International
Labour
Organization

ILO-IPEC Interactive Sampling Tools No. 3

Selection of Primary Sampling Units (PSUs) by systematic PPS sampling with constraints

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. 3 - Selection of Primary Sampling Units (PSUs) by systematic PPS sampling with constraints / 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.

The designations employed in ILO publications, which are in conformity with United Nations practice, and the presentation of material therein do not imply the expression of any opinion whatsoever on the part of the International Labour Office concerning the legal status of any country, area or territory or of its authorities, or concerning the delimitation of its frontiers.
The responsibility for opinions expressed in signed articles, studies and other contributions rests solely with their authors, and publication does not constitute an endorsement by the International Labour Office of the opinions expressed in them.
Reference to names of firms and commercial products and processes does not imply their endorsement by the International Labour Office, and any failure to mention a particular firm, commercial product or process is not a sign of disapproval.
ILO publications and electronic products can be obtained through major booksellers or ILO local offices in many countries, or direct from ILO Publications, International Labour Office, CH-1211 Geneva 22, Switzerland. Catalogues or lists of new publications are available free of charge from the above address, or by email: pubvente@ilo.org or visit our website: www.ilo.org/publns.

Visit our website: www.ilo.org/ipec

Available in electronic PDF format only.

Photocomposed by ILO-IPEC Geneva.

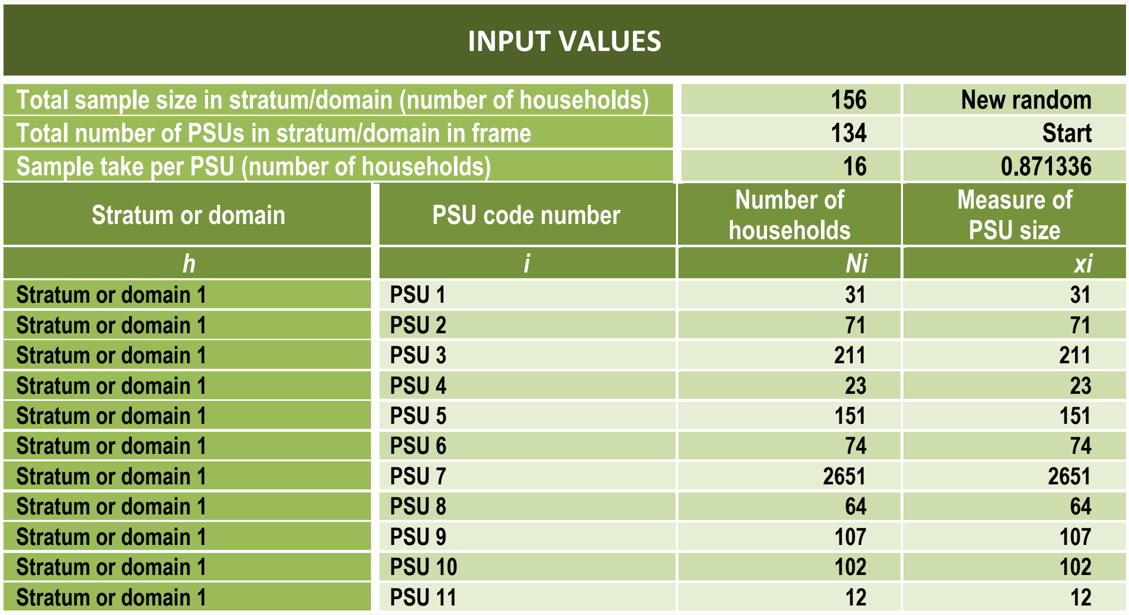
1. Introduction

This document describes the use of the template “PSU Selection” of the SIMPOC Interactive Sampling Tools. The template assists the user to select sample PSUs within a given stratum or domain by systematic sampling with probabilities proportional to size (pps) with constraints. The constraints specify the treatment of very large and very small PSUs. As many PSUs can be inserted in the template as there are PSUs in the sampling frame.

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

There are two types of input values: input parameters and input data as shown in the top and bottom panels below:



INPUT VALUES			
Total sample size in stratum/domain (number of households)		156	New random
Total number of PSUs in stratum/domain in frame		134	Start
Sample take per PSU (number of households)		16	0.871336
Stratum or domain	PSU code number	Number of households	Measure of PSU size
h	*i*	*Ni*	*xi*
Stratum or domain 1	PSU 1	31	31
Stratum or domain 1	PSU 2	71	71
Stratum or domain 1	PSU 3	211	211
Stratum or domain 1	PSU 4	23	23
Stratum or domain 1	PSU 5	151	151
Stratum or domain 1	PSU 6	74	74
Stratum or domain 1	PSU 7	2651	2651
Stratum or domain 1	PSU 8	64	64
Stratum or domain 1	PSU 9	107	107
Stratum or domain 1	PSU 10	102	102
Stratum or domain 1	PSU 11	12	12

The input parameters are few and specified manually in the top rows of the template, while the input data are generally numerous and transferred from another file in the first four columns of the template.

- ### Input parameters

Sample allocation is based on five input parameters:

n_h = Number of sample households to be selected for the specified domain or stratum h

M_h = Total number of PSU in the specified domain or stratum h in the sampling frame

b = Sample take per PSU, i.e., number of households to be sampled per PSU in the specified domain or stratum h

w_{max} = Maximum weight any PSU should receive. The pps sampling scheme may result in very large weights for units of very small size. The specification of w_{max} serves to limit the maximum weight for very small PSUs

ε_o = Random start for systematic pps sampling of PSU. Random number between 0 and 1

- ## Input data

The input data are in the form of four columns:

Col A h = The domain or stratum name. In the present version, the template concerns a unique domain or stratum. The template is applied separately for each domain or stratum. In a later version of the template, it is envisaged to use the template simultaneously to all domains and strata

Col B i = PSU code number. The list of all PSUs in the frame for the specified domain or stratum identified by their PSU code number

Col C N_i = Total number of households in the PSU i according to information in the sampling frame

Col D x_i = Measure of size to be used for systematic pps sampling. The measure of size may be the same as the number of households (Ni) or it may be specified differently, for example, the number of children 5 to 11 years old according to the information in the sampling frame

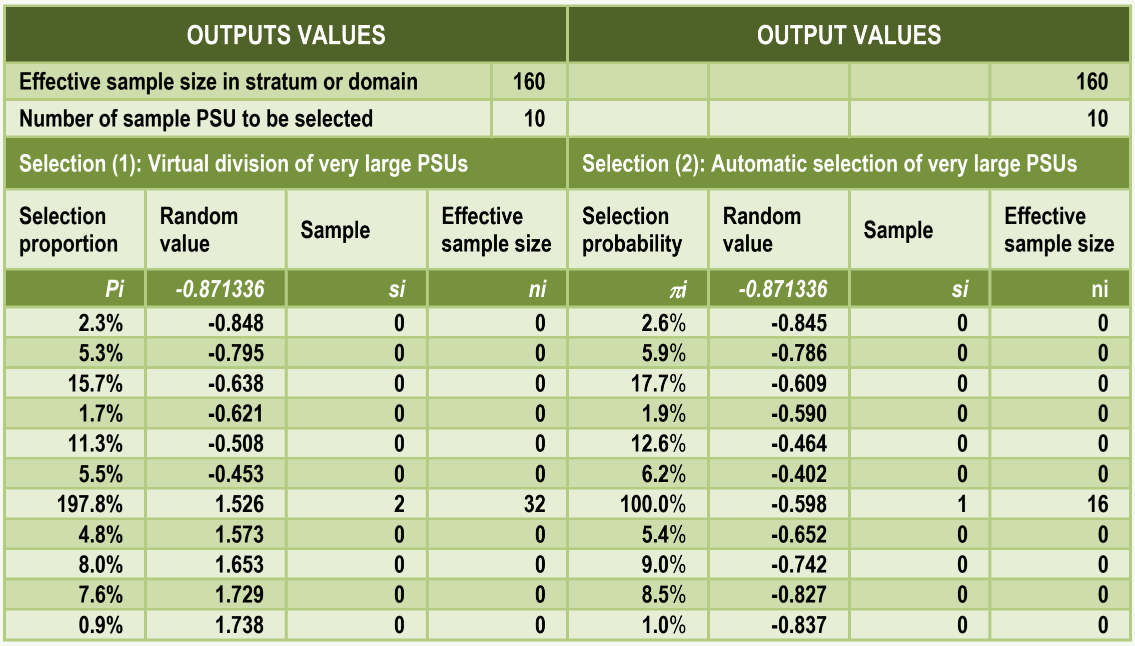
3. Output values

There are two sets of output values depending upon the method of selection for very large PSUs:

Selection (1): Virtual division of very large PSUs

Selection (2): Automatic selection of very large PSUs

In each case, there are also two types of output values: output parameters and output data as shown in the top and bottom panels below:



OUTPUTS VALUES				OUTPUT VALUES			
Effective sample size in stratum or domain			160				160
Number of sample PSU to be selected			10				10
Selection (1): Virtual division of very large PSUs				Selection (2): Automatic selection of very large PSUs			
Selection proportion	Random value	Sample	Effective sample size	Selection probability	Random value	Sample	Effective sample size
Pi	*-0.871336*	*si*	*ni*	*πi*	*-0.871336*	*si*	ni
2.3%	-0.848	0	0	2.6%	-0.845	0	0
5.3%	-0.795	0	0	5.9%	-0.786	0	0
15.7%	-0.638	0	0	17.7%	-0.609	0	0
1.7%	-0.621	0	0	1.9%	-0.590	0	0
11.3%	-0.508	0	0	12.6%	-0.464	0	0
5.5%	-0.453	0	0	6.2%	-0.402	0	0
197.8%	1.526	2	32	100.0%	-0.598	1	16
4.8%	1.573	0	0	5.4%	-0.652	0	0
8.0%	1.653	0	0	9.0%	-0.742	0	0
7.6%	1.729	0	0	8.5%	-0.827	0	0
0.9%	1.738	0	0	1.0%	-0.837	0	0

The output parameters are shown in the top rows of the output values block of the template. The output data are given in the columns below the output parameters.

• Output parameters

For each method of selection, there are two output parameters:

(i) The effective sample size in the specified domain or stratum h. The effective number of sample households may differ from the number of sample households n_h specified as input parameter due to rounding or due to very small PSUs in which the number of households in the frame is lower than the specified sample take

(ii) m_h = The number of sample PSU to be selected for the specified domain or stratum h

In the case of selection method (2), an additional output parameter is calculated:

(iii) p_{min} = The minimum probability of selection of any PSU.

Output data

For each method of selection, there are four sets of output data. In the case of method of selection (1), the output data are given in columns F to I:

Col F p_i = Selection proportion or more precisely the proportion of total measure of size in PSU i

$$p_i = \frac{m_h \times x_i}{\sum_i x_i}$$

p_i is not a probability. Its value can be greater than one, or for that matter greater than two, three, or more depending on the size of the PSU.

Col G e_i = Random number for PSU i obtained systematically from the selection proportion p_i and the random number of the preceding PSU i-1

$$e_i = p_i + e_{i-1}$$

where e_o is given by the random start specified in the input parameters

$$e_o = -randomstart(0,1)$$

Col H s_i = sample inclusion indicator specifying whether PSU_i is included in the sample or not, and if included in how many virtual divisions

$$s_i = Int(e_i) - Int(e_{i-1})$$

where Int(x) is the integer function returning the largest integer less or equal to x

Col I n_i = effective sample size in PSU_i. It is the minimum value between the corresponding sample take and the total number of households in the PSU

$$n_i = \min(b \times s_i, N_i)$$

where b is the sample take per PSU specified in the input parameters.

In the case of method of selection (2), similar output data are produced and given in columns J to M:

Col J π_i = probability of selection of PSU i. The probability of selection is proportional to the size of the PSU,

$$\pi_i = \min(1, k \times x_i)$$

The proportionality factor k is determined as part of in the intermediary calculations.

Col K e_i = Random number for PSU i obtained systematically from the selection probability π_i and the random number of the preceding PSU i-1

$$e_i = \pi_i + e_{i-1}$$

where e_o is given by the same random start specified in the input parameters

$$e_o = -randomstart(0,1)$$

Col L s_i = sample inclusion indicator specifying whether PSU_i is included in the sample or not, and if included in how many virtual divisions

$$s_i = Int(e_i) - Int(e_{i-1})$$

where Int(x) is the integer function returning the largest integer less or equal to x. In selection method (2), the sample selection indicator s_i takes on only values 1 or 0, $s_i = 1$ if PSU i is selected and $s_i = 0$ if PSU i is not selected in the sample.

Col M n_i = effective sample size in PSU_i. It is the minimum value between the corresponding sample take and the total number of households in the PSU

$$n_i = \min(b \times s_i, N_i)$$

where b is the sample take per PSU specified in the input parameters.

4. Intermediary calculations

The intermediary calculations are for obtaining the proportionality factor used for deriving the probabilities of selection of the PSUs. The steps are shown in the following panel.

INTERMEDIARY CALCULATIONS					
Newton's approximation					
Iteration 1		Iteration 2		Iteration 3	
k = 0.0007463		k = 0.0008373		k = 0.0008373	
f'(k)	f(k)	f'(k)	f(k)	f'(k)	f(k)
31	2.3%	31	2.6%	31	2.6%
71	5.3%	71	5.9%	71	5.9%
211	15.7%	211	17.7%	211	17.7%
23	1.7%	23	1.9%	23	1.9%
151	11.3%	151	12.6%	151	12.6%
74	5.5%	74	6.2%	74	6.2%
0	100.0%	0	100.0%	0	100.0%
64	4.8%	64	5.4%	64	5.4%
107	8.0%	107	9.0%	107	9.0%
102	7.6%	102	8.5%	102	8.5%
12	0.9%	12	1.0%	12	1.0%

The probability of selection of a PSU, i, is expressed as a non-linear function of the proportionality factor k,

$$\pi_i(k) = \max[\, p_{\min}, \min(1, k \times x_i)]$$

where

$$p_{\min} = \frac{1}{w_{\max}}$$

is a constant representing the minimum probability of selection so that the corresponding sampling weight does not exceed a maximum specified value W_{max}, and x_i is the measure of size of PSU$_i$. The value 1 in the expression $\min(1, k \times x_i)$ is to ensure that the calculated probability does not exceed the value 1.

Given that the sample design is with a fixed sample size, n_h, in each stratum, the sum of the probabilities must be equal to n_h:

$$\sum_i \pi_i(k) = n_h$$

The determination of the proportionality factor k may thus be expressed as the solution of the non-linear function,

$$f(k)=\sum_i \pi_i(k)-n_h=0$$

The solution may be obtained iteratively by Newton's method [1]

$$k=k-\frac{f(k)-n_h}{f'(k)}$$

where f and f' are the function and its derivative with respect to k,

$$f(k)=\sum_i \max[p_{\min},\min(1,k\times x_i)]$$

$$f'(k)=(\frac{x_i}{2})\times(1+sign([k\times x_i-p_{\min}][1-k\times x_i]))$$

The calculations of f(k) and f'(k) have been programmed in the columns O and T of the intermediary calculations of the Template for 3 iterations with starting value of k_o

$$k_o=\frac{n_h}{\sum_i x_i}$$

[1] http://en.wikipedia.org/wiki/Newton's_method.