SIAP

Statistical Institute for Asia and the Pacific

Global strategy for agriculture and rural statistics

Regional Training Course on Sampling Methods for Producing Core Data Items for Agricultural and Rural Statistics

Jakarta, Indonesia ,29Sep-10 October 2014.

Probability Proportional to size Sampling

The method of Simple Random Sampling for selection of sample from a population is useful when the units do not vary much in size. In a village if the cultivator fields do not vary much in size the use of SRS is advantageous. In real life situations the units vary considerably in size.

What is pps Sampling?

- **All the villages in a district may not be of same size. Some may be very small, some large and some very large. In such situation SRS will not be able to make a distinction between them and all units will have the same probability of selection. An ideal situation would be to assign probabilities proportional to their size. The larger units are expected to make greater contribution to the population total.**

- **A sample so selected is likely to provide a more efficient estimate of the population total;**
- **One way of increasing the efficiency of the estimates is to assign unequal probabilities of selection to the different units in the population;**
- **More specifically the selection probabilities can be made directly proportional to the total area or total crop area or total crop area of units.**
- **A procedure of sampling in which units are selected with probabilities proportional**
  to their size or pps sampling.

This mechanism is called Sampling with probability proportional to size with replacement. This is simpler to handle as compared to sampling without replacement.

Selecting a Sample With PPS With Replacement

Let there be N units in the population and let $x_1, x_2, \ldots, x_N$ be the corresponding sizes of the units. The sizes are proportional to the probabilities assigned to the N units in the population. Here the natural numbers 1 to x_1 are associated with the first unit, x_1+1 to x_1+x_2 with the second unit and so on. We draw a number at random from 1 to $S_N = \sum_{i=1}^{N} x_i$, say R, and select that i-th unit in the population for which

$$x_1 + x_2 + \cdots + x_{i-1} < R \leq x_1 + x_2 + \cdots + x_{i-1} + x_i$$

where x_0 is to be interpreted as zero. It is evident that this procedure of selection gives to the i-th unit in the population a probability of selection proportional to x_i. The procedure is to be repeated n times if a sample of size n is required.

An example of selection of units by Probability Proportional to Size with replacement mechanism is given below for illustration purpose.

EXAMPLE

Example

A village has 10 orchards containing 150, 50, 80, 100, 200, 160, 40, 220, 60 and 140 trees respectively. Select a sample of 4 orchards with replacement and with probability proportional to the number of trees in the orchard.

The total number of trees in all the 10 orchards in the village is 1200. The first step in the selection of orchards is to form successive cumulative totals as shown below.

Sl. No. of the Orchard	Size x_i	Cumulative Total
1	150	150
2	50	200
3	80	280
4	100	380
5	200	580
6	160	740
7	40	780
8	220	1000
9	60	1060
10	140	1200

- From the table of Random Numbers a draw is made in such a way that the selected Random Number does not exceed 1200. Let the selected Random Number be 600. It can be easily seen from the successive cumulative totals that this is one of the numbers from 581 to 740 associated with the 6^{th} orchard. The 6^{th} orchard is therefore, selected corresponding to the Random Number 600. Next, another Random Number is drawn in the same way as earlier, it is matched with the class of successive cumulative totals and the corresponding orchard is selected. Likewise two more Random Numbers are drawn and the procedure is repeated. Let the 3 selected numbers be 650, 850 and 300. Then the orchards selected corresponding to these Random Numbers are the 6^{th}, 8^{th} and 4^{th} respectively. It may be seen that the 6^{th} orchard is selected twice.

- **The main drawback of this procedure is that one has to write down the successive cumulative totals. When the number of units in the population is large, the procedure becomes time consuming and tedious. Lahiri (1951) has suggested an alternative procedure which does not require writing down cumulative totals.**

Lahiri's Method

consists in selecting a pair of random numbers, say (i, j) such that $1 \leq i \leq N$ and $1 \leq j \leq M$, where M is the maximum of the sizes of the N units in the population. If $j \leq x_i$, the i-th unit is selected; otherwise it is rejected and another pair of random numbers is chosen. For selecting a sample of n units with probability proportional to size and with replacement, the procedure is to be repeated till n units are selected. It can be seen that the procedure leads to the required probabilities of selection.

ESTIMATION OF POPULATION MEAN, VARIANCE & ESTIMATE OF VARIANCE

Let there be 'N' units in the population and let Y_i represents the characteristic under study for the i-th population unit (i=1,2,...,N). Let the units be selected by varying probability with replacement sampling design and P_i ; (i=1,2,...,N) be the probability of selecting the i-th unit in the population.

We have,

$$\sum_i^N P_i = 1$$

Define

$$z_i = \frac{Y_i}{NP_i}; \quad i=1,2,...,N.$$

The objective is to estimate the population mean

$$\bar{Y} = \frac{1}{N}\sum_i^N Y_i \qquad (3.1)$$

Then the simple arithmetic mean of z values in the sample is given by

CONTINUED

$\bar{z}_n = \frac{1}{n}\sum_{i=1}^{n} z_i$ which is an unbiased estimator of population mean with variance

$V(\bar{z}_n) = \frac{\sigma_z^2}{n}$; where (3.2)

$\sigma_z^2 = \sum_i^N P_i\,(z_i - \bar{Y})^2$;

An unbiased variance estimator is given by

$\hat{V}(\bar{z}_n) = \frac{s_z^2}{n}$; (3.3)

Where

$s_z^2 = \frac{1}{(n-1)}\sum_i^n (z_i - \bar{z}_n)^2$;

If the objective is to estimate the population total i.e. $\hat{Y} = N\bar{Y}$ then its estimator along with variance and the estimate of variance are given by

$\hat{Y} = N \times \bar{z}_n$ (3.4)

$V(\hat{Y}) = N^2 \times V(\bar{z}_n)$ (3.5)

$\hat{V}(\hat{Y}) = N^2 \times \frac{s_z^2}{n}$ (3.6)

The theory explained above is being illustrated with the help of an example as follows:

EXAMPLE

4. AN EXAMPLE

A sample survey to study fertilizer practices for different crops was carried out by the Institute of Agricultural Research Statistics (I.C.A.R.) in Raipur district of Madhya Pradesh in 1958-59. The table below gives the cultivated area (a_i) and area under rice (y_i) for a sample of 25 villages selected from 892 villages of Baloda Bazar teshil of the district. The sample was selected with replacement and with probability proportional to cultivated area. The total cultivated area in the tehsil in 1958-59 was 568,565 acres. Estimate the area under rice in the tehsil with its standard error. Table 1 gives the required data.

SOLUTION

Here A= Total Cultivated Area in the Tehsil = 56865

TABLE 1

S.No.	Area	Y_i	Pi=Ai/A	Yi/Pi	(Yi/Pi)^2
1	1232	688	0.002166859	317510.3247	1.008E+11
2	327	231	0.000575132	401646.8349	1.613E+11
3	1346	768	0.002367363	324411.5305	1.052E+11
4	1285	898	0.002260076	397331.8054	1.578E+11
5	428	417	0.000752772	553952.3481	3.068E+11
6	871	697	0.001531927	454982.5545	2.070E+11
7	1042	785	0.001832684	428333.5173	1.834E+11
8	1262	1190	0.002219623	536127.0602	2.874E+11
9	497	338	0.00087413	386669.9598	1.495E+11
10	1016	745	0.001786955	416910.3593	1.738E+11
11	651	392	0.001144988	342361.7204	1.172E+11
12	1170	1055	0.002057812	512680.406	2.628E+11
13	2630	2400	0.00462568	518842.5856	2.692E+11
14	515	330	0.000905789	364323.2039	1.327E+11
15	895	810	0.001574138	514567.2067	2.648E+11
16	1055	1026	0.001855549	552936.1991	3.057E+11
17	2110	1666	0.003711097	448923.8341	2.015E+11
18	979	929	0.001721879	539526.951	2.911E+11
19	671	565	0.001180164	478746.9821	2.292E+11
20	120	101	0.000211058	478542.2083	2.290E+11
21	541	516	0.000951518	542291.2015	2.940E+11
22	1331	1036	0.002340981	442549.4666	1.958E+11
23	842	568	0.001480921	383545.0356	1.471E+11
24	162	137	0.000284928	480823.4877	2.312E+11
25	206	107	0.000362316	295322.5971	8.721E+10
Total	23184	18395	0.04077634	11113859.38	5.092E+12

Continued

An unbiased estimate of the total area under rice is given by

$$\hat{Y} = N\bar{Z}_n = \frac{1}{n}\sum_{L=1}^{n}\frac{y_i}{P_i} = \frac{11113859.38}{25} = 444554.4 \text{ acres}$$

An unbiased estimate of the variance of the estimate $\hat{Y}$ is obtained as

$$\hat{V}(\hat{Y}) = \frac{1}{n(n-1)}\left\{\sum_{i}^{n}\frac{y_i^2}{P_i^2} - n\hat{Y}^2\right\};$$

$$= \frac{1}{25\times24}[50921 \times 10^8 - 25 \times (444554.4)^2]$$

$$= 253 \text{ x } 10^6 \text{ (acres)}^2$$

Hence

Continued

$$\widehat{SE}(\hat{Y}) = sqrt\left(\hat{V}(\hat{Y})\right) = 15{,}900\ acres$$

The coefficient of variation is given as

$$\frac{SE(\hat{Y})}{\hat{Y}} = \frac{sqrt\left(\frac{s_z^2}{n}\right)}{\hat{Y}};$$

Assuming that the desired coefficient of variation is 5% i.e. 0.05. The sample size required for a 5% coefficient of variation is

$$n = \frac{s_z^2}{\hat{Y}^2} \times \frac{1}{0.0025} \qquad (3.7)$$

Substituting, the various values computed above, in (3.7) the sample size works out to 13.

A number of statistical packages are available now for computing the estimate and its estimated variance when the sample selection is done through a PPSWR sampling design. Some of the commonly used statistical packages are SAS, SPSS.