It is with this increased probability that each level-A unit within the larger level-B unit appears in the expanded

sample. The larger units coming into the sample in this manner can then be enumerated to obtain the required information (such as size measures c_i) for selecting a sample from these larger units for the LCS.[27]

In more specific terms, the procedure is as follows.

Let the actual selection probability of an A-level unit in the first sample be P_A/I where P_A is its size measure, and I the PPS selection interval. Each A-level unit in the population is supposed to belong to one particular B-level unit. The procedure involves replacing the actual A-level unit in the sample by the whole of B-level unit to which it belongs. The selection probability of the concerned B-level unit is P_B/I, where P_B is the sum of size measures of all A-level units contained in the B-level units.

Subsequently, sub-sampling for the LCS can be applied to the whole B-level unit "brought into" the first sample in the above manner.

It is not necessary to expand all the units in this manner, provided that the criteria determining which type of units to expand and the procedure for such expansion are decided upon before the sample selection, and are not influenced by which particular A-level units happen to be selected into the sample

While the simple procedure described above may often suffice, more complex procedures can also be devised with the objective that, for instance, the expansion of the area units takes place in such a way that the smaller area originally selected lies at or near the geographical centre of the expanded area.

Adoptive cluster sampling is another, more sophisticated approach which may be useful and feasible in certain circumstance. Its basic principles are summarized in the next section.

5.5 Adaptive cluster sampling

Adaptive sampling

Adaptive sampling is particularly useful when the population of interest is rare, unevenly distributed, hidden, or hard to reach. In conventional sampling, the sampling design is based entirely on *a priori* information, and is fixed before the study begins. By contrast, in adaptive sampling, the sampling design is adapted on the basis of observations made during the survey.[28]

The objective of such a sampling procedure is to take advantage of population characteristics so as to obtain, relative to conventional designs, more precise estimates of population values for a given sample size or cost. The secondary objective is to

[27] If, occasionally, a B-level unit contains two A-level units selected into the original sample, then it can be considered as having been selected twice into the enlarged sample; similarly for a unit with more than two selections.

[28] Basic references:
Thompson, S.K. (1990). "Adaptive cluster sampling". *Journal of the American Statistical Association*, No. 85, pp. 1050-1059.
Thompson, S.K. (1992). *Sampling*. John Wiley & Sons, Inc., New York, 339pp.
Thompson, S.K., Seber. G.A.F. (1996*). Adaptive Sampling*. John Wiley & Sons, Inc., New York, 265pp.

increase the yield of observations of interest which may result in better estimates of other parameters for them. This can be an extremely useful feature in capturing large-enough samples of, for instance, non-national populations.

In contrast to conventional sampling designs, adaptive sampling makes use of values observed in the sample. Although sequential sampling also looks at the data, the information so obtained is used to decide how many more units to sample and whether or not to stop sampling. In contrast, adaptive designs tell *which* units to include in the sample.

Special estimation procedures taking the sampling design into account are needed when adaptive sampling has been used. These procedures can yield estimates that are considerably better than conventional estimates. For rare and clustered populations adaptive designs can offer substantial gains in efficiency over conventional designs, and for hidden populations link-tracing and other adaptive procedures may be the only practical way to obtain a sample large enough for the study objectives.

Adaptive cluster sampling

Actually "adaptive sampling" is a whole class of special approaches. Link-tracing designs such as snowball sampling, random walk methods, network sampling, adaptive allocation and adaptive cluster sampling are all various forms of adaptive sampling designs.

The type of adaptive sampling most often referred to is *adaptive cluster sampling*. Adaptive cluster sampling is useful in situations where the characteristic of interest is sparsely distributed but highly concentrated. In environment sampling, for instance, examples of such populations can be found in the study of rare and endangered species, or of pollution concentrations. In human populations, such techniques may be useful in investigating the unconditional worst forms of child labour, the epidemiology of rare diseases, concentrations of immigrant populations, etc.

Adaptive cluster sampling is most useful when a quick turnaround of analytical results is possible, since at any moment further sampling depends on analysis of the information already collected.

Adaptive cluster sampling involves the specification of:

1. the initial sampling design and size (prior to any additions from adaptive sampling);
2. the definition of what constitutes the "neighbourhood" for a sampling unit;
3. the condition that triggers or initiates adaptive sampling from a unit in the initial sample;
4. any restrictions on the repeated application of the rules for adding new units to the sample (so as to control the resulting sample size);
5. estimation procedures (including variance estimation).

We begin with an initial sample 1, selected according to conventional procedures, and the units constituting the neighbourhood of each selected unit identified according to condition 2. If the characteristics of a sampled unit satisfy condition 3 for adaptive sampling, then all units in the already selected unit's neighbourhood are added to the

sample. If any of the newly added units in that neighbourhood satisfy condition 3, then the neighbourhood of that unit defined according to condition 2 is also added to the sample. The process continues until a cluster of units is obtained that contains a boundary of "edge" units that do not satisfy condition 2, or which are subject to restriction 4. The final sample consists of (not necessarily distinct) clusters, one for each unit selected in the initial sample. Estimation procedures 5 take into account the probabilities with which the original and the subsequently added unit appear in the sample, and also the structure of the resulting sample.

Strengths

1. First, unlike traditional designs which focus only on one objective, adaptive sampling aims to address simultaneously the objective of estimating the mean concentration and that of determining the pattern and extent of concentrations of the phenomenon of interest.

2. Adaptive cluster sampling concentrates resources in areas and types of units/events of special interest: i.e. in areas of higher concentration of those units or events. It directs the selection of additional sampling units to these high concentration areas, provided that the initial sample "hits" the areas of interest.

3. In addition, additional characteristics can be observed, adding to the overall usefulness of the study. For instance, in studies on the presence or absence of rare animal populations, measurements on size, weight, etc. can be made on the animals that are found. The same applies in social surveys of human populations, such as children engaged in the unconditional *worst forms of child labour*.

Limitations

1. The iterative nature of adaptive cluster sampling introduces some limitations. With adaptive cluster sampling the process of sampling, testing, re-sampling and again testing may take considerable time. If quick and inexpensive field measurements are not readily available, the total sampling cost could quickly become substantial.

2. Because the sampling process stops only when no more units are found to have the characteristic of interest (or some other specified restriction becomes applicable), the final overall sample size is an unknown quantity. This feature makes the total cost also an unknown quantity.

3. Although it is possible to budget for the sampling process by using the expected total cost, the expected total cost in turn depends strongly on the validity of the assumption about how widely the characteristic of interest is spread. If it is very widely spread, the resulting sample size may "explode" to the point where it becomes unmanageable. Let us consider the case of a survey of a rare population where only a few small areas of high concentration are assumed to exist. Let us also suppose that this assumption is not valid, i.e. that the concentration is more widespread, almost throughout the entire study area. The initial sample has a high probability of "hitting" an area of concentration. Because the concentration areas are widespread, the follow-up sample size will be large, and the total sample size may even become close to the number of sampling units in the whole population.

4. On the other hand, the sample size may turn out to be inadequate for the purpose of the survey if the phenomenon of interest is more concentrated than expected and the initial sample is not large enough to capture it adequately.

5. Generally, estimation procedures are more (often much more) complex than those with a conventional design.

Example

Consider the following scenario for the study of a sub-population which is very unevenly distributed in the general population. Assume that in most places sampled, the concentration of this sub-population is light or negligible, but a few scattered pockets of high concentration are encountered. There are two questions of interest. First, what is the average level of concentration for the whole area – that is, the proportion belonging to this sub-population? Second, where are the pockets of concentration located and what are their characteristics?

Using the traditional statistical approach, a random or systematic sample of sites or units would be selected and the concentration measured at each selected site. The average of these measurements provides an unbiased estimate of the population average. The individual observations can be used to create a contour map to locate peaks of concentration. However, with this pattern of concentration, the traditional statistical approach can have problems. If concentration is negligible over most of the area, the majority of the measurements will be zero or have levels that are not detectable. Furthermore, random sampling may miss most of the pockets of higher concentration.

Thus, even though the sample average is still an unbiased estimator of the population mean, it will be less precise than an unbiased estimator that takes into account the unevenness in the distribution of the sub-population over the entire area. Furthermore, the contour map from a simple random sample design may not be as accurate in the areas of higher concentration because the areas are not well represented in the sample.

Adaptive cluster sampling could provide a better approach in situations similar to the one described above. For populations where the characteristic of interest is sparsely distributed but highly concentrated, adaptive cluster sampling can produce substantial gains in precision over traditional designs using the same sample sizes.

Adaptive sampling is particularly useful when the population of interest is rare, unevenly distributed, hidden, or hard to reach.[29] Examples of such populations are injection drug users, individuals at high risk for HIV/AIDS and young adolescents who are nicotine dependent. In conventional sampling, the sampling design is based entirely on *a priori* information and is fixed before the study begins. By contrast, in adaptive sampling the sampling design adapts to observations made during the survey; for example, drug users may be asked to refer other drug users to the researcher. Special estimation procedures taking the sampling design into account are needed when adaptive sampling has been used.

[29] Thompson SK, Collins LM (2002). Adaptive sampling in research on risk-related behaviours. *Drug and Alcohol Dependence* vol.. 68 (2002), pp. 57-67. This article introduces adaptive sampling designs to substance-use researchers. The text is the abstract of this article.

5.6 Numerical illustrations

The objective of these simulations is to illustrate numerically some aspects of the sample design and selection procedures discussed in this chapter.

The numerical illustrations are based on a population of area units, introduced in Section 3.11.1. As noted, a small data set was generated statistically to provide a reasonably realistic example of a set of areas with their associated measures of size. The original population consisted of some P=50,000 ultimate units (households), in nearly 800 areas of an average size of around 65 households. The areas varied considerably in size, from the smallest consisting of fewer than 10 units to the largest with over 1,000 units. The distribution is fairly typical of real situations, strongly skewed to the left with many small areas and a few very large ones.

A systematic-PPS sample of $a=100$ areas was selected, the sampling interval for the selection of areas being $I=(P/a)=500$.

The number of working children (c_i) for each area in the set was also simulated. In an actual survey this sort of information would be obtained in a CLS preceding the LCS. In the design being discussed here, c_i values (or more precisely, their ratio c_i/p_i to the population of the area) determine the probabilities of selection of area from CLS to LCS.

Columns [6]-[7] of Table 3.3 showed simulated values of c_i, the number of working children in area *i*, and the ratio of this to p_i, the population of children in the relevant age group exposed to the risk of child labour. These values illustrate a wide range of the degree of concentration of child labour in the areas. The proportion of children in child labour varies from 0 to 97 per cent, while the overall average is 22 per cent.

5.6.1 Selection of areas from the base sample

The sample of 100 areas shown in Table 3.3 was ordered according to p_i values (simulated measure of population or number of households). The sample areas were assumed to have been selected with a probability proportional to this measure of size. Areas were numbered sequentially 1 to 100 in the above-mentioned order. The same 100 areas are sorted in different orders in Tables 5.1 to 5.3. The serial number, "S.No" in the first column of each of these tables identifies the original numbering in Table 3.3, so that information on any given area can be linked across the tables.

In the original selection of areas, special treatments were given to areas classified as being "too large" or "too small", as explained in the notes to Table 3.3 in Section 3.11. Areas with size measures exceeding the interval for systematic sampling (I=500) were given a modified size measure of 500 and a probability of selection of 1.0. Areas with size measures smaller than the assumed target sample-take (b=20 households) were given a modified size measure of 20, and their selection probability was taken to be proportional to this modified size measure.

These adjustments have been carried forward to the next stages of selection from CLS to LCS illustrated in Tables 5.1 to 5.3. They do not affect the procedures applied in those tables in any other way.

The cumulation of c_i/p_i over the 100 areas (25.8), divided by the number of areas to be selected (50) gives the interval to be applied for systematic selection of areas, I=0.52.

Areas with $(c_i/p_i)>I$ are kept in the LCS sample with probability 1.0 (a total of 18 areas turned out to be "very large" in this sense – see Table 5.1). For these areas c_i is redefined as $c'_i=I\cdot p_i$ (Table 5.1, col.[7]), i.e. as $(c'_i/p_i)=I$, so as to ensure that the selection probability does not exceed 1.0. In all other areas there is no change: c_i= original c_i.

The final area selection probabilities from the base to the LCS sample (f1) are shown in column [6]; these are proportional to (c'_i/p_i). The overall probability of selection of an area is $f0_i\cdot f1_i$, the former coming from column [5] of Table 3.3.

Note that this is not the final step in the selection of areas for the hypothetical LCS being discussed. The values of c'_i/p_i or c_i in the table indicate a common practical problem: many areas contain no working children or only very few, and the issue is whether it is affordable to let such areas into the sample. These "very small" areas need special treatment.[30] The details of such a treatment depend on the procedure adopted for the second stage of sampling – namely, for the selection of ultimate units within sample areas. This is illustrated in Table 5.2.

[30] Note that "small" here does not refer to the population size (p_i) of the area, but its smallness in terms of the variable of interest in the LCS, namely the number of working children (c_i).

Table 5.1. Selection of areas from the base sample

Parameters:

Base survey, a = 100 [number of areas in base sample]
LCS, a = 50 [number of areas in LCS sample]
I = 1.9 [interval for systematic selection of areas]
n = 500 [expected sample size]

[1]	[2]	[3]	[4]	[5]	[6]	[7]
	(modified)		(sorted by)		(modified)	(modified)
S.No	pi	ci	ci/pi	(ci/pi)/I	f1	ci
73	196	190	0.97	1.88	1.00	101
54	100	93	0.92	1.79	1.00	52
92	425	382	0.90	1.74	1.00	219
84	299	268	0.90	1.74	1.00	154
49	83	74	0.90	1.74	1.00	43
35	60	53	0.89	1.73	1.00	31
42	70	61	0.87	1.69	1.00	36
10	22	18	0.81	1.57	1.00	11
97	500	383	0.77	1.48	1.00	258
65	157	117	0.74	1.44	1.00	81
55	103	73	0.71	1.38	1.00	53
45	73	51	0.69	1.34	1.00	38
86	331	220	0.66	1.29	1.00	171
7	20	12	0.62	1.20	1.00	10
25	43	27	0.62	1.20	1.00	22
60	115	68	0.60	1.15	1.00	59
93	500	292	0.58	1.13	1.00	258
47	79	41	0.52	1.01	1.00	41
26	44	23	0.51	0.99	0.99	23
28	48	23	0.48	0.94	0.94	23
44	72	35	0.48	0.93	0.93	35
78	229	107	0.47	0.90	0.90	107
24	43	20	0.46	0.90	0.90	20
12	25	11	0.45	0.87	0.87	11
20	38	17	0.45	0.86	0.86	17
89	368	162	0.44	0.85	0.85	162
3	20	9	0.44	0.85	0.85	9
1	20	9	0.44	0.85	0.85	9
27	46	18	0.39	0.75	0.75	18
67	163	51	0.31	0.61	0.61	51
18	35	11	0.31	0.61	0.61	11
46	74	20	0.27	0.52	0.52	20
2	20	5	0.26	0.51	0.51	5
76	213	55	0.26	0.50	0.50	55
8	20	5	0.24	0.47	0.47	5
52	97	22	0.23	0.44	0.44	22
70	181	40	0.22	0.42	0.42	40
58	114	25	0.22	0.42	0.42	25
61	124	27	0.21	0.42	0.42	27

[1]	[2]	[3]	[4]	[5]	[6]	[7]
	(modified)		(sorted by)		(modified)	(modified)
S.No	pi	ci	ci/pi	(ci/pi)/I	f1	ci
51	92	20	0.21	0.42	0.42	20
21	39	8	0.21	0.41	0.41	8
15	30	6	0.21	0.40	0.40	6
32	53	11	0.20	0.40	0.40	11
57	112	21	0.19	0.37	0.37	21
94	500	92	0.18	0.35	0.35	92
79	229	42	0.18	0.35	0.35	42
37	65	12	0.18	0.35	0.35	12
59	114	20	0.18	0.35	0.35	20
36	64	11	0.17	0.34	0.34	11
38	67	11	0.17	0.33	0.33	11
33	57	9	0.16	0.31	0.31	9
53	97	15	0.16	0.31	0.31	15
16	33	5	0.15	0.30	0.30	5
56	107	16	0.15	0.29	0.29	16
29	50	7	0.15	0.28	0.28	7
5	20	3	0.14	0.27	0.27	3
68	171	22	0.13	0.25	0.25	22
71	186	23	0.12	0.24	0.24	23
96	500	60	0.12	0.23	0.23	60
85	304	36	0.12	0.23	0.23	36
19	36	4	0.11	0.21	0.21	4
72	189	20	0.11	0.21	0.21	20
64	156	16	0.11	0.20	0.20	16
48	79	7	0.09	0.18	0.18	7
100	500	39	0.08	0.15	0.15	39
88	358	26	0.07	0.14	0.14	26
69	175	11	0.06	0.12	0.12	11
43	72	3	0.04	0.08	0.08	3
40	69	3	0.04	0.08	0.08	3
75	204	8	0.04	0.07	0.07	8
80	237	9	0.04	0.07	0.07	9
39	68	2	0.03	0.07	0.07	2
82	240	7	0.03	0.06	0.06	7
6	20	1	0.03	0.05	0.05	1
63	138	3	0.03	0.05	0.05	3
13	28	1	0.03	0.05	0.05	1
99	500	9	0.02	0.03	0.03	9
41	69	1	0.02	0.03	0.03	1
14	28	0	0.02	0.03	0.03	0
30	52	1	0.01	0.03	0.03	1
22	40	1	0.01	0.03	0.03	1
11	24	0	0.01	0.02	0.02	0
50	92	1	0.01	0.02	0.02	1
81	238	2	0.01	0.02	0.02	2

[1]	[2]	[3]	[4]	[5]	[6]	[7]
	(modified)		(sorted by)		(modified)	(modified)
S.No	pi	ci	ci/pi	(ci/pi)/l	f1	ci
17	34	0	0.01	0.02	0.02	0
66	159	1	0.01	0.01	0.01	1
90	383	2	0.01	0.01	0.01	2
77	213	1	0.01	0.01	0.01	1
31	52	0	0.01	0.01	0.01	0
87	345	1	0.00	0.01	0.01	1
23	40	0	0.00	0.01	0.01	0
83	246	1	0.00	0.01	0.01	1
74	200	1	0.00	0.00	0.00	1
34	59	0	0.00	0.00	0.00	0
4	20	0	0.00	0.00	0.00	0
95	500	0	0.00	0.00	0.00	0
98	500	0	0.00	0.00	0.00	0
91	424	0	0.00	0.00	0.00	0
62	128	0	0.00	0.00	0.00	0
9	21	0	0.00	0.00	0.00	0
Mean	150.0	37.5	25.79	50.00		29.6

5.6.2 Selection of ultimate units and treatment of "very small" areas

In order to illustrate the procedure, let us assume that the ultimate units are selected within a sample area with universe-PPS so as to obtain a self-weighting sample[31]. Alternative designs are of course possible. With the PPS design, let us assume that an average of b=10 such children are to be selected per sample area, giving an expected total sample size of n=a*b=50*10=500 labouring children.

Column [5] shows the c_i values; these are the same as column [7] of Table 5.1, except that now the data are sorted by (increasing) c_i values since our objective is to identify areas which are "very small" in terms of this measure. Column [5] identifies areas which are "very small" in the sense that they contain fewer than b=10 working children. There are many (45 out of 100) such areas in the illustrative population. This is not unrealistic, since labouring children are often geographically concentrated and in many areas there may be none or very few .

Various procedures dealing with the selection of "very small" areas were described in Section 3.10. Method (5) in that section appears to be a good procedure and has been applied in this illustration.

The method involves sorting the set of very small areas by the measure of size c_i, and dividing it into two parts. The first part consists of the largest among this set of very small areas. The number of areas (say c) to be included in this part is determined in such a way that, if each of these units were given a size measure b (=10, the target sample taken per area), that would account for the total measure of size (say $\Sigma_s c_i$) for all the units in the "very small" set. That is;

$c=\Sigma_s c_i/b=138/10=14$ in our illustration.

For the purpose of sample selection, each of these 14 units is assigned a size measure of 10. The remaining 45-14=31 areas in the "very small" set are indeed very small, and as such are excluded from the sampling altogether by being assigned a zero measure of size. The adjusted size measures are shown in column [6].

The final sample weight for the 14 units retained, if they are selected, can be adjusted upwards as a compensation for the 31 of the smallest units that were excluded. This will be illustrated in Table 5.3.

Table 5.2 also shows selection probabilities at various stages:

$f0_i$ for the original selection of CLS areas (from Table 3.3)

$f1_i$ for the selection of areas from CLS to LCS (from Table 5.1)

$f2_i$ for the selection of ultimate units in an area $=10/c_i$, with the size measure c_i modified as explained above.

[31] It is not relevant to the discussion here that the ultimate units may be defined in different ways. They may, for instance, refer to households containing one or more working children, or to individual working children. Some non-working children or households without any working children may also be included for comparison.

Table 5.2. Selection of ultimate units from selected areas

Parameters:

Base survey, a = **100** [number of areas in base sample]
LCS, a = **50** [number of areas in LCS sample]
I = **1.9** [interval for systematic selection of areas]
n = **500** [expected sample size]

[1]	[2]	[3]	[4]	[5]	[6]	[7]
	Base sample		Area selection	Sorted by	Modified-2	hh selection
S.No	pi	f0	f1	ci (modified-1)	ci	f2
9	21	0.04	0.00	0	0	
62	128	0.26	0.00	0	0	
91	424	0.85	0.00	0	0	
4	20	0.04	0.00	0	0	
98	500	1.00	0.00	0	0	
34	59	0.12	0.00	0	0	
23	40	0.08	0.01	0	0	
95	500	1.00	0.00	0	0	
31	52	0.10	0.01	0	0	
11	24	0.05	0.02	0	0	
17	34	0.07	0.02	0	0	
14	28	0.06	0.03	0	0	
74	200	0.40	0.00	1	0	
6	20	0.04	0.05	1	0	
22	40	0.08	0.03	1	0	
13	28	0.06	0.05	1	0	
83	246	0.49	0.01	1	0	
30	52	0.10	0.03	1	0	
66	159	0.32	0.01	1	0	
50	92	0.18	0.02	1	0	
87	345	0.69	0.01	1	0	
41	69	0.14	0.03	1	0	
77	213	0.43	0.01	1	0	
39	68	0.14	0.07	2	0	
90	383	0.77	0.01	2	0	
81	238	0.48	0.02	2	0	
40	69	0.14	0.08	3	0	
5	20	0.04	0.27	3	0	
43	72	0.14	0.08	3	0	
63	138	0.28	0.05	3	0	
19	36	0.07	0.21	4	0	
8	20	0.04	0.47	5	10	1.00
16	33	0.07	0.30	5	10	1.00
2	20	0.04	0.51	5	10	1.00
15	30	0.06	0.40	6	10	1.00
48	79	0.16	0.18	7	10	1.00
29	50	0.10	0.28	7	10	1.00
82	240	0.48	0.06	7	10	1.00
75	204	0.41	0.07	8	10	1.00

[1]	[2]	[3]	[4]	[5]	[6]	[7]
	Base sample		Area selection	Sorted by	Modified-2	hh selection
S.No	pi	f0	f1	ci (modified-1)	ci	f2
21	39	0.08	0.41	8	10	1.00
80	237	0.47	0.07	9	10	1.00
1	20	0.04	0.85	9	10	1.00
3	20	0.04	0.85	9	10	1.00
99	500	1.00	0.03	9	10	1.00
33	57	0.11	0.31	9	10	1.00
7	20	0.04	1.00	10	10	0.97
69	175	0.35	0.12	11	11	0.94
32	53	0.11	0.40	11	11	0.92
18	35	0.07	0.61	11	11	0.92
36	64	0.13	0.34	11	11	0.91
38	67	0.13	0.33	11	11	0.89
10	22	0.04	1.00	11	11	0.89
12	25	0.05	0.87	11	11	0.88
37	65	0.13	0.35	12	12	0.85
53	97	0.19	0.31	15	15	0.65
56	107	0.21	0.29	16	16	0.63
64	156	0.31	0.20	16	16	0.61
20	38	0.08	0.86	17	17	0.60
27	46	0.09	0.75	18	18	0.56
51	92	0.18	0.42	20	20	0.51
46	74	0.15	0.52	20	20	0.51
24	43	0.09	0.90	20	20	0.51
72	189	0.38	0.21	20	20	0.50
59	114	0.23	0.35	20	20	0.49
57	112	0.22	0.37	21	21	0.47
52	97	0.19	0.44	22	22	0.46
68	171	0.34	0.25	22	22	0.45
25	43	0.09	1.00	22	22	0.45
26	44	0.09	0.99	23	23	0.44
71	186	0.37	0.24	23	23	0.43
28	48	0.10	0.94	23	23	0.43
58	114	0.23	0.42	25	25	0.41
88	358	0.72	0.14	26	26	0.39
61	124	0.25	0.42	27	27	0.38
35	60	0.12	1.00	31	31	0.32
44	72	0.14	0.93	35	35	0.29
42	70	0.14	1.00	36	36	0.28
85	304	0.61	0.23	36	36	0.28
45	73	0.15	1.00	38	38	0.26
100	500	1.00	0.15	39	39	0.26
70	181	0.36	0.42	40	40	0.25
47	79	0.16	1.00	41	41	0.25
79	229	0.46	0.35	42	42	0.24
49	83	0.17	1.00	43	43	0.23

[1]	[2]	[3]	[4]	[5]	[6]	[7]
	Base sample		Area selection	Sorted by	Modified-2	hh selection
S.No	pi	f0	f1	ci (modified-1)	ci	f2
67	163	0.33	0.61	51	51	0.20
54	100	0.20	1.00	52	52	0.19
55	103	0.21	1.00	53	53	0.19
76	213	0.43	0.50	55	55	0.18
60	115	0.23	1.00	59	59	0.17
96	500	1.00	0.23	60	60	0.17
65	157	0.31	1.00	81	81	0.12
94	500	1.00	0.35	92	92	0.11
73	196	0.39	1.00	101	101	0.10
78	229	0.46	0.90	107	107	0.09
84	299	0.60	1.00	154	154	0.06
89	368	0.74	0.85	162	162	0.06
86	331	0.66	1.00	171	171	0.06
92	425	0.85	1.00	219	219	0.05
93	500	1.00	1.00	258	258	0.04
97	500	1.00	1.00	258	258	0.04
Mean	150.0			29.6	29.7	

5.6.3 Overall selection probabilities and expected sample size

Table 5.3 shows the final results. It is sorted in the same order as the previous table, by c_i. The illustration does not actually select a particular sample of a=50 areas, but shows the probabilities with which any of the 100 areas would be selected, and the sample an area would contribute if selected.

Column [4] is the overall sampling rate f, which is the product of selection probabilities at all the stages shown in columns [1]-[3]: the selection of the base sample areas (f0); from them, the selection of LCS areas (f1); and within the latter, the selection of ultimate units, e.g. labouring children (f2).

Column [5] shows the special weights for the largest 14 of the areas retained for sample selection from the "very small" set of 45 units discussed above in relation to Table 5.2. These weights are larger than 1.0 as a compensation for the exclusion of the remaining 31 smallest areas in the "very small" set. The weights in fact equal $(b/c_i)=(10/c_i)$ where c_i is the unit measure of size, <10 by definition for any unit in the "very small" set. The excluded 31 areas receive a zero special weight. The special weight by definition is 1.0 for all the remaining areas which are not in the "very small" set.

Column [6] shows the sample-take (b_i) which would be obtained if the unit were selected into the sample, assuming that the number of ultimate units (say c'_i) actually found in the area at the time of the survey was exactly the same as the number (c_i) used in the sample selection.

For most of the areas, the number in column [6] equals 10, as the final stage selection equation is $f2_i=10/c_i$, giving the expected number of selections as $f2_i \cdot c_i=10$.

The number differs from 10 for two groups of units:

- "very large" areas identified in Table 5.1, whose size measures (c_i/p_i) exceeded the area sampling interval (I=0.52); the ratio of these two quantities is the factor by which their final contribution exceeds 10;
- any "very small" area retained for sample selection, which has by definition fewer than 10 units to contribute; in fact, each such area contributes (10/D) ultimate units, where D is the special weight shown in column [5]. Note that this makes the *weighted* contribution $(10/D)\cdot D=10$ in the case of all such units.

Finally, column [7] shows the "expected" value of the sample contributed by each area. This is the contribution the area would make on the average, over all possible samples. It is smaller than the actual contribution which an area would make once it actually appears in the sample (col. [6]), but then there are also many occasions when the area makes no actual contribution because it does not get selected into a sample. The expected values shown in column [7] permit us to examine the average characteristics of the samples which are possible with a given design, without actually having to select individual samples.

The sum over all 100 areas in the "population" of values in column [7] shows, for instance, that the expected number of ultimate units in the sample is 500.

Table 5.3. Overall selection probability and expected sample size

Parameters:

Base survey, a = 100 [number of areas in base sample]
LCS, a = 50 [number of areas in LCS sample]
I = 1.9 [interval for systematic selection of areas]
n = 500 [expected sample size]

[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
	Sorted by	Selection probailities by stage					No. of final units selected	
	original				final	special	if in sample	expected value
S.No	ci	f0*	f1*	f2	=f	weight (D)	bi	bi*D*f1
9	0	0.04	0.00					0
62	0	0.26	0.00					0
91	0	0.85	0.00					0
4	0	0.04	0.00					0
98	0	1.00	0.00					0
34	0	0.12	0.00					0
23	0	0.08	0.01					0
95	0	1.00	0.00					0
31	0	0.10	0.01					0
11	0	0.05	0.02					0
17	0	0.07	0.02					0
14	0	0.06	0.03					0
74	1	0.40	0.00					0
6	1	0.04	0.05					0
22	1	0.08	0.03					0
13	1	0.06	0.05					0
83	1	0.49	0.01					0
30	1	0.10	0.03					0
66	1	0.32	0.01					0
50	1	0.18	0.02					0
87	1	0.69	0.01					0
41	1	0.14	0.03					0
77	1	0.43	0.01					0
39	2	0.14	0.07					0
90	2	0.77	0.01					0
81	2	0.48	0.02					0
40	3	0.14	0.08					0
5	3	0.04	0.27					0
43	3	0.14	0.08					0
63	3	0.28	0.05					0
19	4	0.07	0.21					0
8	5	0.04	0.47	1.00	0.019	2.05	5	5
16	5	0.07	0.30	1.00	0.020	1.99	5	3
2	5	0.04	0.51	1.00	0.021	1.89	5	5
15	6	0.06	0.40	1.00	0.024	1.60	6	4
48	7	0.16	0.18	1.00	0.028	1.40	7	2

[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
	Sorted by	Selection probabilities by stage					No. of final units selected	
	original				final	special	if in sample	expected value
S.No	ci	f0*	f1*	f2	=f	weight (D)	bi	bi*D*f1
29	7	0.10	0.28	1.00	0.029	1.36	7	3
82	7	0.48	0.06	1.00	0.029	1.35	7	1
75	8	0.41	0.07	1.00	0.030	1.30	8	1
21	8	0.08	0.41	1.00	0.032	1.20	8	4
80	9	0.47	0.07	1.00	0.033	1.16	9	1
1	9	0.04	0.85	1.00	0.034	1.15	9	9
3	9	0.04	0.85	1.00	0.034	1.14	9	9
99	9	1.00	0.03	1.00	0.035	1.12	9	0
33	9	0.11	0.31	1.00	0.036	1.09	9	3
7	12	0.04	1.00	0.97	0.039	1.00	12	12
69	11	0.35	0.12	0.94	0.039	1.00	10	1
32	11	0.11	0.40	0.92	0.039	1.00	10	4
18	11	0.07	0.61	0.92	0.039	1.00	10	6
36	11	0.13	0.34	0.91	0.039	1.00	10	3
38	11	0.13	0.33	0.89	0.039	1.00	10	3
10	18	0.04	1.00	0.89	0.039	1.00	16	16
12	11	0.05	0.87	0.88	0.039	1.00	10	9
37	12	0.13	0.35	0.85	0.039	1.00	10	4
53	15	0.19	0.31	0.65	0.039	1.00	10	3
56	16	0.21	0.29	0.63	0.039	1.00	10	3
64	16	0.31	0.20	0.61	0.039	1.00	10	2
20	17	0.08	0.86	0.60	0.039	1.00	10	9
27	18	0.09	0.75	0.56	0.039	1.00	10	8
51	20	0.18	0.42	0.51	0.039	1.00	10	4
46	20	0.15	0.52	0.51	0.039	1.00	10	5
24	20	0.09	0.90	0.51	0.039	1.00	10	9
72	20	0.38	0.21	0.50	0.039	1.00	10	2
59	20	0.23	0.35	0.49	0.039	1.00	10	4
57	21	0.22	0.37	0.47	0.039	1.00	10	4
52	22	0.19	0.44	0.46	0.039	1.00	10	4
68	22	0.34	0.25	0.45	0.039	1.00	10	3
25	27	0.09	1.00	0.45	0.039	1.00	12	12
26	23	0.09	0.99	0.44	0.039	1.00	10	10
71	23	0.37	0.24	0.43	0.039	1.00	10	2
28	23	0.10	0.94	0.43	0.039	1.00	10	10
58	25	0.23	0.42	0.41	0.039	1.00	10	4
88	26	0.72	0.14	0.39	0.039	1.00	10	1
61	27	0.25	0.42	0.38	0.039	1.00	10	4
35	53	0.12	1.00	0.32	0.039	1.00	17	18
44	35	0.14	0.93	0.29	0.039	1.00	10	9
42	61	0.14	1.00	0.28	0.039	1.00	17	17
85	36	0.61	0.23	0.28	0.039	1.00	10	2

[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
	Sorted by	Selection probailities by stage					No. of final units selected	
	original				final	special	if in sample	expected value
S.No	ci	f0*	f1*	f2	=f	weight (D)	bi	bi*D*f1
45	51	0.15	1.00	0.26	0.039	1.00	13	14
100	39	1.00	0.15	0.26	0.039	1.00	10	2
70	40	0.36	0.42	0.25	0.039	1.00	10	4
47	41	0.16	**1.00**	0.25	0.039	1.00	10	10
79	42	0.46	0.35	0.24	0.039	1.00	10	4
49	74	0.17	**1.00**	0.23	0.039	1.00	17	18
67	51	0.33	0.61	0.20	0.039	1.00	10	6
54	93	0.20	1.00	0.19	0.039	1.00	18	18
55	73	0.21	1.00	0.19	0.039	1.00	14	14
76	55	0.43	0.50	0.18	0.039	1.00	10	5
60	68	0.23	1.00	0.17	0.039	1.00	12	12
96	60	1.00	0.23	0.17	0.039	1.00	10	2
65	117	0.31	1.00	0.12	0.039	1.00	14	15
94	92	1.00	0.35	0.11	0.039	1.00	10	4
73	190	0.39	1.00	0.10	0.039	1.00	19	19
78	107	0.46	0.90	0.09	0.039	1.00	10	9
84	268	0.60	1.00	0.06	0.039	1.00	17	18
89	162	0.74	0.85	0.06	0.039	1.00	10	9
86	220	0.66	1.00	0.06	0.039	1.00	13	13
92	382	0.85	1.00	0.05	0.039	1.00	17	18
93	292	1.00	1.00	0.04	0.039	1.00	11	12
97	383	1.00	1.00	0.04	0.039	1.00	15	15
Total								500

Chapter 6
An illustration of sample design and selection procedures

6.1 Introduction

By taking an actual survey on child labour in a developing country as the basis, in this chapter we will illustrate aspects of sample design and selection procedures in concrete, numerically detailed terms. This will hopefully provide some clarification and deeper understanding of these procedures.

As is commonly the case, the total survey on child labour involves two conceptually distinct components. These, using the terminology introduced earlier, are: (1) a *child labour survey* (CLS), with the primary objective of measuring the prevalence of child labour, i.e. the proportion of children who are working according to certain specified definitions of the terms; and (2) a labouring children survey (LCS), with the primary objective of investigating in depth the causes, characteristics, circumstances and consequences of child labour. The two components are related to each other, and often also to two other operations, namely: (3) a labour force survey (LFS), aimed at assessing the economic activity of the entire working-age population; and (4) a household listing operation, with the primary objective of creating a sampling frame of households or similar units within sample areas.

These components can be related in various ways, and various types of substantive and operational links are possible. Some may have more than one objective and/or may be combined with some other operation(s), depending on the specific circumstances and the requirements of the survey system. A common arrangement is the sequence

Listing—LFS—CLS—LCS

where the components of the sequence progressively involve smaller sample sizes and more intensive and detailed data collection. This type of sequence of surveys, including some variations of it, underlines the discussion in Chapters 4 and 5.

In this chapter we will provide a detailed illustration of the sample design and procedures for an alternative arrangement of these components, which is also possible and appropriate in certain circumstances. This arrangement involves two phases: Phase 1, consisting of a combined listing-CLS operation on a relatively large sample; followed by Phase 2, combining the LCS-LFS operations on a substantially smaller sample. This type of arrangement suggests itself in situations where neither household lists nor a major survey such as the LFS are available on which to base the CLS sample and, furthermore, where the opportunity of conducting a child labour survey can be used to fill – albeit partially – gaps in information on the general labour force.

One important difference should be noted between the LCS model discussed in Chapter 5, and the one being discussed as Phase 2 in the present chapter. In the former case, the objective was to obtain a reasonably representative sample of working children in households. Therefore households not containing a working child were not the main interest. Furthermore, for practical and cost reasons, areas containing no households with working children, or even those containing only a very small number of such households, were considered for exclusion from the LCS sample altogether. By contrast, in the present illustration, Phase 2 is required to produce data not only on labouring children but also on the labour force characteristics of the general population. Therefore, the sample must remain representative of the general population. Areas, even if not containing any labouring children according to the previous phase, must be given a non-zero chance of being included in the Phase 2 sample.

The specific illustration in the following sections is based on some proposals for a planned national child labour survey in Yemen. The data are derived from an existing frame of census enumeration areas (EAs)[32], and the survey context and objectives are based on the actual situation. But it should be pointed out that this is not a design which has been implemented as yet, and the survey when conducted may in fact use a somewhat different design.

The design of a sample must of course begin from a specification of the context and objectives of the survey. This is provided in Section 6.2 for the survey in this illustration. The section also provides a summary of the main features of a recommended design, which are then developed with technical details in the following sections.

Section 6.3-6.5 illustrate sampling procedures for, respectively:

- the listing and the CLS in Phase 1,
- the LCS in Phase 2, and
- additional sampling aspects in Phase 2, in particular for the LFS module.

In Section 6.6 we construct a hypothetical frame with some characteristics similar to the population being considered, and illustrate in numerical terms the application of various sampling procedures developed in the previous sections. Much of this demonstration does not require the selection of particular samples. An actual sample selection is illustrated in Section 6.7.

6.2 Context and objectives of the survey in the illustration

As noted, this illustration is based on certain recommendations developed for a planned national child labour survey. For the sake of convenience, we shall assume that the survey is to be conducted as described.

[32] We are thankful to the Government of Yemen for making this information available.

It is assumed that the survey will be carried out in two phases:

- Phase I, listing-CLS component: the listing of households and the measurement of prevalence of child labour in the sample areas (EAs);
- Phase II, LCS-LFS component: the conduct of detailed interviews on labouring children, and also on labour force, for the entire working-age population.

For Phase 1, the results are required by urban and rural parts of individual governorates. The country consists of 21 governorates, forming 41 reporting domains.

6.2.1 Objectives

1. Conducting a large-scale survey for the estimation of the prevalence of child labour at the governorate level, separately for urban and rural areas.
2. A detailed survey of labouring children in households identified as containing such children during Phase 1.
3. Identification and detailed survey of labouring children in other households, even though not identified as containing such children during Phase 1. Labouring children are expected to exist primarily (and in large numbers) in households containing children of school age but not at school. Some working children – but probably a much smaller number – may also exist in other households.
4. Obtaining information on a "control group" of non-working children, for a deeper analysis of the causes, characteristics, circumstances and consequences of child labour.
5. Using the opportunity of the survey also to collect data on the labour force characteristics of the entire working-age population (the LFS "module").

6.2.2 Main features of the sampling design

It is useful to summarize briefly the main features of the proposed design at the outset.

1. For Phase 1 a large sample is required. However, the sample size must be limited to permit quality control of the data collected. All households in selected EAs will be listed and information obtained on the presence of child labour. The recommendation is that the sample size not exceed 800 EAs, amounting to a total of 100,000 households for listing and for the identification of working children.
2. The sample will be allocated to the 41 reporting domains, comprising the urban and rural part of individual governorates, in such a way as to give larger samples to larger domains – but in no way in proportion to the domain size. The allocation is a compromise between the objective of producing national as opposed to domain-level estimates. With minor exceptions, a certain minimum sample size is allocated to any domain irrespective of its size.
3. The frame of EAs will be stratified by type of place and administrative division, and also by area population size and possibly other relevant criteria. Within each stratum an equal probability sample of EAs will be selected independently, using the systematic sampling procedure.

4. Phase 2 will be conducted over a sub-sample of Phase 1 EAs. In the selected areas, household lists as prepared in Phase 1 will provide the frame for the selection of households. Phase 2 therefore uses a two-stage design.

5. Sub-sampling of EAs from Phase 1 to Phase 2 will be with probability proportional to measures of size, with the size measures defined so as to give a higher chance of selection to EAs where child labour is, or is likely to be, high according to the findings of Phase 1. This will help to concentrate the sample for the labouring children survey geographically.

6. Such a procedure makes it essential that there be an adequate time lag between the two phases, so that the necessary frame can be compiled from Phase 1 for the selection of the Phase 2 sample.

7. The list of households in each EA selected for Phase 2 will be divided into two categories: category C households with one working child or more or children not at school; and other, category E households. Category (C) households have a high chance of containing working children, and category (E) households have a low chance.

8. The recommendation is that the sample size not exceed 6,000 category C households and an additional 4,000 category E households. These may come from 350-400 sample EAs. The primary focus is on results for various population groups at the national level, with some breakdown by urban-rural classification and by gender, or by major geographical division of the country.

9. All category C households will be taken into the sample up to a certain maximum number (such as 25-35). That will be the number selected if the available number in the area exceeds that limit.

10. As to the selection of category E households in the same EAs, the sampling rate will vary inversely to the probability with which the EAs were selected from Phase 1, subject to lower and upper limits on the number of such households selected from any area. This procedure will make the probabilities of selection of category E households nearly uniform within domains, and the number of households selected less variable.

Sections 6.3-6.5 below describe, respectively:

- sampling for listing and the CLS in Phase 1;
- sampling for the LCS in Phase 2;
- additional aspects of sampling, primarily from the standpoint of requirements of the LFS component in Phase 2.

Before considering sample design issues, it is necessary to take note of some important aspects of the child labour situation in Yemen that are relevant to the choice of the design.

6.2.3 Some relevant aspects of the child labour situation in Yemen

A rapid assessment report[33] on child labour provides the following picture:[34]

- Child labour is a very serious problem in Yemen and the problem is growing, possibly quite rapidly.
- Of children in the 6-14 age bracket, well over a third (37 per cent) are not at school. This applies to nearly one-in-five (18 per cent) of the boys and to over half (55 per cent) of the girls.

In actual numbers, the above amounts to around 2 million children not at school out of a total of 5.4 million children aged between 6 and 14 years. In a labour force survey which also covered child labour, roughly 0.9 million of these children were reported as working, and the remaining 1.1 million as not working but at the same time not being at school. The latter group appears to be composed entirely of girls, while most boys not at school seem to be reported as working.

The picture appears to be roughly as follows.

Number of children aged 6-14 years (in millions)					
	Total	At school	Not at school	Not at school	
				working	not working
Boys	2.7	2.2	0.5	0.4	0.1
Girls	2.7	1.2	1.5	0.5	1.0
All	5.4	3.4	2.0	0.9	1.1

It is very likely that a great deal of the economic activity of girls is erroneously perceived as "domestic chores" and is not reported as "work" in the surveys. It should be an important objective of the CLS to identify and quantify this phenomenon.

It may be noted that 90 per cent of the working children (i.e. children reported as "working") are reported as working for the family without pay – this applies to boys as well as to girls. Most (over 90 per cent) of this work is concentrated in agriculture and related sectors, in other words in rural areas. (The prevalence of child labour may be 5-7 times higher in rural areas compared to urban). While boys and girls form a nearly identical proportion (50 per cent each) of the children reported as working, girls account for 55 per cent of children working in agriculture and related sectors while boys account for 70-95 per cent of children working in the main non-agriculture sectors such as sales, services, handicraft, unskilled labour and, especially, street work.

Child labour seems to amount to almost a sixth of the total labour force!

A few demographic facts are important for the sample design. With a total population of around 20 million, there are 2.7 million households in the country, giving the average household as containing 7.4 persons. The total labour force is of the order of 5.5 million, i.e. an average of 2.0 economically active persons per household. On the average there are 0.35 working children (aged 5-14) per household and 0.40 other

[33] Shaik, Khaled Rajeh. 2000-2001. *Child Labour in Yemen* (Sana'a, Ministry of Labour of Yemen).

[34] All figures in the report are approximate, and there are inconsistencies in some of the figures reported. Nevertheless they provide an overall picture that is quite clear.

children not at school. This means that, for instance, in a sample listing of 75,000 households, we may get 15,000-20,000 households with a working child, 20,000-25,000 with a child not at school (but not reported as working), and the remaining 30,000-40,000 with all children at school or with no children in the relevant age group.

6.3 Sampling for listing and CLS

6.3.1 Introduction

The Phase 1 listing and CLS operations form part of the first main objective of the planned survey noted above, namely, the conducting a large-scale survey for the estimation of the prevalence of child labour at the governorate level, separately for urban and rural areas.

This calls for a survey on a major scale because of the large number of domains (41) for which sufficiently reliable separate estimates are required. Although this necessitates the use of brief and very simple questions in the survey, a major problem in such a operation is the under-reporting of child labour, and this can be reduced only by using more careful and targeted questioning.

Survey objectives and sample size

A compromise is required therefore between a large sample size, on the one hand, and more accurate measurement on the cases included, on the other.

Above all, the compromise means reconsidering the stated objective of producing estimates for as many as 41 domains, many of which are quite small in terms of population. As to the listing survey to be used for the measurement of the level of child labour, it is likely to be subject to a significant bias of under-reporting. Choosing a sample size simply on the basis of requirements of *sampling precision* is not appropriate. Rather, the primary concern should be to obtain reliable estimates first at the national level, then by urban and rural area and by gender, possibly followed by those for four or five major geographical regions of the country. Estimates for larger governorates will be the next step. Further breakdown at this stage can only be indicative, to identify any really large differences. It is also possible to use some simple "synthetic" methods to achieve a breakdown of the results more reliably than can be done only on the basis of purely "direct" estimates from samples of relatively small size. (See Section 7.5 for an introductory discussion of such estimation procedures.)

Some of these points will be elaborated further in the following sections.

Assessment of the prevalence of child labour during listing

In Yemen many children are not at school, and yet they are not reported in surveys as being engaged in any form of economic activity or child labour. This is particularly the case with girls. This situation undoubtedly results in gross under-reporting of child labour.

A simple question, asking for example "whether there are any working children in the household", is not sufficient. At least a small set of questions should be included in the operation in order to classify households into the following types:

1. households reporting one or more working children;
2. household reporting no working children but reporting one or more school-age children who are not at school, i.e. not in full-time education;
3. households reporting all children as being at school (full-time education) and none as being at work;
4. households with no children in the relevant age group.

As will be described later, for the Phase 2 labouring children survey (plus the LFS module), groups (1) and (2) – termed "category C" households – will need to be sampled in a different way from groups (3) and (4), termed – "category E" households.

The minimum set of questions to identify the above categories could be something like the following:

- How many children (in the relevant age group) are there in the household?
- How many of them are at school in full-time education, and how many are not at school or in full-time education?
- How many of the second group (not at school) are engaged in work, including unpaid work in the household's economic activity?
- Are any children in the first group (at school) engaged in such work?

The implication of the above is that the listing operation cannot (and must not) be simply cursory; it must yield reliable data on these four questions at least. It is therefore extremely important not to make the operation too large in terms of the number of households to be listed. The operation should be seen more like a "mini interview".

6.3.2 Sampling frame

Basic requirements

Census enumeration areas (EAs) will form the primary sampling units (PSUs). The sampling frame should contain the following information at the minimum:

- List of EAs covering the whole country, with accompanying maps and description.
- Codes identifying the geographical-administrative units (governorate, sub-governorate, etc) to which the EA belongs, and whether it is classified as urban, rural or nomadic.
- The population size and, preferably, also the number of households in the EA.

If available, additional characteristics potentially useful for more efficient stratification may also be included. One particular variable – classification of EAs into broad groups in terms of the average illiteracy rate – should be mentioned as a possibility.

Basic figures

In choosing the sampling design, it is useful to keep in view the numbers of units (EAs, households, persons) involved, and their basic characteristics such as average size. These figures should be broken down by administrative division, urban-rural area, etc.

On the basis of the available frame, Table 6.2 shows the basic figures by governorate and by type of area (urban, rural, nomadic). More detailed tables for smaller administrative divisions are available, and of course are also available for individual EAs in the full frame.

Table 6.1. Size of major domains

	Urban	Rural	Nomadic	Total
% of the total in the country				
Persons	28.2	71.3	0.5	100.0
Hhs	29.1	70.4	0.5	100.0
EAs	25.6	72.4	2.0	100.0
Relative EA size (=1.0 for the whole country)				
Hhs/EA	1.14	0.97	0.24	1.00

First, Table 6.1 summarizes the distribution of the population by type of area. Urban areas account for 28 per cent of the population, but nomadic areas for only 0.5 per cent. Urban EAs are somewhat larger than rural EAs in population (by 17 per cent), but nomadic EAs are much smaller – on average 25 per cent of the size of other EAs.

Table 6.2. Number of persons, households and EA's by urban-rural and Governorate. (From census frame)

	Gover.	Persons	HH's	EA's	persons/HH	HHs/EA
URBAN	11	368,726	51,339	371	7.2	138
	12	109,514	15,566	106	7.0	147
	13	1,654,933	244,693	1637	6.8	149
	14	105,551	13,459	91	7.8	148
	15	528,601	80,349	621	6.6	129
	16	55,232	7,687	44	7.2	175
	17	132,934	17,493	135	7.6	130
	18	748,697	111,903	803	6.7	139
	19	468,542	59,049	428	7.9	138
	20	182,272	24,766	183	7.4	135
	21	73,729	8,670	74	8.5	117
	22	101,605	13,666	91	7.4	150
	23	25,103	3,483	29	7.2	120
	24	567,133	89,695	633	6.3	142
	25	60,150	9,057	65	6.6	139
	26	29,700	3,760	29	7.9	130
	27	35,119	4,597	35	7.6	131
	28	35,241	5,468	40	6.4	137
	29	146,308	19,117	152	7.7	126
	30	60,081	8,196	62	7.3	132
	31	3,652	556	5	6.6	111

	Gover.	Persons	HH's	EA's	persons/HH	HHs/EA
RURAL	11	1,752,020	250,331	2,057	7.0	122
	12	311,994	40,979	356	7.6	115
	13	39,994	4,647	41	8.6	113
	14	472,696	53,205	441	8.9	121
	15	1,860,061	284,078	2,286	6.5	124
	16	365,511	48,300	290	7.6	167
	17	1,327,485	175,416	1,503	7.6	117
	18	1,397,351	235,482	1,878	5.9	125
	19	533,210	60,892	537	8.8	113
	20	1,136,526	161,502	1,416	7.0	114
	21	383,309	41,216	397	9.3	104
	22	580,991	70,650	608	8.2	116
	23	885,392	112,865	913	7.8	124
	24					
	25	654,953	93,758	756	7.0	124
	26	202,490	23,744	208	8.5	114
	27	457,181	63,951	531	7.1	120
	28	36,835	5,732	67	6.4	86
	29	720,503	86,269	753	8.4	115
	30	411,296	51,120	409	8.0	125
	31	385,959	54,780	480	7.0	114
TOTAL URBAN		5,492,823	792,569	5,634	6.9	141
TOTAL RURAL		13,915,757	1,918,917	15,927	7.3	120
TOTAL NOMADIC		96,492	13,397	443	7.2	30
TOTAL COUNTRY		19,505,072	2,724,883	22,004	7.2	124

The variation in the size of EAs is an important factor determining the appropriate procedure for selecting a sample of those units. Table 6.3 gives some indicators of the variability in EA sizes, first among averages for governorates and then among averages for smaller administrative divisions. In the latter case, a large part of the variability is accounted for by 20 or so sub-governorates at each end of the distribution.

The variability in population size among individual EAs is likely to be considerably greater than that indicated by the above figures averaged over administrative divisions. This can be examined from the full frame providing listings at the EA level.

Table 6.3. Mean number of household per EA: variation by Governorate			
	Urban	Rural	Nomadic
Number of Governorates	21	20	15
Mean number of hhs per EA			
minimum value	111	86	1
maximum value	175	167	69
mean ove- governorates	136	119	25
median value	137	116	22
coefficient of variation (cv)	0.10	0.12	0.74
Mean number of household per EA: variation by sub-Governorate			
	Urban	Rural	Nomadic
Number of sub-governorates			
Mean number of hhs per EA	239	309	88
minimum value	51	12	1
maximum value	298	311	151
mean over sub-governorates	134	119	30
median value	134	120	21
coefficient of variation (cv)	0.22	0.22	0.92
Above indicators, leaving out 20 values at the top and 20 at the bottom of the distribution			
Mean number of hhs per EA			
minimum value	100	88	
maximum value	164	137	
mean over sub-governotaes	133	119	
median value	134	120	
coefficient of variation (cv)	0.12	0.08	

6.3.3 Sample size and allocation

In a national survey, the most important practical consideration is the total size of the sample at the country level – irrespective of the requirements of reporting for sub-national domains.

Listings, and consequently also the basic CLS questions, need to cover all households in each selected EA – sub-sampling of households within an EA is not an option in a listing operation[35]. Very large EAs may first be segmented before listing, and the listing operation confined to a sample of one or more segments in the EA. However, this operation can be expensive and difficult, though nevertheless necessary some times.

As a consequence of taking all households in each selected area into the sample, the number of households in the sample is directly determined by the number of EAs selected and the average EA size.

Table 6.4 shows the recommended maximum sample size as 100,000 households to be canvassed in Phase 1. For the specific allocation assumed there, the corresponding number of EAs in the sample will be around 870. It may be desirable to make the

[35] Sub-sampling is of course possible for the CLS if the CLS is operationally separated from the household listing operation. This can be meaningful for a more elaborate survey than the present CLS involving only a handful of questions for screening.

sample size somewhat smaller still –perhaps by 20 per cent - given the need for emphasis on data quality noted earlier.

Reporting domains

A total of 41 reporting domains have been indicated: urban areas of each of the 21 governorates, and rural areas in the 20 governorates which have such areas.

A little refinement of this specification is necessary.

Firstly, all nomadic areas together should be seen as forming an important reporting domain in their own right. (The breakdown of nomadic areas by governorate is not possible because of their small size.)

Secondly, some domains may be so small that a sufficient sample size cannot be reasonably allocated to them for separate estimation. This applies specifically to the urban sector of governorate 31 (see Table 6.2 above). While very small groups like this cannot realistically form domains for separate reporting, it is generally convenient and efficient to retain them as separate strata for the purpose of sample selection.

Sample allocation

For similar levels of precision different domains, irrespective of their size, require *equal* sample sizes. By contrast, for maximum precision of estimates at the national level, the required allocation is close to *proportionate*. Neither of these extremes is appropriate when the objective is to produce both domain-level and national-level estimates.

A common and convenient compromise is to allocate the sample as proportional to the square-root of the population size, i.e. as

$$n_i = n_0 \cdot M_i^{1/2}$$

where $M_i = (P_i/\overline{P})$ is the relative population size of domain i, i.e. its population divided by the average population per domain; n_0 is a constant determined so as to obtain the target total sample size n, that is

$$n_0 = n/\Sigma M_i^{1/2}.$$

In order to ensure a minimum sample size for every domain, including the smallest, the above may be generalized as follows:

$$n_i = n_0[k^2 + (1-k^2)M_i^{\alpha}]^{1/2}$$

Here k is a factor determined by the relative importance given to the domain-level estimation as opposed to the national-level estimation. It determines the minimum sample size given to any domain, which is $n_{min} = n_0 \cdot k$ for a domain with M_i close to zero.

Parameter α determines how big the sample for large domains is allowed to become. A small value of α places a more strict limit on the maximum sample allocated; that limit is raised with increasing α.

Table 6.4 shows the allocation by domain by taking the following values for the parameters:

k=0.5, which corresponds to giving equal importance to the two objectives of domain-level and national-level estimation

α=1.0 which corresponds to the widely used square-root allocation, except for the minimum allocation guaranteed by k

n=100,000 households, which is the maximum recommended sample size in the present case. This determines the required constant n_0 in the allocation formula.[36]

The various domains differ to some extent in terms of average household size and the number of households or persons per EA. To obviate the effect of this variation, in Table 6.4 the allocation has been worked out consistently in terms of number of persons. The following steps clarify the procedure applied.

1. For each domain i, the relative measure of size is defined in terms of the number of *persons* in it

   $M_i = P_i/\overline{P}$.

2. Relative sample size allocation p_i in terms of numbers of persons is determined by taking an arbitrary value (such as 1000) for the scaling factor n_0, and k=0.5, α=1.

3. The sampling rate $f_i = p_i/P_i$ thus obtained is then applied to the number of households, H_i, in the domain to obtain the expected number of households, h_i, in the sample (corresponding to the arbitrary value of the scaling factor n_0), $h_i = f_i \cdot H_i$.

4. The sampling rates f_i are scaled uniformly so that the target sample size, h_{target}=100,000 households, is obtained:

   $f_i' = f_i \cdot h_{target}/\Sigma h_i$.

5. Finally, the number a_i to be selected from A_i EAs in the domain is

   $a_i = f_i' \cdot A_i$.

6.3.4 Stratification and selection of areas

For the selection of areas, stratification will be applied at various levels.

1. Stratification by type of area (urban, rural, nomadic).
2. Within those, by governorate.
3. Within governorates, by the next administrative level.
4. By EA population size group.

[36] Some adjustment has been made in Table 6.4 in the case of a couple of domains. Domain 31-urban is so small that the above allocation would mean a full census. The allocation for the nomad domain has been increased by 50 per cent because of its special relevance.

5. Possibly followed by arranging EAs within the strata defined above by population size and then applying systematic sampling.

Is it possible and desirable to introduce additional stratification?

One possibility which has been considered is stratification of areas according to the level of illiteracy in the area. If used, this can be introduced in place of step 4 above, possibly followed by 5, ordering by EA population size and systematic selection.

However, this option of stratification by literacy level may not be possible because of the lack of data for individual EAs. Also, it is not certain that it will be an effective stratification variable. Literacy may be a factor working primarily at the individual rather than the area level. In any case, retaining good control over the variation in EA population size is important in sampling.

Within each stratum, EAs can be selected with the required uniform probabilities using the standard circular systematic procedure. (See Section 6.7 for a numerical illustration of the procedure.)

Table 6.4. Illustration of sample allocation and sampling rates for the selection of EAs by urban-rural and by governorate (target sample size 100,000 households)

		Number in population			Sampling rate	Expected	EAs
	Governorate	Persons	Hhs	EAs	(f %)	n (hhs)	selected
URBAN	11	368,726	51,339	371	4.65	2,388	17
	12	109,514	15,566	106	11.13	1,732	12
	13	1,654,933	244,693	1637	1.93	4,715	32
	14	105,551	13,459	91	11.46	1,542	10
	15	528,601	80,349	621	3.71	2,979	23
	16	55,232	7,687	44	19.67	1,512	9
	17	132,934	17,493	135	9.57	1,673	13
	18	748,697	111,903	803	3.01	3,368	24
	19	468,542	59,049	428	3.99	2,359	17
	20	182,272	24,766	183	7.55	1,870	14
	21	73,729	8,670	74	15.37	1,333	11
	22	101,605	13,666	91	11.81	1,615	11
	23	25,103	3,483	29	37.42	1,303	11
	24	567,133	89,695	633	3.55	3,186	22
	25	60,150	9,057	65	18.27	1,655	12
	26	29,700	3,760	29	34.28	1,289	10
	27	35,119	4,597	35	29.42	1,352	10
	28	35,241	5,468	40	29.32	1,603	12
	29	146,308	19,117	152	8.89	1,700	14
	30	60,081	8,196	62	18.29	1,499	11
	31	3,652	556	5	37.42	208	2

		Number in population			Sampling rate	Expected	EAs
	Governorate	Persons	Hhs	EAs	(f %)	n (hhs)	selected
RURAL	11	1,752,020	250,331	2057	1.87	4,677	38
	12	311,994	40,979	356	5.19	2,127	18
	13	39,994	4,647	41	26.16	1,216	11
	14	472,696	53,205	441	3.97	2,114	18
	15	1,860,061	284,078	2286	1.81	5,139	41
	16	365,511	48,300	290	4.68	2,260	14
	17	1,327,485	175,416	1503	2.17	3,814	33
	18	1,397,351	235,482	1878	2.11	4,977	40
	19	533,210	60,892	537	3.69	2,246	20
	20	1,136,526	161,502	1416	2.37	3,828	34
	21	383,309	41,216	397	4.54	1,870	18
	22	580,991	70,650	608	3.50	2,473	21
	23	885,392	112,865	913	2.73	3,082	25
	24						
	25	654,953	93,758	756	3.26	3,054	25
	26	202,490	23,744	208	7.00	1,662	15
	27	457,181	63,951	531	4.06	2,594	22
	28	36,835	5,732	67	28.17	1,615	19
	29	720,503	86,269	753	3.08	2,656	23
	30	411,296	51,120	409	4.34	2,217	18
	31	385,959	54,780	480	4.52	2,474	22
NOMAD	total	96,492	13,397	443	22.59	3,026	**100**
TOTAL		19,505,072	2,724,883	22,004	3.67	100,000	869

Allocation with parameters (see text) k=0.5, alpha=1.

Sizes and sampling rates of major domains

	Number in population			Expected	EAs	Average
	Persons	Hhs	EAs	n (hhs)	selected	sampling rate (%)
URBAN	5,492,823	792,569	5,634	40,882	297	5.158
RURAL	13,915,757	1,918,917	15,927	56,092	472	2.923
NOMADIC	96,492	13,397	443	3,026	100	22.587
TOTAL	19,505,072	2,724,883	22,004	100,000	869	3.670

Relative sizes and sampling rates

	Number in population			Expected	EAs	Average
	Persons	Hhs	EAs	n (hhs)	selected	sampling rate
URBAN	28.2	29.1	25.6	40.9	34.1	1.41
RURAL	71.3	70.4	72.4	56.1	54.3	0.80
NOMADIC	0.5	0.5	2.0	3.0	11.5	6.15
TOTAL	100.0	100.0	100.0	100.0	100.0	1.00

Sample weights to be applied at the estimation stage will be inversely proportional to the sampling rate f_i' defined above in Table 6.4. A small correction may however be noted. This arises from the fact that the number of EAs to be selected in any stratum has to be fixed as a whole number. Rounding to the nearest whole number involves a small adjustment to the sampling rate.

Let us assume that f_i is the sampling rate determined above for domain i, and that j indicates a stratum within domain i. Additional levels of stratification, such as 3 and 4 above, mean that a domain may contain several separate strata.

From A_{ij} EAs in stratum j of domain i, the number to be selected is

$a_{ij}=f_i \cdot A_{ij}$.

Let a'_{ij} be the above quantity rounded to the nearest integer, that being the actual number to be selected. The effective sampling rate in the stratum is therefore

$f_{ij}=a'_{ij}/A_{ij}=f_i\ (a'_{ij}/a_{ij})$.

The "design weights" are the inverse of this f_{ij}, being uniform for all selected EAs and households in the stratum.

In the discussion on Phase 2 in the next section, we will refer to these sampling rates as the Phase 1 sampling probabilities, denoted as $f1_i$, for *sample area* (EA) i. (explicit identification of stratum and domain is not needed for that discussion).

6.4 Sampling for the labouring children survey (LCS)

The determination of the sample design involves the following major steps;

1. Specification or clarification of the objectives and main characteristics of the design. This is linked to the substantive objectives of the survey.
2. Preparation of the sampling frame, including stratification, assigning measures of size for the selection with PPS if required, and estimating parameters from the frame as needed for sample design and selection. This primarily concerns the area sampling stage(s).
3. Choice of the sample size in terms of the ultimate sampling units, e.g. households.
4. Specification of the sampling stages and of the number of clusters (ultimate area units) to be selected. Given 3, this also determines the average number of ultimate units (households) per cluster.
5. Description of the method of selecting area units, in particular the sampling probabilities.
6. Description of the method of selecting ultimate units (households) within sample clusters, in particular the sampling rates and the final sampling probabilities.
7. Specification of the weighting and estimation procedure, including variance estimation.

These are discussed below in turn.

6.4.1 Objective and main characteristics of the design

Objectives

Five main objectives of the planned survey were noted earlier. Objective 1, namely, estimation of the prevalence of child labour by urban-rural domain of individual governorates, is the concern of Phase 1 (Listing and CLS) described above. The objectives of Phase 2 (LCS, with LFS "module") were stated as the following.

A detailed survey of labouring children in households identified as containing such children during Phase 1.

Identification and detailed survey of labouring children in other households, even though not identified as containing such children during Phase 1. Labouring children are expected to exist primarily (and in large numbers) in households containing children of school age but not at school. Some working children – but probably a much smaller number – may also exist in other households.

Obtaining information on a "control group" of non-working children as needed for a deeper analysis of the causes, characteristics, circumstances and consequences of child labour.

Using the opportunity of the survey also to collect data on the labour force characteristics of the entire working-age population (the LFS "module").

Given the context of the whole survey, objective 2 and 3 are the primary ones, while objective 4 and 5, even though important, should be seen as secondary. These priorities must be reflected in the sample design as well. In other words, the design is to be determined mainly to satisfy 2 and 3, but to be modified to accommodate 4 and 5 to the extent possible. Among the latter, it is 5 which mainly determines the sampling requirements.

Main characteristics

From a substantive point of view, therefore, Phase 2 may be seen as consisting of two components: the "LCS" part concerning objective 2 and 3; and the "LFS" part concerning mainly objective 5.

This conceptual distinction does not imply any operational separation: the two components will take the form of a unified interview in the same sample areas and households, with the applicable sections of the questionnaire(s) being used in each household. This distinction is fundamental from the standpoint of the sample design. The primary determinant of the sampling design is the LCS part, and this will be discussed first in the following sections.

Interviewing labouring children (and/or their guardians) is an intensive and time-consuming operation. More detailed questioning will be required to discover labouring children not easily identified, or missed altogether during the Phase 1 listing and CLS. The emphasis therefore has to be on data quality and control of non-sampling errors, rather than on high sampling precision and a large sample size. Furthermore, the sample design has to be as efficient as possible so as to minimize the necessary sample sizes for a given level of precision.

Secondly, for the same reasons as above, it is necessary at this stage to focus on national level estimates. Little is known in Yemen about the conditions of labouring children. Some of the more important indicators may be reported for at most 3-5 major domains, for instance by urban and rural area and by gender, or by major geographical division of the country. A more detailed breakdown should wait till the basic figures become available with sufficient reliability.

Thirdly, the sample should be focused on areas (EAs) with a high prevalence of child labour, as indicated by the results of Phase 1. Otherwise, the sample will be too scattered over areas with very little or no child labour. The number of sample areas in Phase 2 should be much smaller than the number in Phase 1. This is necessary to ensure quality control.

6.4.2 Sampling frame

The sample EAs and listings from Phase 1 form the sampling frame for Phase 2. For the sake of convenience, we shall here refer to the former as the "population" and to the latter as the "sample".

There should be a time lag between the two phases so that a frame can be compiled for the selection of the Phase 2 sample. The frame should contain at least the following information.

1. List of EAs, with accompanying maps and description for their location and identification.
2. Phase I (first sampling stage) selection probabilities, $f1_i$ for EA i. These vary by domain and stratum as already discussed.
3. List of households in the EAs, with information for each household on which of the two categories it belongs to (i.e. belonged to at the time of the Phase 1 listings):

   Category "C": the household reports one or more working children,[37] or reports no working children but one or more children not at school (in full-time education): or

   Category "E": the household reports that all children are at school (in full- time education) and none are working; or that there are no children in the specified age bracket in the household.[38]
4. Based on the above, the computation of the following parameters:

   c_i number of households in category C in EA i

   e_i number of households in category E in EA i

   c_m, e_m respective means of the above over all areas in the "population" (i.e. in the Phase 1 sample).

[37] This refers to "working children" as defined for the survey with eecified age limits. For sampling, these definitions and their implication have to be taken as given.

[38] Symbol "C" is used here to indicate households which actually contain, or have a high probability of containing, working children. These households are the main providers of sample cases for the labouring children component of the survey. Symbol "E" indicates other households which do not contain, or have a low probability of containing, working children as reported during the listing operation. These households are primarily for the labour force (employment) module.

5. The computation of a "measure of size" M_i for each EA i as follows

$$M_i = g1*(c_i/c_m)+(1-g1)*(e_i/e_m)$$

where g1 is a constant parameter determined by the relative importance to be given, in the selection of areas, to the number of households containing or likely to contain working children, compared to other households. As explained below, such a measure of size can be used for PPS selection so as to focus the Phase 2 sample on areas with high concentrations of child labour.

It appears reasonable to recommend the value of $g1=0.8$[39] for the particular case being considered. Values appreciably larger than 0.5 would be appropriate in most situations.

6.4.3 Sample size

As emphasized above, the sample size for Phase 2 should be modest, such as indicated below, with appropriate additions to allow for non-response.

Households in category C	$n_c=6,000$
Households in category E	$n_e=4,000$
Total achieved sample size	$n=n_c+n_e=10,000$ households.

These two categories of households will of course come from the same sample of EAs. The rationale for this recommendation on sample size is as follows.

Most of the households in category C will contain one or more working or potentially working children, so that the number of such children in the survey would be of the order of 6,000, and probably more. This will permit a sample size of the order of 3,000 children each for urban and rural areas, which are the major reporting domains. With four reporting domains, by urban and rural area and by gender, or by main geographic region, a sample size of the order of 1,500 children per domain should be obtained. On the basis of experience, such sizes are sufficient for quite elaborate analysis.

For the LFS module, both parts of the sample are relevant. The total available sample size will be around 10,000 households. Given that there are on average two economically active persons per household in the country, around 20,000 such persons can be expected in the sample. This should permit their breakdown by gender, main age group and urban and rural area.

The choice of n_e can be made almost independently of the choice of n_c being discussed here. It is a matter of how much of the budget and effort can be devoted to the LFS component in a survey primarily targeted at child labour.

[39] Such a recommendation is justified on the basis of pra›ical judgement and past experience. With actual data, it should be possible to adjust this value so as to obtain a sample with the desired characteristics.

6.4.4 Sampling stages and number of clusters

The Phase 2 sample is to be selected in two stages from Phase 1: the selection of areas, followed by the selection of households. This does not preclude "compact cluster" sampling in certain cases, i.e. taking into the sample all the households in the cluster. All persons in selected households will be included in the survey.

How many clusters to select is a very important question, as it determines survey costs, sampling precision and data quality.

In the case of an LCS it is also important to ensure that a large enough number of clusters is selected so as to meet the sample size target of working children. If the prevalence of child labour is low, a large number of clusters will have to be selected for this purpose. In Yemen, however, child labour appears to be very prevalent, so that a "modest" number of clusters may suffice.

The procedure for determining the number of clusters, a, to be selected to meet the target sample size n_c is as follows.

1. We shall take it that the selection of EAs from Phase 1 to Phase 2 will be with PPS, probability proportional to the measure of size M_i as defined earlier.

2. We shall assume first that, in each area selected, all category C households are taken into the sample. With this procedure we obtain the minimum possible number of clusters, a_{min}, that can meet the sample size target. This minimum value depends on the mean number of category C households per EA in the sample, $c_{m(s)}$:

   $$a_{min}=n_c/c_{m(s)}.$$

3. The average per cluster in the Phase 2 sample is not the same as the known average c_m per cluster in the population (i.e. in the Phase 1 sample). The former is larger than the latter in accordance with the variability of the size measures M_i used for selecting the areas.

   Since the Phase 2 sample has not been selected at this stage, it cannot be used to compute the mean $c_{m(s)}$. However, we can compute its *expected value* from the information in the frame

   $$c_{m(s)} =\Sigma M_i \cdot c_i /\Sigma M_i ,$$

   where the sum is over all clusters in the frame (i.e. Phase 1 sample).

   Note that the average over clusters in the population is, by contrast, the unweighted mean

   $$c_m=\Sigma c_i/\Sigma 1.$$

4. Some clusters may contain too many households with working children and it is generally advisable to put a limit, say c_{max}, on the maximum number of category C households which will be selected in any EA. Putting such an upper limit will in general increase the number of clusters, a, required to obtain the target sample size

n_c. This number is smallest when the upper limit noted above equals or exceeds the largest of c_i values in the population.

The important point is that, for a given target sample size n_c, there is a relationship between the quantities *a* and c_{max}; they cannot be chosen independently. The actual relationship depends on the empirical distribution of the c_i values in the sample.

The following procedure is recommended. Take different values of c_{max} and compute the implied number of clusters, *a*. Then, on the basis of practical considerations and what is found in the actual frame, make a choice of parameter *a* and obtain the c_{max} implied by this choice.

This can be done from the frame, without actually selecting the Phase 2 sample. The algorithm is as follows.

With a chosen value of c_{max}, say c'_{max}, compute for each EA i

$$c'_i=\min(c_i,c'_{max}).$$

The expected value of the mean number of category C households in the sample is

$$c'_{m(s)}=\Sigma M_i \cdot c'_i/\Sigma M_i.$$

From the practical point of view, it would be desirable to keep the number of clusters in the sample from becoming too large. Obviously, we need to choose $a \geq a_{min}$ defined in procedure 2 above.

Tentatively, a value such as a=400 could be considered, if it is found sufficient to yield the required sample size for category C households.

Then the implied c_{max} corresponding to the chosen value of *a* can be computed as described above and applied later at the household selection stage. In the present illustration, with n_c=6,000 and a=400, c_{max} may turn out to be of the order of 20-30 households.

6.4.5 Selection of area units

The sample for Phase 2 is obtained from the areas in Phase 1 by selecting units with probabilities proportional to the measure of size M_i defined above. The standard circular systematic sampling procedure can be used for this purpose.

With *a* as the number of clusters to be selected, the sampling interval to be applied to cumulated size measures is

$$I=\Sigma M_i/a,$$

where the sum is over all clusters i in the frame. The probability of selection of an area is

$$f2_i=a\cdot(M_i/\Sigma M_i)=M_i/I.$$

This is the conditional selection probability from Phase 1 to Phase 2. With Phase 1 probability for the area as $f1_i$, the total probability of an area appearing in Phase 2 sample is

$$f1_i * f2_i.$$

6.4.6 Selection of households

As explained above, up to a certain upper limit c_{max} all category C households in the selected cluster will be included in the sample. The probability of selection of a C category household in an area with $c_i \leq c_{max}$ is the same as that of its EA.

In clusters where the number found of C category households, c_i, exceeds c_{max}, only c_{max} will be included in the sample. Such selection of households can be done using the standard equal probability circular systematic sampling procedure. The last-stage sampling rate is c_{max}/c_i.

In either case, using the definition $c'_i = \min(c_i, c_{max})$,

the overall selection probability of a C category household in area i is

$$f_i = f1_i * f2_i * (c'_i / c_i).$$

6.4.7 Weighting and estimation

Weighting and estimation procedures, as well as practical procedures for variance estimation, are discussed in detail in Chapter 7. A few salient points are noted here.

As in any sample, it is necessary to ensure that for each sample case a weight has been appropriately computed and recorded in the micro data. The weights are inversely proportional to selection probabilities f_i, adjusted for non-response, calibration, etc. as required. For producing "primary" statistics such as means and ratios, no further reference is required to the sample structure such as the stratum and cluster to which the ultimate unit belongs.

Practical procedures are available for variance estimation for surveys of the size and type being discussed. For the application of these procedures, the essential sample structure information that should by available, and ideally coded in the micro data, includes specification of the following for each ultimate unit: (i) its estimation domain; (ii) sampling stratum; (iii) PSU; (iv) order of selection of the PSU if selected systematically; and (v) sample weight.

One important objective of the Phase 2 survey is to improve the Phase 1 estimates of the prevalence of child labour by providing correction factors for its likely under-estimation.

The results of Phase 1 are subject to the error of "false positives" (child labour reported when none exists in the household), but even more to that of "false negatives" (failure to report existing child labour). Phase 2 can be used to estimate these error rates separately for different classes of children, and possibly also for a few major domains. These factors can then be applied to Phase 1 results to produce improved estimates. For smaller reporting domains, adjustment factors estimated for the major domain to which they belong may be used.

6.5 Sampling for the LFS module

This module is simply a part of an integrated interview and is applied in the same areas as above.

To meet the sample size requirements of the LFS module in the survey, additional households are selected from the category E households defined earlier (Section 6.2.2).[40]

6.5.1 Sample size

The main parameter to be chosen is the sample size n_e for category E households. This is determined primarily by the requirements of the LFS module. In principle, this choice can be made independently of the choice of n_c. Normally, most EAs will contain sufficient numbers of category E households to yield the required sample number of cases.

However, moderation is desirable in the choice of this sample size. Note that for the LFS module, both the parts of the sample are used, giving a total sample of n_c+n_e households, and almost twice as many economically active persons.

6.5.2 Household selection probabilities

As detailed in Section 6.4, The sample areas have been selected with probabilities $f2_i$ proportional to the measure of size M_i which is heavily weighted in favour of areas with a high prevalence of child labour (as reflected by c_i). This procedure is appropriate for category C households from which most of the LCS interviews will come.

The over-sampling of areas with a higher concentration of child labour is retained at the household level for category C households by using "take-all" sampling in all clusters with $c_i \leq c_{max}$ defined earlier. (Section 6.4).

Such over-sampling of households with a high level of child labour is, however, not efficient for the LFS component. It is obviated by selecting households within selected clusters with probabilities inversely proportional to $f2_i$, thus cancelling out the over-sampling at the previous stage.

Let us now consider a sampling rate

$$f3_i = g2/f2_i$$

as applied in the selection of category E households within Phase 2 sample areas. Constant g2 is determined by the sample size requirements. The expected value of the category E sample size can be estimated prior to actually selecting the Phase 2 sample areas as follows.

[40] The two parts of the sample, from categories C and E respectively in each area, are selected separately and possibly using different procedures, but they can then be put together to form a single sample to be treated uniformly at the implementation stage. The two parts are selected at different rates within each sample cluster, since they are expected to differ greatly in the prevalence of child labour.

The total expected sample size is

$$n_e=\Sigma e_i \cdot f2_i \cdot f3_i=\Sigma e_i \cdot f2_i \cdot (g2/f2_i)=g2 \cdot \Sigma e_i$$

where the sum is over all areas in the frame (Phase 1 sample). The above gives the required constant g2 as

$$g2=n_e/\Sigma e_i.$$

The overall final selection probability of a category E household in area i is

$$f_i=f1_i \cdot f2_i \cdot f3_i=g2 \cdot f1_i.$$

In other words, the uniform selection probabilities of areas within domains of the Phase 1 sample are restored, in relative terms, in the final sample of category E households.

6.5.3 Fine-tuning

It is usually desirable for practical reasons to put upper and lower limits on the number of category E households to be selected from any EA.

A lower limit, e_{min}, may be considered desirable in order to avoid having to go into the area for only one or two interviews. This is done by taking the number of households to be selected as

$$t_i=\max (e_{min}, e_i \cdot f3_i),$$

or more precisely as

$$t_i=\min [\max (e_{min}, e_i \cdot f3_i), e_i].$$

Similarly, we can place a limit, e_{max}, on the maximum number of category E households to be taken into the sample in any area.

Note that the expected impact of such adjustments can be estimated from the frame prior to selecting the Phase 2 sample by examining the distribution of e_i values in the frame after weighting their distribution by M_i.

Obviously, the final-stage selection probabilities in the presence of such adjustments become

$$f3_i=t_i/e_i.$$

The appropriate choice of the limits e_{min} and e_{max} is a matter of practical judgement in the light of empirical information on the extent to which cases requiring such adjustment exist.

Reasonable values in the present case may perhaps be something like

$$e_{min}=3,\ e_{max}=20 \text{ to } 30.$$

6.6 Numerical illustration of the sampling procedures

Let us consider one hypothetical domain in the country for which separate estimates are required on various characteristics of child labour. The domain consists of 100,000 households in 1,000 census EAs, each of exactly 100 households. In reality, EAs invariably vary in size but the above assumption is made in the illustration for the sake of simplicity.

Phase I of the survey operation consists of the selection of a sample of EAs and the complete listing of households in each selected EA, and also the identification of households which:

- contain one or more children in the specified age group engaged in child labour, or
- contain no children reported as working, but do contain one or more such children not at school.

These are our category C households. The remainder in the EA have been termed category E households.

The total national Phase 1 sample has been allocated among the domains in some appropriate manner, and we will assume that our domain receives a sample size of 100 EAs, i.e. 10 per cent of its total, containing 10,000 households to be listed by means of a "mini interview" to identify category C households.

6.6.1 Selection of EAs for Phase 1

The process involves selecting 100 out of 1,000 EAs in the domain with equal probability, i.e. with a constant

$$f1=0.1.$$

The EAs in this frame are arranged by urban and rural sub-domain, with nomadic EAs forming a separate sub-domain. Urban localities are arranged according to size, and villages into three groups of nearly equal size according to the average level of illiteracy, assumed to be available from the population census. Beyond that, the EAs are arranged according to geographical location within the groups defined above.

The selection of EAs is to be done using "circular systematic sampling with constant probability". Referring back to the three sub-domains – urban, rural, nomadic - we have two slightly different options.

1. The first option is to assign to each sub-domain a sample size proportional to the number of EAs in it, rounded to the nearest integer, and to selecti the number so assigned independently within each sub-domain. This may result in slight differences in the sampling rates f1 among the domains due to the rounding of the numbers to be selected. For instance, if we suppose that the urban sub-domain contains 233 EAs, giving its sample allocation as 23.3 EAs, which is rounded to 23 to be selected, then the f1 for the domain is reduced by the factor 23/23.3.

2. The second option is to place one sub-domain after another in a combined list and to select a sample of 100 EAs in a single systematic operation. The sampling rate f1 remains uniform throughout, but the actual number selected from a particular sub-domain may vary slightly.

The actual procedure used should be recorded as it is relevant to the variance computation procedures (see Chapter 7). Even more important is to keep a record of the selection order of EAs when systematic sampling has been used.

Both options have their merits, but let us assume here for the sake of simplicity that procedure 2 has been followed. The selection procedure involves the following steps.

1. The EAs ordered in the list as described above are numbered sequentially, in this case from 1 to 1,000.
2. A three-digit random number is selected in the range 000-999. If we assume this number to be 086, then the EA with sequence number R=86 is the first one to be selected.
3. The sampling interval I is computed as the ratio of the number of EAs in the frame to the number to be selected, in this case I=1,000/100=10. Here this interval happens to be a whole number, but this is not always the case – see step 7 below.
4. The next EA selected is the one with sequence number equal to R+I=96, the one after that equal to 96+10=106, and so on.
5. When the number to be selected exceeds 1,000 for the first time, reduce it by 1,000 and use the result to select the next unit, starting from the beginning of the list. In this example, when we reach the number 1006, unit number 6 is selected, and thereafter number 16, etc.[41]
6. Continue the process till the required number (here 100) of units are selected.
7. Neither the selection interval I, nor the initial random number R, needs to be integers. If any of these is non-integral, use exactly the same procedure as above to construct the sequence R, R+I, R+2·I, …, and each time round up the result to specify the sequence number of the unit to be selected next.

6.6.2 Frame for Phase 2 sampling

The sample for Phase 1 forms the "population" for Phase 2. The results of Phase 1 provide the corresponding sampling frame for the selection of the Phase 2 sample. This frame is hierarchical, with data at two levels:

1. A list of all EAs in the Phase 1 sample (with accompanying maps and description, etc.), with information for each EA i on c_i, the number of category C households in it, and on e_i, the number of remaining, category E households.

[41] Note that if original random number R is 000, it is treated as 1000, so the first unit to be selected is the one at the very end of the list. Then we continue from the beginning of the list – the second unit selected being R+I=0+10=10, i.e. unit number 10.

2. For each EA, a listing of all households in it (including address, location map, description, etc.), with information for each household on whether it belongs to category C or to category E.

Actually, once the sample of Phase 2 EAs has been selected, the above lists of households are used only for those selected areas to draw samples of households, as far as the LCS-LFS survey is concerned. However, the full lists may be useful for other surveys.

Table 6.7 later in this chapter provides an example of the Phase 2 frame at EA level, synthetically constructed for the present illustration.

It contains 100 EAs, numbered i=1-100. For each EA assumed values are also given of variables c_i and e_i defined above.[42] The EAs have been ordered according to c_i values – which in fact is a good order for the selection of units with probabilities depending on c_i.

Each EA is assigned a measure of size defined as

$$M_i=0.8(c_i/c_m)+0.2(e_i/e_m).$$

Obviously, the average value of M_i so defined is 1.0. The units will be selected with probabilities proportional to this measure of size.

6.6.3 Maximum and minimum numbers of EAs selected

Generally, a two-stage design is appropriate in the type of situation being discussed: the selection of a sub-sample of EAs, followed by the selection of households within each selected EA.

However, Table 6.7 also illustrates the two "extreme" sampling schemes, each involving a selection only at one stage: take-all sampling, and retaining all EAs.

Take-all sampling

This refers to selecting a sample of EAs, but then retaining all C category households in the sample. Clearly, to meet the sample size target $n_c=600$ category C households in our example, the required number of EAs to be selected (with specified probabilities proportional to M_i) is the minimum possible, since each EA selected makes the maximum possible contribution (all its category C households) to the sample. In Table 6.7, column [6] shows the number of such households contributed by an EA *if it were selected*. This is simply equal to column [3], the c_i value for the EA.

In order to appreciate the characteristics of samples resulting from such a design, it is not necessary actually to select different samples. We can examine what each EA will contribute to the sample on the average (i.e. over all possible samples which can be drawn). This *expected value* for an EA is given by its actual value (c_i) multiplied by its

[42] These values have been simulated for the illustration using the following procedures.
The units have been assign x_i values starting from 0 up to 0.99 in steps of 0.01, in the order in which they are listed. The required variables are computed as: $c_i=100\cdot x_i^2$, $e_i=100-c_i$.
With this distribution, the mean value c_m is slightly under 100/3. (In fact, there are 32.8 per cent households with working or not-at-school children in the present example.)

probability of selection (M_i).[43] This is shown in column [7]. The average of these expected values is the mean $c_{m(s)}$ we expect in a sample. This equals 51.5 category C households in our illustration, which is considerably larger than the mean over all 100 EAs in the population (32.8). The minimum number of EAs to be selected which can be expected to yield a sample of 600 category C households (if all such households in each area selected are taken) is therefore

$$a=n_c/c_{m(s)}=600/51.5=11.7.$$

Selecting 12 EAs and then taking all category C households in the selected EAs into the sample should suffice.

This procedure of identifying the sample properties from the frame without actually selecting any particular samples is very convenient in working out consequences or estimating parameters for different sample design options. In a sample the contributions of individual units, such as those given in column [6], are realized provided the unit concerned actually happens to be selected. Therefore, their averages or sums, etc. are computed by summing over actually selected unit in a *particular sample*. By contrast, all units in the *population* contribute to quantities of the type shown in column [7]. These are determined by unit values (such as c_i) and the probability (M_i) of their appearing in any sample. Therefore averages or sums, etc. over all units in the population show what these statistics would be for a sample on the average under a given sample design.

In the above example, "take all" sampling applies to category C households. Category E households are more numerous and would generally require sub-sampling – see Sections 6.6.5 and 6.6.6.

Retaining all EAs

This is the other extreme: maximizing the number of EAs in the sample, and minimizing the number of households taken per EA. To obtain 600 category C households from 100 EAs, we need an average of 6.0 per EA selected into the sample. The cut-off point, c_{max}, i.e. the maximum number of households to be taken from any EA, has to be a little higher than 6.0 because some EAs have fewer than this number of category C households to contribute and the others EAs have to make-up for it.

In order to obtain the required cut-off point, we can work with the *expected* contribution of each EA in the sense explained above, but now of course subject to the constraint that it cannot exceed the imposed c_{max} in any sample EA. An EA can contribute only the smaller of its actual value c_i and the limit c_{max}, i.e.

$$c'_i=\min(c_i, c_{max}).$$

[43] Actually, this probability equals c_iM_i multiplied by a constant, $a/\Sigma M_i=a/100$ in our case.

The expected value, as before, is $M_i \cdot c'_i$ which now depends on the c_{max} value chosen. The average of these is required to be 6.0. Starting with $c_{max}=6.0$ say, we can iteratively adjust c_{max} upwards till this average is achieved.[44]

In our illustration, this value comes out as 6.4. Column [8] shows the actual number contributed by an EA when it happens to appear in the sample, with the estimated cut-off applied.

6.6.4 Two stage sampling from Phase 1 to Phase 2

Table 6.8 shows examples of selecting the sample in two stages: first a sample of EAs, and then a sample of households within each EA.

It is assumed that within each selected EA all category C households are kept in the sample up to a certain limit c_{max}. If the number of available category C households in the area (c_i) exceeds that limit, only c_{max} are retained in the sample.

In the first example in Table 6.8, we take $c_{max}=30$ as given, and work out the number of EAs which need to be selected to obtain, on the average, a sample size of n_c category C households. Thus an area in the sample contributes the smaller of the two values: c_i and c_{max}.

We simply work out the expected contribution of each area in the population (which isproportional to column [6]). From these, the required number of clusters is computed directly as explained in the footnote below. [45]

In the second example, we consider the reverse situation: it has been decided to select a certain number of clusters (a=35), and the objective is to determine the maximum take per cluster c_{max} to achieve the target sample size n_c. This requires an iteration: we adjust values of c_{max} till the result gives a=35 as required. Normally this can be done easily on a spread-sheet. The results are shown in columns [9] and [10] of Table 6.8. For $n_c=600$ and a=35, the required c_{max} equals 19.3. By taking $c_{max}=20$, a sample of a=34 EAs would suffice.

6.6.5 Sampling of category E households

The illustrations assume the following design. The selection of EAs is determined primarily by the requirements of the sample of category C households. Once the EAs are selected, sampling for category E households is performed to meet the required sample size n_e. Within selected areas, the sampling rate for this category of households is inversely proportional to the measure of size M_i with which the area was selected from Phase 1. *Within an area*, households are normally selected with equal probabilities, such as by using constant probability circular systematic sampling.

[44] Of course in practice we have to work with whole numbers. In any case a number such as 6.4 can be imposed in the mean, for instance by taking it as 7 in 0.4 proportion of the cases, and as 6 in the remaining 1-0.4=0.6 proportion of the cases. There is an example in Section 6.7 of a procedure to round a set of small number in such a way that cumulation of errors in the resulting total is avoided.

[45] The actual values of the expected numbers are (a/100) times those shown in that column. This form of presentation is convenient in numerical work since the resulting figures do not depend on *a*, the number of clusters to be selected, which is to be determined. The required sample size n_c divided by the sum of the contributions shown in column [9] gives the required value of (a/100).

This design is applied to the two examples in Table 6.8 in columns [7] and [8] and columns [11] and [12].

Note that, because of the particular design adopted, the number in column [8] has become proportional to the number of E households that the EA contributes on average.[46]

6.6.6 Imposing limits on the number of E households to be selected from an area

Such limits are often required for practical reasons. Table 6.9 works out the sample design parameters in a situation where it has been decided to take at least $e_{min}=3$ category E households into the sample from an EA once it has been selected.

Columns [5]-[8] of Table 6.9 show the results for a design with parameters as specified in the table ($c_{max}=10$, resulting in a sample with required *a* around 65). These are constructed in the same way as in Table 6.8. Columns [9]-[12] work out the results when the limit $e_{min}=3$ has been imposed. This does not alter the sample of clusters in any way, as that has been already determined by the C sample. Imposing a lower limit e_{min} means that areas which would have contributed fewer than that number of E household now contribute more. This increases the resulting total sample size above the required n_e (columns [9]-[10]). By simply reducing proportionately the household sampling rates for E households throughout, the sample size can be reduced to the required size. This is shown in columns [11]-[12]. In our example, the effect of imposing e_{min} was quite large, requiring the final sampling rate for category E households to be reduced by the factor 0.72.

Table 6.10 shows a similar example when there is an upper limit, e_{max}, on the number of E households in any area. In fact, the case shown in the table concerns the upper limit automatically imposed simply by the condition that the maximum number of E households to be selected from an area cannot exceed the number of such households available in the area. The procedure is the same irrespective of the particular reason for the upper limit.

In Table 6.10, columns [7]-[8], apply the usual procedure without considering the e_{max} limit. Columns [9]-[10] rework this with the limit imposed. Some households now contribute fewer households, and this reduces the resulting sample size to below the target (columns [9]-[10]). The solution is to increase the final sampling rate (nearly) proportionately to obtain the required sample size. This is shown in columns [11]-[12].

6.6.7 Summary of the simulations

Table 6.5 shows the main results of the simulations described above: how with various constraints the resulting parameters of the design change. These are simply some key figures from the detailed Tables 6.7-6.11 at the end of this chapter.

A final comment on the e_{max} values. Each design implies an upper limit on the number of category E household selected from any area, even when such a limit is not explicitly

[46] As before, the actual contribution is a/100 times the values shown. Therefore, the sum of the values shown over all EAs in the population gives the required sample size n_c times 100/a. Given n_c, the required parameter *a* can be computed. The procedure is used for the convenience of making the values shown in column [8] independent of the particular value of *a*.

imposed. This simply derives from the fact that in our examples the areas are selected mainly for meeting the sampling requirements of category C households, and any overall adjustments to the E household sampling rates have to be made within the areas already selected to meet the sample size (n_e) requirements.[47]

Table 6.5. Summary characteristics of various illustrations of the sampling scheme

	Number of clusters	cmax	emax	Mean 'C' hhs per EA	Mean 'E' hhs per EA	Total number of 'C' hhs (n_c)	Total number of 'E' hhs (n_e)
Population	100	98.0	100.0	32.8	67.2	3,284	6,717
Take-all sampling	11.7	98.0	100.0	51.5	25.7	600	300
All EAs kept in sample	100	6.4	15.0	6.0	3.0	600	300
cmax=30	23.7	30.0	63.2	25.3	12.6	600	300
EAs in sample, a=35	35.0	19.3	42.9	17.14	8.6	600	300
cmax=10	64.7	10	23.2	9.3	4.6	600	300
cmax=10, emin=3	64.7	10	16.6	9.3	4.6	600	300

6.7 Illustration of sample selection (PPS circular systematic sampling)

Table 6.11 shows the commonly used "circular systematic sampling with probabilities proportional to size" method for selecting area units. We apply this to our frame of 100 EAs. A specified number *a* of EAs will be selected with probability proportional to the measures of size M_i shown in column [4] of the table.

In any systematic sampling, the first step is to arrange the units in the required order so as to maximize the advantages of implicit stratification provided by systematic sampling. In Table 6.11 the areas are arranged according to increasing concentration of households with working children (variable c_i).

In order to demonstrate the use of fractional selection intervals and starting points for systematic selection, let us assume that a sample of a=22 cluster out of a total of 100 has to be selected. The measures of size M_i add up to a total of T=100 in our illustration, the average per EA being 1.0. We assume that individual M_i values are available and used to 2 decimal places.

Column [5] shows the cumulation T_i of the M_i values. With $T_1=M_1$, we have $T_2=T_1+M_2$, $T_3=T_2+M_3$, ... $T_i=T_{i-1}+M_i$, up to $T_{100}=T$.

To start the circular systematic procedure, we select a 4-digit random number R in the range 0000-9999. (There is an implied decimal point before the last 2 digits.)

This defines the first "selection point" $S_1=R$. A particular unit, say j, is identified for selection by this selection point. It is the unit for which the cumulative size measures T_j meet the following condition (there can be only one such unit in the list):

[47] With the measures of size for the selection of EAs as assumed in these illustrations, namely $M_i=0.8\cdot(c_i/c_m)+0.2\cdot(e_i/e_m)$, it can be shown that e_{max} does not exceed 5 (=1/0.2) times c_m, the mean number of E household per area in the population.

$T_{j-1} < R \leq T_j$.

In our example R=64.91, and unit number 84 meets the above conditions.[48] For this unit

$T_{83}=64.74$, $T_{85}=66.51$.

It is very important to keep a record of the *order of selection* of units into the sample, as shown in column [7]. For this unit the order of selection is 1.

The required sampling interval is I=T/a=4.55 in the example.

The second selection point is given by

$S_2=R+I=69.46$.

This identifies the next unit to be selected as unit number 86 with cumulated size measure $T_{86}=70.16 \geq S_2$, while $T_{85}=68.32 < S_2$.

Similarly the third selection point is

$S_3=S_2+I=R+2I=74.00$

and this identifies unit number 89 for selection.

In general, for the k-th selection point

$S_k=S_{(k-1)}+I=R+(k-1)\cdot I$.

When the selection point exceeds the cumulative total T for the first time it is reset by reducing it by T: when ST, reset S>S-T, and continue the process.

In our example

$S_9=S_8+I=96.73+4.55=101.27$

reset $S_9=101.27-100.00=1.27$

which identified unit number 5 to be the next (9^{th}) selection. The process stops after the selection of unit number 81, when the required sample of 22 units has been obtained.

Incidentally, circular systematic sampling with constant probability is just a special instance of the above. In the above example, we can assign numbers 0001-0100 to unit 1, 0101-0200 to unit 2, etc., and 9901-(1)0000 to unit 100. As in the above case, random start R in the range 0000-9999 and sampling interval I=10,000/22 give an equal probability sample of 22 units.

[48] A minor point: if R happens to be "0000", this implies the selection o the *last* unit in the list. Thereafter, we continue the section from the top of the list.

Rounding

In any application the number of units selected must be a whole number. This is not always important, but it can make some difference when the number of units involved is small.

When small fractional numbers in a long sequence have to be rounded, their total can be affected significantly if the numbers are rounded independently of each other. Table 6.6 shows how the total can be controlled better. The numbers in column [1] are first cumulated in column [2]. The *cumulated* numbers are then rounded in column [3]. The rounded numbers are then "de-cumulated" in column [4] to give the results. De-cumulation simply means reversing the cumulation process

$$T_i = T_{i-1} + M_i$$

to

$$M_i = T_i - T_{i-1}.$$

Table 6.6. Example of the procedure for rounding a sequence of small numbers

(1)	(2)	(3)	(4)
Fractional numbers	Cumulation	Rounded cumulation	'De-cumulation'
0.3	0.3	0.0	0.0
0.4	0.6	1.0	1.0
0.5	1.1	1.0	0.0
0.6	1.7	2.0	1.0
0.8	2.6	3.0	1.0
1.0	3.6	4.0	1.0
1.2	4.8	5.0	1.0
1.4	6.2	6.0	1.0
1.7	7.9	8.0	2.0
2.0	9.9	10.0	2.0
2.3	12.1	12.0	2.0
2.6	14.7	15.0	3.0
2.9	17.6	18.0	3.0
3.2	20.8	21.0	3.0
3.6	24.4	24.0	3.0
4.0	28.4	28.0	4.0
4.4	32.8	33.0	5.0
4.8	37.7	38.0	5.0
5.3	42.9	43.0	5.0
5.8	48.7	49.0	6.0
total 48.7			49.0

The 'de-cumulation' gives a series of rounded values of the numbers such that their total is nearly preserved.

Table 6.7. Illustrative frame of EAs for Phase 2 sampling

Frame Parameters g1= 0.8 nc= 600					Two 'extreme' sampling schemes			
					Take-all sampling		All EAs kept in sample	
					No maximum limit cmax		cmax=	6.4
					Number of C hhs selected:		Number of C hhs selected:	
					If EA in sample	Expected value	If EA in sample	Expected value
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
i	xi	ci=100*x2	ei=100-ci	Mi	ci'	ci' * Mi	ci'	ci' * Mi
1	0.00	0.0	100.0	0.30	0.0	0.0	0.0	0.0
2	0.01	0.0	100.0	0.30	0.0	0.0	0.0	0.0
3	0.02	0.0	100.0	0.30	0.0	0.0	0.0	0.0
4	0.03	0.1	99.9	0.30	0.1	0.0	0.1	0.0
5	0.04	0.2	99.8	0.30	0.2	0.0	0.2	0.0
6	0.05	0.3	99.8	0.30	0.3	0.1	0.3	0.1
7	0.06	0.4	99.6	0.31	0.4	0.1	0.4	0.1
8	0.07	0.5	99.5	0.31	0.5	0.2	0.5	0.2
9	0.08	0.6	99.4	0.31	0.6	0.2	0.6	0.2
10	0.09	0.8	99.2	0.32	0.8	0.3	0.8	0.3
11	0.10	1.0	99.0	0.32	1.0	0.3	1.0	0.3
12	0.11	1.2	98.8	0.32	1.2	0.4	1.2	0.4
13	0.12	1.4	98.6	0.33	1.4	0.5	1.4	0.5
14	0.13	1.7	98.3	0.33	1.7	0.6	1.7	0.6
15	0.14	2.0	98.0	0.34	2.0	0.7	2.0	0.7
16	0.15	2.3	97.8	0.35	2.3	0.8	2.3	0.8
17	0.16	2.6	97.4	0.35	2.6	0.9	2.6	0.9
18	0.17	2.9	97.1	0.36	2.9	1.0	2.9	1.0
19	0.18	3.2	96.8	0.37	3.2	1.2	3.2	1.2
20	0.19	3.6	96.4	0.37	3.6	1.4	3.6	1.4
21	0.20	4.0	96.0	0.38	4.0	1.5	4.0	1.5
22	0.21	4.4	95.6	0.39	4.4	1.7	4.4	1.7
23	0.22	4.8	95.2	0.40	4.8	1.9	4.8	1.9
24	0.23	5.3	94.7	0.41	5.3	2.2	5.3	2.2
25	0.24	5.8	94.2	0.42	5.8	2.4	5.8	2.4
26	0.25	6.3	93.8	0.43	6.3	2.7	6.3	2.7
27	0.26	6.8	93.2	0.44	6.8	3.0	6.4	2.8
28	0.27	7.3	92.7	0.45	7.3	3.3	6.4	2.9
29	0.28	7.8	92.2	0.47	7.8	3.6	6.4	3.0
30	0.29	8.4	91.6	0.48	8.4	4.0	6.4	3.0
31	0.30	9.0	91.0	0.49	9.0	4.4	6.4	3.1
32	0.31	9.6	90.4	0.50	9.6	4.8	6.4	3.2
33	0.32	10.2	89.8	0.52	10.2	5.3	6.4	3.3
34	0.33	10.9	89.1	0.53	10.9	5.8	6.4	3.4
35	0.34	11.6	88.4	0.55	11.6	6.3	6.4	3.5
36	0.35	12.3	87.8	0.56	12.3	6.9	6.4	3.6
37	0.36	13.0	87.0	0.57	13.0	7.5	6.4	3.7
38	0.37	13.7	86.3	0.59	13.7	8.1	6.4	3.8

Frame Parameters g1= 0.8 nc= 600					Two 'extreme' sampling schemes			
					Take-all sampling No maximum limit cmax		All EAs kept in sample cmax= 6.4	
					Number of C hhs selected:		Number of C hhs selected:	
					If EA in sample	Expected value	If EA in sample	Expected value
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
i	xi	ci=100*x2	ei=100-ci	Mi	ci'	ci' * Mi	ci'	ci' * Mi
39	0.38	14.4	85.6	0.61	14.4	8.8	6.4	3.9
40	0.39	15.2	84.8	0.62	15.2	9.5	6.4	4.0
41	0.40	16.0	84.0	0.64	16.0	10.2	6.4	4.1
42	0.41	16.8	83.2	0.66	16.8	11.0	6.4	4.2
43	0.42	17.6	82.4	0.68	17.6	11.9	6.4	4.3
44	0.43	18.5	81.5	0.69	18.5	12.8	6.4	4.4
45	0.44	19.4	80.6	0.71	19.4	13.8	6.4	4.5
46	0.45	20.3	79.8	0.73	20.3	14.8	6.4	4.6
47	0.46	21.2	78.8	0.75	21.2	15.9	6.4	4.8
48	0.47	22.1	77.9	0.77	22.1	17.0	6.4	4.9
49	0.48	23.0	77.0	0.79	23.0	18.2	6.4	5.0
50	0.49	24.0	76.0	0.81	24.0	19.5	6.4	5.2
51	0.50	25.0	75.0	0.83	25.0	20.8	6.4	5.3
52	0.51	26.0	74.0	0.85	26.0	22.2	6.4	5.4
53	0.52	27.0	73.0	0.88	27.0	23.7	6.4	5.6
54	0.53	28.1	71.9	0.90	28.1	25.2	6.4	5.7
55	0.54	29.2	70.8	0.92	29.2	26.9	6.4	5.9
56	0.55	30.3	69.8	0.94	30.3	28.6	6.4	6.0
57	0.56	31.4	68.6	0.97	31.4	30.4	6.4	6.2
58	0.57	32.5	67.5	0.99	32.5	32.3	6.4	6.3
59	0.58	33.6	66.4	1.02	33.6	34.2	6.4	6.5
60	0.59	34.8	65.2	1.04	34.8	36.3	6.4	6.6
61	0.60	36.0	64.0	1.07	36.0	38.4	6.4	6.8
62	0.61	37.2	62.8	1.09	37.2	40.7	6.4	7.0
63	0.62	38.4	61.6	1.12	38.4	43.0	6.4	7.1
64	0.63	39.7	60.3	1.15	39.7	45.5	6.4	7.3
65	0.64	41.0	59.0	1.17	41.0	48.1	6.4	7.5
66	0.65	42.3	57.8	1.20	42.3	50.8	6.4	7.6
67	0.66	43.6	56.4	1.23	43.6	53.6	6.4	7.8
68	0.67	44.9	55.1	1.26	44.9	56.5	6.4	8.0
69	0.68	46.2	53.8	1.29	46.2	59.5	6.4	8.2
70	0.69	47.6	52.4	1.32	47.6	62.7	6.4	8.4
71	0.70	49.0	51.0	1.35	49.0	65.9	6.4	8.6
72	0.71	50.4	49.6	1.38	50.4	69.4	6.4	8.7
73	0.72	51.8	48.2	1.41	51.8	72.9	6.4	8.9
74	0.73	53.3	46.7	1.44	53.3	76.6	6.4	9.1
75	0.74	54.8	45.2	1.47	54.8	80.4	6.4	9.3
76	0.75	56.3	43.8	1.50	56.3	84.4	6.4	9.5
77	0.76	57.8	42.2	1.53	57.8	88.5	6.4	9.7

Frame Parameters g1= 0.8 nc= 600					Two 'extreme' sampling schemes			
					Take-all sampling No maximum limit cmax Number of C hhs selected:		All EAs kept in sample cmax= 6.4 Number of C hhs selected:	
					If EA in sample	Expected value	If EA in sample	Expected value
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
i	xi	ci=100*x2	ei=100-ci	Mi	ci'	ci' * Mi	ci'	ci' * Mi
78	0.77	59.3	40.7	1.57	59.3	92.8	6.4	10.0
79	0.78	60.8	39.2	1.60	60.8	97.3	6.4	10.2
80	0.79	62.4	37.6	1.63	62.4	101.9	6.4	10.4
81	0.80	64.0	36.0	1.67	64.0	106.7	6.4	10.6
82	0.81	65.6	34.4	1.70	65.6	111.6	6.4	10.8
83	0.82	67.2	32.8	1.74	67.2	116.7	6.4	11.0
84	0.83	68.9	31.1	1.77	68.9	122.0	6.4	11.3
85	0.84	70.6	29.4	1.81	70.6	127.5	6.4	11.5
86	0.85	72.3	27.8	1.84	72.3	133.2	6.4	11.7
87	0.86	74.0	26.0	1.88	74.0	139.0	6.4	11.9
88	0.87	75.7	24.3	1.92	75.7	145.1	6.4	12.2
89	0.88	77.4	22.6	1.95	77.4	151.3	6.4	12.4
90	0.89	79.2	20.8	1.99	79.2	157.8	6.4	12.7
91	0.90	81.0	19.0	2.03	81.0	164.4	6.4	12.9
92	0.91	82.8	17.2	2.07	82.8	171.3	6.4	13.2
93	0.92	84.6	15.4	2.11	84.6	178.4	6.4	13.4
94	0.93	86.5	13.5	2.15	86.5	185.7	6.4	13.7
95	0.94	88.4	11.6	2.19	88.4	193.3	6.4	13.9
96	0.95	90.3	9.8	2.23	90.3	201.1	6.4	14.2
97	0.96	92.2	7.8	2.27	92.2	209.1	6.4	14.4
98	0.97	94.1	5.9	2.31	94.1	217.4	6.4	14.7
99	0.98	96.0	4.0	2.35	96.0	225.9	6.4	14.9
100	0.99	98.0	2.0	2.39	98.0	234.6	6.4	15.2
	Mean C hhs per EA							
		32.8	67.2	1.00	32.8	51.5		6.0
	Number of EAs selected					11.7		100.0
	Expected sample size (nc)					600.0		600.0

Table 6.8. Examples of two-stage sampling from Phase 1 to Phase 2

Sampling parameters nc= 600 ne= 300				Specified: cmax= 30.0 Result: EAs in sample, a= 23.7				Specified: EAs in sample, a= 35.0 Result: cmax= 19.3			
				Number of C hhs selected:		Number of E hhs selected:		Number of C hhs selected:		Number of E hhs selected:	
				If EA in sample	Expected value	If EA in sample	Expected value	If EA in sample	Expected value	If EA in sample	Expected value
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
i	ci	ei	Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi
1	0.0	100.0	0.3	0.0	0.0	63.2	18.8	0.0	0.0	42.9	12.8
2	0.0	100.0	0.3	0.0	0.0	63.2	18.8	0.0	0.0	42.8	12.8
3	0.0	100.0	0.3	0.0	0.0	63.0	18.8	0.0	0.0	42.7	12.8
4	0.1	99.9	0.3	0.1	0.0	62.8	18.8	0.1	0.0	42.5	12.8
5	0.2	99.8	0.3	0.2	0.0	62.4	18.8	0.2	0.0	42.3	12.7
6	0.3	99.8	0.3	0.3	0.1	61.9	18.8	0.3	0.1	42.0	12.7
7	0.4	99.6	0.3	0.4	0.1	61.4	18.8	0.4	0.1	41.6	12.7
8	0.5	99.5	0.3	0.5	0.2	60.8	18.7	0.5	0.2	41.2	12.7
9	0.6	99.4	0.3	0.6	0.2	60.1	18.7	0.6	0.2	40.7	12.7
10	0.8	99.2	0.3	0.8	0.3	59.3	18.7	0.8	0.3	40.2	12.7
11	1.0	99.0	0.3	1.0	0.3	58.4	18.6	1.0	0.3	39.6	12.6
12	1.2	98.8	0.3	1.2	0.4	57.5	18.6	1.2	0.4	39.0	12.6
13	1.4	98.6	0.3	1.4	0.5	56.5	18.6	1.4	0.5	38.3	12.6
14	1.7	98.3	0.3	1.7	0.6	55.4	18.5	1.7	0.6	37.6	12.5
15	2.0	98.0	0.3	2.0	0.7	54.3	18.5	2.0	0.7	36.8	12.5
16	2.3	97.8	0.3	2.3	0.8	53.2	18.4	2.3	0.8	36.1	12.5
17	2.6	97.4	0.4	2.6	0.9	52.0	18.3	2.6	0.9	35.3	12.4
18	2.9	97.1	0.4	2.9	1.0	50.8	18.3	2.9	1.0	34.5	12.4
19	3.2	96.8	0.4	3.2	1.2	49.6	18.2	3.2	1.2	33.6	12.3
20	3.6	96.4	0.4	3.6	1.4	48.4	18.1	3.6	1.4	32.8	12.3
21	4.0	96.0	0.4	4.0	1.5	47.1	18.1	4.0	1.5	32.0	12.3
22	4.4	95.6	0.4	4.4	1.7	45.9	18.0	4.4	1.7	31.1	12.2
23	4.8	95.2	0.4	4.8	1.9	44.6	17.9	4.8	1.9	30.3	12.1
24	5.3	94.7	0.4	5.3	2.2	43.4	17.8	5.3	2.2	29.4	12.1
25	5.8	94.2	0.4	5.8	2.4	42.1	17.7	5.8	2.4	28.6	12.0
26	6.3	93.8	0.4	6.3	2.7	40.9	17.6	6.3	2.7	27.7	12.0
27	6.8	93.2	0.4	6.8	3.0	39.7	17.6	6.8	3.0	26.9	11.9
28	7.3	92.7	0.5	7.3	3.3	38.5	17.5	7.3	3.3	26.1	11.8
29	7.8	92.2	0.5	7.8	3.6	37.3	17.3	7.8	3.6	25.3	11.8
30	8.4	91.6	0.5	8.4	4.0	36.1	17.2	8.4	4.0	24.5	11.7
31	9.0	91.0	0.5	9.0	4.4	34.9	17.1	9.0	4.4	23.7	11.6
32	9.6	90.4	0.5	9.6	4.8	33.8	17.0	9.6	4.8	22.9	11.5
33	10.2	89.8	0.5	10.2	5.3	32.7	16.9	10.2	5.3	22.2	11.5
34	10.9	89.1	0.5	10.9	5.8	31.6	16.8	10.9	5.8	21.4	11.4
35	11.6	88.4	0.5	11.6	6.3	30.5	16.6	11.6	6.3	20.7	11.3
36	12.3	87.8	0.6	12.3	6.9	29.5	16.5	12.3	6.9	20.0	11.2
37	13.0	87.0	0.6	13.0	7.5	28.5	16.4	13.0	7.5	19.3	11.1
38	13.7	86.3	0.6	13.7	8.1	27.5	16.2	13.7	8.1	18.7	11.0

Sampling parameters nc= 600 ne= 300				Specified: cmax= 30.0 Result: EAs in sample, a= 23.7				Specified: EAs in sample, a= 35.0 Result: cmax= 19.3			
				Number of C hhs selected:		Number of E hhs selected:		Number of C hhs selected:		Number of E hhs selected:	
				If EA in sample	Expected value	If EA in sample	Expected value	If EA in sample	Expected value	If EA in sample	Expected value
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
i	ci	ei	Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi
39	14.4	85.6	0.6	14.4	8.8	26.6	16.1	14.4	8.8	18.0	10.9
40	15.2	84.8	0.6	15.2	9.5	25.6	16.0	15.2	9.5	17.4	10.8
41	16.0	84.0	0.6	16.0	10.2	24.7	15.8	16.0	10.2	16.8	10.7
42	16.8	83.2	0.7	16.8	11.0	23.8	15.7	16.8	11.0	16.2	10.6
43	17.6	82.4	0.7	17.6	11.9	23.0	15.5	17.6	11.9	15.6	10.5
44	18.5	81.5	0.7	18.5	12.8	22.1	15.3	18.5	12.8	15.0	10.4
45	19.4	80.6	0.7	19.4	13.8	21.3	15.2	19.3	13.8	14.5	10.3
46	20.3	79.8	0.7	20.3	14.8	20.5	15.0	19.3	14.1	13.9	10.2
47	21.2	78.8	0.8	21.2	15.9	19.8	14.8	19.3	14.5	13.4	10.1
48	22.1	77.9	0.8	22.1	17.0	19.0	14.7	19.3	14.9	12.9	9.9
49	23.0	77.0	0.8	23.0	18.2	18.3	14.5	19.3	15.3	12.4	9.8
50	24.0	76.0	0.8	24.0	19.5	17.6	14.3	19.3	15.7	12.0	9.7
51	25.0	75.0	0.8	25.0	20.8	17.0	14.1	19.3	16.1	11.5	9.6
52	26.0	74.0	0.9	26.0	22.2	16.3	13.9	19.3	16.5	11.1	9.4
53	27.0	73.0	0.9	27.0	23.7	15.7	13.7	19.3	16.9	10.6	9.3
54	28.1	71.9	0.9	28.1	25.2	15.1	13.5	19.3	17.4	10.2	9.2
55	29.2	70.8	0.9	29.2	26.9	14.5	13.3	19.3	17.8	9.8	9.0
56	30.3	69.8	0.9	30.0	28.3	13.9	13.1	19.3	18.3	9.4	8.9
57	31.4	68.6	1.0	30.0	29.1	13.3	12.9	19.3	18.7	9.0	8.8
58	32.5	67.5	1.0	30.0	29.8	12.8	12.7	19.3	19.2	8.7	8.6
59	33.6	66.4	1.0	30.0	30.5	12.3	12.5	19.3	19.7	8.3	8.5
60	34.8	65.2	1.0	30.0	31.3	11.8	12.3	19.3	20.1	8.0	8.3
61	36.0	64.0	1.1	30.0	32.0	11.3	12.0	19.3	20.6	7.6	8.2
62	37.2	62.8	1.1	30.0	32.8	10.8	11.8	19.3	21.1	7.3	8.0
63	38.4	61.6	1.1	30.0	33.6	10.3	11.6	19.3	21.6	7.0	7.9
64	39.7	60.3	1.1	30.0	34.4	9.9	11.4	19.3	22.2	6.7	7.7
65	41.0	59.0	1.2	30.0	35.2	9.5	11.1	19.3	22.7	6.4	7.5
66	42.3	57.8	1.2	30.0	36.0	9.0	10.9	19.3	23.2	6.1	7.4
67	43.6	56.4	1.2	30.0	36.9	8.6	10.6	19.3	23.8	5.9	7.2
68	44.9	55.1	1.3	30.0	37.7	8.2	10.4	19.3	24.3	5.6	7.0
69	46.2	53.8	1.3	30.0	38.6	7.9	10.1	19.3	24.9	5.3	6.9
70	47.6	52.4	1.3	30.0	39.5	7.5	9.9	19.3	25.4	5.1	6.7
71	49.0	51.0	1.3	30.0	40.4	7.1	9.6	19.3	26.0	4.8	6.5
72	50.4	49.6	1.4	30.0	41.3	6.8	9.3	19.3	26.6	4.6	6.3
73	51.8	48.2	1.4	30.0	42.2	6.4	9.1	19.3	27.2	4.4	6.1
74	53.3	46.7	1.4	30.0	43.1	6.1	8.8	19.3	27.8	4.1	6.0
75	54.8	45.2	1.5	30.0	44.1	5.8	8.5	19.3	28.4	3.9	5.8
76	56.3	43.8	1.5	30.0	45.0	5.5	8.2	19.3	29.0	3.7	5.6
77	57.8	42.2	1.5	30.0	46.0	5.2	8.0	19.3	29.6	3.5	5.4

Sampling parameters nc= 600 ne= 300				Specified: cmax= 30.0 Result: EAs in sample, a= 23.7				Specified: EAs in sample, a= 35.0 Result: cmax= 19.3			
				Number of C hhs selected:		Number of E hhs selected:		Number of C hhs selected:		Number of E hhs selected:	
				If EA in sample	Expected value	If EA in sample	Expected value	If EA in sample	Expected value	If EA in sample	Expected value
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
i	ci	ei	Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi
78	59.3	40.7	1.6	30.0	47.0	4.9	7.7	19.3	30.3	3.3	5.2
79	60.8	39.2	1.6	30.0	48.0	4.6	7.4	19.3	30.9	3.1	5.0
80	62.4	37.6	1.6	30.0	49.0	4.3	7.1	19.3	31.5	2.9	4.8
81	64.0	36.0	1.7	30.0	50.0	4.1	6.8	19.3	32.2	2.8	4.6
82	65.6	34.4	1.7	30.0	51.0	3.8	6.5	19.3	32.9	2.6	4.4
83	67.2	32.8	1.7	30.0	52.1	3.6	6.2	19.3	33.5	2.4	4.2
84	68.9	31.1	1.8	30.0	53.1	3.3	5.9	19.3	34.2	2.2	4.0
85	70.6	29.4	1.8	30.0	54.2	3.1	5.5	19.3	34.9	2.1	3.8
86	72.3	27.8	1.8	30.0	55.3	2.8	5.2	19.3	35.6	1.9	3.5
87	74.0	26.0	1.9	30.0	56.4	2.6	4.9	19.3	36.3	1.8	3.3
88	75.7	24.3	1.9	30.0	57.5	2.4	4.6	19.3	37.0	1.6	3.1
89	77.4	22.6	2.0	30.0	58.6	2.2	4.2	19.3	37.8	1.5	2.9
90	79.2	20.8	2.0	30.0	59.8	2.0	3.9	19.3	38.5	1.3	2.7
91	81.0	19.0	2.0	30.0	60.9	1.8	3.6	19.3	39.2	1.2	2.4
92	82.8	17.2	2.1	30.0	62.1	1.6	3.2	19.3	40.0	1.1	2.2
93	84.6	15.4	2.1	30.0	63.2	1.4	2.9	19.3	40.7	0.9	2.0
94	86.5	13.5	2.1	30.0	64.4	1.2	2.5	19.3	41.5	0.8	1.7
95	88.4	11.6	2.2	30.0	65.6	1.0	2.2	19.3	42.3	0.7	1.5
96	90.3	9.8	2.2	30.0	66.8	0.8	1.8	19.3	43.1	0.6	1.2
97	92.2	7.8	2.3	30.0	68.1	0.7	1.5	19.3	43.8	0.4	1.0
98	94.1	5.9	2.3	30.0	69.3	0.5	1.1	19.3	44.6	0.3	0.8
99	96.0	4.0	2.4	30.0	70.6	0.3	0.7	19.3	45.4	0.2	0.5
100	98.0	2.0	2.4	30.0	71.8	0.2	0.4	19.3	46.3	0.1	0.3
Mean per EA											
	32.8	67.2	1.0		25.3		12.6		17.14		8.6
Number of EAs selected					23.7		Same EAs		35.00		Same EAs
Expected sample size				nc=	600.0	ne=	300.0		600.00		300.0

Table 6.9. Specifying a minimum number (e_{min}) of type E households to be taken from any sample EA

nc= 600 ne= 300 cmax= 10.0 required a= 64.7				no emin limit				emin= 3.0			
				Number of C hhs selected:		Number of E hhs selected:		Number of 'E' hhs selected:		Number of 'E' hhs selected:	
				If EA in sample	Expected value	If EA in sample	Expected value	If EA in sample	Expected value	If EA in sample	Expected value
								before adjustment of f3		after adjustment f3 reduced by= 0.72	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
i	ci	ei	Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi
1	0.0	100.0	0.3	0.0	0.0	23.2	6.9	23.2	6.9	16.6	4.9
2	0.0	100.0	0.3	0.0	0.0	23.1	6.9	23.1	6.9	16.6	4.9
3	0.0	100.0	0.3	0.0	0.0	23.1	6.9	23.1	6.9	16.6	4.9
4	0.1	99.9	0.3	0.1	0.0	23.0	6.9	23.0	6.9	16.5	4.9
5	0.2	99.8	0.3	0.2	0.0	22.9	6.9	22.9	6.9	16.4	4.9
6	0.3	99.8	0.3	0.3	0.1	22.7	6.9	22.7	6.9	16.3	4.9
7	0.4	99.6	0.3	0.4	0.1	22.5	6.9	22.5	6.9	16.1	4.9
8	0.5	99.5	0.3	0.5	0.2	22.3	6.9	22.3	6.9	16.0	4.9
9	0.6	99.4	0.3	0.6	0.2	22.0	6.9	22.0	6.9	15.8	4.9
10	0.8	99.2	0.3	0.8	0.3	21.7	6.8	21.7	6.8	15.6	4.9
11	1.0	99.0	0.3	1.0	0.3	21.4	6.8	21.4	6.8	15.3	4.9
12	1.2	98.8	0.3	1.2	0.4	21.1	6.8	21.1	6.8	15.1	4.9
13	1.4	98.6	0.3	1.4	0.5	20.7	6.8	20.7	6.8	14.8	4.9
14	1.7	98.3	0.3	1.7	0.6	20.3	6.8	20.3	6.8	14.6	4.9
15	2.0	98.0	0.3	2.0	0.7	19.9	6.8	19.9	6.8	14.3	4.8
16	2.3	97.8	0.3	2.3	0.8	19.5	6.7	19.5	6.7	14.0	4.8
17	2.6	97.4	0.4	2.6	0.9	19.1	6.7	19.1	6.7	13.7	4.8
18	2.9	97.1	0.4	2.9	1.0	18.6	6.7	18.6	6.7	13.4	4.8
19	3.2	96.8	0.4	3.2	1.2	18.2	6.7	18.2	6.7	13.0	4.8
20	3.6	96.4	0.4	3.6	1.4	17.7	6.6	17.7	6.6	12.7	4.8
21	4.0	96.0	0.4	4.0	1.5	17.3	6.6	17.3	6.6	12.4	4.7
22	4.4	95.6	0.4	4.4	1.7	16.8	6.6	16.8	6.6	12.1	4.7
23	4.8	95.2	0.4	4.8	1.9	16.4	6.6	16.4	6.6	11.7	4.7
24	5.3	94.7	0.4	5.3	2.2	15.9	6.5	15.9	6.5	11.4	4.7
25	5.8	94.2	0.4	5.8	2.4	15.4	6.5	15.4	6.5	11.1	4.7
26	6.3	93.8	0.4	6.3	2.7	15.0	6.5	15.0	6.5	10.7	4.6
27	6.8	93.2	0.4	6.8	3.0	14.5	6.4	14.5	6.4	10.4	4.6
28	7.3	92.7	0.5	7.3	3.3	14.1	6.4	14.1	6.4	10.1	4.6
29	7.8	92.2	0.5	7.8	3.6	13.7	6.4	13.7	6.4	9.8	4.6
30	8.4	91.6	0.5	8.4	4.0	13.2	6.3	13.2	6.3	9.5	4.5
31	9.0	91.0	0.5	9.0	4.4	12.8	6.3	12.8	6.3	9.2	4.5
32	9.6	90.4	0.5	9.6	4.8	12.4	6.2	12.4	6.2	8.9	4.5
33	10.2	89.8	0.5	10.0	5.2	12.0	6.2	12.0	6.2	8.6	4.4
34	10.9	89.1	0.5	10.0	5.3	11.6	6.1	11.6	6.1	8.3	4.4
35	11.6	88.4	0.5	10.0	5.5	11.2	6.1	11.2	6.1	8.0	4.4
36	12.3	87.8	0.6	10.0	5.6	10.8	6.1	10.8	6.1	7.8	4.3
37	13.0	87.0	0.6	10.0	5.7	10.4	6.0	10.4	6.0	7.5	4.3

nc= 600 ne= 300 cmax= 10.0 required a= 64.7				no emin limit				emin= 3.0			
				Number of C hhs selected:		Number of E hhs selected:		Number of 'E' hhs selected:		Number of 'E' hhs selected:	
				If EA in sample	Expected value	If EA in sample	Expected value	If EA in sample	Expected value	If EA in sample	Expected value
								before adjustment of f3		after adjustment f3 reduced by= 0.72	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
i	ci	ei	Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi
38	13.7	86.3	0.6	10.0	5.9	10.1	6.0	10.1	6.0	7.2	4.3
39	14.4	85.6	0.6	10.0	6.1	9.7	5.9	9.7	5.9	7.0	4.2
40	15.2	84.8	0.6	10.0	6.2	9.4	5.8	9.4	5.8	6.7	4.2
41	16.0	84.0	0.6	10.0	6.4	9.1	5.8	9.1	5.8	6.5	4.2
42	16.8	83.2	0.7	10.0	6.6	8.7	5.7	8.7	5.7	6.3	4.1
43	17.6	82.4	0.7	10.0	6.8	8.4	5.7	8.4	5.7	6.0	4.1
44	18.5	81.5	0.7	10.0	6.9	8.1	5.6	8.1	5.6	5.8	4.0
45	19.4	80.6	0.7	10.0	7.1	7.8	5.6	7.8	5.6	5.6	4.0
46	20.3	79.8	0.7	10.0	7.3	7.5	5.5	7.5	5.5	5.4	3.9
47	21.2	78.8	0.8	10.0	7.5	7.2	5.4	7.2	5.4	5.2	3.9
48	22.1	77.9	0.8	10.0	7.7	7.0	5.4	7.0	5.4	5.0	3.9
49	23.0	77.0	0.8	10.0	7.9	6.7	5.3	6.7	5.3	4.8	3.8
50	24.0	76.0	0.8	10.0	8.1	6.5	5.2	6.5	5.2	4.6	3.8
51	25.0	75.0	0.8	10.0	8.3	6.2	5.2	6.2	5.2	4.5	3.7
52	26.0	74.0	0.9	10.0	8.5	6.0	5.1	6.0	5.1	4.3	3.7
53	27.0	73.0	0.9	10.0	8.8	5.7	5.0	5.7	5.0	4.1	3.6
54	28.1	71.9	0.9	10.0	9.0	5.5	5.0	5.5	5.0	4.0	3.6
55	29.2	70.8	0.9	10.0	9.2	5.3	4.9	5.3	4.9	3.8	3.5
56	30.3	69.8	0.9	10.0	9.4	5.1	4.8	5.1	4.8	3.7	3.4
57	31.4	68.6	1.0	10.0	9.7	4.9	4.7	4.9	4.7	3.5	3.4
58	32.5	67.5	1.0	10.0	9.9	4.7	4.7	4.7	4.7	3.4	3.3
59	33.6	66.4	1.0	10.0	10.2	4.5	4.6	4.5	4.6	3.2	3.3
60	34.8	65.2	1.0	10.0	10.4	4.3	4.5	4.3	4.5	3.1	3.2
61	36.0	64.0	1.1	10.0	10.7	4.1	4.4	4.1	4.4	3.0	3.2
62	37.2	62.8	1.1	10.0	10.9	4.0	4.3	4.0	4.3	3.0	3.3
63	38.4	61.6	1.1	10.0	11.2	3.8	4.2	3.8	4.2	3.0	3.4
64	39.7	60.3	1.1	10.0	11.5	3.6	4.2	3.6	4.2	3.0	3.4
65	41.0	59.0	1.2	10.0	11.7	3.5	4.1	3.5	4.1	3.0	3.5
66	42.3	57.8	1.2	10.0	12.0	3.3	4.0	3.3	4.0	3.0	3.6
67	43.6	56.4	1.2	10.0	12.3	3.2	3.9	3.2	3.9	3.0	3.7
68	44.9	55.1	1.3	10.0	12.6	3.0	3.8	3.0	3.8	3.0	3.8
69	46.2	53.8	1.3	10.0	12.9	2.9	3.7	3.0	3.9	3.0	3.9
70	47.6	52.4	1.3	10.0	13.2	2.7	3.6	3.0	3.9	3.0	3.9
71	49.0	51.0	1.3	10.0	13.5	2.6	3.5	3.0	4.0	3.0	4.0
72	50.4	49.6	1.4	10.0	13.8	2.5	3.4	3.0	4.1	3.0	4.1
73	51.8	48.2	1.4	10.0	14.1	2.4	3.3	3.0	4.2	3.0	4.2
74	53.3	46.7	1.4	10.0	14.4	2.2	3.2	3.0	4.3	3.0	4.3
75	54.8	45.2	1.5	10.0	14.7	2.1	3.1	3.0	4.4	3.0	4.4

nc= 600 ne= 300 cmax= 10.0 required a= 64.7				no emin limit				emin= 3.0			
				Number of C hhs selected:		Number of E hhs selected:		Number of 'E' hhs selected:		Number of 'E' hhs selected:	
				If EA in sample	Expected value	If EA in sample	Expected value	If EA in sample	Expected value	If EA in sample	Expected value
								before adjustment of f3		after adjustment f3 reduced by= 0.72	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
i	ci	ei	Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi
76	56.3	43.8	1.5	10.0	15.0	2.0	3.0	3.0	4.5	3.0	4.5
77	57.8	42.2	1.5	10.0	15.3	1.9	2.9	3.0	4.6	3.0	4.6
78	59.3	40.7	1.6	10.0	15.7	1.8	2.8	3.0	4.7	3.0	4.7
79	60.8	39.2	1.6	10.0	16.0	1.7	2.7	3.0	4.8	3.0	4.8
80	62.4	37.6	1.6	10.0	16.3	1.6	2.6	3.0	4.9	3.0	4.9
81	64.0	36.0	1.7	10.0	16.7	1.5	2.5	3.0	5.0	3.0	5.0
82	65.6	34.4	1.7	10.0	17.0	1.4	2.4	3.0	5.1	3.0	5.1
83	67.2	32.8	1.7	10.0	17.4	1.3	2.3	3.0	5.2	3.0	5.2
84	68.9	31.1	1.8	10.0	17.7	1.2	2.1	3.0	5.3	3.0	5.3
85	70.6	29.4	1.8	10.0	18.1	1.1	2.0	3.0	5.4	3.0	5.4
86	72.3	27.8	1.8	10.0	18.4	1.0	1.9	3.0	5.5	3.0	5.5
87	74.0	26.0	1.9	10.0	18.8	1.0	1.8	3.0	5.6	3.0	5.6
88	75.7	24.3	1.9	10.0	19.2	0.9	1.7	3.0	5.7	3.0	5.7
89	77.4	22.6	2.0	10.0	19.5	0.8	1.6	3.0	5.9	3.0	5.9
90	79.2	20.8	2.0	10.0	19.9	0.7	1.4	3.0	6.0	3.0	6.0
91	81.0	19.0	2.0	10.0	20.3	0.6	1.3	3.0	6.1	3.0	6.1
92	82.8	17.2	2.1	10.0	20.7	0.6	1.2	3.0	6.2	3.0	6.2
93	84.6	15.4	2.1	10.0	21.1	0.5	1.1	3.0	6.3	3.0	6.3
94	86.5	13.5	2.1	10.0	21.5	0.4	0.9	3.0	6.4	3.0	6.4
95	88.4	11.6	2.2	10.0	21.9	0.4	0.8	3.0	6.6	3.0	6.6
96	90.3	9.8	2.2	10.0	22.3	0.3	0.7	3.0	6.7	3.0	6.7
97	92.2	7.8	2.3	10.0	22.7	0.2	0.5	3.0	6.8	3.0	6.8
98	94.1	5.9	2.3	10.0	23.1	0.2	0.4	3.0	6.9	3.0	6.9
99	96.0	4.0	2.4	10.0	23.5	0.1	0.3	3.0	7.1	3.0	7.1
100	98.0	2.0	2.4	10.0	23.9	0.1	0.1	3.0	7.2	3.0	7.2
Mean per EA											
	32.8	67.2	1.0		9.3		4.6		5.7		4.6
Number of EAs slected					64.7		Same EAs		Same EAs		Same EAs
Expected sample size					600.0		300.0		370.3		300.0

Table 6.10. Maximum limt (e_{max}) on the number of type E households to be taken from any sample EA

nc= 600 ne= 300 cmax= no limit (take-all sampling for category C)				Number of category 'E' hhs selected:							
				Original calculation		Limit: sample<=ei		Adjusted to get ne= 300			
			Number of C hhs selected:	Number selected:		Number selected:		f3 increased by= 1.07 Number selected:			
			If EA in sample	Expected value	If EA in sample	Expected value	If EA in sample	Expected value	If EA in sample	Expected value	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
i	ci	ei	Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi
1	0.0	100.0	0.3	0.0	0.0	128.7	38.3	100.0	29.8	100.0	29.8
2	0.0	100.0	0.3	0.0	0.0	128.6	38.3	100.0	29.8	100.0	29.8
3	0.0	100.0	0.3	0.0	0.0	128.3	38.3	100.0	29.9	100.0	29.9
4	0.1	99.9	0.3	0.1	0.0	127.8	38.3	99.9	29.9	99.9	29.9
5	0.2	99.8	0.3	0.2	0.0	127.1	38.3	99.8	30.1	99.8	30.1
6	0.3	99.8	0.3	0.3	0.1	126.1	38.2	99.8	30.2	99.8	30.2
7	0.4	99.6	0.3	0.4	0.1	125.0	38.2	99.6	30.4	99.6	30.4
8	0.5	99.5	0.3	0.5	0.2	123.7	38.1	99.5	30.7	99.5	30.7
9	0.6	99.4	0.3	0.6	0.2	122.3	38.1	99.4	30.9	99.4	30.9
10	0.8	99.2	0.3	0.8	0.3	120.7	38.0	99.2	31.3	99.2	31.3
11	1.0	99.0	0.3	1.0	0.3	118.9	37.9	99.0	31.6	99.0	31.6
12	1.2	98.8	0.3	1.2	0.4	117.0	37.9	98.8	32.0	98.8	32.0
13	1.4	98.6	0.3	1.4	0.5	115.0	37.8	98.6	32.4	98.6	32.4
14	1.7	98.3	0.3	1.7	0.6	112.8	37.7	98.3	32.8	98.3	32.8
15	2.0	98.0	0.3	2.0	0.7	110.6	37.6	98.0	33.3	98.0	33.3
16	2.3	97.8	0.3	2.3	0.8	108.3	37.5	97.8	33.8	97.8	33.8
17	2.6	97.4	0.4	2.6	0.9	105.9	37.3	97.4	34.3	97.4	34.3
18	2.9	97.1	0.4	2.9	1.0	103.5	37.2	97.1	34.9	97.1	34.9
19	3.2	96.8	0.4	3.2	1.2	101.0	37.1	96.8	35.5	96.8	35.5
20	3.6	96.4	0.4	3.6	1.4	98.5	36.9	96.4	36.1	96.4	36.1
21	4.0	96.0	0.4	4.0	1.5	96.0	36.8	96.0	36.8	96.0	36.8
22	4.4	95.6	0.4	4.4	1.7	93.4	36.6	93.4	36.6	95.6	37.5
23	4.8	95.2	0.4	4.8	1.9	90.9	36.5	90.9	36.5	95.2	38.2
24	5.3	94.7	0.4	5.3	2.2	88.3	36.3	88.3	36.3	94.3	38.7
25	5.8	94.2	0.4	5.8	2.4	85.8	36.1	85.8	36.1	91.6	38.5
26	6.3	93.8	0.4	6.3	2.7	83.3	35.9	83.3	35.9	88.9	38.3
27	6.8	93.2	0.4	6.8	3.0	80.8	35.7	80.8	35.7	86.2	38.1
28	7.3	92.7	0.5	7.3	3.3	78.3	35.5	78.3	35.5	83.6	37.9
29	7.8	92.2	0.5	7.8	3.6	75.9	35.3	75.9	35.3	81.0	37.7
30	8.4	91.6	0.5	8.4	4.0	73.5	35.1	73.5	35.1	78.4	37.5
31	9.0	91.0	0.5	9.0	4.4	71.1	34.9	71.1	34.9	75.9	37.2
32	9.6	90.4	0.5	9.6	4.8	68.8	34.6	68.8	34.6	73.5	37.0
33	10.2	89.8	0.5	10.2	5.3	66.6	34.4	66.6	34.4	71.1	36.7
34	10.9	89.1	0.5	10.9	5.8	64.4	34.2	64.4	34.2	68.7	36.5
35	11.6	88.4	0.5	11.6	6.3	62.2	33.9	62.2	33.9	66.4	36.2
36	12.3	87.8	0.6	12.3	6.9	60.1	33.6	60.1	33.6	64.1	35.9

nc= 600 ne= 300 cmax= no limit (take-all sampling for category C)						Number of category 'E' hhs selected:					
						Original calculation		Limit: sample<=ei		Adjusted to get ne= 300	
				Number of C hhs selected:		Number selected:		Number selected:		f3 increased by= 1.07 Number selected:	
				If EA in sample	Expected value	If EA in sample	Expected value	If EA in sample	Expected value	If EA in sample	Expected value
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
i	ci	ei	Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi
37	13.0	87.0	0.6	13.0	7.5	58.0	33.4	58.0	33.4	61.9	35.6
38	13.7	86.3	0.6	13.7	8.1	56.0	33.1	56.0	33.1	59.8	35.3
39	14.4	85.6	0.6	14.4	8.8	54.1	32.8	54.1	32.8	57.7	35.0
40	15.2	84.8	0.6	15.2	9.5	52.2	32.5	52.2	32.5	55.7	34.7
41	16.0	84.0	0.6	16.0	10.2	50.3	32.2	50.3	32.2	53.7	34.4
42	16.8	83.2	0.7	16.8	11.0	48.5	31.9	48.5	31.9	51.8	34.0
43	17.6	82.4	0.7	17.6	11.9	46.8	31.6	46.8	31.6	49.9	33.7
44	18.5	81.5	0.7	18.5	12.8	45.1	31.2	45.1	31.2	48.1	33.3
45	19.4	80.6	0.7	19.4	13.8	43.4	30.9	43.4	30.9	46.3	33.0
46	20.3	79.8	0.7	20.3	14.8	41.8	30.6	41.8	30.6	44.6	32.6
47	21.2	78.8	0.8	21.2	15.9	40.3	30.2	40.3	30.2	43.0	32.3
48	22.1	77.9	0.8	22.1	17.0	38.8	29.9	38.8	29.9	41.4	31.9
49	23.0	77.0	0.8	23.0	18.2	37.3	29.5	37.3	29.5	39.8	31.5
50	24.0	76.0	0.8	24.0	19.5	35.9	29.1	35.9	29.1	38.3	31.1
51	25.0	75.0	0.8	25.0	20.8	34.5	28.7	34.5	28.7	36.9	30.7
52	26.0	74.0	0.9	26.0	22.2	33.2	28.4	33.2	28.4	35.4	30.3
53	27.0	73.0	0.9	27.0	23.7	31.9	28.0	31.9	28.0	34.1	29.8
54	28.1	71.9	0.9	28.1	25.2	30.7	27.6	30.7	27.6	32.7	29.4
55	29.2	70.8	0.9	29.2	26.9	29.5	27.2	29.5	27.2	31.4	29.0
56	30.3	69.8	0.9	30.3	28.6	28.3	26.7	28.3	26.7	30.2	28.5
57	31.4	68.6	1.0	31.4	30.4	27.2	26.3	27.2	26.3	29.0	28.1
58	32.5	67.5	1.0	32.5	32.3	26.1	25.9	26.1	25.9	27.8	27.6
59	33.6	66.4	1.0	33.6	34.2	25.0	25.4	25.0	25.4	26.7	27.1
60	34.8	65.2	1.0	34.8	36.3	24.0	25.0	24.0	25.0	25.6	26.7
61	36.0	64.0	1.1	36.0	38.4	23.0	24.5	23.0	24.5	24.5	26.2
62	37.2	62.8	1.1	37.2	40.7	22.0	24.1	22.0	24.1	23.5	25.7
63	38.4	61.6	1.1	38.4	43.0	21.1	23.6	21.1	23.6	22.5	25.2
64	39.7	60.3	1.1	39.7	45.5	20.2	23.1	20.2	23.1	21.5	24.7
65	41.0	59.0	1.2	41.0	48.1	19.3	22.6	19.3	22.6	20.6	24.2
66	42.3	57.8	1.2	42.3	50.8	18.4	22.1	18.4	22.1	19.7	23.6
67	43.6	56.4	1.2	43.6	53.6	17.6	21.6	17.6	21.6	18.8	23.1
68	44.9	55.1	1.3	44.9	56.5	16.8	21.1	16.8	21.1	17.9	22.5
69	46.2	53.8	1.3	46.2	59.5	16.0	20.6	16.0	20.6	17.1	22.0
70	47.6	52.4	1.3	47.6	62.7	15.3	20.1	15.3	20.1	16.3	21.4
71	49.0	51.0	1.3	49.0	65.9	14.5	19.5	14.5	19.5	15.5	20.9
72	50.4	49.6	1.4	50.4	69.4	13.8	19.0	13.8	19.0	14.7	20.3
73	51.8	48.2	1.4	51.8	72.9	13.1	18.5	13.1	18.5	14.0	19.7
74	53.3	46.7	1.4	53.3	76.6	12.5	17.9	12.5	17.9	13.3	19.1

nc= 600 ne= 300 cmax= no limit (take-all sampling for category C)						Number of category 'E' hhs selected:					
						Original calculation		Limit: sample<=ei		Adjusted to get ne= 300	
				Number of C hhs selected:		Number selected:		Number selected:		f3 increased by= 1.07 Number selected:	
				If EA in sample	Expected value	If EA in sample	Expected value	If EA in sample	Expected value	If EA in sample	Expected value
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
i	ci	ei	Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi
75	54.8	45.2	1.5	54.8	80.4	11.8	17.3	11.8	17.3	12.6	18.5
76	56.3	43.8	1.5	56.3	84.4	11.2	16.8	11.2	16.8	11.9	17.9
77	57.8	42.2	1.5	57.8	88.5	10.6	16.2	10.6	16.2	11.3	17.3
78	59.3	40.7	1.6	59.3	92.8	10.0	15.6	10.0	15.6	10.6	16.7
79	60.8	39.2	1.6	60.8	97.3	9.4	15.0	9.4	15.0	10.0	16.0
80	62.4	37.6	1.6	62.4	101.9	8.8	14.4	8.8	14.4	9.4	15.4
81	64.0	36.0	1.7	64.0	106.7	8.3	13.8	8.3	13.8	8.8	14.7
82	65.6	34.4	1.7	65.6	111.6	7.7	13.2	7.7	13.2	8.3	14.1
83	67.2	32.8	1.7	67.2	116.7	7.2	12.6	7.2	12.6	7.7	13.4
84	68.9	31.1	1.8	68.9	122.0	6.7	11.9	6.7	11.9	7.2	12.7
85	70.6	29.4	1.8	70.6	127.5	6.2	11.3	6.2	11.3	6.7	12.0
86	72.3	27.8	1.8	72.3	133.2	5.8	10.6	5.8	10.6	6.2	11.4
87	74.0	26.0	1.9	74.0	139.0	5.3	10.0	5.3	10.0	5.7	10.7
88	75.7	24.3	1.9	75.7	145.1	4.9	9.3	4.9	9.3	5.2	9.9
89	77.4	22.6	2.0	77.4	151.3	4.4	8.6	4.4	8.6	4.7	9.2
90	79.2	20.8	2.0	79.2	157.8	4.0	8.0	4.0	8.0	4.3	8.5
91	81.0	19.0	2.0	81.0	164.4	3.6	7.3	3.6	7.3	3.8	7.8
92	82.8	17.2	2.1	82.8	171.3	3.2	6.6	3.2	6.6	3.4	7.0
93	84.6	15.4	2.1	84.6	178.4	2.8	5.9	2.8	5.9	3.0	6.3
94	86.5	13.5	2.1	86.5	185.7	2.4	5.2	2.4	5.2	2.6	5.5
95	88.4	11.6	2.2	88.4	193.3	2.0	4.5	2.0	4.5	2.2	4.8
96	90.3	9.8	2.2	90.3	201.1	1.7	3.7	1.7	3.7	1.8	4.0
97	92.2	7.8	2.3	92.2	209.1	1.3	3.0	1.3	3.0	1.4	3.2
98	94.1	5.9	2.3	94.1	217.4	1.0	2.3	1.0	2.3	1.0	2.4
99	96.0	4.0	2.4	96.0	225.9	0.6	1.5	0.6	1.5	0.7	1.6
100	98.0	2.0	2.4	98.0	234.6	0.3	0.8	0.3	0.8	0.3	0.8
Mean per EA			32.8	32.8	51.5						25.7
Number of EA's slected					11.7						11.7
Expected sample size					600.0		300.0		286.3		300.0

Table 6.11. Example of circular systematic sampling with probability proportional to size

Sampling parameters

nc= 600
ne= 300

EA sampling frame			Measure of size (MoS)	Cumulative MoS	Selection point	Selection order
(1)	(2)	(3)	(4)	(5)	(6)	(7)
i	ci	ei	Mi	Ti	Si	
1	0.0	100.0	0.30	0.30		
2	0.0	100.0	0.30	0.60		
3	0.0	100.0	0.30	0.89		
4	0.1	99.9	0.30	1.19		
5	0.2	99.8	0.30	1.50	1.27	9
6	0.3	99.8	0.30	1.80		
7	0.4	99.6	0.31	2.10		
8	0.5	99.5	0.31	2.41		
9	0.6	99.4	0.31	2.72		
10	0.8	99.2	0.32	3.04		
11	1.0	99.0	0.32	3.36		
12	1.2	98.8	0.32	3.68		
13	1.4	98.6	0.33	4.01		
14	1.7	98.3	0.33	4.34		
15	2.0	98.0	0.34	4.68		
16	2.3	97.8	0.35	5.03		
17	2.6	97.4	0.35	5.38		
18	2.9	97.1	0.36	5.74		
19	3.2	96.8	0.37	6.11	5.82	10
20	3.6	96.4	0.37	6.48		
21	4.0	96.0	0.38	6.87		
22	4.4	95.6	0.39	7.26		
23	4.8	95.2	0.40	7.66		
24	5.3	94.7	0.41	8.07		
25	5.8	94.2	0.42	8.49		
26	6.3	93.8	0.43	8.92		
27	6.8	93.2	0.44	9.37		
28	7.3	92.7	0.45	9.82		
29	7.8	92.2	0.47	10.29		
30	8.4	91.6	0.48	10.76	10.36	11
31	9.0	91.0	0.49	11.25		
32	9.6	90.4	0.50	11.76		
33	10.2	89.8	0.52	12.27		
34	10.9	89.1	0.53	12.80		
35	11.6	88.4	0.55	13.35		
36	12.3	87.8	0.56	13.91		
37	13.0	87.0	0.57	14.48		
38	13.7	86.3	0.59	15.07	14.91	12
39	14.4	85.6	0.61	15.68		
40	15.2	84.8	0.62	16.30		

Sampling parameters
nc= 600
ne= 300

EA sampling frame			Measure of size (MoS)	Cumulative MoS	Selection point	Selection order
(1)	(2)	(3)	(4)	(5)	(6)	(7)
i	ci	ei	Mi	Ti	Si	
41	16.0	84.0	0.64	16.94		
42	16.8	83.2	0.66	17.60		
43	17.6	82.4	0.68	18.28		
44	18.5	81.5	0.69	18.97		
45	19.4	80.6	0.71	19.68	19.46	13
46	20.3	79.8	0.73	20.41		
47	21.2	78.8	0.75	21.16		
48	22.1	77.9	0.77	21.93		
49	23.0	77.0	0.79	22.72		
50	24.0	76.0	0.81	23.53		
51	25.0	75.0	0.83	24.37	24.00	14
52	26.0	74.0	0.85	25.22		
53	27.0	73.0	0.88	26.10		
54	28.1	71.9	0.90	27.00		
55	29.2	70.8	0.92	27.92		
56	30.3	69.8	0.94	28.86	28.55	15
57	31.4	68.6	0.97	29.83		
58	32.5	67.5	0.99	30.82		
59	33.6	66.4	1.02	31.84		
60	34.8	65.2	1.04	32.88		
61	36.0	64.0	1.07	33.95	33.09	16
62	37.2	62.8	1.09	35.04		
63	38.4	61.6	1.12	36.16		
64	39.7	60.3	1.15	37.31		
65	41.0	59.0	1.17	38.48	37.64	17
66	42.3	57.8	1.20	39.68		
67	43.6	56.4	1.23	40.91		
68	44.9	55.1	1.26	42.17		
69	46.2	53.8	1.29	43.46	42.18	18
70	47.6	52.4	1.32	44.77		
71	49.0	51.0	1.35	46.12		
72	50.4	49.6	1.38	47.50	46.73	19
73	51.8	48.2	1.41	48.90		
74	53.3	46.7	1.44	50.34		
75	54.8	45.2	1.47	51.81	51.27	20
76	56.3	43.8	1.50	53.31		
77	57.8	42.2	1.53	54.84		
78	59.3	40.7	1.57	56.41	55.82	21
79	60.8	39.2	1.60	58.01		
80	62.4	37.6	1.63	59.64		

Sampling parameters

nc= 600

ne= 300

EA sampling frame			Measure of size (MoS)	Cumulative MoS	Selection point	Selection order
(1)	(2)	(3)	(4)	(5)	(6)	(7)
i	ci	ei	Mi	Ti	Si	
81	64.0	36.0	1.67	61.31	60.36	22
82	65.6	34.4	1.70	63.01		
83	67.2	32.8	1.74	64.74		
84	68.9	31.1	1.77	66.51	64.91	1
85	70.6	29.4	1.81	68.32		
86	72.3	27.8	1.84	70.16	69.46	2
87	74.0	26.0	1.88	72.04		
88	75.7	24.3	1.92	73.96		
89	77.4	22.6	1.95	75.91	74.00	3
90	79.2	20.8	1.99	77.91		
91	81.0	19.0	2.03	79.94	78.55	4
92	82.8	17.2	2.07	82.00		
93	84.6	15.4	2.11	84.11	83.09	5
94	86.5	13.5	2.15	86.26		
95	88.4	11.6	2.19	88.45	87.64	6
96	90.3	9.8	2.23	90.68		
97	92.2	7.8	2.27	92.94	92.18	7
98	94.1	5.9	2.31	95.25		
99	96.0	4.0	2.35	97.61	96.73	8
100	98.0	2.0	2.39	100.00		

Chapter 7

Estimation from sample data

This chapter addresses selected issues relating to estimation from sample survey data. As in Chapter 3 this is done from a general perspective so as to be useful for any population-based survey (including the LFS), though the assumed context is, of course, that of child labour surveys.

We first consider issues and statistical procedures relating to the weighting of sample data, and the production of estimates from the surveys (Sections 7.1-7.3). Then we consider practical procedures for the computation and analysis of information on sampling errors, suitable for large-scale application in large complex surveys (Sections 7.4-7.7). Finally, some implications are drawn regarding an issue of great practical importance – determination of the appropriate number of clusters (or sample-takes per cluster, for a given sample size).

7.1 Weighting of sample data

In most situations (though not necessarily in all) sample data have to be weighted to produce estimates for the population of interest. When the sample data *have 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.* 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 simply require weighted aggregates at the level of primary sampling units, along with the identification of PSUs and the strata in which they lie as defined for the purpose of computing sampling errors.

It is necessary, however, to begin with a basic question which arises in producing estimates from a survey: *whether or not the sample data need to be weighted.* The objective of weighting sample data is to make the sample more representative in terms of the size, distribution and characteristics of the study population. For example, when sample units have been selected with differing probabilities, it is necessary to assign to each selected unit a weight inversely proportional to its selection probability, so as to reflect the actual situation in the population. In a survey that has been selected from a good frame and well implemented with a high response rates, these “design weights” may be all that is required.

In practice, however, the situation is usually more complex because of shortcomings in the selection and implementation of the sample, which introduce biases in the results, and also because of the need for (and possibilities of) introducing improved estimation procedures to reduce variances. The need for more complicated weighting and estimation procedures tends to be greater in surveys suffering from high non-response and coverage errors: inconsistencies in the definitions of units used at different stages in

the survey operation; departure from representative (probability) sampling; small sample sizes; and having to produce estimates for many separate sub-populations. The need (and the opportunity) is also greater in the presence of more extensive and more reliable external information for the intended purpose. These factors must be considered when conducting child labour surveys.

7.1.1 Self-weighting samples

A self-weighting sample means that each elementary unit in the population has the same, non-zero chance of being included in the sample. Higher stage units may of course be selected with differing probabilities, but differences in probabilities of selection at various stages then cancel out. With self-weighting samples, sample estimates can be prepared from unweighted data, and the results can then be inflated, if necessary, by a constant factor throughout. There are a number of arguments in favour of self-weighting designs:

- Weighting increases the complexity of the survey analysis.
- Haphazard weights which are not related to population variances increase the variance of the results.
- Weighting may reduce the flexibility and ease with which the same sample may be used for diverse purposes and different surveys. To meet the requirements of different topics and different surveys, the compromise of a multipurpose allocation may turn out to be nearly self-weighting.
- Self-weighting samples are more readily understood and accepted by the non-statistical user and the general public.
- It should be noted that moderate departures from self-weighting have a small effect on variance. This means that over-sampling for optimal allocation, or weighting for other reasons at the analysis stage, is worthwhile only if it involves a relatively large departure from self-weighting.

There can, however, be good reasons for a departure from a self-weighting design:

- While the above considerations favouring self-weighting apply in particular to broadly-based household surveys of the general population, surveys aiming to represent separately a number of sub-populations (such as labouring children in different types of activities) often need disproportionate allocations, such as over-sampling of small sub-populations of interest.
- Departures from self-weighting are often required in surveys of labouring children to meet the sample size target and ensure sufficient numbers of working children in individual sample areas. This arises from the target population being relatively small and unevenly distributed.
- Even in general household surveys, it is sometimes necessary to use different sampling rates in different domains, for example to represent adequately smaller, more important or more variable domains.
- Practical constraints such as the need to have a fixed sample size within each sample area may result in varying probabilities of selection.

- Even if in design the elements are selected with equal probabilities, in selection and implementation the resulting sample may still turn out to be non-self-weighting, due to defects in the frame, errors in selection, non-response, etc.
- Many surveys are inadequate for producing good estimates of population aggregates (as distinct from means and ratios) without weighting with external standards. Also, a consistent system of weighting helps to ensure consistency among estimates from different sources.

7.1.2 When to weight?

In deciding to weight sample data, a balance is required between the various costs and benefits of weighting. The former includes increased complexity, inconvenience, programming and analysis work and cost, possibilities of making errors or misusing the data, increased variance with haphazard weighting, and even increased bias if inappropriate standards are used to modify (re-weight) the survey results. Benefits include a reduction in bias, and possibly also a reduction in variance which may be achieved with more elaborate estimation procedures. As a practical rule, *weighting should be considered when departures from self-weighting, due to the combined effect of differing sampling probabilities, frame defects, non-response or whatever, are significant* (say outside the range +/- 20 per cent). It is normally not worthwhile applying different weights when their range of variation is smaller.

7.1.3 Effect of weights on variance and bias

There are many situations in which different parts of the population are sampled at different rates, determined by reporting requirements (for example, the production of estimates with specified minimum precision for small domains). However, variations in selection probabilities and the associated sample weights are often introduced which are largely independent of variances, costs and various characteristics of the population. In this sense, the required weights to compensate for such differential sampling rates may be considered arbitrary or haphazard. Their effect is generally to increase the variance beyond what would be expected in a corresponding self-weighting sample. A close approximation to the factor by which *variance* is increased is the following:

$$D_W^2 = \frac{n * \Sigma_s w_i^2}{(\Sigma_s w_i)^2} = \left(1 + cv^2(w_i)\right),$$

where $cv(w_i)$ is the coefficient of variation of the individual weights w_i, and the sum is calculated over the n units in the sample.

The important thing about the above is that it gives the magnitude of the effect by which *all variances for different survey estimates (different variables over different subclasses, comparisons between subclasses) are inflated more or less uniformly* as a result of haphazard weighting. Herein lies the practical utility of isolating this effect.

The *bias* resulting from ignoring weights depends on the difference in both the mean values and the size of the groups with different weights, and it is not the same for

different types of statistics. Hence the relative magnitude of the bias in relation to the effect of weighting on variance can vary according to the type of statistic considered. Unfortunately, for this very reason we are not able to give specific guidance on the appropriate balance between a possible increase in variance but a decrease in bias as a result of introducing sample weights.

Avoiding extreme weights

One important aspect, however, is clear.

The basic concern in determining the choice of weights is to maximize their contribution in reducing the total error due to variance and bias in the resulting estimates. In practice this makes it desirable to avoid the introduction of extreme weights, especially very large weights. The use of highly variable (large) weights, even if affecting only a small part of the sample cases, can result in a substantial increase in variance, while their contribution to reducing the bias may be small.

It is a common practice therefore to trim extreme weights to within a specified range so as to limit the associated increase in variance. Though quite sophisticated approaches are possible, many organizations have found it adequate to rely on simple "rules of thumb" for trimming extreme values of weights, at least for the routine production of their statistical information. A practical recommendation is that, apart from design weights, extreme weights should be trimmed so that *the ratio of the largest to the smallest case weights introduced should not exceed 5 or so*.

7.2 Computing sample weights: A systematic approach

The issues involved in sample weighting tend to be complex, and the "best" solution may be that which meets the specific situation, depending on the nature of the data at hand, the sources of error that need to be controlled, knowledge of the specific circumstances and limitations of the survey, and existing practices and preferences of the survey organization. Nevertheless, there are major advantages to following certain basic standards and a systematic approach.

7.2.1 Sources of information for weighting

In applying the weights, the best use has to 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. Various types of information sources may be cited which can be used in a systematic manner to apply weights in a step-by-step process:

- sample design, i.e. the design probabilities of selecting each ultimate unit (e.g. household);
- the sampling frame, which may provide additional information on sample areas and on all responding and non-responding units;

- sample implementation, i.e. response rates and information on non-respondents;
- other, significantly larger surveys with better coverage, higher response rates and more reliable information on certain characteristics of households and/or persons; an example is the use of a recent large-scale survey with good coverage and response rates (such as the LFS in many countries) to improve the representative nature of a more complex and difficult survey with a smaller sample size (such as a survey of labouring children);
- current registers or population projections from the census, providing information on size, characteristics and distribution of the population.

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 introducing corrections based on external data*. In using external information, it is necessary to ensure that the information is significantly more reliable than available information that is internal to the survey, that the items of information used are defined and measured comparably in the two sources, and that the coverage and scope are the same. For instance, if the survey is confined to the population residing in private households, external information on the population used for sample weighting should be similarly restricted.

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

These observations apply to statistics such as rates, ratios and distributions estimated from the sample. Generally sample surveys are much better at estimating such statistics than at estimating population aggregates. For the latter, alternative sources such as the census, population projections and possibly registers and other administrative sources are often more reliable. (See Section 7.3.)

Step-by step procedure

To achieve common standards, as well as for the sake of clarity and convenience, it is desirable that a step-by-step procedure be adopted which distinguished between 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 subsections describe the main steps.

7.2.2 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*.

Relative values matter more than absolute values of the selection probabilities. With a multi-stage sampling design, the reference is to the overall selection probabilities of households. In many household surveys, the ultimate units are selected with uniform probabilities, so that the design weights are all the same (e.g. =1.0). In some other cases the samples may have been designed to be self-weighting within major sampling or reporting domains, with variable selection rates across domains; in such cases, all units within a given domain receive the same design weight, and the weights vary across domains in inverse proportion to the domain probabilities of selection.

If p_i is the overall sampling probability of household i (and of all persons within that household), and n the number of households successfully enumerated in the sample, the design weights are

$$w_i = \left(\frac{n}{\Sigma(1/p_i)} \right) . \frac{1}{p_i},$$

where the factor in parenthesis is simply a constant, chosen to ensure that the average value of the weight per household (successfully enumerated in the survey) is 1.0. Such scaling is usually convenient in practice.

7.2.3 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 because of defects in the sampling frame or other reasons to do with the sample selection and implementation procedures. One way to apply such a correction would be to reduce the design probabilities of selection (that is, to increase the weights correspondingly) in proportion to the coverage rates in the affected domains, or to incorporate compensation in other covered domains similar to the one(s) excluded. *This can be important in child labour surveys 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.

Here is a simple example to illustrate the point.

Suppose two of the various survey domains consist of very small area units: 1. the smallest, and 2. the next smallest. Suppose that domain 1 accounts for 2 per cent of the total population, but due to cost and practical reasons this domain cannot be included in the sample. Suppose that domain 2 accounts for 3 per cent of the total population. It is the domain most resembling domain 1 and it has therefore been decided to increase its weight to compensate for the non-inclusion of domain 1. To represent the total of 5 per cent for the two domains, the design weights in domain 2 should all be multiplied by the factor 5/3.

7.2.4 Non-response weights

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

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*, in an attempt to "make up" for the non-responding cases in that class.

Of course, weighting for non-response cannot take into account the effect of the absolute levels of non-response – at best it corrects only for differential non-response across the classes. The effectiveness of the procedure depends on the extent to which non-responding units within each class are similar to the responding units in that class on important variables. Differences in unit characteristics and in response rates should be maximized across the weighting classes chosen. It is obvious that weighting classes can be defined only on the basis of those characteristics that are available for both the responding and non-responding units.

Given this requirement, it is still necessary to choose the appropriate number and size of classes to be used for this 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 the characteristics of units.

Often, weighting classes of an average size of *at least* 100 sample cases may be a reasonable choice. In a sample with 5,000 households for instance, this would mean creating 50 or fewer weighting classes.

Weighting classes can be defined only on the basis of those characteristics that are available for both the responding and the non-responding units. In a multi-stage sample of households, the characteristics involved may be of two types: those pertaining to sample areas, and those concerning individual households, including both responding and non-responding households.

Area-level characteristics 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 always available from the sample design itself (specifically, the geographic location and other information used for stratification of sampling areas). Additional information may also come from the sampling frame or external sources such as local area statistics from the census or administrative sources.

Household-level characteristics refer to characteristics relating to individual households, such as tenure of accommodation, household size and type, socio-economic status and other characteristics of the household head or reference person, the number of working persons, and other characteristics which may be related to the level of household living conditions, income and economic activity. We may also mention similar characteristics at the personal level. Several sources of such information are possible. 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, an added source of information can be the linkage of sample units with external sources such as the population census or administrative records, where the necessary access to micro-level data is possible. When such external information is used for the present purpose, the same source should be used for the classification of both non-responding and responding households, 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. In cases where the current sample is based on some previous survey, the latter can provides information on non-respondents to the current survey. 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.

In principle, the computation of the response rate is straightforward; it is the ratio of the number of units successfully enumerated to the number originally selected. For instance, in a sample of households, response rate R_j is computed as the ratio of the number of households interviewed (e.g. m_j) to the number selected (e.g. n_j) in the weighting class:

$$R_j = m_j / n_j$$

The required non-response weight is

$$w_j = \overline{R} / R_j$$

where the inverse of R_j values have been multiplied by their mean value so as to scale the weights to average 1.0, which is convenient.

In practice complications can arise for several reasons. Firstly, the type of units for sampling may not be the same as the type of units finally enumerated in the survey. For instance, we may select a sample of dwellings for the enumeration of households, or a sample of dwellings or households for the enumeration of persons. In such situations the exact number of the enumeration units selected (denominator of the response rate) may not be known.

Secondly, in the actual computation care has to be taken to define the denominator in the above expression correctly. It should include only valid units, e.g. addresses in the sample; in other words it should exclude empty, non-existent, inaccessible or lost addresses, or those otherwise containing no eligible household.

In addition, if the households within a class have different design weights, then R_j is more appropriately computed as the ratio of the weighted number of households interviewed to the weighted number of households selected.

At a minimum, geographic location and other information used for stratification of the sample areas should be used to define appropriately weighting classes for non-response adjustment in household surveys. 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 of, let us say, 100 households on the average. (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.) Then all sample households within a group j are given the same non-response weight, inversely proportional to the response rate R_j in the group.

Other classifications based on additional information of the type described above could be created where possible. When additional variables (whether at the area or the household level) are available, each variable (or each cross-classification of two or more variables) may be used to divide the whole sample into parallel sets of weighting classes. Each set will provide the distribution, according to the classification variable(s) concerned, of the number of households selected and the number successfully interviewed, on the basis of which response rates can be computed in each category of the classification.

7.2.5 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 distribution 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.

Taking this step does not require matching the sample and the external source at the level of individual households or persons. The weighting adjustments are made on the basis of a comparison of sample and external distribution at the aggregate level.

> **Example.** Suppose that for a certain characteristic proportion p_i lies in class i of the classification in the sample.[49] Let the same proportion from a more reliable external source be P_i. Then the additional weights to be applied to all sample units in this class are simply $=(P_i/p_i)$. The resulting weighted sample proportion will then agree with P_i.

However, a number of requirements should 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 have been obvious shortcomings in its design and implementation – especially serious departures from probability sampling. These considerations may often apply to intensive labouring children surveys.

[49] The data may already have been weighted, by design and non-response weights for instance, in computing this proportion

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

3. It is important that the external information be clearly more *reliable* than the information available from the survey itself.

4. The external information should be *consistent* with the survey, i.e. the variables used in the adjustment should be defined and measured in the same way in the sample and in the external source(s). Consistency is also required between the external sources when more than one is involved.

5. To the extent possible, the external information should *cover diverse variables.*

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 for a number of *marginal distributions* simultaneously rather than to insist on controlling too detailed a cross-classification of different variables.

A convenient method of adjusting the sample distribution to a number of external controls simultaneously is the classical iterative proportional fitting or "raking" method originally proposed by Deming. 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 till sufficiently close agreement is obtained for all the variables concerned simultaneously.

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 applying 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.

7.2.6 Overall inflation factor

This refers to the factor required to inflate the sample results to the corresponding population aggregate.

The objective of this factor is to isolate the effect of the overall (average) sampling rate, thereby reducing the scale of all other factors involved in the weighting so as to average some convenient number such as 1.0. Also, multiplying all survey results by F will compensate for the gross effects of any under-coverage, overall non-response, and random variation on the achieved sample size.

This scaling does not affect the survey estimate of proportions, means or other ratios, but only the estimate of totals or aggregates.

However, proper scaling is essential if the results from different domains or countries have to be put together. In such an aggregation, the weights should be scaled in such a way that for each domain the sum of weights for the sample cases is proportional to the population size of the domain.

In principle, different inflation factors may be involved in the estimation of aggregates for different types of quantities, and for different types of units such as households and persons. As noted below, the estimation of totals (population aggregates) from sample surveys normally requires control information on the size and relevant aggregates of the base population from more reliable sources outside the survey.

7.3 Estimating ratios and totals

7.3.1 Proportions, means, ratios

As noted at the beginning of this chapter, computing appropriate weighting factors where required and *attaching them with the data for each analysis unit in the survey* makes the process of estimating proportions, means, ratios and even more complex statistics very straightforward, rendering it unnecessary to evoke the complexities of the sampling design explicitly. The most common type of estimator encountered in surveys takes the form of a ratio of two sample aggregates, say *y* and *x*:

$$y = \sum_i y_i = \sum_i \sum_j w_{ij} \cdot y_{ij}$$

$$x = \sum_i x_i = \sum_i \sum_j w_{ij} \cdot x_{ij}.$$

$$r = y/x.$$

Both the numerator (y) and denominator (x) may be substantive variables – as, for example, in the estimation of per capita income 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} refers to its income and x_{ij} to its size (the number of persons, in this example). Quantity w_{ij} is the weight associated with the unit.

Ordinary means, percentages and proportions are just special cases of the ratio estimator. In a mean, the denominator is a count variable; that is to say, x_{ij} is identically equal to 1 for all elements in the sample. This gives

$$\bar{y} = \frac{\sum_i \sum_j w_{ij} \cdot y_{ij}}{\sum_i \sum_j w_{ij}}.$$

For a proportion (or percentage) the additional condition is that y_{ij} is a dichotomy equal to 1 or 0, depending on whether or not unit j possesses the characteristic whose proportion is being estimated. On the other hand, the survey may also involve more complex statistics such as differences, weighted sums, ratios or other functions of ratios. These can be estimated in an analogous way.

One very important and convenient point should be noted. 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 i, the 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 stratified samples, the normal practice is to use "combined" ratio estimates computed from results aggregated across strata to achieve greater stability (lower mean-squared error):

$$r = \frac{y}{x} = \frac{\sum_h y_h}{\sum_h x_h} = \frac{\sum_h \sum_i y_{hi}}{\sum_h \sum_i x_{hi}} = \frac{\sum_h \sum_i \sum_j w_{hij} \cdot y_{hij}}{\sum_h \sum_i \sum_j w_{hij} \cdot x_{hij}}$$

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 over the whole sample (or a domain of interest).

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 types of 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 sample weights of individual units.

7.3.2 Estimating totals

The ratio estimates of the above type, while often suitable for means and other ratios, usually require modification when the objective is to estimate population aggregates. This is especially true of surveys with a multi-stage design and small sample size. This is because with multi-stage sampling design, the resulting sample size varies at random, and therefore aggregates estimated directly from the survey can have a large sampling error. The problem tends to become more serious when estimates are required for population subclasses whose selection is not explicitly controlled in the multi-stage design.

An equally important problem arises from the fact that *estimates of aggregates are directly biased in proportion to the magnitude of the coverage errors*. By contrast, this effect on estimates of proportions, means, other ratios and more analytical statistics is often much less marked. Generally aggregates estimated from sample surveys are subject to under-estimation. This is particularly true for populations which are difficult to survey. Labouring children is an obvious example.

The appropriate procedure for estimating population aggregates is as follows. In place of simple inflation of the form $\hat{Y} = F.y$, i.e. instead of inflating the sample aggregate y by a constant scaling factor F, the inverse of the overall sampling fraction, the required aggregate may be expressed in the form of a ratio-type estimate

$$\hat{Y}_r = \frac{y}{x}.X$$

where y and x are estimated totals from the sample – y being the variable of interest, and x an auxiliary variable for which a more reliable population aggregate value X is available from some external source.

Example. A child labour survey has been used to estimate the *proportion* of children engaged in child labour. The *number* of labouring children is estimated not directly from the above estimate from the survey, but by multiplying this proportion by the total number of relevant children in the population, the latter being estimated from a more reliable external source such as population projections.

The value and applicability of this procedure depends on several factors. Firstly, the correlation coefficient between y and x must be positive and preferably large – greater than 0.6 or 0.7 at least. Secondly, X should be available with higher precision than the simple estimate x of the population aggregate that can be directly produced from the sample itself. Thirdly, X in the population and x in the sample should be based on essentially similar measurement on the same population; a difference between the two would introduce a bias into the estimate.

7.3.3 Usefulness of preparing simple unbiased estimates

Nevertheless, it is very important to be able to prepare simple unbiased estimates of the form $\hat{Y} = F.y$ from the survey data, even though these may be refined and modified subsequently in the production of the final estimates. The term "simple unbiased estimates" is used in the sense that the estimates are produced directly from the survey results without recourse to data external to the survey, by weighting each observation in inverse proportion to its probability of selection into the sample. Such an estimate for a population aggregate is produced by weighting each sample value simply by the inverse of its selection probability.[50]

Such estimates can be prepared only with probability sampling, i.e. for samples selected in such a way that each element in the population has a known and non-zero probability of being selected. To prepare good estimates, it is also necessary that problems of sample implementation, such as non-response and under-coverage, do not significantly distort these probabilities. Good simple estimates would also imply that any adjustments which may have to be made subsequently to improve their precision will not turn out to be large. In short, being able to produce good simple unbiased estimates indicates that the survey has been well designed and well implemented.

7.4 Note on small area estimation

7.4.1 The context

There is an increasing demand for statistics, including child labour statistics, for small areas, small sub-populations, and other small domains. In practice it is not possible – and generally not even desirable – to expand the sample sizes to meet these reporting requirements using direct estimates from the survey. Indirect methods need to be developed that can combine different types of information from different sources to produce more reliable estimates than would be possible on the basis of any of the sources used individually.

[50] More precisely, we write the estimate $\hat{Y} = F \cdot y$ mentioned above as $\hat{Y} = \Sigma y_i/f_i$, where the sum is over all elements i in the sample and f_i is the actual value of the unit's selection probability.

This section aims to provide only a few introductory ideas about the complex and quite rapidly developing techniques of small-area estimation. The terminology and description are those of the conventional combination of census and sample survey data. To put this in the context of child labour surveys, it is helpful to think of the census as a large-scale survey such as has been described in this manual.

Let us suppose that, from a large-scale operation such as the LFS, estimates X_{igh} have been obtained of the number of working children by type of activity (i), gender-age group (g) and administrative areas (h) of the country. These estimates are based on a brief set of questions implementing only a simplified definition of child labour employed in a large-scale and less intensive survey, with the main focus on topics other than child labour. Let us assume that these estimates provide a good picture of the structure of variation (for example, by region) but that the actual magnitudes of the estimates need adjustment. Suppose also that a more intensive but smaller survey provides improved estimates Y_{ig} but only by type of activity (i) and gender age group (g); the sample is too small to permit a regional breakdown (by h).

On the assumption that, for any given i,g, the large survey provides a reasonable picture of the regional variation, we may estimate the variation by activity (i) and region (h) as:

$$Y_{i.h} = \sum_{g} \left(\frac{X_{igh}}{X_{ig.}} \right) . Y_{ig.}$$

Theoretically, we could also estimate the full breakdown as

$$Y_{igh} = \left(\frac{X_{igh}}{X_{ig.}} \right) . Y_{ig.},$$

but such extension generally involves a much greater degree of uncertainty.

7.4.2 Estimates for small domains based on the combined use of census and survey data

The need for small domain estimates

The census can provide geographically detailed but less frequent and less up to date data, while sample surveys can be more frequent and up to date but their results cannot provide sufficient geographical detail because of insufficient sample size. Sample survey data are widely used to derive reliable estimates for totals and means for large areas and domains. However, despite major developments in survey capability and practice, the usual "direct" survey estimators for a small domain, drawn only from sample units in the domain, are likely to yield unacceptably large sampling errors on account of the smallness of the sample sizes involved.

Yet demands are growing everywhere, not only in developed but also in developing countries, for more timely and varied statistics for lower-level administrative units and other small domains. These needs can be met only by producing estimates which are more frequent and more up to date than can be provided by censuses conducted at long

intervals but which can be classified in much greater detail than is possible from sample surveys of limited size. Small area estimation methods, aimed at providing current yet detailed information, have been developed for this purpose. The basic idea of these procedures is to borrow and combine the relative strength of more than one sources in the production of more accurate, and therefore more useful, estimates.

In the past few decades, there have been major and sophisticated developments in the methodology of small domain estimation procedures. The objective of this section is not to review or evaluate those procedures, but merely to provide an introduction to the issues and some basic approaches.[51]

What are small domains?

Firstly, it should be noted that the term "domain" or "estimation domains" is used to refer to the population or any part of the population for which separate statistics are required. The term "small area" or "local area" is commonly used to denote a geographical area which is considered "small" in some sense, such as provinces, counties, districts, other smaller administrative divisions, localities, or even census divisions or enumeration areas. The term domain is more general and may refer to geographical areas or to other sub-populations of interest, such as specific groups by age, gender, nationality, race or other characteristics.

Secondly, it is important to note that "small" does not refer to the smallness of the population size of the domain of interest, but to the smallness of the sample (number of observations) available on it. Small in one context may be very different from small in another context. For example, in statistically more developed countries where large and frequent sample surveys and/or administrative source are available, the reference may be to very local areas or small population groups. In many less developed countries, the production of useful statistics even for large provinces or districts may require special small domain estimation techniques because of the impossibility of expanding the sample size of national surveys sufficiently for the purpose. The possibilities and suitability of specific small domain estimation procedures may differ greatly between these two situations.

Classification of domains according to size category can be useful in keeping the distinction clear. Adapting a rough classification proposed by Kish[52], we may distinguish between:

- **major domains**, e.g. 1-10 divisions of the population, such as major regions of the country, major groups by occupation, industry, gender and age, etc.; sample surveys are usually designed to provide useful direct estimates for such divisions;
- **minor domains**, e.g. 10-100 divisions of the population, such as individual provinces in Indonesia and Thailand or individual districts in Kenya, two-way classification by occupation and gender-age group; *in many developing countries, the primary interest is to extend the available statistics to this level of detail;*

[51] A comprehensive text on the subject is Rao, J.N.K. (2003). *Small area estimation*. Wiley.

[52] Kish, L. (1980). "Design and estimation for domains". *The Statistician*, vol. 29(4), pp. 209-222.

- **mini domains**, e.g. 100-1000 divisions of the population, such as individual counties in the United States or China; the production of reliable estimates at this level of detail is often beyond the capacity of many developing counties, while some more developed countries are able to produce estimates at this level or even at lower levels of detail; and
- **rare domains**, which may be used to refer to, say, 1000 or more subdivisions, such as rare populations or multiple classifications.

7.4.3 Diversity of methods

Earlier methods of small area estimation focused on demographic methods used for the estimation of population characteristics in post-censal years, or on estimation of the size of population categories by labour force status and other similar characteristics. Many of these methods used current data from administrative registers in conjunction with related data from the latest population census. Essentially, these "symptomatic estimation" methods exploit the relationships between "symptomatic" variables (such as the locally recorded number of births, deaths, school enrolment, housing units, etc.) and variables of interest to be estimated (such as the local population size). Generally, such methods are very specific to the situation and depend on the quality and type of administrative data available for the purpose.

A second class of method, the so-called "synthetic estimation", borrows the structural relationships between variables available in detail (but which are not current) in the census and impose these relationships in some appropriate manner on the less detailed but more current survey data to produce estimates that are current and detailed at the same time. The quality of the resulting estimates depends on the validity of the imported relationships.

It can also be useful to combine indirect estimates with direct estimates from the sample, to construct "composite estimates" as an appropriately weighted sum of direct and indirect estimates.

More recent and sophisticated techniques include procedures such as "empirical Bayes", "hierarchical Bayes", and "empirical best linear unbiased prediction" (EBLUP) procedures. These methods have made a significant impact on small area estimation practice in recent years.

Some aspects and forms of the "synthetic estimation" techniques referred to above appear both feasible and practical for application in developing-country circumstances. The basic idea behind this sort of approach will be outlined below. Beyond that, the procedures mentioned above will not be reviewed here any further. Rather, for the purposes of this practical sampling manual, it is useful to emphasize the correct approach to the development and use of small domain estimation procedures. The suitability of any technique is situation-specific and is determined by several criteria:

- availability of data;
- accuracy of the estimates;
- practicality;
- acceptability among the users

The following extract summarize a number of practical lessons in the application of such techniques:

> "From a variety of available methods several lessons may be learnt. First, that one may find among them a better method than the one he is arbitrarily using at present for small domains; this is often the passive "null" method of continued reliance on the last decennial census which may be 12 years out-of-date.
>
> Second, there is no single method that is best for all situations. Great differences between countries exist in the sources and quality of data available; the scope and quality of its census; the extents, contents and sizes of its sample surveys; and especially the scope and quality of it administrative registers. However, passive and negative attitudes are generally unjustified since every country has some resources, and ingenuity and effort can find unused resources of data. These may be of apparently different origins, but potentially useful because of high correlation with population sizes …
>
> Furthermore the choice of sources and methods should vary with and depend on the nature of the statistics, on the desired estimates, and also on the domains to which they pertain. Note also the effects of the lapse of time since the last census … More generally, the balance between biases of a census and the variance of a large sample survey will move in favour of the latter during the 10 years between decennial censuses. The balance will also move in favour of less accurate but more current registers. The balance of sample surveys versus censuses or registers should depend on the sample size, but a fixed sample size has relative advantages in smaller populations.…
>
> Finally, the choice between methods is more difficult because the "best" is often not clear even after the event. Errors in the estimates arise from biases chiefly, and "true values" are usually not available for measuring the biases directly. Tests must be combinations of empirical and model bases, often depending eventually (and uncertainly) on results from decennial censuses. Better methods and criteria must be pursued with several methods and over the long range with an evolutionary approach and with patience".[53]
>
> "We may … sketch the bare outline of present and future developments. (1) Methods are useful and used now for post-censal estimates for local area statistics. (2) These methods will be used for other statistics also and in other situations. (3) The present methods can and will be improved. (4) The relative strength of different methods is difficult to predict and it depends on specific circumstances; they may be discovered in specific empirical trials. (5) Success depends on first using better data and second on better methods. … Good data sources are the principal means to better statistics … [We need to] work towards the collection of other and better data [and also consider] strategies for cumulating data from samples for small areas".[54]

[53] Purcell, N.J., and L. Kish. (1980). "Postcensal estimates for local areas (or domains)". *International Statistical Review*, vol. 48, pp. 3-18.

[54] Kish, L. (1987). *Small area statistics: An international symposium*, Platek et al (ed.). Wiley.

7.4.4 Illustration of a procedure

Below we provide a simple illustration of a potentially useful synthetic approach. As a simple illustration of the procedure, let us consider that a sample provides current estimates of some quantity or count (such as numbers by activity status) $y_{g.}$ by gender-age group (g). From the census, the detailed but less current distribution of the population Y_{gh} by gender-age (g) and small area (h) is known. This latter provides an estimate of the distribution of a gender-age group by small area:

$$\left[Y_{gh} / Y_{g}\right]$$

On the assumption that this distribution (the "structural relationship") is still valid, small area estimates of variable of interest y are given by summing over all gender-age groups g:

$$y_{.h} = \sum_{g} \left(Y_{gh} / Y_{g}\right).y_{g.}.$$

The basic idea of a more general form of the procedure, termed "structure preserving estimation" (SPREE) by its authors,[55] is illustrated in Figure 2 below.

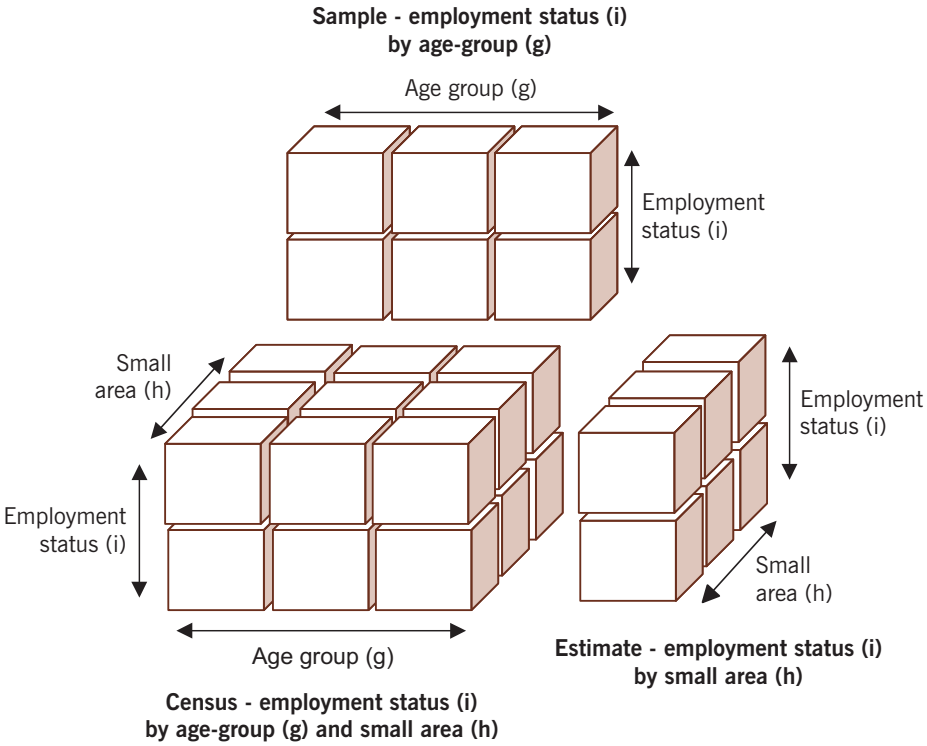
Suppose that in the census, detailed but less current distribution X_{igh} of activity status (i) by age-group (g) and small area (h) is available. Current but less detailed distribution $Y_{ig.}$ of activity status (i) by age-group (g) is available from the sample. This corresponds to the marginal distribution of the full census distribution, summed across small areas in the country. The interest is in estimating the distribution $Y_{i.h}$ of activity status (i) by small area (h), i.e. the other margin of the full census distribution as shown in the diagram. This estimate is given by

$$Y_{i.h} = \sum_{g} \left(\frac{X_{igh}}{X_{ig.}} \right).Y_{ig.}.$$

This is based on the assumption that the term in the parentheses, namely the distribution of employment status by age-group (i by g) across small domains (h) has remained unchanged since the census. The corresponding distribution is "imposed" on each current (sample) value by i and g, and then summed over all age-groups to obtain small area estimates of employment status (without classification by age-group of course).

[55] Purcell, N.J., and L. Kish. (1980). op. cit.

Figure 2. Illustration of a small-area estimation procedure



Different forms of the procedure can be developed according to the completeness of the information available from the census and the types of constraints implied by the available information from the survey, which must be satisfied by the resulting small area estimates.

7.5 Importance of information on sampling errors

The particular units that happen to be selected for a particular sample depend on chance, the possible outcome being determined by the procedures specified in the sample design. This means that, even if the required information on every selected unit is obtained entirely without error, the results from the sample are subject to a degree of uncertainty as a result of these chance factors affecting the selection of units. Sampling variance (sampling error, standard error) is a measure of this uncertainty.

While survey data are subject to errors from various sources, information on sampling errors is of crucial importance to a proper interpretation of the survey results, and to the rational design of sample surveys. The importance of including information on sampling errors in survey reports cannot be over-emphasized.

Of course, sampling error is only one component of the total error in survey estimates, and not always the most important. By the same token, it is the lower (and the more

easily estimated) boundary of the total error: *a survey will be useless if this component alone becomes too large for the survey results to add useful information with any measure of confidence to what is already known prior to the survey*.

Furthermore, survey estimates are normally required not only for the whole population but also separately for many subgroups in the population. Generally, the relative magnitude of sampling error compared to that of other types of errors increases as we move from estimates for the total population to estimates for individual subgroups and comparisons between subgroups. *Information on the magnitude of sampling errors is therefore essential in deciding the degree of detail with which the survey data may be meaningfully tabulated and analyzed*.

Similarly, sampling error information is needed for sample design and evaluation. While the design is also determined by many other considerations (such as costs, availability of sampling frames, the need to control measurement errors), rational decisions on the choice of sample size, allocation, clustering, stratification, estimation procedures, etc. can only be made on the basis of detailed knowledge of their effect on the magnitude of sampling errors in statistics obtained from the survey.

Various practical methods and computer software have been developed for computing sampling errors, and there is no justification in most situations for the continued failure to include information on sampling errors in the presentation of survey results.

7.6 Practical procedures for computing sampling errors

Practical procedures for estimating sampling errors must take into account the actual structure of the design, but need to be flexible enough to be applicable to diverse designs. They should be suitable and convenient for large-scale application (producing results for various statistics and subclasses), and economical in terms of the effort and cost involved.

Three types of method are commonly used in practice for surveys based on complex multi-stage designs:

1. computation from comparisons among certain aggregates for “primary selections” within each stratum of the sample;

   The other methods are based on the fact that sampling errors may also be computed from comparisons among estimates for ‘replications’ of the sample. We can distinguish between the following two very different types of approaches in this class;

2. variance estimates (and possibly other error estimates) based on comparisons of *independent* replications of the full sample, and

3. variance estimates based on comparisons of overlapping and non-independent replications generated by repeated re-sampling from the parent sample according to a specified scheme.

The term "primary selection" refers to the set of ultimate units obtained by applying a certain specified sub-sampling procedure to each selection of a primary sampling unit. (The terms "ultimate cluster" and "replicate" are also used to describe the same concept.) Normally, it simply means the set of ultimate units coming from the same PSU in the sample – for instance, units selected from the same cluster in a two-stage design. There are designs in which more than two independent sub-samples may be selected from the same physical PSU; in that case each such sub-sample constitutes a "primary selection".

The term "replications" refers to parts of the sample each of which reflects the structure (clustering, stratification, allocation, etc.) of the full sample, differing from it only in size. For example, suppose we select five samples independently from the same frame, each with the same design and size, and then we put these together to make our final, total sample. Each of the original five parts is a "replication" of the total sample. Each has the same structure as the total sample, from which it differs only in sample size. In this particular example, the replications are independent.

7.6.1 Method 1. Estimating variance from comparison among primary selections

This method is based on the comparison of estimates for independent primary selections within each stratum of a multi-stage design. It is perhaps the simplest approach to the computation of sampling errors in common statistics such as proportions, means, rates and other ratios, and the method can easily be extended to more complex functions of differences or ratios, double ratios, indices, etc. The basic equations are as follows.

For a population total Y obtained by summing individual values Y_{hij} for elements j over PSU i, and then over all PSU's and strata h in the population:

$$Y = \sum_h Y_h = \sum_h \sum_i Y_{hi} = \sum_h \sum_i \sum_j Y_{hij}.$$

The above is estimated by summing appropriately *weighted* values in the units in the sample:

$$y = \sum_h y_h = \sum_h \sum_i y_{hi} = \sum_h \sum_i \sum_j w_{hij} \cdot y_{hij}.$$

For the combined ratio estimator of two aggregates y and x

$$r = \frac{y}{x} = \frac{\sum_h y_h}{\sum_h x_h} = \frac{\sum_h \sum_i y_{hi}}{\sum_h \sum_i x_{hi}} = \frac{\sum_h \sum_i \sum_j w_{hij} \cdot y_{hij}}{\sum_h \sum_i \sum_j w_{hij} \cdot x_{hij}},$$

the general expression for variance is

$$\mathrm{var}(r) = \sum_h \left[(1 - f_h) . \frac{a_h}{a_h - 1} . \sum_i (z_{hi} - \frac{z_h}{a_h})^2 \right],$$

where a_h is the number of primary selections in stratum h, f_h the sampling rate in it, and the computational variable z defined as

$$z_{hi} = \frac{1}{x}.(y_{hi} - r.x_{hi}); \quad z_h = \sum_i z_{hi}$$

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 valid use of the ratio estimator and the approximation involved in the expression for its variance.
5. The quantities x_{hi} in the denominator (which often correspond to sample sizes per PSU) are reasonably uniform in size within strata.[56]

The above variance estimation formulae are very simple despite the complexity of the design, as they are based only on weighted aggregations for the primary selection, and identification of the strata.

A most important point is that the complexity of sampling within PSUs does not appear explicitly to complicate the estimation procedure. No separate computation of variance components is required. This gives the method great flexibility in handling diverse sampling designs, which is one of its major strengths and the reason for its widespread use in survey work.

On the other hand, the method requires the development of different variance estimation formulae for different types of statistics, and is not easily extended to very complex statistics. In descriptive reporting of survey results, most of the statistics of interest are in the form of proportions, means, rates, ratios, or some functions of those. For these, the basic formulae noted above apply for variance estimation, making the procedure very straightforward. Other more complex measures may be involved in analytical reporting of survey results. Variance computation can be more difficult for such measures using the method described above.

7.6.2 Method 2. Variance from independent replications

This is in principle a very simple procedure capable of handling estimates of any complexity. It has been used extensively for official surveys in India and elsewhere. The method is based on the assumption that the parent sample can be regarded as consisting of a number of *independent replications* or sub-samples, each reflecting the full complexity of the parent sample and differing from it only in size. With these assumptions the replicated estimates can be regarded as independent and identically distributed (IDD) random variables, so that variability among them gives, in a very simple form, a measure of variance of the overall sample estimator.

[56] The last mentioned requirement is concerned with keeping the bias of the ratio estimator small. In practical terms the requirement, according to Kish, is that relative variance, var(x)/x, should ideally be below 0.1, and anyway should not exceed 0.2 when ratio estimation is used.

The limitation of the method is that in many situations the total sample cannot be divided into a sufficient number of independent replications of adequate size for the method to be applicable. Its strength is its simplicity when applicable.

The procedure

The basic requirement of the method is that, by design or subsequent to sample selection, it should be possible to divide the parent sample into more or less *independent* replications, each with essentially the same design as the parent sample. In a multi-stage design, for example, the parent sample has to be divided at the level of the PSUs, i.e. divided exhaustively into a number of non-overlapping replications each consisting of a separate, independently selected (and ideally also independently enumerated) set of PSUs. For a combined estimation over a number of strata, each replication must itself be a stratified sample covering all the strata.

With independent replications, each providing a valid estimate of the same population parameter of interest, the results of the theory of "independent replicated variance estimator" can be directly applied. The variance of the simple average of n replicated estimates y_j

$$\bar{y} = \frac{\Sigma_j y_j}{n}; \quad \text{var}(\bar{y}) = \frac{1}{n}.\left[\frac{\Sigma_j (y_j - \bar{y})^2}{(n-1)}\right]$$

provides an estimate of the variance of the same estimate from the full sample, exactly for a linear estimate y or approximately for a non-linear estimate $\tilde{y}$. A somewhat conservative estimator (giving a higher value) is obtained by replacing $\bar{y}$ in the summation by $\tilde{y}$; that is, by writing

$$\text{var}(\tilde{y}) = \frac{1}{n}.\left[\frac{\Sigma_j (y_j - \tilde{y})^2}{(n-1)}\right]$$

The above may be modified to incorporate the finite population correction if that is important.

Constructing independent replications

It is useful to begin by stating the requirements which should be met ideally in applying the procedure. The basic requirement is that the parent (full) sample be composed of a number of independent sub-samples or replications, each with the same design and procedures but selected and implemented independently. The requirement of common and independent procedures in constructing the replications from the sample applies to sample selection as well as to data collection and estimation.

Sample selection. The replications should be designed according to the same sample design, on the basis of the same frame and type of units, system of stratification, sampling stages and selection methods, etc. as the parent sample. In drawing several replications from the same population, independence requires that each replication is replaced into the frame before the next is drawn and the randomized selection procedure is applied separately for each selection.

Data collection. Following sample selection, the procedure for data collection should be the same and applied independently for each replication. Data collection refers to various steps in the whole measurement process, including questionnaire design, staff recruitment and training, mode and procedures for data-collection, fieldwork organization, supervision and control, recording and coding of responses, data entry, and so on. Independent application of common data collection procedures requires, for example, that independent sets of field staff (supervisors, interviewers, coders, etc.), drawn in principle from a common pool, are used for different replications.

Estimation. A common estimation procedure refers not only to the mathematical form of the particular estimator used, but also to all the other steps involved in computing the final estimates from the survey data - steps such as data editing, imputation, treatment of outliers, weighting and other adjustments. Independent application means that all steps in the estimation procedure are applied separately to each replication. For example, if the sample data are weighted to agree with certain population control totals, it is implied that the relevant weights are determined independently for each replicated sub-sample, as distinct from using a common set of weights determined on the basis of the full data set.

Brief comments on various aspects of the application of the procedure in practice follow.

Approximation to independent replications

In practice the above requirements are rarely met exactly. For instance, if the estimation procedure is complex, repeating all its steps for each replication can be too expensive and time consuming. Thus, while separate estimates are produced for each replication (as must be done for the variance estimation procedure to be applied), some steps in the procedure - imputation, weighting, adjustment of the results against external control totals, etc. - are applied only once to the sample as a whole. The results can be different if these steps are applied to the sample results from each replication separately and independently. Strict independence of data collection procedures is even more difficult to implement. That would require the organization and implementation of numerous steps in the measurement process independently for each replication, possibly involving a great increase in cost and inconvenience. (Some such separation in an appropriate form can of course be useful in the assessment of non-sampling errors.)

Perhaps the most critical requirement is that the independent selection of replications should follow the same design. Ideally, the full sample may be formed by *combining* independent sub-samples. Usually, however, it is a matter of *partitioning* an existing sample into more or less independent sub-samples. It is important to note that in a multi-stage design the partitioning of the sample should be done at the primary selection level, i.e. all sample elements within a PS should be assigned to the same replication. To estimate the total variance across strata in a stratified sample, each replication must itself be a stratified sample paralleling the parent sample.

The sample may be divided into replications at the time of selection or subsequently, after selection. Consider for instance a systematic sample in which primary units are to be selected with interval I; the sample may be selected in the form of n replications, each selected systematically with a distinct random start and selection interval (n.I).

A more convenient and common alternative is to select the full sample in one operation, but in such a way that it can subsequently be divided into sub-samples which are by and large independent and reflect the design of the full sample. As an example, let us take a systematic sample of 500 PSUs to be divided into 20 replications each of size 500/20 = 25 PSUs. One may imagine an ordered list of the 500 sample units divided into 25 "zones", each comprised of 20 adjacent units. A replication would consist of one unit taken from each zone: for instance, the first unit from each zone forming the first replication, the second unit from each zone forming the second replication, and so on. In fact, such a simple scheme can be applied with great flexibility and permits many straightforward variations. The units in the full sample may have been selected with uniform or varying probabilities; the above sub-sampling scheme retains the original relative probabilities of selection. If the original sample is stratified, one may order the selected units stratum after stratum and divide the entire list into equal zones for the application of the above procedure. The effect of original stratification will be reflected in the replications if the number of units to be selected is large enough for all or most strata to be represented in each replication. Alternatively, units may be cross-classified by zone and stratum, i.e. each stratum is divided into a number of zones and each zone while each zone is formed by linking units (sample PSUs) across a number of strata. Deming (1960) provides many examples and extensions of such procedures.[57]

Choice of the number of replications

In most multi-stage designs the number of primary selections involved is limited, which constrains the number of replications into which the sample may be divided. There are of course samples in which the total number of PSs available is so inadequate that the number of replications and the number of units per replication both have to be rather small. In that situation the method of independent replication is inappropriate for variance estimation. However, when the sample design permits, choice still has to be made between the extremes of having many small replications, or having only a few but large replications. If many replications are created, the number of PSs per replication may become too small to reflect the structure of the full sample. This will tend to bias the variance estimation. On the other hand, variances estimated from only a small number of replications tend to be unstable, i.e. they themselves are subject to large variance. There is no agreement as to the most appropriate choice in general terms. Kish (1965, Section 4.4), for example, summarizes the situation as follows:

> "Mahalanobis .. and Lahiri .. have frequently employed 4 replicates... Tukey and Deming .. have often used 10 replicates... Jones .. presents reasons and rules for using 25 to 50 replicates. Generally I too favour a large number, perhaps between 20 and 100."[58]

The primary argument in favour of having many replications (each necessarily comprised of a correspondingly small number of units) is that the variance estimator is more precise and the statistic $\bar{y}$ (average of the replicated estimates) is more nearly normally distributed. The precision of the variance estimator decreases as the number of replications is reduced. Furthermore, for a given value of variance or standard error, the

[57] Deming, W.E. (1960). *Sample design in business research*. Wiley.
[58] Kish, L. (1965). *Survey sampling*. Wiley.

interval associated with any given level of confidence becomes wider. Another consideration is that with a small number of replications it is necessary to assume that the individual replicated estimates y_j are normally distributed, though mild departures from normality are generally not important; fortunately the assumption of normality of y_j improves as the number of primary selections per replication is increased. In any case, when the number of replications is large, it is necessary to assume only that the mean $\bar{y}$ is normally distributed.

On the other hand, having fewer and larger replications also has some statistical advantages.

1. With large sample size per replication, the individual replicated estimates y_j are more stable and more normally distributed. This helps in inference.

2. The replicated estimates y_j and even more so their average $\bar{y}$ are closer to the estimate $\tilde{y}$ based on the full sample for non-linear statistics as well. This facilitates extension of the method of variance estimation to non-linear statistics, which is the main justification for its use.

3. Most importantly, increasing the number of primary units per replication makes it easier to reflect the structure of the full sample in each replication, which reduces the bias in the variance estimator.

   Some practical considerations and wider objectives when opting for a smaller number of replications, each consequently larger and potentially more complex in design, should also be noted.

4. With fewer replications, there is less disturbance of the overall design as a result of the need to select the sample in the form of independent replications.

5. The additional cost and difficulty involved in separate measurement and estimation is less.

6. With a smaller number of replications, it is more feasible to randomize appropriately the work allocation of interviewers, coders, etc. in measuring the non-sampling components of variance

7. Replicated or “interpenetrating“ designs can be useful for more general checking of survey procedures and results. These objectives are better served when the number of replications to be dealt with is small.

8. The same is true of displaying the survey results separately by replication to convey to the user a vivid impression of the variability in sample survey results.

It is of course also possible to have a larger number of replications for more stable sampling error estimation, and collapse them into a smaller number for objectives 6, 7 and 8.

In view of the conflicting considerations and opinions noted above, it is not possible to make specific recommendations as to the appropriate choice of number and size of replications. With for example 100-1000 PSUs in the sample, a simple rule which has been found reasonable is to begin by making both the number of replications and the number of sample

PSUs per replication equal to the square-root of the given total number of PSUs in the sample. For example, with somewhat over 200 sample PSUs, this calculation would result in around 15 replications, each with 15 PSUs. Similarly, with 600 or so PSUs, one would begin by considering around 25 replications, each with 25 PSUs.

7.6.3 Method 3. Comparison among replications of the full sample based on repeated re-sampling

These procedures are more complex and computer intensive, but they can be applied to statistics of any complexity. 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. *Each replication is designed to reflect the full complexity of the parent sample.*

Unlike the procedure described in Section 7.6.2, the replications in themselves are not independent (in fact they overlap), and special procedures are required to control the bias in the variance estimates generated from comparisons among such replications.

Compared to the method described in Section 7.5.1, repeated re-sampling methods have the disadvantage of greater complexity and increased computational work. They also tend to be less flexible in the sample designs handled.

However, they do have the advantage of not requiring an explicit expression for the variance of each particular statistic. The same basic expression (given below) applies to any statistic.

Once the replications have been created, no further explicit reference is required even to the structure of the sample.

The procedures are also more encompassing; 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, as explained below.

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 "jack-knife repeated replication", the "balanced repeated replication" and the "bootstrap exit".

The first of these generally provides the most versatile and convenient method and is outlined below.

Jack-knife repeated replication (JRR)

The jack-knife repeated replication (JRR) is one method for estimating sampling errors from comparisons among sample replications that are generated through repeated resampling of the same parent sample. Each replication needs to be a representative

sample in itself and to reflect the full complexity of the parent sample. However, inasmuch as the replications are not independent, special procedures are required in constructing them to avoid bias in the resulting variance estimates.

Originally introduced as a technique of bias reduction, the jack-knife method has by now been widely tested and used for variance estimation. We prefer the JRR to similar methods such as the Balanced Repeated Replication because the JRR is generally simpler and more flexible.

The basic model of the JRR for application in the context described above may be summarized as follows. Consider a design in which two or more primary units have been selected independently from each stratum in the population. Within each primary selection unit (PSU), sub-sampling of any complexity may be involved, including weighting of the ultimate units.

In the "standard" version, each JRR replication can be formed by eliminating one PSU from a particular stratum at a time, and increasing the weight of the remaining PSU's in that stratum appropriately so as to obtain an alternative but equally valid estimate to that obtained from the full sample.

The above involves creating as many replications as the number of primary units in the sample. The computational work involved is sometimes reduced by reducing the number of replications required. For instance, the PSU may be grouped within strata, and JRR replications formed by eliminating a whole group at a time. This is possible only when a stratum contains several units. Alternatively, or in addition, the grouping of units may cut across strata. Thirdly, it is possible to define the replications in the standard way ("delete one-PSU at a time jack-knife") but actually construct and use only a subset of them.

One situation in which some grouping of units is unavoidable is when the sample or a part of it is a direct sample of ultimate units or of small clusters, so that the number of replications under "standard" JRR is too large to be practical. Normally, the appropriate procedure to reduce this number would be to form new computational units by means of random grouping of the units within strata.

Apart from the above, often in real social surveys using multi-stage sample designs, there is little need or motivation for introducing this type of shortcut or approximation to the standard JRR.

Briefly, the standard JRR involves the following.

Theoretical basis

The theoretical basis involves generalizing from linear statistics, such as a simple aggregate, to statistics of any complexity.

If we consider a replication formed by dropping a particular PSU i in stratum h and compensate by increasing the weight of the remaining (a_h-1) PSUs in that stratum appropriately, the estimate for a simple aggregate (total) for this replication is

$$y_{(hi)} = \sum_{k \neq h} y_k + \frac{a_h}{a_h - 1}.(y_h - y_{hi}) = y - \frac{a_h}{a_h - 1}.(y_{hi} - \frac{y_h}{a_h}).$$

With the average of these estimates over the stratum, $y_{(h)} = \sum_i y_{(hi)} / a_h$,

and the average over all a=Σah replications, $\tilde{y} = \frac{\sum_h \sum_i y_{(hi)}}{\sum_h a_h}$,

the expression for variance of any statistic can be written in a number of statistically equivalent forms:

$$var_1(y) = \sum_h [(1 - f_h).\frac{a_h - 1}{a_h}.\sum_i (y_{(hi)} - y_{(h)})^2]$$

$$var_2(y) = \sum_h [(1 - f_h).\frac{a_h - 1}{a_h}.\sum_i (y_{(hi)} - \tilde{y})^2]$$

$$var_3(y) = \sum_h [(1 - f_h).\frac{a_h - 1}{a_h}.\sum_i (y_{(hi)} - y)^2]$$

The standard variance form for simple aggregate y

$$var(y) = \sum_h \left[(1 - f_h).\frac{a_h}{a_h - 1}.\sum_i (y_{hi} - \frac{y_h}{a_h})^2 \right]$$

is replaced in the JRR method by one of the above three expressions (usually the last of the three, var_3, as it is more conservative). Being based on nearly the full sample, estimates like $y_{(hi)}$, $y_{(h)}$ and even more so their overall average $\tilde{y}$ are expected to be close to the full sample estimate y, even for complex statistics. In this way their variance, expressed by any of the three forms, provides a measure of variance of y as well. *This applies to statistics y of any complexity*, not only to a simple aggregate. By contrast, the standard form var(y) applies only when y is a simple aggregate, in which case all the forms mentioned above are equivalent.

Application to complex statistics

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 an appropriate factor g_h (see below). Let $z_{(h)}$ be the simple average of the $z_{(hi)}$ over the a_h sample units in h. The variance of z is then estimated as

$$var(z) = \Sigma_h \left[(1 - f_h).\frac{a_h - 1}{a_h}.\Sigma_i (z_{(hi)} - z_{(h)})^2 \right]. \quad (1)$$

A major advantage of a procedure like the JRR is that, under general conditions for the application of the procedure, the same and relatively simple variance estimation formula holds for z of any complexity.

Normally, the factor g_h is taken as 2.a:

$$g_h = a_h / (a_h - 1), \quad (2.a)$$

but for reasons noted below it is preferable to use the 2.w version:

$$g_h = w_h / (w_h - w_{hi}), \quad (2.b)$$

where $w_h = \Sigma_i w_{hi}$, $w_{hi} = \Sigma_j w_{hij}$, the sum of sample weights of ultimate units j in primary selection i.

The 2.w form is proposed because it retains the total weight of the sample cases unchanged across the replications created. With the sample weights scaled in such a manner that their sum is equal (or proportional) to some more reliable external population total, population aggregates from the sample can be estimated more efficiently, often with the same precision as proportions or means.

Another possible variation which may be mentioned is to replace $z_{(h)}$, the simple average of the $z_{(hi)}$ over the a_h sample units in h, by the *full-sample* estimate z_h for stratum h. In this case, if $z_{(h)}$ is the simple average of the $z_{(hi)}$ over the a_h sample units in h, then the variance of z is estimated as:

$$\mathrm{var}(z) = \Sigma_h \left[(1 - f_h) \cdot \frac{a_h - 1}{a_h} \cdot \Sigma_i (z_{(hi)} - z_h)^2 \right]. \quad (3)$$

Version (1) tends to provide a "conservative" estimate of variance, but normally the difference with this last version is small.

The JRR variance estimates take into account the effect on variance of aspects of the estimation process which are allowed to vary from one replication to another. In principle this can include complex effects of imputation and weighting, for example. Often in practice it is not possible to repeat such operations from the start at each replication.

7.7 Application of the methods in practice

Though the basic assumptions regarding the structure of the sample for applying the method are met reasonably well in many large-scale household surveys, often they are not met exactly. Practical solutions can however be found in most situations. Here are some of the most common ones.

Problem	Common practical solution
Systematic sampling of primary units, which is a common and convenient procedure, does not strictly give a minimum of two independent primary selections per stratum.	Pairing of adjacent units to form strata to be used in the computations is the usual practice. It is assumed that the units paired have been selected independently within the stratum so defined.
Similarly, stratification is often carried out to a point where only one or even less than one primary unit is selected per stratum.	This requires "collapsing" of similar strata to define new strata, so that each contains at least two selections which are then assumed to be independent.
Sometimes the primary units are too small, variable or otherwise inappropriate to be used directly in the variance estimation formulae.	More suitable computational units may be defined by such techniques as the random grouping of units within strata, and the linking or combining of units across strata.
Samples are normally selected without replacement.	For populations of large size sampled at a low rate, this is hardly ever a problem in population-based surveys.

How to ensure that sampling errors can be and are computed?

1. *Use probability and measurable samples.* Probability sampling means the use of randomized procedures in sample selection that assign a calculable non-zero probability to each element in the population. Measurability refers to a set of practical criteria that allow the computation, from the sample itself, of valid estimates or approximations of its sampling variability.

2. *Define the sample structure appropriately to ensure "measurability" in the practical sense.* Generally, the available practical variance estimation procedures require the presence of two or more primary selections in each stratum, selected independently and with replacement. In practical application, it is sometimes necessary to redefine the actual structure to fit this model. Examples have been given in the table above. Such redefinition of computing units necessarily requires access to sampling expertise and experience.

3. *Ensure that all information on the structure of the sample needed for variance computation is readily available. It is most desirable to code this as an integral part of the survey data file.*

4. *Use general and simple procedures where possible, which can provide useful variance estimates for many different types of design, variable and subclass.* The use of more specific and complex procedures, even if more accurate, is secondary to the production of at least approximate estimates of sampling errors for many different types of estimate and subgroup.

5. *Aim at large-scale computation, analysis and "modelling" of sampling errors.* It is generally insufficient to confine the computations to a few arbitrarily selected statistics from among the thousand or so that may be produced for different variables, measures and subgroups. The pattern and magnitude of sampling errors can vary greatly for different statistics in the same survey. Hence, routine and large-scale computation is normally desirable. At the same time, however, it is necessary to explore the *patterns of variation of the sampling errors* (standard error as well as derived measures such as coefficient of variation, design effect and intra-cluster correlation) for various statistics, computed over different sampling domains and sub-populations.

7.8 Portable measures of sampling error

The magnitude of the standard error of a statistic depends on a variety of factors such as:

- the nature of the estimate
- its units of measurement (scale) and magnitude
- variability among elements in the population (population variance)
- sample size
- the nature and size of sampling units
- sample structure, sampling procedures
- estimation procedures.

Consequently, the value of the standard error for a particular statistic is specific to the statistic concerned. In order to relate the standard error of one statistic to that of another, it is necessary to decompose the error into components from which the effect of some of the above factors has been removed, i.e. components that are more stable or "portable" from one type of statistic or design to another. The standard error of a statistic such as a mean is written in several forms, in terms of measures which are more portable in the above sense. These are described below.

Relative standard error

$$se(\bar{y}) = \bar{y}.rse(\bar{y})$$

This refers to the standard error of an estimate, divided by the value of the estimate. It removes the effect on the standard error of the magnitude and scale of measurement of the estimate, but it still depends on other factors such as sample size and design.

Standard error in an equivalent simple random sample (SRS); Population variance

The standard error of a statistic estimated from a complex sample can be factorized into two parts:

(1) sr The standard error which would have been obtained in a simple random sample of the same size

(2) deft The design factor, summarizing the effect of design complexities

$$se(\bar{y}) = deft.sr(\bar{y})$$

The second component *(sr)* is independent of the sample design and relates to the sample size in a very simple way:

$$sr(\bar{y}) = s/\sqrt{n},$$

where *s*, the standard deviation, is a measure of variability in the population, independent of sample design or size. The scale of measurement can also be removed by considering the coefficient of variation, *cv*:

$$s = \bar{y}.cv.$$

Standard deviation is a useful and highly portable measure. Furthermore, it can be estimated in a simple way irrespective of complexities of the design in most practical situations. For example, for a proportion *p*,

$$s^2 = \frac{n}{n-1}.p(1-p) \approx p(1-p),$$

while more generally, for a weighted ratio *r*, we have

$$s^2 = \tfrac{n}{n-1}.\textstyle\sum_i w_i z_i^2 / \sum_i w_i;\ z_i = (y_i - r.x_i)/\bar{x},$$

where

$$r = \sum_i w_i y_i / \sum_i w_i x_i; \quad \bar{x} = \sum_i w_i x_i / \sum_i w_i.$$

The coefficient of variation is more portable, but it is not so useful when the denominator in its definition is close to zero, as may happen for estimates of differences between subclasses. Also, there is generally no advantage to going from s to *cv* in the case of proportions; in fact the former is preferable since it is symmetrical (the same) for a proportion p and its complement 1-p.

The design effect and rate of homogeneity

The *design effect*, deft^2 (or its square-root, deft, which is sometimes called the *design factor*), is a comprehensive summary measure of the effect on sampling error of various complexities in the design. By taking the ratio of an actual standard error to a simple random sample (SRS) standard error, deft removes the effect of factors common to both, such as the size of the estimate and the scale of measurement, the population variance and the overall sample size. However, for a given variable, its magnitude still depends on other features of the design.

A major factor determining the deft value is the size of the sample taken per PSU. When the PSU sample sizes do not vary greatly and the sample is essentially self-weighting, the effect of these sample sizes can be isolated by considering the more portable measure "roh":

$$\text{deft}^2 = 1 + (\bar{b} - 1).\text{roh}.$$

In practice, the design effect for a statistic is computed by estimating its variance (i) under the actual sample design, and (ii) assuming a simple random sample of the same size. The ratio of these two quantities gives deft^2. Parameter roh can be estimated from this deft^2 and the average number of ultimate units selected per sample PSU, using the formula given above.

The above process can be refined by isolating some other sources of variation. For example, in the presence of variable PSU sample sizes, it is more appropriate to replace their simple average in the above expression by the quantity

$$\bar{b'} = \bar{b}.(1 + cv^2) = \sum_i b_i^2 / \sum_i b_i.$$

Another useful refinement is to isolate the effect on deft of D_w, the inflation in variance resulting from arbitrary departures from a self-weighting design, introduced earlier (see Section 7.1.3):

$$\text{deft}^2 = D_w^2.[1 + (\bar{b'} - 1).\text{roh}].$$

It is most important to accumulate information on design effects, as such information is essential for improving sample design for future surveys.

7.9 How big should the sample-take per cluster be?

It was noted in Chapter 2 that, from an examination of the types of sample hitherto used in child labour surveys (see Tables 2.1 and 2.2), we find a surprisingly large range of variation in the cluster sizes (sample-takes per area) used in child labour surveys. The range of variation is more than 10 to 50 households per area. While some of this variation may reflect differing national circumstances and differences in the type of units involved, it is likely that much of it is not based on real statistical or cost differences.

By computing sampling errors and design effects and collecting information on the distribution of field costs between and within areas, the sample designs for child labour surveys need to be made more efficient.

An appropriate choice of the number of ultimate units (households, children) per sample area depends on the structure of survey costs and the degree of heterogeneity (of the variable of interest) within sample clusters. By the cost structure, we mean essentially the relative cost of obtaining the information (including data collection and processing) per ultimate unit, compared to the average cost of enumerating one sample area (additional to the cost of enumerating ultimate units). The degree of heterogeneity depends on a host of factors, among them the type of units used as the primary sampling units, the method of sub-sampling within those PSU's, and, most important, the particular survey variable being considered.

The following simple expression for the "optimum cluster size" is very useful as an indication of the factors on which the choice of the number of ultimate units per sample area depends, and of the manner of that dependence:

$$b_{OPT} = \left(\frac{C.(1 - roh)}{c. \cdot roh} \right)^{0.5}.$$

Here C is the cost per sample area, and c the cost per ultimate unit in the sample; roh, defined earlier, is the "intra-cluster correlation", measuring the relative degree of heterogeneity within sample areas. For surveys with high interview costs (large C), we should have smaller takes per cluster. When the cost of bringing in new areas into the sample is high (large c), the optimum sample take is increased – in other words, it is better to have fewer bigger clusters. The parameter roh is normally a small number (0.02-0.20) for most variables, so that for variables with large roh the optimum sample take is smaller. Note that only the square-root of these parameters is involved in the above expression. This means that, fortunately, the result is not too sensitive to the value of the parameters chosen. Often good compromises suffice in practice.

References

- Adames, Yadira, Aquino Cornejo, Margarita, Castillo, Roberto and Rodriguez, Alexis. 2003. *National report on the results of the child labour survey in Panama* (San José, ILO).
- Arnold-Talbert, Elizabeth and Constanza-Vega, Leticia. 2003. *Child Labour in Belize: A Statistical Report 2001* (San José, ILO and Central Statistical Office, Government of Belize).
- Deming, W.E. 1960. *Sampling design in business research* (New York, John Wiley).
- Department of Census and Statistics of Sri Lanka. 1999. *Child Activity Survey Sri Lanka 1999* (Colombo, Department of Census and Statistics, Ministry of Finance and Planning).
- Dobles Trejos, Cecilia and Pisoni, Rodolfo. 2003. *National Report on the Results of the Child and Adolescent Labour Survey in Costa Rica* (San José, ILO).
- Fox, Kristin. 2004. *Report of Youth Activity Survey 2002* (Kingston, Statistical Institute of Jamaica).
- Ghana Statistical Service. 2003. *Ghana Child Labour Survey* (Accra, Ghana Statistical Service).
- IPEC. 2004. Child labour statistics: *Manual on methodologies for data collection through surveys* (Geneva, ILO).
- IPEC. 2004. *National Report on the Results of the Child and Adolescent Labour Survey in Costa Rica, 2003* (San José, ILO).
- Kish. L. 1965. *Survey sampling* (New York, John Wiley).
- Kish, L. 1980. "Design and estimation for domains" *The Statistician,* vol. 29.
- Kish, L. 1987. *Small area statistics: An international symposium* (New York, John Wiley).
- Marschatz, Astrid. 2004. *Report on the results of the National Child labour Survey in the Dominican Republic* (San José, ILO and State Department of Labour of the Dominican Republic).
- Ministry of Public Service, Labour and Social Welfare. 1999. *National Child Labour Survey, Country Report* (Harare, Ministry of Public Service, Labour and Social Welfare, Zimbabwe).
- National Institute of Statistics of Cambodia. 1997. *Report on Child Labour in Cambodia 1996* (Phnom Penh, National Institute of Statistics, Cambodia).
- National Statistical Office of Mongolia. 2004. *Report on National Child Labour Survey 2002-2003* (Ulaanbaatar, National Statistical Office of Mongolia).
- Orkin, FM. 2000. *Survey of Activities of Young People* (SAYP) (Pretoria, Statistics South Africa).

- Portugal Ministry of Labour and Solidarity (MTS).1998. *Child labour in Portugal: Social Characterisation of School Age Children and Their Families* (Lisbon, Portugal Ministry of Labour and Solidarity (MTS).
- Purcell, N.J. and L. Kish. 1980. "Postcensal estimates for local areas (or domains)". *International Statistical Review*, vol. 48.
- Rao, J.N.K. 2003. *Small area estimation* (New York, John Wiley).
- Shaik, Khaled Rajeh. 2000-2001. *Child Labour in Yemen* (Sana'a, Ministry of Labour of Yemen).
- SIS. 1997. *Child Labour 1994* (Ankara, State Institute of Statistics Prime Ministry, Republic of Turkey, 1997).
- State Statistics Committee of Ukraine. 2001. *Child Labour Ukraine1999, Statistical Bulletin* (Kyiv, ILO and State Statistics Committee of Ukraine).
- Thompson, S.K. 1990. *Adaptive cluster sampling* (Journal of the American Statistical Association).
- —. 1992. *Sampling* (New York, John Wiley).
- Thompson, S.K., Seber. G.A.F. 1996. *Adaptive Sampling* (New York, John Wiley).
- Thompson SK. Collins LM. 2002. "Adaptive sampling in research on risk-related behaviours". *Drug and Alcohol Dependence* vol. 68.