VIVEKANANDA COLLEGE
THAKURPUKUR
KOLKATA-700063

NAAC ACCREDITED 'A' GRADE

Topic: Theory of Cluster sampling

Course Title: Sample Survey

Paper: CC-8

Unit: 2

Semester: 4

Name of the Teacher: Riddhi Das Majumder

Name of the Department: Statistics

Cluster Sampling

Several references have been made to surveys in which the sampling unit consists of a group or cluster of small units that we have called elements or subunits. There are two main reasons for the widespread application of cluster sampling:

(i) It is found in many surveys that no reliable list of elements in the population is available and that it would be prohibitively expensive to construct such a list.

(ii) Even if such a list existed, it would not be economical to base the enquiry on a SRS of persons because this would require interviewers to visit almost every commune in the country and resource do not permit it.

For example, a simple random sample of 600 houses covers a town more evenly than 20 city blocks containing an average of 30 houses a piece. But greater field costs are incurred in locating 600 houses and in travel between them than in locating 20 blocks and visiting all the houses in these blocks. Though, for a given size of sample, a small unit usually gives more precise results than a large unit, but all these considerations point to the need of selecting larger units or clusters, rather than elements (individuals in this case) directly from the population.

A simple cluster sampling plan is a sampling plan in which (a) the elementary units of the population to be sampled are grouped into clusters, such that each elementary unit is associated with one and only one cluster; and (b) a sample is drawn by using the clusters as sampling units and selecting a simple random sample of the clusters. The clusters are referred to as primary sampling units (psu's) or as first-stage sampling units or single-stage cluster sampling.

If all elementary units in the selected clusters are included in the sample, the sampling plan is a one-stage sampling plan. If subsample is selected from each of the selected psu's, with a uniform fraction of the second-stage sampling units selected from each primary unit included in the sample, the sampling plan is referred to as

Now, $\sum_{i=1}^{N}(\bar{y}_i - \bar{\bar{y}})^2 = \sum_{i=1}^{N}\left(\sum_{j=1}^{M} y_{ij} - M\bar{\bar{y}}_c\right)^2 = \sum_{i=1}^{N}\left\{\sum_{j=1}^{M}(y_{ij} - \bar{\bar{y}}_c)\right\}^2$

$= \sum_i\sum_j (y_{ij} - \bar{\bar{y}}_c)^2 + \sum_i \sum_j \sum_{k(\neq j)} (y_{ij} - \bar{\bar{y}}_c)(y_{ik} - \bar{\bar{y}}_c)$

$= (NM-1)S_y^2 + (M-1)(NM-1)\rho S_y^2 = (MN-1)S_y^2\{1+(M-1)\rho\}$

Hence, $V(\hat{\bar{y}}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{NM-1}{N-1}S_y^2\{1+(M-1)\rho\}$.

Corollary: For estimating $\bar{\bar{y}}_c$, the average per element, an unbiased estimator is $\bar{y}_c = \sum_{i=1}^{n}\bar{y}_i/n$ and its variance is $V(\bar{y}_c) = \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}(\bar{y}_i - \bar{\bar{y}}_c)^2$.

Remark: If, instead of sampling in clusters, a SRSWOR of nM elements is taken directly from the population, the estimator is $\hat{\bar{y}}' = N\sum_i\sum_j y_{ij}/n$ and $V(\hat{\bar{y}}') = \frac{(NM)^2}{nM}\left(1-\frac{nM}{NM}\right)S_y^2 = \frac{N^2}{n}\left(1-\frac{n}{N}\right)MS_y^2$.

Again, $V(\hat{\bar{y}}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{NM-1}{N-1}S_y^2\{1+(M-1)\rho\}$

$\simeq \frac{N^2}{n}\left(1-\frac{n}{N}\right)MS_y^2\{1+(M-1)\rho\}$.

Hence $\frac{V(\hat{\bar{y}})}{V(\hat{\bar{y}}')} \simeq \{1+\overline{M-1}\rho\}$.

Generally ρ is found to be positive, since clusters are usually formed by putting together geographically contiguous farms, states, families, etc. Thus, for the same number of elements in the sample, cluster sampling will give a higher variance than sampling elements directly. But the real point is that it is far cheaper to collect information on a per-element basis if sampling is done in clusters. If $\rho < 0$, both cost and the variance point to the use of clusters.

(B) Two-stage cluster sampling or subsampling:

Suppose that a sample of n clusters has been selected from a popln containing N clusters. If elementary units within a selected cluster give similar results, it seems uneconomical to measure them all. A common practice is to select and measure

a simple two-stage cluster sampling plan.

(i) Single-stage cluster sampling:

No new principles are involved in making estimates when a probability sample of clusters has been taken and each sample cluster is enumerated completely (i.e. there is no subsampling). A problem to be considered is the optimum size of the cluster. This will naturally depend upon the cost of collecting information from clusters of different size and the resulting variance.

Assume that the popln. contains N clusters ($U_1, \ldots, U_N$) each containing M elements. The average of y per cluster is $\bar{Y} = \sum_{i=1}^{N} Y_i/N = \sum_{i=1}^{N}\sum_{j=1}^{M} Y_{ij}/N$ and the average per element is $\bar{Y}_e = \sum_i \sum_j Y_{ij}/NM = \bar{Y}/M$; where Y_{ij} be the y value for the j^{th} element within the i^{th} cluster, and Y_i be the cluster total.

Theorem: In SRSWOR of n clusters, each containing M elements, from a population of N clusters, an unbiased estimate of the population total Y is given by

$$\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} y_i = \frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij}.$$

and $$V(\hat{Y}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\cdot\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y})^2$$

$$= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\cdot\frac{MN-1}{N-1}S_y^2[1+(M-1)\rho].$$

where $S_y^2 = \sum_i\sum_j (Y_{ij}-\bar{Y}_e)^2/(NM-1)$ and ρ is the intra-cluster correlation coefficient as $\rho = \dfrac{E(Y_{ij}-\bar{Y}_e)(Y_{iK}-\bar{Y}_e)}{E(Y_{ij}-\bar{Y}_e)^2}$

$$= \frac{2\sum_i \sum_{j<K}(Y_{ij}-\bar{Y}_e)(Y_{iK}-\bar{Y}_e)}{(M-1)(NM-1)S_y^2}.$$

Proof: Under SRSWOR, $E(\hat{Y}) = N\cdot E\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = N\cdot\bar{Y} = Y$.

$\therefore$ $V(\hat{Y}) = \left(1-\frac{n}{N}\right)\frac{S_y^2}{n}$ $\quad$ $Var\left(N\cdot\frac{1}{n}\sum_i y_i\right) = N^2\, Var\left(\frac{1}{n}\sum_i y_i\right)$

$$= N^2\left(1-\frac{n}{N}\right)\cdot\frac{S_y^2}{n} = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y})^2.$$

Scanned with CamScanner

a sample of elementary units from any selected cluster. This technique is called Sub-sampling. Since the cluster selected is not measured completely but is itself sampled or two-stage sampling. because the sample is taken in two steps — first step is to select a sample of clusters (often, called primary sampling units (psu)) and the second is to select a sample of elementary units from each selected clusters. (Second-stage units (ssu)).

Here, we consider the simplest case in which every cluster contains M elementary units, of which m are chosen from each selected cluster.

The principal advantage of two-stage sampling is that it is more flexible than one-stage sampling. It reduces to one-stage sampling when $m = M$ but, unless this is the best choice of m, we have the opportunity of taking some smaller value that appears more efficient. As usual, the issue reduces to a balance between statistical precision and cost. When the elementary units (ssu's) in the same cluster agree very closely, considerations of precision suggest a small value of m. On the other hand, it is sometimes almost as cheap to measure the whole of a cluster as to subsample it, for example, when the cluster is a household and a single respondent can give accurate data about all members of the household.

Two stage Sampling with equal-sized psu's and subsampling with equal-sized s.s.u's.

Here all psu's have the same number M of second-stage units and a constant number 'm' of them are sampled from every selected psu.

The following notation is used:

Y_{ij} = Value obtained for the j^{th} subunit in the i^{th} primary unit. $\bar{y}_i = \frac{\sum_{j=1}^{m} y_{ij}}{m}$ = Sample mean per subunit. $\bar{y}_0 = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i$

$Y_i = \sum_{j=1}^{M} Y_{ij}$ = total over-all subunits in the i^{th} psu (or cluster)

$S_b^2 = \frac{\sum_{i=1}^{n} (\bar{y}_i - \bar{Y})^2}{n-1}$ = Variance among primary unit means or Variance between the psu's

$S_w^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} (Y_{ij} - \bar{y}_i)^2}{n(m-1)}$ = Variance among subunits within primary units.

Scanned with CamScanner

Estimation of Variance:

Under the conditions of Theorem just given, an unbiased estimator of $V(\hat{\bar{y}})$ is

$$v(\hat{\bar{y}}) = M^2N^2\left(\frac{1}{n}-\frac{1}{N}\right)s_b^2 + \frac{N}{n}M^2\left(\frac{1}{m}-\frac{1}{M}\right)s_w^2$$

where $s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_i - \bar{\bar{y}})^2$, $s_w^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-\bar{y}_i)^2/n(m-1)$.

proof: that is $\boxed{v(\hat{\bar{y}}) = M^2N^2\left\{\frac{1-f_1}{n}s_b^2 + \frac{f_1(1-f_2)}{mn}s_w^2\right\}}$.

Now, $(n-1)s_b^2 = \sum_{i=1}^{n}\bar{y}_i^2 - n\bar{\bar{y}}^2$.

Hence, $E[(n-1)s_b^2] = (n-1)E_1[E_2(s_b^2)] = \cdots$

$= E_1\left[\sum_{i=1}^{n}E_2(\bar{y}_i^2) - nE_2(\bar{\bar{y}}^2)\right]$

$= E_1\left[\sum_{i=1}^{n}\{V_2(\bar{y}_i)+E_2^2(\bar{y}_i)\} - n\{V_2(\bar{\bar{y}})+E_2^2(\bar{\bar{y}})\}\right]$

$= E_1\left[\left\{\sum_{i=1}^{n}\left(\frac{1}{m}-\frac{1}{M}\right)s_{wi}^2 + \sum_{i=1}^{n}\bar{Y}_i^2\right\} - n\left\{\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{m}-\frac{1}{M}\right)s_{wi}^2 - \left(\frac{\sum Y_i}{n}\right)^2\right\}\right]$

$= E_1\left[\sum_{i=1}^{n}(\bar{Y}_i - \bar{\bar{Y}}_n)^2 + \frac{(n-1)\left(1-\frac{m}{M}\right)}{mn}\sum_{i=1}^{n}s_{wi}^2\right]$, where $\bar{\bar{Y}}_n = \sum_{i=1}^{n}Y_i/n$.

$= (n-1)\left\{s_b^2 + \frac{\left(1-\frac{m}{M}\right)}{mn}\cdot n\sum_{i=1}^{N}s_{wi}^2\right\}$, taking expectation w.r.t. the first stage simple random sampling.

Hence, $E\left[\left(\frac{1}{n}-\frac{1}{N}\right)s_b^2\right] = \left(\frac{1}{n}-\frac{1}{N}\right)s_b^2 + \frac{\left(1-\frac{n}{N}\right)\left(1-\frac{m}{M}\right)}{mn}s_w^2$.

Again, $E_1(s_w^2) = E_1E_2(s_w^2) = \frac{1}{n}E_1\sum_{i=1}^{n}\left[E_2\left\{\sum_{j=1}^{m}(y_{ij}-\bar{y}_i)^2/m-1\right\}\right]$

$= E_2\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{M}\frac{(y_{ij}-\bar{y}_i)^2}{M-1}\right] = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{\bar{y}}_i)^2$

$= s_w^2$

Therefore, $E_1(v(\hat{\bar{y}})) = M^2N^2\left[\left\{\left(\frac{1}{n}-\frac{1}{N}\right)s_b^2 + \frac{\left(1-\frac{n}{N}\right)\left(\frac{1}{m}-\frac{1}{M}\right)}{mn}s_w^2\right\} + \frac{\frac{n}{N}\left(1-\frac{m}{M}\right)}{mn}s_w^2\right]$

$= M^2N^2\left\{\left(\frac{1}{n}-\frac{1}{N}\right)s_b^2 + \left(\frac{1-\frac{m}{M}}{mn}\right)s_w^2\right\}$

$= M^2N^2\left\{\left(\frac{1}{n}-\frac{1}{N}\right)s_b^2 + \frac{1}{n}\cdot\left(\frac{1}{m}-\frac{1}{M}\right)s_w^2\right\}$.

$= \frac{(MN)^2}{n}\left\{\left(1-\frac{n}{N}\right)s_b^2 + \frac{1}{m}\left(1-\frac{m}{M}\right)s_w^2\right\}$

Scanned with CamScanner

Theorem: If n units (p.s.u's) and m subunits (s.s.u's) from each selected p.s.u are selected by SRSWOR, $\hat{Y} = \frac{NM}{n}\sum_{i=1}^{n} y_i$ is an unbiased estimate of Y with variance

$$V(\hat{Y}) = M^2N^2\left\{\left(1-\frac{n}{N}\right)\frac{S_b^2}{n} + \frac{1}{n}\left(1-\frac{m}{M}\right)\frac{S_w^2}{m}\right\}$$

Proof: With simple random sampling at both stages,

$$E(\hat{Y}) = E_1E_2(\hat{Y}) = E_1\left[\frac{NM}{n}\sum_{i=1}^{n} E_2(y_i)\right] = E_1\left[\frac{NM}{n}\sum_{i=1}^{n}\bar{Y}_i\right]$$

$$= N\,E_1\left[\frac{1}{n}\sum_{i=1}^{n} M_i\bar{Y}_i\right] = N\,E_1\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = N\bar{Y} = Y.$$

Again, we have $V(\hat{Y}) = V_1[E_2(\hat{Y})] + E_1[V_2(\hat{Y})]$

Here E_2 and V_2 represent the conditional expectation and variance over all selections of sizes of m from the p.s.u's which are fixed (like strata); E_1 and V_1 denote similarly the expectation and variance over all possible samples of n p.s.u's from the N p.s.u's.

Thus $E_2(\hat{Y}) = N\cdot\frac{1}{n}\sum_{i=1}^{n} Y_i$, $V_1[E_2(\hat{Y})] = M^2N^2\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2$

$V_2(\hat{Y}) = \frac{N^2}{n^2}\sum_{i=1}^{n} M^2\left(\frac{1}{m}-\frac{1}{M}\right)S_{wi}^2$, where $S_{wi}^2 = \frac{\sum_{j=1}^{M}(y_{ij}-\bar{Y}_i)^2}{M-1}$

is the variance among subunits for the i^{th} primary unit.

and $E_1[V_2(\hat{Y})] = \frac{N^2}{n^2}\sum_{i=1}^{n} M^2\left(\frac{1}{m}-\frac{1}{M}\right)E(S_{wi}^2)$

$= \frac{N^2}{n^2}\cdot M^2\left(\frac{1}{m}-\frac{1}{M}\right)\cdot E\left(\frac{1}{n}\sum_{i=1}^{n} S_{wi}^2\right) = \frac{N^2}{n}\cdot M\left(\frac{M-m}{m}\right)\cdot\left(\frac{\sum_{i=1}^{N} S_{wi}^2}{N}\right)$

$= \frac{N^2}{n}\cdot\frac{M^2}{m}\left(1-\frac{m}{M}\right)S_w^2.$

Hence, $V(\hat{Y}) = \frac{N^2}{n}\left\{M^2\left(1-\frac{n}{N}\right)S_b^2 + \frac{M^2}{m}\left(1-\frac{m}{M}\right)S_w^2\right\}$.

$= \frac{(NM)^2}{n}\left\{\left(1-\frac{n}{N}\right)S_b^2 + \frac{1}{m}\left(1-\frac{m}{M}\right)S_w^2\right\}$.

If $f_1 = \frac{n}{N}$, $f_2 = \frac{m}{M}$ are sampling fractions in the first and second stages, an alternative form of the result is

$$V(\hat{Y}) = \frac{N^2}{n}\left\{M^2(1-f_1)S_b^2 + \frac{M^2}{m}(1-f_2)S_w^2\right\} = \frac{(NM)^2}{n}\left\{(1-f_1)S_b^2 + \frac{1-f_2}{m}S_w^2\right\}$$

Corollary: $\bar{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$ is an unbiased estimator $\bar{\bar{Y}}$ with

$V(\bar{\bar{y}}) = \left(1-\frac{n}{N}\right)\frac{S_b^2}{n} + \left(1-\frac{m}{M}\right)\frac{S_w^2}{mn} = \left(\frac{1-f_1}{n}\right)S_b^2 + \frac{(1-f_2)}{mn}S_w^2$.

Scanned with CamScanner

Corollary : An unbiased estimator of $V(\bar{y})$ is.

$$v(\bar{y}) = \frac{1-f_1}{n} s_b^2 + \frac{f_1(1-f_2)}{nm} s_w^2$$

Optimum Sampling and Subsampling fractions

(i) **Cost functions :** Determination of optimum sampling and subsampling fractions depend on the type of cost function used.

If travel costs between units are unimportant, one form that has proved useful is $C = c_1 n + c_2 nm$, in which the two components are proportional to the number of psu's and the no. of s.s.u's.

If travel between psu's is substantial, a more relevant cost function appears to be $C = c_0\sqrt{n} + c_1 n + c_2 nm$. Since travel between n points is represented more appropriately by a quantity proportional to $\sqrt{n}$ (Mahalanobis, Hansen).

(ii)

□ Assume the simple cost function $C = c_1 n + c_2 nm$.

We have $V(\bar{\bar{y}}) = \frac{1}{n}\left(S_b^2 - \frac{S_w^2}{M}\right) + \frac{1}{mn} S_w^2 - \frac{1}{N} S_b^2$. The last term on the RHS does not depend on the choice of n and m. Minimizing V for fixed C or C for fixed V, is equivalent to minimizing the product

$$\left(V + \frac{1}{N} S_b^2\right)(c_1 n + c_2 nm) = \left\{\left(S_b^2 - \frac{S_w^2}{M}\right) + \frac{1}{m} S_w^2\right\}\{c_1 + c_2 m\}$$

$$\geq \left\{\sqrt{S_b^2 - \frac{1}{M} S_w^2}\sqrt{c_1} + \frac{1}{\sqrt{m}} S_w \sqrt{c_2 m}\right\}^2, \text{ by Cauchy-Schwartz}$$

inequality. Equality holds $\Leftrightarrow$ attains the minimum iff

$$\frac{S_w}{\sqrt{m}} \Big/ \sqrt{c_2 m} = \frac{\sqrt{S_b^2 - \frac{1}{M} S_w^2}}{\sqrt{c_1}} \Leftrightarrow m = \frac{S_w}{\sqrt{S_b^2 - \frac{S_w^2}{M}}} \times \sqrt{\frac{c_1}{c_2}}$$

provided $S_b^2 > \frac{S_w^2}{M}$. Now, the value of 'n' is found by solving either the cost equation or the variance equation, depending on which has been preassigned.

□ Assume the cost function $C = c_0\sqrt{n} + c_1 n + c_2 nm$. If a desired value of $V(\bar{\bar{y}})$ has been specified, pairs of values of (n, m) that give this variance are easily computed from $V(\bar{\bar{y}}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right)$

Scanned with CamScanner

The costs for different combinations are then computed from $C = c_0 n + c_1 n + c_2 mn$ and the combination giving the smaller cost is found

~~This stage sampling with equal sized p.s.u's but subsampling~~ with unequal sized s.s.u's.

Remark: Let n psu's be selected at random from a pop. of N psu's. Let random samples of m_i, $i = 1(1)n$, second-stage units (s.s.u's) be taken from the M s.s.u's in the selected psu's. Then $\hat{\bar{y}} = \frac{MN}{n} \sum_{i=1}^{n} \bar{y}_i$.

$$V(\hat{\bar{y}}) = (MN)^{-2} \left\{ \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \frac{1}{nN} \sum_{i=1}^{N} \left(\frac{1}{m_i} - \frac{1}{M}\right) S_{wi}^2 \right\}$$

and $$v(\hat{\bar{y}}) = (MN)^{-2} \left\{ \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2 + \frac{1}{nN} \sum_{i=1}^{n} \left(\frac{1}{m_i} - \frac{1}{M}\right) s_{wi}^2 \right\}.$$

Derivation of the result is similarly to the previous case with subsampling with equal sized s.s.u's

---

Problem: To estimate the total number ticketless commuters travelling through a station on a day n trains are chosen randomly by an SRSWOR out of N trains passing through the station on that day, and from each chosen train (each train containing M compartments) m compartments are chosen following SRSWOR procedure.

Let y_{ij} = total number of ticketless passengers in the j^{th} compartment of i^{th} train in the sample, $i = 1(1)n$, $j = 1(1)m$

On the basis of $\{y_{ij}\}$ suggest an unbiased estimator of the total number of ticketless passengers travelling on the particular day. Find the variance of the estimator and also find an unbiased estimator of the variance.

Assuming the cost function as $C = c_1 n + c_2 mn$, find the optimum choices of 'm' and 'n' for given cost.

Scanned with CamScanner