PROBABILITY-PROPORTIONAL-TO-SIZE (PPS) SAMPLING

Probability-proportional-to-size (PPS) sampling is one technique that uses auxiliary data and yields unequal probabilities of inclusion. If population units vary in size and these sizes are known, such information can be used during sampling to increase the statistical efficiency.

The main **advantage** of PPS sampling is that it can improve the statistical efficiency of the sampling strategy by using auxiliary information. This can result in a dramatic reduction in the sampling variance compared with SRS or even stratified sampling.

The **disadvantages** of PPS sampling are:

- It requires a survey frame that contains good quality, up-to-date auxiliary information for all units on the frame that can be used as size measures.
- It is inappropriate if the size measures are not accurate or stable. In such circumstances, it is better to create size groupings and perform stratified sampling.
- It is not always applicable, since not every population has a stable size measure that is correlated with the main survey variables.

Methods of PPS Sampling

There are many PPS sampling schemes, however, three commonly used techniques are the random method, the systematic method and the randomised systematic method. (The following assumes that the size measures are integer values.)

a) The random method for PPS sampling

- for each unit in the population, cumulate the size measures for units up to and including itself.
- determine the range corresponding to each unit in the population, that is, from (but not including) the cumulative sum for the previous unit to the cumulative sum for the current unit.
- select a random number between 0 (if dealing with non-integer size measures) or 1 (for integer size measures) and the total cumulative size and select the unit whose range contains the random number.
- repeat previous step until *n* units have been selected.

To illustrate using the farm example:

Table 5: PPS Sampling using the Random Method

Farm	Size	Cumulative Size	Range
1	50	50	1-50
2	1000	1050	51-1050
3	125	1175	1051-1175
4	300	1475	1176-1475

5	500	1975	1476-1975
6	25	2000	1976-2000

For a sample containing three units, three random numbers between 1 and 2000 are selected. Suppose these numbers are: 1697, 624 and 1109. Then the farms selected are: farm 5, farm 2 and farm 3.

In the case of the random method for PPS sampling without replacement, if more than one unit is selected, complications arise both in attempting to keep probabilities directly proportional to size and in estimating the sampling variances of survey estimates. This becomes even more complicated when more than two or three units are selected with PPS without replacement.

b) The systematic method

This involves the following steps;

- for each unit in the population, cumulate the size measures for units up to and including itself.
- determine the range corresponding to each unit in the population, that is, from (but not including) the cumulative sum for the previous unit to the cumulative sum for the current unit.
- determine the sampling interval, *k*=(total cumulative size)/n.
- determine a random start, *r*, between 0 (if dealing with non-integer size measures) or 1 (for integer size measures) and *k.*
- select those units whose range contains the random numbers *r, r+k, r+2k, ... r+(n-1)k*.

c) It can result in a sampling strategy that is less statistically efficient than SRS for survey variables that are not correlated with the size variables.
d) Estimation of the sampling variance of an estimate is more complex.
e) Frame creation is more costly and complex than SRS or SYS, since the size of each unit in the population needs to be measured and stored.

CLUSTER SAMPLING

Definition: Cluster sampling is the process of randomly selecting complete groups (clusters) of population units from the survey frame. It is usually a less statistically efficient sampling strategy than SRS and is performed for several reasons. The first reason is that sampling clusters can greatly reduce the cost of collection, particularly if the population is spread out and personal interviews are conducted. The second reason is that it is not always practical to sample individual units from the population. Sometimes, sampling groups of the population units is much easier (e.g., entire households). Finally, it allows the production of estimates for the clusters themselves (e.g., average revenue per household).

Cluster sampling is a two-step process. First, the population is grouped into clusters (this may consist of natural clustering, e.g., households, schools). The second step is to select a sample of clusters and interview all units within the selected clusters.

The **advantages** of cluster sampling are:

- It can greatly reduce the cost of collection by having a less dispersed sample than SRS.
- It is easier to apply than SRS or SYS to populations that are naturally clustered (e.g., households, schools).
- It allows the production of estimates for the clusters themselves. For example, estimates of the average number of teachers per school (where schools are clusters).
- It can be more statistically efficient than SRS if the units within the clusters are heterogeneous (different) with respect to the study variables and the clusters are homogeneous.
- Easy to administer because there is no need for frames as they are available or can be constructed easily.
- It leads to savings on costs and time of travel between clusters and elements within clusters.
- Single stage cluster sampling leads to simple field instructions and training. It therefore leads to a decline in non-sampling errors because it results in better supervision and organization of field staff.
- It is possible to get information from neighbours in some surveys.

The **disadvantages** of cluster sampling are:

- It can be less statistically efficient than SRS if the units within the clusters are homogeneous with respect to the study variables.
- its final sample size is not usually known in advance, since it is not usually known how many units are within a cluster until after the survey has been conducted.
- Its survey organisation can be more complex than for other methods.
- Its variance estimation will be more complex than for SRS if clusters are sampled without replacement.
- It is generally less efficient than direct sampling.
- Its efficiency reduces as the number of stages increases.

<u>SIMPLE SINGLE STAGE CLUSTER SAMPLING</u>

Clusters are selected from a frame of clusters using simple random sampling and all the elements within the selected clusters are enumerated. They can either be equal sized clusters or unequal sized clusters.

a) Equal sized clusters. Let A be the number of clusters in the population, B is the size of each cluster, N = AB is the total population size. let 'a" be the random sample of clusters, n =aB is the sample size , f_1= a/A is the first stage sampling fraction, Y_{ij} is the value of the study variable for the j^{th} unit in the i^{th} cluster.
    - The population mean $\bar{Y}_N = \frac{1}{AB}\sum_{i=1}^{A}\sum_{j=1}^{B} Y_{ij}$
    - Unbiased estimator of the population mean $\bar{y}_n = \frac{1}{a}\sum_{i=1}^{a}\bar{y}_{i.} = \frac{1}{aB}\sum_{i=1}^{a}\sum_{j=1}^{B} y_{ij}$.

- Estimator of the variance for the mean $v(\bar{y}) = \frac{1}{a-1}\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y})^2$ $OR\ V\ (\bar{y}) = (\frac{1}{a} - \frac{1}{A})S_b^2$ $where\ S_b^2 = \frac{1}{a-1}[\sum_{i=1}^{a} y_i^2 - a\bar{y}^2]$. OR v($\bar{y}$) $= \frac{1-f_1}{a(a-1)B^2}[\sum_{i=1}^{a} y_i^2 - \frac{y^2}{a}]$. Where y is the overall sample total or cluster total and for equal sized clusters, B=1.
- Estimator of the total $\hat{Y} = N\bar{y}$ and its variance is $v(\hat{Y}) = N^2 v(\bar{y})$

EXAMPLE

A random sample of 14 zones out of 90 zones were selected for the sample of graduates in each zone where the zones are the clusters as below.

Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14
graduates	33	18	54	35	22	84	39	42	09	12	34	99	22	10

Estimate;

i. The average number of graduates.
ii. The total number of graduates.
iii. The variance of the mean estimate.
iv. The standard error of the estimator for the total number of graduates.

b) Unequal sized clusters. In practice clusters are usually of unequal sizes e.g enumerator areas whose sizes vary. Let A represent the number of clusters in a population, B_i number of elements in the i[th] cluster, N=$\sum_{i=1}^{A} B_i$ $as\ the\ total\ number\ of\ elements\ in\ the\ population.$

- Population mean $\bar{Y}_N = \frac{\sum_{i=1}^{A}\sum_{j=1}^{B_i} Y_{ij}}{\sum_{i=1}^{A} B_i}$
- Let a be the number of clusters in the sample, n is the sample size where $n = \sum_{i=1}^{a} B_i$. the sample mean $\bar{y} = \frac{\sum_{i=1}^{a} y_i}{\sum_{i=1}^{a} B_i} = r = \frac{y}{x}$. where y and x are total of y_i and B_i. y is the sample total for the study variable and x is the sample size.
- Estimator of the variance for r is $v(r) = \frac{1-f_1}{a-1}[\sum_{i=1}^{a} y_i^2 + r^2\sum_{i=1}^{a} x_i^2 - 2r\sum_{i=1}^{a} x_i y_i]$

DESIGN EFFECT (Deff)

Is a measure of relative efficiency of a design compared with what it would have been had the sample been selected by simple random sampling design.

Deff $= \frac{variance\ (cluster)}{variance\ (simple\ random\ sampling)} X\ 100\% = 1 + \rho(B-1), where\ \rho\ is\ the\ intra - cluster\ correlation\ coefficient\ which\ measures\ the\ degree\ of\ homogeneity\ of\ the\ units\ within\ clusters.$

Deductions

- ✓ If deff=1, it implies that cluster sampling design is equal to SRS design. This is a case of perfect heterogeneity.
- ✓ If B =1, deff =1. This means that each cluster contains only one unit or there is no clustering.

✓ The closer ρ is to one, the greater the deff which implies that cluster sampling is less precise.

COMPARISON BETWEEN STRATIFICATION AND CLUSTERING

STRATIFICATION	CLUSTERING
• Fraction of the population used. • Each stratum investigated • Within each stratum a sample is fixed in advance • Higher precision than SRS • Higher costs than SRS • To improve precision of estimates, strata should be internally homogeneous	• Fraction of the population used. • Only a sample of clusters investigated • The sample sizes vary with the size of the clusters • Lower precision than SRS • Lower costs than SRS • To improve precision of estimates, clusters should be internally heterogeneous

Multi-Stage Sampling

Multi-stage sampling is the process of selecting a sample in two or more successive stages. The units selected at the first stage are called primary sampling units (PSU's), units selected at the second stage are called second stage units (SSU's), etc. The units at each stage are different in structure and are hierarchical (for example, people live in dwellings, dwellings make up a city block, city blocks make up a city, etc.). In two-stage sampling, the SSU's are often the individual units of the population.

A common multi-stage sample design involves two-stage cluster sampling using an area frame at the first stage to select regions (the PSU's) and then a systematic sample of dwellings (the SSU's) within a region at the second stage. With the one-stage cluster sampling presented earlier, every unit within a sampled cluster is included in the sample. In two-stage sampling, only some of the units within each selected PSU are subsampled.

Each stage of a multi-stage sample can be conducted using any sampling technique. Consequently, one of the chief advantages of a multi-stage sample is its flexibility. For example, within one PSU drawn at the first stage, an SRS sample may be drawn. For another PSU, there may be a measure of size that is correlated with the key survey variables, so PPS may be used within this PSU.

The **advantages** of multi-stage sampling are:

i. It can result in a more statistically efficient sampling strategy than a one-stage cluster design when clusters are homogeneous with respect to the variables of interest (i.e., a sample size reduction).

ii. It can greatly reduce the travel time and cost of personal interviews as a result of the sample being less dispersed than for other forms of sampling, such as SRS.

iii. It is not necessary to have a list frame for the entire population. All that is needed is a good frame at each stage of sample selection.

The **disadvantages** of multi-stage sampling are:

i. It is usually not as statistically efficient as SRS (although it can be more efficient than a one-stage cluster strategy).
ii. The final sample size is not always known in advance, since it is not usually known how many units are within a cluster until after the survey has been conducted. (The sample size can be controlled, however, if a fixed number of units are selected per cluster.)
iii. Its survey organisation is more complex than for one-stage cluster sampling.
iv. Its formulas for calculating estimates and sampling variance can be complex.