International Labour Organization

ILO-IPEC Interactive Sampling Tools No. 6

Calculation of sampling weights

Version 1

August 2014

International Programme on the Elimination of Child Labour (IPEC)

Fundamental Principles and Rights at Work (FPRW) Branch
Governance and Tripartism Department

ILO-IPEC

ILO-IPEC Interactive Sampling Tools No. 6 - Calculation of sampling weights / International Labour Office, International Programme on the Elimination of Child Labour (IPEC) - Geneva: ILO, 2014

ACKNOWLEDGEMENTS

This publication was elaborated by Mr. Farhad Mehran, consultant, for ILO-IPEC and coordinated by Mr. Federico Blanco Allais from IPEC Geneva Office.

Funding for this ILO publication was provided by the United States Department of Labor (Projects GLO/13/21/USA & GLO/10/55/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

Available in electronic PDF format only.

Photocomposed by ILO-IPEC Geneva.

1. Introduction

This document describes the use of the template “Calculation of sampling weights”. The template assists the user to calculate the sampling weights for extrapolation of the sample results to the population totals. The calculation of the sampling weights takes into account the probabilities of selection according to the sample design, the extent of non-response among the sample households and any population aggregates to which the survey results should conform.

The template is divided into three parts: Input values, Output values and Intermediary calculations. The contents and use of each part is described in turn below.

2. Input values

The input values provide information on the sample design and response rate of the survey. They also provide data on auxiliary variables for use in calibration of the sampling weights. The input values are provided for each sample PSU separately and are entered by column as shown in Diagram 1 below.

The first two columns are for identifiers. The identifier of the sample PSU is specified in Column (1) and that of the PSU stratum in Column (2). In the next 3 columns, data on the first stage of sampling are provided: Column (3) specifies the total number of sample PSUs that were selected in the stratum; Column (4) the total number of households in all PSUs of the stratum according to the sampling frame; and Column (5) the number of households in the specific sample PSU also according to the sampling frame.

Diagram 1. Input values on sample design

INPUT VALUES						
Identifiers		**First stage sampling**			**Second stage sampling**	
PSU Code	Stratum code	Number of sample PSUs in stratum (frame)	Total number of hslds in stratum (frame)	Number of hslds in PSU (frame)	Number of hslds in PSU (listing)	Number of sample hslds in PSU
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1111	111	7	52069	137	180	16
1112	111	7	52069	173	271	16
1113	111	7	52069	147	179	16
1114	111	7	52069	124	144	16
1115	111	7	52069	140	143	16
1116	111	7	52069	157	192	16
1117	111	7	52069	163	204	16
2111	211	9	253183	126	152	16
2112	211	9	253183	121	132	16
2113	211	9	253183	131	159	16
2114	211	9	253183	121	123	16
2115	211	9	253183	105	143	16

The next two columns provide data on the second stage of sampling. In Column (6), the number of households in the PSU obtained from the listing operations is entered. If the PSU is not freshly listed to update the sampling frame, column (6) should in principle be equal to the total number of households in the PSU according to the sampling frame as given in column (5). In the numerical example given in Diagram 1, the first row indicates that the listing operations found 180 households in the PSU against 137 households in the sampling frame.

The last column in Diagram 1 indicates the sample-take or the number of sample households drawn in the sample PSU. This value is generally a fixed number, constant for all PSUs. In the numerical example here the value is 16 households.

In the next diagram (Diagram 2 below), data on households' response and auxiliary variables for adjustments of the design weights are entered. The number of responding sample households are recorded for each sample PSU in column (8). In the numerical example of diagram 2, the row of column (8) indicates that there are 15 responding sample households out of the 16 sample take in PSU 001.

The next six columns of Diagram 2 are input values on the auxiliary variables for calibrating the sampling estimates to known aggregates obtained from external sources. Here, six auxiliary variables are envisaged on the sex and age composition of the children population. The sample data for each PSU are entered in columns (9) to (14). Thus, according to the first row of Diagram 2, among the members of the 15 responding sample households, there were 3 boys in the age category 5-11 years old, 1 boy 12-14 years old, and 1 boy 15-17 years old. Similarly, there were 3 girls in the age category 5-11 years old, 2 girls 12-14 years old, and 1 girl 15-17 years old.

Diagram 2. Input values on households' response and auxiliary variables

	KNOWN AGGREGATES OF AUXILIARY VARIABLES					
	689430	433726	359844	656479	397081	383005
Response	Auxiliary variables for calibration					
Number of sample hslds response	Boys 5-11 years	Boys 12-14 years	Boys 15-17 years	Girls 5-11 years	Girls 12-14 years	Girls 15-17 years
(8)	(9)	(10)	(11)	(12)	(13)	(14)
15	3	1	1	3	2	1
14	2	2	2	4	2	1
15	5	3	1	4	2	2
16	2	2	1	3	1	1
14	4	1	2	3	2	2
14	4	2	2	3	2	2
14	4	2	1	3	2	2
16	2	1	2	2	2	2
15	4	2	1	3	2	1
16	3	1	3	2	2	2
16	5	3	1	4	2	2
16	2	2	1	3	2	2

The know aggregates from external sources are recorded in the top of the column headings as shown Diagram 2. In general, the external sources are population estimates of the sex and age composition of the children population projected to the reference period of the survey. Thus in this numerical example, as indicated in the top panel of Diagram 2, sampling weights should be calibrated to the population projections: 689,430 boys in the age category 5-11 years old, 433,726 boys 12-14 years old, and 359,844 boys 15-17 years old, and 656,479 girls in the age category 5-11 years old, 397,081 girls 12-14 years old, and 383,005 girls 15-17 years old.

3. Output values

There are three columns of output values as shown in red in columns (15) to (17) of Diagram 3 below. Column (15) gives the sampling weights defined as the inverse of the probability of selection of households in the given sample PSU. The adjusted sampling weights for non-response are given in column (16) and the final calibrated weights to the known totals in column (17).

In the numerical example of Diagram 3, the first row indicates that sample households in the first PSU 001 represent 611 households in the aggregate population according to the sample design. The weights of the responding households are increased to account for the non-responding households in the sample PSU. The final sampling weight of the responding households in PSU 001 is further adjusted to 660 for ensuring agreement to the population totals obtained from population projections mentioned earlier.

Diagram 3. Output values: sampling weights

OUTPUT VALUES		
Weights		
Design weight	Weight adjusted for non-response	Final calibrated weight
(15)	(16)	(17)
611	652	660
728	832	1054
566	604	781
540	540	754
475	543	749
569	650	826
582	665	589
2121	2121	1257
1918	2046	2401
2134	2134	2665
1787	1787	2311
2395	2395	826

Note: Values in columns 15, 16 and 17 are calculated values and should not be altered.

It should be mentioned that the template calculates the same sampling weight for all households and for that matter all household members and individual records in the given sample PSU.

4. Intermediary calculations

The intermediary calculations involve the calculations of the selection probabilities, the response rates and the calibration factors of the auxiliary variables.

The selection probabilities and the response rates are calculated in columns (18) to (21) as shown in Diagram 4 below. In the first stage of sampling, the probability of drawing a particular PSU i in stratum h is given by

$$p_{ih} = m_h \frac{x_{ih}}{\sum_{j \varepsilon s_h} x_{jh}}$$

where m_h is the number of sample PSU selected from stratum h and x_{ih} is the measure of size of PSU i in stratum h. Column (18) gives the results calculated from the input values,

$$col(18) = \frac{col(4) \times col(6)}{col(5)}$$

Diagram 4. Intermediate calculations of sampling and response probabilities

INTERMEDIATE CALCULATIONS

Probabilities			
Probability of selection of sample PSU	Probability of selection of sample hsld in PSU	Overall probability of selection of sample hsld	Overall response rate
(18)	(19)	(20)	(21)
0.0184	0.0889	0.0016	0.9375
0.0233	0.0590	0.0014	0.8750
0.0198	0.0894	0.0018	0.9375
0.0167	0.1111	0.0019	1.0000
0.0188	0.1119	0.0021	0.8750
0.0211	0.0833	0.0018	0.8750
0.0219	0.0784	0.0017	0.8750
0.0045	0.1053	0.0005	1.0000
0.0043	0.1212	0.0005	0.9375
0.0047	0.1006	0.0005	1.0000
0.0043	0.1301	0.0006	1.0000
0.0037	0.1119	0.0004	1.0000

In the second stage of sampling, the probability of drawing a particular household k from sample PSU I in stratum h is given by

$$p_{k|ih} = \frac{b}{n'_{ih}}$$

where b is the fixed number of households drawn in a PSU (the sample-take) and n'_{ih} the number of households listed in PSU i of stratum h. The calculation is made in column (19) on the basis of the input values as follows,

$$col(19) = \frac{col(7)}{col(6)}$$

The overall probability of selection of a sample household k in PSU i of stratum h is then obtained as the product of the first-stage and second-stage probabilities of selection,

$$p_k = p_{ih} \times p_{k|ih}$$

or, equivalently,

$$col(20) = col(18) \times col(19)$$

The design weights in the output values are then obtained as the inverse of the overall probability of selection

$$DesignWeight = \frac{1}{p_k}$$

or, equivalently,

$$col(15) = \frac{1}{col(20)}$$

The next intermediary calculation is the calculation of the response rate in each PSU,

$$RR_{ih} = \frac{b'_{ih}}{b}$$

where b'_{ih} is the number of responding sample households in PSU i in stratum h and b is the sample take in the PSU. The calculations are made in column (21) as follows,

$$col(21) = \frac{col(8)}{col(7)}$$

The resulting response rate is then used to adjust the design weight for non-response households. Thus,

$$DesignWeightAdjustedForNon-response = \frac{DesignWeight}{RR}$$

or, equivalently,

$$col(16) = \frac{col(15)}{col(21)}$$

The design weights adjusted for non-response are further adjusted to conform to known results on auxiliary variables. This process of adjustment is called calibration. Calibration means using calibrated weights such that the application of these weights to the auxiliary variables will give estimates exactly equal to the known population totals on those auxiliary variables.

Suppose that associated to each population element k, there is a vector of J auxiliary variables x_k with values $x_k = (x_{k1}, \ldots, x_{kj}, \ldots, x_{kJ})'$. The corresponding population total is given by the known vector $t_x = \Sigma_{k\varepsilon U}\, x_k$ Applying the extrapolation weights d_k to the sample values of the auxiliary variables gives

$$t_{x\pi} = \Sigma_{k\varepsilon s}\, x_k/\pi_k = \Sigma_{k\varepsilon s}\, d_k\, x_k.$$

which may differ from the known population values t_x.

Deville and Särndal (1992)[1] have shown that the extrapolation weights may be adjusted by minimizing the expected average distance between the adjusted weights (w_k) and the original weights (d_k) to obtain the following adjusted weights that conform to the known population totals of the auxiliary variables

$$w_k = d_k (1 + q_k x_k'\lambda)$$

where

$$\lambda = T_s^{-1}(t_x - t_{x\pi})$$

and

$$T_s = \Sigma_s d_k q_k x_k.x_k'.$$

where the parameters q_k are generally set to one (q_k=1).

The sample values of the auxiliary variables x_k are entered as input values in columns (9) to (14) of Diagram 2. The known population totals, t_x, are also given as input values in Diagram 2(values in at the top of the panel). In the intermediary calculations, the auxiliary variables are first weighted by the square-root of the design weights adjusted for non-response. The weighted auxiliary variables are calculated in columns (22) to (27) in Diagram 5 below,

$$col(8+j) = \sqrt{col(16)} \times col(21+j) \qquad j = 1,\ldots,6.$$

[1] Särndal, Carl-Erik, and Jean-Claude Deville, "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association*, June 1992, Vol. 87, No. 48, pp. 376-382.

Diagram 5. Intermediary calculations on auxiliary variables

INTERMEDIARY CALCULATIONS					
Weighted auxiliary variables					
Sqrt(weight$_k$ adjusted for non-response)X$_k$					
(22)	(23)	(24)	(25)	(26)	(27)
77	26	26	77	51	26
58	58	58	115	58	29
123	74	25	98	49	49
46	46	23	70	23	23
93	23	47	70	47	47
102	51	51	76	51	51
103	52	26	77	52	52
92	46	92	92	92	92
181	90	45	136	90	45
139	46	139	92	92	92
211	127	42	169	85	85
98	98	49	147	98	98

Then, the square matrix of cross-classified weighted auxiliary variables (T) is calculated in Excel array format as shown in the top rows of columns (28) to (33) of Diagram 6 below. The inverse of the matrix (T^{-1}) is then calculated as reproduced in the next six rows of columns (28) to (33) of Diagram 6.

Diagram 6. Intermediary calculations of calibrated factor

INTERMEDIARY CALCULATIONS							
Matrix of cross-classified weighted auxiliary variables $T = \sum_k w_k X_k X'_k$						**Difference between known and estimated aggregates**	**Calibration factor**
(28)	(29)	(30)	(31)	(32)	(33)	(34)	(35)
2289381	1406020	1061459	2079121	1309416	1204870		
1406020	1001247	719950	1351063	849162	809523		
1061459	719950	692063	1016355	686164	672696		
2079121	1351063	1016355	2031191	1235256	1151951		
1309416	849162	686164	1235256	811458	739861		
1204870	809523	672696	1151951	739861	747757		
T^{-1}						t_x-t_w	λ
0.0000086	-0.0000001	0.0000034	-0.0000038	-0.0000089	-0.0000021	40632	0.2701
-0.0000001	0.0000125	0.0000011	-0.0000036	-0.0000033	-0.0000054	-2	-0.1101
0.0000034	0.0000011	0.0000163	0.0000028	-0.0000133	-0.0000125	8158	0.3857
-0.0000038	-0.0000036	0.0000028	0.0000113	-0.0000065	-0.0000035	32510	0.2477
-0.0000089	-0.0000033	-0.0000133	-0.0000065	0.0000389	0.0000013	-2190	-0.7637
-0.0000021	-0.0000054	-0.0000125	-0.0000035	0.0000013	0.0000259	503	-0.2884
$\lambda = T^{-1}(t_x-t_w)$						Quartile	
0.2701	-0.1101	0.3857	0.2477	-0.7637	-0.2884	0	0

Next, the difference between the known population totals and the corresponding sample estimates based on the design weights adjusted for non-response are calculated as shown in column (34). Finally, the vector of calibration factors (λ) is calculated as shown in column (35) and in reproduced in row form at the last row of Diagram 6.

The final calibrated weights calculated as output values in column (17) are obtained as

$$col(17) = col(16) \times \sum_{j=9}^{j=14} 1 + col(j) \times col(19+j)_{ofrow(22)}$$

Generally, the size distribution of the final calibrated weights should be reviewed to identify and adjust for any extreme weight. Very large weights may over influence the survey estimates and introduce bias and additional sampling errors in the results.