ILO-IPEC Interactive Sampling Tools No. 4

Selection of Sample Households by systematic sampling with equal probability

Version 1

August 2014

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

ACKNOWLEDGEMENTS

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 "Households Selection". The template assists the user to select a fixed number of sample households within each sample PSU by systematic sampling with equal probability. The template covers the particular cases of very small and very large PSUs. In the case of very small PSUs, the sample take is all households in the PSU, which is equivalent to a census. In the case of very large PSUs, the template allows for two possibilities: very large PSUs virtually divided into smaller parts or very large PSUs treated as any other PSU.

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 are entered in the form of a rectangular array as shown in the numerical illustration, Diagram 1. Each row corresponds to a sample PSU for which the input values are provided in five columns:

Column A   h   Stratum in which the sample PSU is located

Column B   i   Serial number or identification code of sample PSU

Column C v_i Indicator for virtually divided sample PSUs (default = 1)

Column D N_i Number of households listed in sample PSU

Column E b_i Number of sample households to be selected in sample PSU

For example the first row of Diagram 1 indicates that sample PSU 001 (Column B) is in stratum 1 (Column A). It appears only once which means it is not virtually divided (Column C). The PSU has 163 listed households (Column D), among which 16 sample households are to be selected (Column E).

Diagram 1.   Input values: Numerical illustration

INPUT VALUES				
Stratum	PSU number	PSU virtual division	Number of households listed	Number of household to be selected
h	i	v_i	N_i	b_1
1	001	1	163	16
1	002	1	206	16
1	003	1	13	16
1	004	1	291	16
1	005	1	521	16
1	006	1	228	16
1	007	1	261	16
1	007	2	260	16
1	008	1	162	16
1	009	1	306	16
1	010	1	191	16
1	011	1	139	16
1	012	1	69	16

The parameter v_i keeps track of very large PSUs in the case they are virtually divided into smaller parts. As mentioned in the earlier section on selection of sample PSUs, very large PSUs may be virtually divided into two or more parts according to the value of the selection proportion. In such cases, there should be as many consecutive rows as input values as there are virtual divisions of the sample PSU. The default value of v_i is 1. A value greater than 1 say $v_i=2$ means the row corresponds to the second part of a virtually divided PSU. Similarly, if $v_i=3$ it means that the row corresponds to the third part of an extremely large virtually divided sample PSU. And, so on.

An example is given in the seventh row of Diagram 1. It refers to PSU 007 in stratum 1. The sample PSU is represented in two rows with $v_i=1$ and $v_i=2$. It corresponds to a very large PSU virtually divided into two smaller parts, one with 261 households and the other with 260 households. In each part, a sample of 16 households is to be drawn. In sample PSU 007, therefore, in total 32 sample households will be selected.

By contrast, the fifth row corresponding to sample PSU 005 in stratum 1 is also a very large PSU with 521 listed households, but it is not virtually divided as the PSU is represented in only one row with default value $v_i=1$. In this sample PSU only 16 households will be selected in the sample.

ILO's International Programme on the Elimination of Child Labour (IPEC)

A final point should be noted. In general N_i should be equal or greater than b_i, except in the case of very small PSUs for which $N_i < b_i$. An example is given in the third row of Diagram 1. It refers to sample PSU 003 in stratum 1. It corresponds to a very small PSU. There are 13 listed households among which 16 should be drawn in the sample (13<16). In this sample PSU, all 13 listed households will be drawn in the sample.

3. Output values

The output values of the template are the serial number of the listed households selected in the sample in each sample PSU. The output values corresponding to the input values of the numerical example presented in the previous section are shown in Diagram 2. The output values are generated by a random value between 0 and 1 specified at the top right of Diagram 2.

Diagram 2. ***Output values: Numerical illustration corresponding to input values in Diagram 1***

OUTPUT VALUES																								
RANDOM NUMBER										**0.632297**														
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
SELECT HOUSEHOLD WITH THE FOLLOWING NUMBERS																								
7	17	27	37	48	58	68	78	88	98	109	119	129	139	149	160	0	0	0	0	0	0	0	0	0
9	21	34	47	60	73	86	99	112	124	137	150	163	176	189	202	0	0	0	0	0	0	0	0	0
1	2	3	4	5	6	7	8	9	10	11	12	13	0	0	0	0	0	0	0	0	0	0	0	0
12	30	48	66	85	103	121	139	157	176	194	212	230	248	266	285	0	0	0	0	0	0	0	0	0
21	54	86	119	151	184	216	249	281	314	347	379	412	444	477	509	0	0	0	0	0	0	0	0	0
9	24	38	52	66	81	95	109	123	138	152	166	180	195	209	223	0	0	0	0	0	0	0	0	0
11	27	43	60	76	92	109	125	141	157	174	190	306	223	239	255	0	0	0	0	0	0	0	0	0
272	288	304	320	337	353	369	385	402	418	434	450	467	483	499	515	0	0	0	0	0	0	0	0	0
7	17	27	37	47	57	68	78	88	98	108	118	128	138	149	159	0	0	0	0	0	0	0	0	0
12	32	51	70	89	108	127	146	165	185	204	223	242	261	280	299	0	0	0	0	0	0	0	0	0
8	20	32	44	56	68	80	91	103	115	127	139	151	163	175	187	0	0	0	0	0	0	0	0	0
6	15	23	32	41	49	58	67	75	84	93	101	110	119	127	136	0	0	0	0	0	0	0	0	0

The first row gives the serial number of the households to be selected in sample PSU 001 which as noted in Diagram 1 has 163 listed households. Thus in this PSU, the 16 households in the list with the following serial numbers are selected for interviewing:

Four households with household numbers ending with 7:
7   17   27   37

Six households with household numbers ending with 8:
48   58   68   78   88   98

Six households with household numbers ending with 9:
109   119   129   139   149

And, finally, one household with household number ending with 0
160

The template allows for selection of up to 25 sample households per PSU. In the case of very large sample PSUs that have been virtually divided into parts, the output values are given in two or more consecutive rows, as there are virtual divisions. For example, in the case the very large sample PSU 007 with 521 listed households virtually divided into two parts, the output values are given in the two consecutive rows 7 and 8 of Diagram 2. In this PSU, there are 32 households selected in the sample. Their household numbers are:

11	27	43	60	76	92	109	125	141
157	174							

190	206	223	239	255	273	288	304	320
337	353							

369	385	402	418	434	450	467	483	499
515								

For very small PSUs as PSU 003 in the third row, the sample take is all households in the PSU, i.e., household numbers:

1	2	3	4	5	6	7	8	9	10
11	12	13							

4. Intermediary calculations

The intermediary calculations determine the sampling interval of the systematic sampling scheme and the random start. In systematic sampling with equal probabilities from a list of N units ordered in some way, one unit is selected from every I units in the list, where

$$I = \frac{N}{n}$$

where n is the sample size. Diagram 3 below shows the sampling interval for each of the sample PSUs of the numerical illustration with input values and output values given in Diagrams 1 and 2, respectively.

ILO's International Programme on the Elimination of Child Labour (IPEC)

Diagram 3.   Intermediary calculations

INTERMEDIARY	
Selection interval	0.632297 Random number
I	r
10.19	6.81
12.88	8.51
0.81	0.88
18.19	11.87
32.56	20.96
14.25	9.38
16.31	10.68
16.25	10.64
10.13	6.77
19.13	12.46
11.94	7.92
8.69	5.86

For example, for PSU 001 with 163 listed households (N_i=163) and sample take of 16 households (b_i=16), the sampling interval is given by

$$I = \frac{N_i}{b_i} = \frac{163}{16} = 10.19$$

The next column generates a random number r between 1 and I

$$r = rand() * (I - 1) + 1$$

The random number r identifies the sequence number of the first unit selected. Then starting with r, every I-th unit is selected, i.e. the sequence numbers selected is r, r+I, r+2I, ... , r+(n-1)I. If the sampling interval I is not an integer, fractional intervals are used. The resulting selection numbers are rounded at the end to the nearest integer to identify the particular units selected.

For the first PSU 001, the random value rand()=0.632297 leads to the random start 6.81, a value between 1 and I=10.19 as shown in the first row of the second column of Diagram 3. In this example, the 16 systematic random numbers are

6.81   17.00   27.18   37.37   47.56   57.75

67.93   78.12   88.31   98.50   108.68   118.87

129.06   139.25   149.43   159.62

When rounded to the first digit, they give the serial numbers of the selected sample households

7	17	27	37	48	58	68	78	88
98	109							

119	129	139	149	160

Widespread use of systematic sampling is due to the great convenience of the method in many situations and to the fact that the resulting sample is equivalent to a simple random sample to the extent that the units in the original list appear in a random order. It should be noted, however, that the objective of systematic sampling is generally 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 an implicit stratification. It can be regarded as stratification of the population into zones of size I, the interval for systematic selection, and the selection of one unit per zone or `implicit stratum'.