CLUSTER SAMPLING

Sampling Scheme :

Its advantages and disadvantages

Cluster Sampling consists in forming suitable clusters of units and surveying all the units in a sample of clusters selected according to an appropriate sampling scheme. The advantages of cluster sampling from the point of view of cost arise mainly due to the fact that collection of data for nearby units is easier, faster, cheaper and more convenient than observing units scattered over a region. For instance, in ~~application~~ a population survey it may be cheaper to collect data from all the persons in a sample of households than from a sample of the same no. of persons selected directly from all the persons. Similarly, it would be operationally more convenient to survey all households situated in a sample of villages than to survey a sample of the same no. of households selected at random from a list of all households.

Another example of utility of cluster sampling is provided by crop survey where locating a randomly selected plot requires a considerable part of total time taken for the survey. But once the plot is located, the time taken for identifying and surveying a few neighbouring plots will be only marginal.

Because of its operational convenience and the possible reduction of cost, cluster sampling is resorted to in many surveys using mutually exclusive non-overlapping clusters formed by grouping nearby units which can be conveniently observed together. For a given total number of sampling units, cluster sampling is less efficient than sampling of individual units from the view point of sampling variance as the latter expected to provide a better cross section of the population than the former due to the usual tendency of units in a cluster to be similar. In fact the sampling efficiency of cluster sampling is likely to decrease with increasing cluster size. (However, cluster sampling is operationally more convenient and less costly than sampling of units directly due to the possible saving in time for journey, identification, contact etc. and hence in many principal situation the loss in sampling efficiency is likely to be offset by the reduction

– Why cluster sampling is used instead of its disadvantages?

Total : NM units

N Clusters (mutually exclusive)

M units within each cluster

<u>Case-I</u> : <u>One cluster selected (SRS)</u>

Y_{ij} : j^{th} unit in the i^{th} cluster , $i = 1(1)N$, $j = 1(1)M$

$\bar{Y}_i$: $\frac{1}{M} \sum_{j=1}^{M} Y_{ij}$ = i^{th} cluster mean

$\bar{Y}$: $\frac{1}{MN} \sum_{i=1}^{N} \sum_{j=1}^{M} Y_{ij} = \frac{1}{N} \sum_{i=1}^{N} \bar{Y}_i$ = Population Mean

$\hat{\bar{Y}}_c$ = Mean of Selected Cluster

<u>Claim</u>: $\hat{\bar{Y}}_c$ is unbiased for $\bar{Y}$.

<u>Proof</u>: $\hat{\bar{Y}}_c$ can take values $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_N$ each with probability $\frac{1}{N}$.

$$\therefore E(\hat{\bar{Y}}_c) = \frac{1}{N} \sum_{i=1}^{N} \bar{Y}_i = \bar{Y}$$

<u>Variance</u>: $\mathrm{Var}(\hat{\bar{Y}}_c) = \frac{1}{N} \sum_{i=1}^{N} (\bar{Y}_i - \bar{Y})^2 = \sigma_b^2$

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

$$= \frac{1}{NM^2} \sum_{i=1}^{N} \left[\sum_{j=1}^{M} (Y_{ij} - \bar{Y})^2 + \sum_{j \neq j'=1}^{M} \sum (Y_{ij} - \bar{Y})(Y_{ij'} - \bar{Y}) \right]$$

$$= \frac{1}{NM^2} \left[\sum_{i=1}^{N} \sum_{j=1}^{M} (Y_{ij} - \bar{Y})^2 + \sum_{i=1}^{N} \sum \sum_{j \neq j'=1}^{M} (Y_{ij} - \bar{Y})(Y_{ij'} - \bar{Y}) \right]$$

Total Variance of all observations : $\sigma^2 = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} (Y_{ij} - \bar{Y})^2$

Between Cluster Variance : $\sigma_b^2 = \frac{1}{N} \sum_{i=1}^{N} (\bar{Y}_i - \bar{Y})^2$

Intra-cluster correlation coefficient : $\rho_c = \dfrac{\sum_{i=1}^{N} \sum \sum_{j \neq j'=1}^{M} (Y_{ij} - \bar{Y}_0)(Y_{ij'} - \bar{Y})}{(M-1) \sum_{i=1}^{N} \sum_{j=1}^{M} (Y_{ij} - \bar{Y})^2}$

Scanned with CamScanner

$$\therefore \; Var(\hat{\bar{Y}}_c) = \frac{1}{NM^2}\left[NM\sigma^2 + (M-1)\rho_c NM\sigma^2\right]$$

$$= \frac{1}{M}\sigma^2\left[1 + (M-1)\rho_c\right]$$

If ~~just~~ instead of cluster sampling, a SRSWR of size M is drawn from the population of size NM, then the unbiased estimator of $\bar{Y}$ will be the sample mean $\bar{y}_r$ and its variance is

$$Var(\bar{y}_r) = \frac{\sigma^2}{M}$$

$\therefore$ Cluster sampling will be more efficient than SRSWR, if

$$Var(\hat{\bar{Y}}_c) < Var(\bar{y}_r)$$

$$\Rightarrow \quad \frac{\sigma^2}{M}\left[1 + (M-1)\rho_c\right] < \frac{\sigma^2}{M}$$

$$\Rightarrow \quad 1 + (M-1)\rho_c < 1 \qquad [\because \sigma^2 > 0,\; M \in \mathbb{N}]$$

$$\Rightarrow \quad (M-1)\rho_c < 0$$

$$\Rightarrow \quad \rho_c < 0 \qquad [M > 1].$$

Therefore, cluster sampling will be more efficient than SRSWR only if ρ_c is negative. But in practice ρ_c is usually positive when nearby units are grouped to form clusters and hence cluster sampling is usually less efficient than SRSWR.

Case-II : Sampling of n-clusters :

$\bar{y}_i$: Mean of the ith sample cluster, $i = 1(1)n$

Claim: U.E. of $\bar{Y}$ is $\frac{1}{n}\sum_{i=1}^{n}\bar{y}_i = \hat{\bar{Y}}_c$

Where, $\bar{y}_i = \frac{1}{M}\sum_{j=1}^{M} y_{ij}$

Proof: $E(\bar{y}_i) = \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i = \bar{Y}$, since, $\bar{y}_i$ can take values $\bar{Y}_1 \ldots \bar{Y}_N$ with equal probability $\frac{1}{N}$.

$\therefore$ $E(\hat{\bar{Y}}_c) = \frac{1}{n}\sum_{i=1}^{n} E(\bar{y}_i) = \bar{Y}$

Variance

$$Var\left(\frac{1}{n}\sum_{i=1}^{n}\bar{y}_i\right) = \frac{1}{n^2}\left[\sum_{i=1}^{n}\left\{Var(\bar{y}_i) + 2\sum_{i \neq i'=1}^{n}\sum Cov(\bar{y}_i, \bar{y}_{i'})\right]\right.$$

$$Var(\hat{\bar{Y}}_c) = \frac{1}{n^2}\sum_{i=1}^{n} Var(\bar{y}_i) \quad \left[\because Cov(\bar{y}_i, \bar{y}_{i'}) = 0 \ \forall i \neq i' = 1(1)n\right.$$

$$= \frac{\sigma_b^2}{n}$$

, the clusters are mutually ~~exclusive~~ independent]

Carefully note that in case of sampling of n clusters, we are actually doing a simple random sampling from a population whose elements are cluster means as $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_N$.

Let, the sample cluster means are $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n$.

Claim: $\hat{\bar{y}}_{cl} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$ is an unbiased estimator of $\bar{Y}$, the population mean.

This is proved previously.

Now,
$$Var(\hat{\bar{y}}_{cl}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(\bar{y}_i) + \frac{1}{n^2}\sum_{i \neq i'=1}^{n}\sum Cov(\bar{y}_i, \bar{y}_{i'})$$

$Cov(\bar{y}_i, \bar{y}_{i'}) = 0 \ \forall i \neq i' = 1(1)n$, since each clusters are mutually independent.

We can directly do this as

$$Var(\hat{\bar{y}}_{cl}) = \frac{N-n}{nN} S_b^2, \text{ where } S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{Y}_i - \bar{Y})^2$$

= Between Cluster SS for population

and
$$\widehat{Var(\hat{\bar{y}}_{cl})} = \frac{N-n}{nN} s_b^2, \text{ where } s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_i - \bar{y})^2$$

= Between Cluster SS for sample

$$\bar{y}_i = \frac{1}{M}\sum_{j=1}^{M} y_{ij} \quad \text{and} \quad \sigma_b^2 = \frac{1}{N}\sum_{i=1}^{N}(\bar{Y}_i - \bar{Y})^2$$

The u.e. of σ_b^2 is $\therefore$ $s_b^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(\bar{y}_i - \hat{\bar{Y}}_c)^2$

= Variance of sample cluster means

$$\therefore \widehat{Var}(\hat{\bar{Y}}_c) = \frac{s_b^2}{n}, \text{ if clusters are selected by SRSWR}$$

If clusters are selected by SRSWOR;

$$\widehat{Var}(\hat{\bar{Y}}_c) = \frac{N-n}{n(N-1)}\hat{\sigma}_b^2 = \frac{s_b^2}{n}\cdot\frac{N-n}{N-1}$$

If a SRSWOR of size nM is selected out of NM units instead of cluster sampling, the variance of the u.e. of $\bar{y}_r$ is

$$Var(\bar{y}_r) = \frac{\sigma^2}{nM}\cdot\frac{NM-nM}{NM-1} = \frac{\sigma^2}{n}\,\frac{N-n}{NM-1}$$

In this case, the efficiency of cluster sampling w.r.to SRSWOR is

$$E = \frac{Var(\bar{y}_r)}{Var(\hat{\bar{Y}}_c)} = \frac{\frac{\sigma^2}{n}\,\frac{N-n}{NM-1}}{\frac{\sigma_b^2}{n}\,\frac{N-n}{N-1}}$$

$$= \frac{\sigma^2}{\sigma_b^2}\cdot\frac{N-1}{NM-1} = \frac{\sigma^2}{\frac{\sigma^2}{M}\left[1+(M-1)\rho_c\right]}\cdot\frac{N-1}{NM-1}$$

$$= \frac{M}{[1+(M-1)\rho_c]}\cdot\frac{(N-1)}{(NM-1)} \qquad \left[\because \sigma_b^2 = \frac{\sigma^2}{M}\left(1+(M-1)\rho_c\right)\right]$$

Remark: $E > 1$, if cluster sampling is more efficient than SRSWOR.

$$\Rightarrow \left(\frac{M}{1+(M-1)\rho_c}\right)\cdot\left(\frac{N-1}{NM-1}\right) > 1$$

$$\Rightarrow NM - M > NM - 1 + \rho_c\,(M-1)\,(NM-1)$$

$$\Rightarrow -(M-1) > \rho_c\,(M-1)\,(NM-1)$$

$$\Rightarrow \rho_c < -\frac{1}{NM-1} \qquad [M > 1]$$

But the ~~intraclass~~ intra cluster correlation coefficient is positive when nearby units are grouped ~~into~~ to form clusters and hence cluster sampling is usually less efficient than SRSWOR.

Case – III : n-clusters selected (each cluster consists of M_i units).:

Y_{ij} : j^{th} unit in the i^{th} cluster $\left(\begin{array}{l} j=1(1)M_i \\ i=1(1)N \end{array}\right)$ in population

$\bar{Y}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} Y_{ij} =$ i^{th} cluster mean $\left(= \frac{Y_i}{M_i}\right)$ in population

$Y_i =$ i^{th} cluster total in population

$\bar{Y} =$ Mean of the total $\sum_{i=1}^{N} M_i$ population $= \frac{\sum_{i=1}^{N}\sum_{j=1}^{M_i} Y_{ij}}{\sum_{i=1}^{N} M_i} = \frac{\sum_{i=1}^{N} M_i \bar{Y}_i}{\sum_{i=1}^{N} M_i}$

We take n-clusters from N-clusters and survey all the units in the selected clusters.

Consider, $y_{ij} =$ j^{th} unit in the i^{th} cluster in sample $\left(\begin{array}{l} j=1(1)M_i \\ i=1(1)n \end{array}\right)$

$\bar{y}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij} =$ i^{th} cluster mean in sample $(i=1(1)n)$

$y_i = \sum_{j=1}^{M_i} y_{ij} =$ i^{th} cluster total in sample $(i=1(1)n)$

We know,

$$\frac{y_1+y_2+\dots+y_n}{n} \text{ estimates } \frac{Y_1+Y_2+\dots+Y_N}{N}.$$

i.e. $\frac{N}{n}\sum_{i=1}^{n} y_i$ " $\sum_{i=1}^{N} Y_i$:

i.e. $\frac{N}{n}\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{N} M_i} = \hat{\bar{Y}}_c$ " $\frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} M_i} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M_i} Y_{ij}}{\sum_{i=1}^{N} M_i} = \bar{Y}$

Claim: $\hat{\bar{Y}}_c$ is an u.e. of $\bar{Y}$, where $\hat{\bar{Y}}_c = \frac{N}{n}\frac{1}{\sum_{i=1}^{N} M_i}\sum_{i=1}^{n}\sum_{j=1}^{M_i} y_{ij}$

$$= \frac{N}{n}\cdot\frac{\sum_{i=1}^{n} M_i\bar{y}_i}{\sum_{i=1}^{N} M_i}$$

~~Proof: Now, $E(\bar{y}_i) = \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i$, since, $\bar{y}_i$ takes any value~~

$\bar{y}_i =$ i^{th} Sample cluster ~~total~~ mean can take any value ~~among $\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_N$~~

~~ie $E(\bar{y}_i) = \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i = \frac{1}{M_i}\sum_{i=1}^{N}\sum_{j=1}^{M_i} Y_{ij}$~~

Proof: Since, $E\left(\frac{y_1+y_2+\dots+y_n}{n}\right) = \frac{1}{N}\sum_{i=1}^{N} Y_i$

$$\Rightarrow E\left(\frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{M_i} y_{ij}\right) = \sum_{i=1}^{N}\sum_{j=1}^{M_i} Y_{ij}$$

$$\Rightarrow E\left[\frac{N}{n\sum_{i=1}^{N}M_i}\sum_{i=1}^{n} M_i\bar{y}_i\right] = \frac{1}{\sum_{i=1}^{N}M_i}\sum_{i=1}^{N}\sum_{j=1}^{M_i} Y_{ij} = \bar{Y}$$

$$\Rightarrow E\left[\hat{\bar{Y}}_c\right] = \bar{Y}$$

$\therefore$ $\hat{\bar{Y}}_c$ is an u.e. of $\bar{Y}$.

Variance:

~~$$Var(\hat{\bar{Y}}_c) = \left(\frac{N}{n\sum_{i=1}^{N}M_i}\right)^2 Var\left(\sum_{i=1}^{n} M_i\bar{y}_i\right)$$~~

~~$$= \left(\frac{N}{n\sum_{i=1}^{N}M_i}\right)^2 \sum_{i=1}^{n} M_i^2\, Var(\bar{y}_i) \quad \left[\because Cov(\bar{y}_i, \bar{y}_{i'}) = 0,\ i \neq i' = 1(1)n\right]$$~~

~~$$= \left(\frac{N}{n\sum_{i=1}^{N}M_i}\right)^2 \left(\sum_{i=1}^{n} M_i^2\right) \frac{\sigma_b^2}{n}\left(\frac{N-n}{N-1}\right)$$~~

~~$$= \left(\frac{N}{n\sum_{i=1}^{N}M_i}\right)^2 \left(\sum_{i=1}^{n} M_i^2\right)\left[\sigma_b^2\left(\frac{N-n}{N-1}\right)\right]$$~~

$$Var(\hat{\bar{Y}}_c) = \left(\frac{N}{\sum_{i=1}^{N}M_i}\right)^2 Var\left(\frac{1}{n}\sum_{i=1}^{n} M_i\bar{y}_i\right) = \left(\frac{N}{\sum_{i=1}^{N}M_i}\right)^2 \frac{1}{n^2}\sum_{i=1}^{n} M_i^2\, Var(\bar{y}_i)\left(\frac{N-n}{N-1}\right)$$

$$= \left(\frac{N}{\sum_{i=1}^{N}M_i}\right)^2 \frac{1}{n^2}\sum_{i=1}^{n} M_i^2\, \sigma_b^2\left(\frac{N-n}{N-1}\right) \quad \left[\because Cov(\bar{y}_i, \bar{y}_{i'}) = 0\ \forall\, i \neq i' = 1(1)n\right]$$

where, $\sigma_b^2 = \frac{1}{\sum_{i=1}^{N} M_i} \sum_{i=1}^{N} (\bar{Y}_i - \bar{Y})^2 M_i = Var(\bar{y}_i)$, $i = 1(1)n$

$$\left[Var(\bar{y}_i) = E\left[M_i(\bar{y}_i - E(\bar{y}_i))\right]^2 = E\left[\bar{y}_i - \bar{Y}_i\right]^2 \right.$$

$$\left. = E\left[\frac{1}{M_i}\sum_{j=1}^{M_i}(y_{ij} - \bar{Y})\right]^2 \right]$$