Sampling Theory

MODULE IX

LECTURE - 30

CLUSTER SAMPLING

DR. SHALABH
DEPARTMENT OF MATHEMATICS AND STATISTICS
INDIAN INSTITUTE OF TECHNOLOGY KANPUR
It is one of the basic assumptions in any sampling procedure that the population can be divided into a finite number of distinct and identifiable units, called **sampling units**. The smallest units into which the population can be divided are called **elements** of the population. The groups of such elements are called **clusters**.

In many practical situations and many types of populations, a list of elements is not available and so the use of an element as a sampling unit is not feasible. The method of cluster sampling or area sampling can be used in such situations.

**In cluster sampling**

- divide the whole population into clusters according to some well defined rule.
- Treat the clusters as sampling units.
- Choose a sample of clusters according to some procedure.
- Carry out a complete enumeration of the selected clusters, i.e., collect information on all the sampling units available in selected clusters.

**Area Sampling**

In case, the entire area containing the populations is subdivided into smaller area segments and each element in the population is associated with one and only one such area segment, the procedure is called as area sampling.
Examples

- In a city, the list of all the individual persons staying in the houses may be difficult to obtain or even may be not available but a list of all the houses in the city may be available. So every individual person will be treated as sampling unit and every house will be a cluster.

- The list of all the agricultural farms in a village or a district may not be easily available but the list of village or districts are generally available. In this case, every farm is sampling unit and every village or district is the cluster.

Moreover, it is easier, faster, cheaper and convenient to collect information on clusters rather than on sampling units.

In both the examples, draw a sample of clusters from houses/villages and then collect the observations on all the sampling units available in the selected clusters.
Conditions under which the cluster sampling is used

Cluster sampling is preferred when

i. No reliable listing of elements is available and it is expensive to prepare it.

ii. Even if the list of elements is available, the location or identification of the units may be difficult.

iii. A necessary condition for the validity of this procedure is that every unit of the population under study must correspond to one and only one unit of the cluster so that the total number of sampling units in the frame may cover all the units of the population under study without any omission or duplication. When this condition is not satisfied, bias is introduced.

Open segment and closed segment

It is not necessary that all the elements associated with an area segment need be located physically within its boundaries. For example, in the study of farms, the different fields of the same farm need not lie within the same area segment. Such a segment is called an open segment.

In a closed segment, the sum of the characteristic under study, i.e., area, livestock etc. for all the elements associated with the segment will account for all the area, livestock etc. within the segment.
Construction of clusters

The clusters are constructed such that the sampling units are heterogeneous within the clusters and homogeneous among the clusters. The reason for this will become clear later. This is opposite to the construction of the strata in the stratified sampling.

There are two options to construct the clusters – equal size and unequal size. We discuss the estimation of population means and its variance in both the cases.

Case of equal clusters

- Suppose the population is divided into $N$ clusters and each cluster is of size $n$.
- Select a sample of $n$ clusters from $N$ clusters by the method of SRS, generally WOR.

So

$$\text{total population size} = NM$$
$$\text{total sample size} = nM.$$  

Let

$y_{ij}$: Value of the characteristic under study for the value of $j^{th}$ element ($j = 1,2,\ldots,M$) in the $i^{th}$ cluster ($i = 1,2,\ldots,N$)

$$\bar{y}_i = \frac{1}{M} \sum_{j=1}^{M} y_{ij} \quad \text{mean per element of } i^{th} \text{ cluster}.$$
**Estimation of population mean**

First select $n$ clusters from $N$ clusters by SRSWOR. Based on $n$ clusters find the mean of each cluster separately based on all the units in every cluster. So we have the cluster means as $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n$. Consider the mean of all such cluster means as an estimator of population mean as

$$\bar{y}_{cl} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i$$

**Bias**

$$E(\bar{y}_{cl}) = \frac{1}{n} \sum_{i=1}^{n} E(\bar{y}_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \bar{Y}$$  

(since SRS is used)

$$= \bar{Y}.$$  

Thus $\bar{y}_{cl}$ is an unbiased estimator of $\bar{Y}$.

**Variance**

The variance of $\bar{y}_{cl}$ can be derived on the same lines as deriving the variance of sample mean in SRSWOR. The only difference is that in SRSWOR, the sampling units are $y_1, y_2, \ldots, y_n$ whereas in case of $\bar{y}_{cl}$, the sampling units are $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n$.

Note that in case of SRSWOR, $Var(\bar{Y}) = \frac{N-n}{Nn} S^2$ and $\widehat{Var}(\bar{Y}) = \frac{N-n}{Nn} S^2$

$$Var(\bar{y}_{cl}) = E(\bar{y}_{cl} - \bar{Y})^2$$

$$= \frac{N-n}{Nn} S_b^2$$

where $S_b^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\bar{y}_i - \bar{Y})^2$ which is the mean sum of square between the cluster means in the population.
**Estimate of variance**

Using again the philosophy of estimate of variance in case of SRSWOR, we can find

$$\widehat{Var}(y_{cl}) = \frac{N - n}{nn} \cdot s^2_b$$

where $s^2_b = \frac{1}{n - 1} \sum_{i=1}^{n} (\bar{y}_i - \bar{y}_{cl})^2$ is the mean sum of squares between cluster means in the sample.

**Comparison with SRS**

If an equivalent sample of $nM$ units were to be selected from the population of $NM$ units by SRSWOR, the variance of the mean per element would be

$$Var(y_{nM}) = \frac{NM - nM}{NM} \cdot \frac{S^2}{nM}$$

$$= \frac{f}{n} \cdot \frac{S^2}{M}$$

where $f = \frac{N - n}{N}$ and $S^2 = \frac{1}{NM - 1} \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{Y})^2$.

Also

$$Var(y_{cl}) = \frac{N - n}{nn} \cdot \frac{S^2_b}{n}$$

$$= \frac{f}{n} \cdot S^2_b.$$
Consider 
\[(NM - 1)S^2 = \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \overline{Y})^2\]
\[= \sum_{i=1}^{N} \sum_{j=1}^{M} [(y_{ij} - \overline{y}_i) + (\overline{y}_i - \overline{Y})]^2\]
\[= \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \overline{y}_i)^2 + \sum_{i=1}^{N} \sum_{j=1}^{M} (\overline{y}_i - \overline{Y})^2\]
\[= N(M - 1)S_w^2 + M(N - 1)S_b^2\]

where
\[\overline{S}_w^2 = \frac{1}{N} \sum_{j=1}^{N} S_i^2\]

is the mean sum of squares within clusters in the population.
\[S_i^2 = \frac{1}{M - 1} \sum_{j=1}^{M} (y_{ij} - \overline{y}_i)^2\]
is the mean sum of squares for the \(i^{th}\) cluster.

The efficiency of cluster sampling over SRSWOR is
\[E = \frac{\text{Var}(\overline{y}_{NM})}{\text{Var}(\overline{y}_{cl})}\]
\[= \frac{S^2}{MS_b^2}\]
\[= \frac{1}{(NM - 1)} \left[ \frac{N(M - 1)S_w^2}{M} + (N - 1) \right].\]

Thus the relative efficiency increases when \(S_w^2\) is large and \(S_b^2\) is small. So cluster sampling will be efficient if clusters are so formed that the variation between the cluster means is as small as possible while variation within the clusters is as large as possible.
**Efficiency in terms of intra class correlation**

The intra class correlation between the elements within a cluster is given by

\[
\rho = \frac{E(y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{E(y_{ij} - \bar{Y})}; \quad -\frac{1}{M-1} \leq \rho \leq 1
\]

\[
= \frac{1}{MN(M-1)} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k(x,j)=1}^{M} (y_{ij} - \bar{Y})(y_{ik} - \bar{Y})
\]

\[
= \frac{1}{MN(M-1)} \left( \frac{MN-1}{MN} \right) S^2
\]

\[
= \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{(MN-1)(M-1)S^2}.
\]

Consider

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

\[
= \sum_{i=1}^{N} \left[ \frac{1}{M^2} \sum_{j=1}^{M} (y_{ij} - \bar{Y})^2 + \frac{1}{M^2} \sum_{j=1}^{M} \sum_{k(x,j)=1}^{M} (y_{ij} - \bar{Y})(y_{ik} - \bar{Y}) \right]^2
\]

\[
\Rightarrow \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k(x,j)=1}^{M} (y_{ij} - \bar{Y})(y_{ik} - \bar{Y}) = M^2 \sum_{i=1}^{N} (\bar{Y}_i - \bar{Y})^2 - \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{Y})^2
\]
or \[ \rho(MN - 1)(M - 1)S^2 = M^2(N - 1)S_b^2 - (NM - 1)S^2 \]

or \[ S_b^2 = \frac{(MN - 1)}{M^2(N - 1)}[1 + \rho(M - 1)S^2]. \]

The variance of \( \bar{y}_{cl} \) now becomes

\[
Var(\bar{y}_{cl}) = \frac{N - n}{N} S_b^2
\]

\[
= \frac{N - n}{Nn} \frac{MN - 1}{N - 1} \frac{S^2}{M^2} [1 + (M - 1)\rho].
\]

For large \( N \), \( \frac{MN - 1}{MN} \approx 1, \frac{N - n}{N} \approx 1 \) and so

\[
Var(\bar{y}_{cl}) \approx \frac{1}{nM} S^2 [1 + (M - 1)\rho].
\]

The variance of sample mean under SRSWOR for large \( N \) is

\[
Var(\bar{y}_{nm}) = \frac{S^2}{nM}.
\]

The relative efficiency for large \( N \) is now given by

\[
E = \frac{Var(\bar{y}_{nm})}{Var(\bar{y}_{cl})}
\]

\[
= \frac{S^2}{nM}
\]

\[
= \frac{S^2}{nM} \left[ 1 + (M - 1)\rho \right]
\]

\[
= \frac{1}{1 + (M - 1)\rho}; \quad 1 \leq \rho \leq \frac{1}{(M - 1)}.
\]
If \( M = 1 \) then \( E = 1 \), i.e., SRS and cluster sampling are equally efficient. Each cluster will consist of one unit, i.e., SRS.

If \( M > 1 \), then cluster sampling is more efficient when

\[ E > 1 \]

or \( (M - 1) \rho < 0 \)

or \( \rho < 0 \).

If \( \rho = 0 \), then \( E = 1 \), i.e., there is no error which means that the units in each cluster are arranged randomly. So the sample is heterogeneous.

In practice, \( \rho \) is usually positive and \( \rho \) decreases as \( M \) increases but the rate of decrease in \( \rho \) is much lower in comparison to the rate of increase in \( M \). The situation that \( \rho > 0 \) is possible when the nearby units are grouped together to form cluster and which are completely enumerated.

There are situations when \( \rho < 0 \).

---

**Estimation of relative efficiency**

The relative efficiency of cluster sampling relative to an equivalent SRSWOR is obtained as

\[
E = \frac{S^2}{MS_b^2}
\]

An estimator of \( E \) can be obtained by substituting the estimates of \( S^2 \) and \( S_b^2 \).

Since \( \bar{y}_{cl} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i \) is the mean of \( n \) means \( \bar{y}_i \) from a population of \( N \) means \( \bar{y}_i, i = 1, 2, ..., N \) which are drawn by SRSWOR, so from the theory of SRSWOR,
\[ E(s_b^2) = E \left[ \frac{1}{n-1} \sum_{i=1}^{n} (\overline{y}_i - \overline{Y})^2 \right] \]

\[ = \frac{1}{N-1} \sum_{i=1}^{N} (\overline{y}_i - \overline{Y})^2 \]

\[ = S_b^2. \]

Thus \( s_b^2 \) is an unbiased estimator of \( S_b^2 \).

Since \( s_w^2 = \frac{1}{n} \sum_{i=1}^{n} S_i^2 \) is the mean of \( n \) mean sum of squares \( S_i^2 \) drawn from the population of \( N \) mean sums of squares \( S_i^2, \ i = 1, 2, ..., N \), so it follows from the theory of SRSWOR that

\[ E(s_w^2) = E \left[ \frac{1}{n} \sum_{i=1}^{n} S_i^2 \right] \]

\[ = \frac{1}{N} \sum_{i=1}^{N} S_i^2 \]

\[ = \overline{S}_w^2. \]

Thus \( \overline{S}_w^2 \) is an unbiased estimator of \( \overline{S}_w^2 \).
Consider

\[ S^2 = \frac{1}{MN-1} \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{Y})^2 \]

or \((MN-1)S^2 = \sum_{i=1}^{N} \sum_{j=1}^{M} \left[ (y_{ij} - \bar{Y}_i) + (\bar{Y}_i - \bar{Y}) \right]^2\]

\[ = \sum_{i=1}^{N} \sum_{j=1}^{M} \left[ (y_{ij} - \bar{Y}_i)^2 + (\bar{Y}_i - \bar{Y})^2 \right] \]

\[ = \sum_{i=1}^{N} (M-1)S_i^2 + M(N-1)S_b^2 \]

\[ = N(M-1)\bar{S}_w^2 + M(N-1)S_b^2. \]

An unbiased estimator of \( S^2 \) can be obtained as

\[ \hat{S}^2 = \frac{1}{MN-1} \left[ N(M-1)\bar{S}_w^2 + M(N-1)S_b^2 \right] \]

so

\[ \widehat{Var} (\bar{Y}_{cl}) = \frac{N-n}{Nn} s_b^2 \]

\[ \widehat{Var} (\bar{Y}_{nM}) = \frac{N-n}{Nn} \frac{\hat{S}^2}{M} \]

where \( s_b^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{Y}_i - \bar{Y}_{cl})^2 \).
An estimate of efficiency \( E = \frac{S^2}{MS_b^2} \) is

\[
\hat{E} = \frac{N(M - 1)\bar{s}_w^2 + M(N - 1)s_b^2}{M(NM - 1)s_b^2}.
\]

If \( N \) is large so that \( M(N - 1) = MN \) and \( MN - 1 = MN \), then

\[
E = \frac{1}{M} + \left( \frac{M - 1}{M} \right) \frac{\bar{s}_w^2}{MS_b^2}
\]

and its estimate is

\[
\hat{E} = \frac{1}{M} + \left( \frac{M - 1}{M} \right) \frac{\bar{s}_w^2}{MS_b^2}.
\]