UNIT-III

Systematic random Sampling:-

Systematic Sampling is a commonly employed technique of the complete and up to data list of the sampling unit is available. Let us suppose that N sampling units are serially numbered from 1 to N in some order and a sample of size n is to be drawn such that

$$N = nk \Rightarrow n = N/n$$

Variance of the estimate mean:-

$$\text{var}(\bar{y}_{sy}) = \frac{N-1}{N} S^2 - \frac{(n-1)}{N} k S^2_{wsy}$$

where

$$S^2_{wsy} = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - y_{i\cdot})^2$$

Proof:-

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{k} \sum_{i=1}^{n} (y_{ij} - \bar{y}_{\cdot\cdot})^2$$

$$(N-1) S^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{\cdot\cdot})^2$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i\cdot} + \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i\cdot}) + \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{i\cdot} - \bar{y}_{\cdot\cdot})$$

$$+ 2 \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i\cdot}) (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})$$

$$(N-1)S^2 = k(n-1)S^2_{wsy} + nk \, var(\bar{y}_{sys}) \quad (\because nk = N)$$

$$var(\bar{y}_{sys}) = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S^2_{wsy}.$$

Theorem :-

$$var(\bar{y}_{sys}) = \frac{nk-1}{nk} \frac{S^2}{n} (1+(n-1)\rho)$$

$$\rho = \sum_{i=1}^{k} \sum_{i+j=1}^{n} \frac{(y_{ij} - \bar{y}_{\cdot\cdot})(y_{ij} - \bar{y}_{\cdot\cdot})}{nk(n-1)S^2}$$

$$= \sum_{i=1}^{k} \sum_{i+j=1}^{n} \frac{(y_{ij} - \bar{y}_{\cdot\cdot})(y_{ij} - \bar{y}_{\cdot\cdot})}{(n-1)(nk-1)S^2}.$$

Proof :-

$$var(\bar{y}_{sys}) = \frac{1}{k} \sum_{i=1}^{k} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

$$= \frac{1}{k} \left[\sum_{i=1}^{k} 1/n \sum_{i=1}^{n} (y_{ij} - \bar{y}_{\cdot\cdot})^2 \right]$$

$$nk^2 \, var(\bar{y}_{sys}) = \sum_{i=1}^{k} \left[\sum_{j=1}^{n} y_{ij} - \bar{y}_{\cdot\cdot} \right]^2$$

$$\Rightarrow \sum_{i=} \left[\sum_{i=1}^{n} (y_{ij} - \bar{y}_{\cdot\cdot})^2 + \sum_{i+j=1}^{n} (y_{ij} - \bar{y}_{\cdot\cdot})(y_{ij} - \bar{y}_{\cdot\cdot}) \right.$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{\cdot\cdot})^2 + \sum_{i=1}^{n} \sum_{i+j=1} (y_{ij} - \bar{y}_{\cdot\cdot})(y_{ij} - \bar{y}_{\cdot\cdot})$$

$$= (nK-1)\,\delta + (n-1)(nK-1)S^2\rho$$

$$= (nK-1)\,S^2[1+(n-1)\rho].$$

$$\text{Var}[\bar{y}_{sys}] = \frac{nK-1}{nK}\,\frac{S^2}{n}\{1+(n-1)\rho\}$$

Theorem:-

$$\text{Var}(\bar{y}_{sys}) = \frac{k-1}{nk}\,S^2_{wsc}[1+(n-1)\rho_{wsc}]$$

Proof:-

$$\text{Var}(\bar{y}_{sys}) = \frac{1}{K}\sum_{i=1}^{K}(\bar{y}_{i.}-\bar{y}_{..})^2$$

$$= \frac{1}{K}\sum_{i=1}^{K}\left[\frac{1}{n}\sum_{i=1}^{n} y_{ij} - \frac{1}{n}\sum_{i=1}^{n} y_{ij}\right]^2$$

$$= \frac{1}{n^2K}\left[\sum_{i=1}^{K}(y_{ij}-\bar{y}_{.j})^2 + \sum_{i=1}^{K}\sum_{j\neq j'=1}^{n}[y_{ij}-\bar{y}_{.j}][y_{ij'}-\bar{y}_{j'.}]\right]$$

$$= \frac{1}{n^2K}\left[n(K-1)S^2_{wsc} + n(n-1)(K-1)\delta_{wsc} + S^2_{wsc}\right]$$

$$= \frac{K-1}{nK}\,S^2[1+(n-1)\rho_{wsc}]$$

Systematic sampling vs Stratified Random sampling:-

$$\text{Var}(\bar{y}_{st}) = \sum_{j=1}^{n} \left[\frac{1}{nj} - \frac{1}{Nj}\right] P_i^2 \sigma_j^2$$

But

$Ni = K$ and $nj = 1$ $[j = 1, 2, \ldots n]$ and

$$P_j = \frac{N_j}{N} = \frac{K}{nK} = \frac{1}{n}$$

$$\text{var}(\bar{y}_{st}) = \sum_{j=1}^{n} (1 - 1/K)\ 1/n^2\ \sigma_j^2 \rightarrow (1)$$

$$= \frac{K}{n^2 K} \sum_{j=1}^{n} \sigma_j^2$$

$$= \frac{K-1}{n^2 K} \sum_{j=1}^{n} \left[\frac{1}{K-1} \sum_{i=1}^{K} (y_{ij} - \bar{y}_{.j})^2\right].$$

$$= \frac{K}{n^2 K} \sum_{j=1}^{n} s_i^2$$

$$= \frac{K-1}{n^2 K} \sum_{j=1}^{n} \left[\frac{1}{K-1} \sum_{i=1}^{K} (y_{ij} - \bar{y}_{.j})^2\right]$$

$$= \frac{1}{n^2 K} \sum_{i=1}^{K} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{.j})^2 -$$

$$\text{var}(\bar{y}_{st}) = \frac{K-1}{nK} S_{wse}^2.$$

Merits and demerits of systematic sampling :-

1) Systematic sampling is operationally more convient than simple random sampling or stratified sampling time and work is involved is also respectively much less.

2) Systematics sampling may be more efficient than simple random sampling provided the frame is arranged wholly at random. the most common approach to randomness is provided by alphabetical list such as name in telephone directory although even those may have certain non-random characteristics.

Demerits :-

the main advantage of systematic sampling is that systematic samples are not in general random sample since the requirement in merit two is rarely fulfilled.

If N is not multiple of n, then

1) The actual sample size is different from that required

2) Sample mean is not an unbiased estimate of the population mean.

However these disadvantage can be overcom by adopting a technique known as circular systemati Sampling [C.S.S]

3) In systematic sampling we have,

$$Var(\bar{y}_{sys}) = \frac{1}{kn} \sum_{i=1}^{n} (\bar{y}_{i.} - \bar{y}_{..})^2$$

Systematic sampling Vs simple random sampling:-

$$Var(\bar{y}_{i.})_n = \left(\frac{N-n}{N}\right) \frac{S^2}{n} \rightarrow (1)$$

$$Var(\bar{y}_{sys}) = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S^2_{wsy} \rightarrow (2)$$

$$Var(y_{i.})_n - Var(\bar{y}_{sys})$$

$$\Rightarrow \left[\frac{N-n}{N}\right] S^2/n - \left[\frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S^2_{wsy}\right]$$

$$\rightarrow \left[\frac{N-n}{Nn}\right] S^2 - \frac{N-1}{N} S^2 + \frac{k(n-1)}{N} S^2_{wsy}$$

$$\Rightarrow \left[\frac{N-n}{n} - (N-1)\right] \frac{S^2}{N} + \frac{k(n-1)}{N} S^2_{wsy}.$$

$$\Rightarrow \left[\frac{N-n}{n} - N + 1\right] \frac{S^2}{N} + \frac{S(n-1)}{N} S^2_{wsy}.$$

$$= \frac{N(1-n)}{n} \frac{S^2}{N} + \frac{K(n-1)}{N} S^2_{wsy}$$

$$= \frac{K(n-1)}{nK} S^2_{wsy} - \frac{(n-1)}{n} S^2$$

$$= \frac{(n-1)}{n} S^2_{sys} - \frac{n-1}{n} S^2$$

$$\text{Var}(y_i) - \text{var}(y_{sys}) = \left[\frac{n-1}{n}\right]\left[S^2_{wsy} - S^2_{sys}\right]$$

$$\text{Var}(\bar{y}_i)_n - \text{var}(\bar{y}_{sys}) > 0$$

$$S^2_{wsy} > S^2$$

Theorem :-

$$\text{Var}(\bar{y}_{st}) \leq \text{var}(\bar{y}_{se}) \leq \text{var}(\bar{y}_n)_R$$

Proof :-

$$Y_i = i \quad (i = 1, 2, \ldots n)$$

$$\sum_{i=1}^{N} Y_i = \sum_{i=1}^{n} i = N(N+1)$$

$$\sum_{i=1}^{N} Y_i^2 = \sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6}$$

$$\bar{Y}_N = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

$$= \frac{1}{N} \frac{N[N+1]}{2}$$

$$\bar{Y}_N = \frac{N+1}{2}$$

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y}_N)^2$$

$$= \frac{1}{N-1} \sum_{i=1}^{N} [Y_i^2 - N\bar{Y}_N^2]$$

$$= \frac{1}{N-1} \left[\frac{N[N+1][2N+1]}{6} - \frac{N[N+1]^2}{4} \right]$$

On simplification :-

$$S^2 = \frac{N(N+1)}{12}$$

$$\text{Var}(\bar{y}_s)_R = \left[\frac{N-n}{N}\right]\left[\frac{S^2}{n}\right] \Rightarrow \left[\frac{1}{n} - \frac{1}{N}\right] S^2$$

$$\text{Var}(\bar{y}_n)_R = \left[\frac{1}{n} - \frac{1}{N}\right] \cdot \frac{N[N+1]}{12} \quad [\because N = nk]$$

$$= \frac{nk-n}{n^2 k} \cdot \frac{nk(nk+1)}{12}$$

$$= \frac{n(k-1)}{n^2 k} \cdot \frac{nk(nk+1)}{12}$$

$$\text{Var}(\bar{y}_n)_R = \frac{k-1(nk+1)}{12} \rightarrow (1)$$

$$\text{var}(\bar{y}_{st}) = \frac{k-1}{n^2 k} \sum_{j=1}^{n} S_i^2$$

$$\Rightarrow \sum_{j=1}^{n} \frac{k(k+1)}{12} = \frac{nk(k+1)}{12} \quad \left[\because S^2 = \frac{N+1}{12}(N)\right]$$

$$\text{var}(\bar{y}_{st}) = \frac{k-1}{n^2 k} \cdot \frac{nk(k+1)}{12}$$

$$\text{var}(\bar{y}_{st}) = \frac{k-1}{n} \cdot \frac{(k+1)}{12} \Rightarrow \frac{k^2+(k-1)}{12n} = \frac{k^2-1}{12n} \rightarrow (2)$$

$$\text{var}(\bar{y}_{sys}) = \frac{1}{k} \sum_{i=1}^{k} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})$$

$$\bar{y}_{i\cdot} = \frac{1}{n} \sum_{i=1}^{n} y_{ij}$$

$$= \frac{1}{n}\left[i + (i+k) + (i+2k) + \cdots + i(n-1)+k\right]$$

$$= \frac{1}{n}\left[ni(1+2+3+\cdots+n-1)k\right]$$

$$\bar{y}_i = \frac{i + (n-1)k}{2}$$

$$\bar{y}_{\cdot\cdot} = \bar{y}_N = \frac{N+1}{2} = \frac{nk+1}{2}$$

$$(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) = \frac{i + (n-1)k}{2} - \frac{nk+1}{2}$$

$$= \frac{i + (nk-k)}{2} - \frac{nk-1}{2}$$

$$= 1 - \frac{(k+1)}{2}$$

$$\text{Var}(\bar{y}_{sys}) = \frac{1}{k}\sum_{i=1}^{k}(y_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

$$= \frac{1}{k}\sum_{i=1}^{k}\left[\frac{i-k+1}{2}\right]^2$$

$$= \frac{1}{k}\sum_{i=1}^{k}\left[i^2 + \left[\frac{k+1}{2}\right]^2 - 2i\frac{[k+1)}{2}\right]$$

$$= \left[\frac{1}{k}\sum_{i=1}^{k} i + \frac{1}{k}\sum_{i=1}^{k}\frac{(k+1)^2}{4} - \frac{1}{k}\sum_{i=1}^{k} 2i\,(k+1)\right]$$

$$= \left[\frac{1}{k}\sum_{i=1}^{k} i + \frac{(k+1)^2}{4} - \frac{k+1}{k}\sum_{i=1}^{k} i\right]$$

$$= \frac{1}{k}\left[\frac{k(k+1)(2k+1)}{6}\right] + \frac{k(+1)^2}{4} - \frac{k+1}{k} - \frac{k(k+1)}{2}$$

$$= \frac{[k+1]\cdot[2k+1]}{6} + \frac{(k+1)^2}{4} - \frac{(k+1)^2}{2}$$

On simplification,

$$\text{Var}(\bar{y}_{sys}) = \frac{k^2-1}{12} \rightarrow (3)$$

From (1), (2) & (3), we get

$$\text{Var}(\bar{y}_{st}) : \text{Var}(\bar{y}_{sys}) : \text{var}(\bar{y}_n)R$$

$$\frac{k^2-1}{12n} : \frac{k^2-1}{12} : \frac{k-1(nk+1)}{12}$$

$\frac{1}{n} : 1 : n$

$Var(\bar{y}_{st}) \leq Var(\bar{y}_{sys}) \leq Var(\bar{y}_n)$ R.

Ratio And Regression Estimators :-

In most of the sample survey information auxiliary variable is available. along with then information on the variable under Study. This information can be suitable used in estimating the mean of the characters under Study mean efficiently. The auxiliary information can be utilized either is the section of the sample or at the estimation of the stage. The common estimation which at the information on the auxiliary variable are ratio and the regression estimates.

Ratio Estimates :-

In the ratio method an auxiliary variable x_i correlated with y_i is obtained

for such unit of the population total X of x_i must be known. In practice x_i is often the value of y_i at some previous time when a complete census was taken. The aim in this method is obtained increased precision by taking advantages of the correlation co-efficient between x_i and y_i.

Sampling Theory

UNIT - IV

Cluster Sampling :-

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 Populations can be divided are called elements of the population. The group of such elements are called clusters.

Equal cluster sampling :-

i) Suppose the population is divided into N clusters and each cluster is of size n.

ii) Select a sample of n clusters from N clusters by the method of SRS, generally WOR, so

total population size = NM

Total sample size = nM.

Estimate of mean and its variance :-

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.

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

Bias :

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

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

$$= \bar{Y}$$

Thus, $\bar{Y}_{et}$ is an unbiased estimates of $\bar{Y}$

Variance :-

$$\text{var}(\bar{y}) = \frac{N-n}{Nn} s^2 \text{ and } \text{var}(\bar{y}) = \frac{N-n}{Nn} s^2$$

$$\text{var}(\bar{Y}_{et}) = E(\bar{Y}_{et} - \bar{Y})^2 = \frac{N-n}{Nn} S_b^2$$

$$S_b^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\bar{Y}_i - \bar{Y})^2$$

In case SRSWOR,

$$\widehat{Var}(\bar{Y}_{ct}) = \frac{N-n}{Nn} s_b^2$$

$$s_b^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{y}_i - \bar{Y}_{ct})^2$$

Multistage Sampling :-

Multistage sampling is defined as a sampling method that divides the populations into groups for conducting research. It is a complex form of cluster sampling sometimes, also known as multistage sampling. During this sampling method, significant clusters of the selected people are split into subgroups at various stages to make it simpler for primary data collection.

Estimation of relative Efficiency :-

$$E = \frac{S^2}{MS_b^2}$$

An estimator of E can be obtained by the

Substituting the estimates of S^2 and S_b^2

$$E(s_b^2) = E\left[\frac{1}{n}\sum_{i=1}^{n}(\bar{y}_i - \bar{y}_e)^2\right]$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}(\bar{Y}_i - \bar{Y})^2$$

$$= S_b^2$$

Y_b^2 is an unbiased estimates of S_b^2

$$S_w^2 = \frac{1}{n}\sum_{i=1}^{n} s_i^2$$

$s_i^2, i = 1, 2, \ldots N$

$$E(S_w^2) = E\left[\frac{1}{N}\sum_{i=1}^{n} s_i^2\right] = \frac{1}{N}\sum_{i=1}^{n} E(s_i^2)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{N}\sum_{i=1}^{N} s_i^2\right]$$

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

$$= S_w^2$$

Consider,

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

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

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

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

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

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

So,

$$\widehat{\text{var}}(\bar{y}_{oi}) = \frac{N-n}{Nn} S_b^2$$

$$\widehat{\text{var}}(\bar{y}_{nM}) = \frac{N-n}{Nn} \frac{\bar{S}_2}{M}$$

where $$S_b^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{y}_i - \bar{y}_{ot})^2$$

$$E = \frac{S^2}{M S_b^2} \text{ ui}$$

$$\bar{E} = \frac{N(M-1) \bar{S}_w^2 + M(N-1) S_b^2}{M(NM-1) S_b^2}$$

$$M(N-1) = MN \quad ; \quad MN - 1 = MN$$

$$E = \frac{1}{M} + \left(\frac{M-1}{M}\right) \frac{\bar{S}_w^2}{MS_b^2}$$

$$\hat{E} = \frac{1}{M} + \left(\frac{M-1}{M}\right) \frac{\bar{S}_w^2}{MS_b^2}$$

Two Stage Sampling With Equal First Stage Units :-

1) The population consist of NM elements.

2) NM elements are grouped into N first stage units of M second stage units each.

3) sample of n first stage units is selected.

4) Units at each stage are selected with SRSWOR.

Two stage sampling.

$$E(\hat{\theta}) = E_1[E_2(\hat{\theta})].$$

$$E(\bar{\theta}) = E_1\left[E_2\{E_3(\bar{\theta})\}\right]$$

$$\text{Var}(\bar{\theta}) = E\left[(\bar{\theta} - \theta)^2\right]$$

$$= E_1 E_2 (\bar{\theta} - \theta)^2$$

Consider,

$$E_2(\bar{\theta} - \theta)^2 = E_2(\bar{\theta}^2) - 2\theta E_2(\bar{\theta}) + \theta^2$$

$$= \left[\{E_2(\bar{\theta})\}^2 + V_2(\bar{\theta})\right] - 2\theta E_2(\bar{\theta}) + \theta^2$$

Average over first stage selection,

$$E_1 E_2(\bar{\theta} - \theta)^2 = E_1\left[E_2(\bar{\theta})\right]^2 + E_1\left[V_2(\bar{\theta}) - 2\theta E_1 E_2(\bar{\theta}) + E_1(\theta^2)\right]$$

$$= E_1\left[\{E_2(\bar{\theta})\}^2 - \theta^2\right] + E_1\left[V_2(\bar{\theta})\right]$$

$$\text{Var}(\bar{\theta}) = V_1\left[E_2(\bar{\theta})\right] + E_1\left(V_2(\bar{\theta})\right).$$

Three stage sampling :-

$$\text{Var}(\bar{\theta}) = V_1\left[E_2\{E_3(\bar{\theta})\}\right] + E_1\left[V_2\{E_3(\bar{\theta})\}\right] + E_1\left[E_2\{V_3(\bar{\theta})\}\right]$$

Estimation of Population mean :-

$$\bar{Y} = \bar{y}_{mn}$$

Bias :-

$$E(\bar{y}) = E_1[E_2(\bar{y}_{mn})]$$

$$= E_1[E_2(\bar{y}_{im}/i)]$$

$$= E_1[E_2(\bar{y}_{im}/i)]$$

$$= E_1\left[\frac{1}{n}\sum_{i=1}^{n}\bar{y}_i\right]$$

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

$$= \bar{Y}.$$

Variance :-

$$Var(\bar{y}) = E_1[V_2(\bar{y}/i)] + V_1[E_2(\bar{y}/i)]$$

$$= E_1\left[V_2\left\{\frac{1}{n}\sum_{i=1}^{n}\bar{y}_i/i\right\}\right] + V_1\left[E_2\left\{\frac{1}{n}\sum_{i=1}^{n}\bar{y}_i/i\right\}\right]$$

$$= E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}V(\bar{y}_i/i)\right] + V_1\left[\frac{1}{n}\sum_{i=1}^{n}E_2(\bar{y}_i/i)\right]$$

$$= E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{m}-\frac{1}{M}\right)S_i^2\right] + V_1\left[\frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\left[\frac{1}{m}-\frac{1}{M}\right]E_1(S_i^2) + V_1(\bar{Y}_c)$$

$$= \frac{1}{n^2}\, n\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + \frac{N-n}{Nn}S_b^2$$

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

where $\bar{S}_w^2 = \frac{1}{N}\sum_{i=1}^{N}S_i^2 = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{ij}-\bar{Y}_i)^2$

$$\bar{S}_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{Y}_i-\bar{Y})^2$$

Two stage sampling with unequal first stage units :-

Let,

Y_{ij} : value of j^{th} second stage unit of j^{th} first stage unit

M_i : number of second stage units in i^{th} first stage unit

M_0 : $\sum_{i=1}^{N} M_i$: total no. of second stage units in the population

m_i : num. of second stage units to be selected from i^{th} first stage unit, if it is in the sample

$m_0 = \sum_{i=1}^{n} m_i$: total no. of second stage units in sample

$$\bar{y}_i(m_i) = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$$

$$\bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij}$$

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} \bar{Y}_i = \bar{Y}_N$$

$$\bar{Y} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}}{\sum_{i=1}^{N} M_i} = \frac{\sum_{i=1}^{N} M_i \bar{Y}_i}{M N} = \frac{1}{N} \sum_{i=1}^{N} U_i \bar{Y}_i$$

$$U_i = \frac{M_i}{M}$$

$$\bar{M} = \frac{1}{N} \sum_{i=1}^{N} M_i$$

Estimation of Mean and Variance :-

$$\hat{\bar{Y}} = \bar{y}_{s2} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i(m_i)$$

Bias :-

$$E(\bar{y}_{s2}) = E\left[\frac{1}{n} \sum_{i=1}^{n} \bar{y}_{t(m_i)}\right]$$

$$= E_1\left[\frac{1}{n} \sum_{i=1}^{n} E_2\, \bar{y}_i(m_i)\right]$$

$$\bar{Y}_{S2} + \frac{N-1}{NM} \frac{1}{N-1} \sum_{i=1}^{n} (M_i - \bar{m}) (\bar{y}_{i(m_i)} - \bar{y}_{S2})$$

Variance :-

$$\text{Var}(\bar{y}_{S2}) = \text{Var}[E(\bar{y}_{S2}/n)] + E[\text{Var}(\bar{y}_{S2}/n)]$$

$$= \text{Var}\left[\frac{1}{n} \sum_{i=1}^{n} \bar{Y}_i\right] + E \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(\bar{y}_{i(m_i)}/i)$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + E\left[\frac{1}{n^2} \sum_{i=1}^{n} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2\right]$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \frac{1}{Nn} \sum_{i=1}^{N} \left[\frac{1}{m_i} - \frac{1}{M_i}\right] S_i^2$$

$$S_b^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\bar{Y}_i - \bar{\bar{Y}}_N)^2$$

$$S_i^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (Y_{ij} - \bar{Y}_i)^2$$

$$MSE(\bar{Y}_{S2}) = \text{Var}(\bar{y}_{S2}) + [\text{Bias}(\bar{y}_{S2})]^2$$

$$= E_1\left[\frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i = \bar{\bar{Y}}_N \quad \pm \bar{Y}.$$

$$\text{Bias}(\bar{y}_{st}) = E(\bar{y}_{st}) - \bar{Y}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i - \frac{1}{NM}\sum_{i=1}^{N}M_i\bar{Y}_i$$

$$= -\frac{1}{NM}\left[\sum_{i=1}^{N}M_i\bar{Y}_i - \frac{1}{N}\left(\sum_{i=1}^{N}\bar{Y}_i\right)\left(\sum_{i=1}^{N}M_i\right)\right]$$

$$= \frac{1}{NM}\sum_{i=1}^{N}(M_i - \bar{M})(\bar{Y}_i - \bar{\bar{Y}}_N)$$

$$\widehat{\text{Bias}}(\bar{y}_{st}) = -\frac{N-1}{NM(n-1)}\sum_{i=1}^{n}(M_i - m)(\bar{y}_{i(m)} - \bar{y}_{st})$$

$$E\left[\widehat{\text{Bias}}(\bar{y}_{st})\right] = -\frac{N-1}{NM}E_1\frac{1}{n-1}\sum_{i=1}^{n}E_2\{(M_i - \bar{m})(\bar{y}_{i(m)} - \bar{y}_{st})\}$$

$$= -\frac{N-1}{NM}E\left[\frac{1}{n-1}\sum_{i=1}^{n}(M_i - \bar{m})(\bar{Y}_i - \bar{\bar{Y}}_n)\right]$$

$$= -\frac{1}{NM}\sum_{i=1}^{N}(M_i - \bar{M})(\bar{Y}_i - \bar{\bar{Y}}_N) = \bar{\bar{Y}}_N - \bar{Y}$$

$$\bar{\bar{y}}_n = \frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i.$$

UNIT-V

Double Sampling :-

The ratio and regression methods of estimation require the knowledge of population mean of an auxiliary variable ($\bar{X}$) to estimate the population mean of study variable ($\bar{Y}$). If the information on the auxiliary variables is not available, then there are two options - one option is to collect a sample only on study variable and use sample mean as an estimator of the population mean.

Optimal allocation :-

optimal allocation is a procedure for dividing the sample survey. A Stratified sample selects separate samples from subgroups called Strata of the population and can often increase the accuracy of survey results.

Double sampling in Ratio method of Estimation :-

If the population mean $\bar{Y}$ is not known, then the double sampling technique is applied,

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

$$\bar{V}_{Rd} = \frac{\bar{Y}}{\bar{X}} \bar{X}'.$$

Let

$$\varepsilon_0 = \frac{\bar{y} - \bar{Y}}{\bar{Y}}, \quad \varepsilon_1 = \frac{\bar{x} - \bar{X}}{\bar{X}}, \quad \varepsilon_2 = \frac{\bar{X}' - \bar{X}}{\bar{X}}$$

$$E(\varepsilon_0) = E(\varepsilon_1) = E(\varepsilon_2) = 0$$

$$E(\varepsilon_1^2) = \left(\frac{1}{n} - \frac{1}{N}\right) C_x^2$$

$$E(\varepsilon_1 \varepsilon_2) = \frac{1}{\bar{X}^2} E(\bar{x} - \bar{X})(\bar{x}' - \bar{X})$$

$$= \frac{1}{\bar{X}^2} E_1\left[E_2(\bar{x} - \bar{X})(\bar{x}' - \bar{X}) \mid n'\right]$$

$$= \frac{1}{\bar{X}^2} E_1\left[(\bar{x}' - \bar{X})^2\right]$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right) \frac{S_x^2}{\bar{X}^2}$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right) C_x^2 = E(\varepsilon_2^2)$$

$$E(\varepsilon_0 \varepsilon_2) = \frac{1}{\bar{X}\bar{Y}} \operatorname{cov}(\bar{y}, \bar{x}')$$

$$= \frac{1}{\bar{X}\bar{Y}} cov\left[E(\bar{y}/n'), E(\bar{x}'/n')\right] + \frac{1}{\bar{X}\bar{Y}} E\left[cov(\bar{y}, \bar{x}')/n'\right]$$

$$= \frac{1}{\bar{X}\bar{Y}} cov\left[\bar{y}, \bar{x}\right] + \frac{1}{\bar{X}\bar{Y}} E\left[cov(\bar{y}', \bar{x}')\right]$$

$$= \frac{1}{\bar{X}\bar{Y}} cov\left[\bar{y}', \bar{x}'\right]$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right) \frac{S_{xy}}{\bar{X}\bar{Y}}$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right) \rho \frac{S_x}{\bar{X}} \cdot \frac{S_y}{\bar{Y}}$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right) \rho C_x C_y$$

where $\bar{y}'$ is the sample mean of y's based on the sample size n'.

$$E(\varepsilon_0 \varepsilon_1) = \frac{1}{\bar{X}\bar{Y}} cov(\bar{y}, \bar{x})$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right) \frac{S_{xy}}{\bar{X}\bar{Y}}$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right) \rho \frac{S_x}{\bar{X}} \frac{S_y}{\bar{Y}}$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right) \rho C_x C_y$$

$$E(\varepsilon_0^2) = \frac{1}{\bar{Y}^2} \text{var}(\bar{y})$$

$$= \frac{1}{\bar{y}^2}\left[V_1\{E_2(\bar{y}/n)\}^2 + E_1\{V_2(\bar{y}_n/n)\}^2\right]$$

$$= \frac{1}{\bar{Y}^2}\left[V_1(\bar{y}/n) + E_1\left\{\left(\frac{1}{n}-\frac{1}{n'}\right)s_y'^2\right\}^2\right]$$

$$= \frac{1}{\bar{Y}^2}\left[\left(\frac{1}{n'}-\frac{1}{N}\right)S_y^2 + \left(\frac{1}{n}-\frac{1}{n}\right)S_y^2\right.$$

$$= \left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_y^2}{\bar{Y}^2}$$

$$= \left(\frac{1}{n}-\frac{1}{N}\right)C_x^2$$

$$E(\varepsilon_1\varepsilon_2) = \frac{1}{\bar{X}^2}\text{cov}(\bar{x}, \bar{x}')$$

$$= \frac{1}{\bar{X}^2}\left[\text{cov}\{E(\bar{x}/n)E(\bar{x}'/n)\}^2 + 0\right]$$

$$= \frac{1}{\bar{X}^2}\text{var}(\bar{x}').$$

where $\text{var}(\bar{x}')$ is the variance of mean of sc based on an initial sample of size n'.

Double sampling in regression method of Estimation :-

When the population mean of the auxiliary variable $\bar{X}$, is not known, then double sampling is used as follows :-

i) A large sample of size n' is taken from the population of SRSWOR from which the population mean $\bar{X}$ is estimated as $\bar{x}'$, (i.e) $\bar{X} = \bar{x}'$

ii) Then a subsample of size n is chosen from the larger sample and both the variables x and y are measured from it by taking $\bar{x}'$ in place of $\bar{X}$ and treat is if it known.

Then $E(\bar{x}') = \bar{X}$, $E(\bar{x}) = \bar{X}$, $E(\bar{y}) = \bar{Y}$.

$$\bar{Y}_{regd} = \bar{y} + \beta(\bar{x}' - \bar{x})$$

where, $\beta = \frac{S_{xy}}{S_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

$\beta = \frac{S_{xy}}{S_x^2}$ based on the sample of size n.

Let,

$$\varepsilon_0 = \frac{\bar{y} - \bar{Y}}{\bar{Y}} \Rightarrow \bar{y} = (1+\varepsilon_0)\bar{Y}$$

$$\varepsilon_1 = \frac{\bar{x} - \bar{X}}{\bar{X}} \Rightarrow \bar{x} = (1+\varepsilon_1)\bar{X}$$

$$\varepsilon_2 = \frac{\bar{x}' - \bar{x}}{\bar{x}} \Rightarrow \bar{x}' = (1+\varepsilon_2)\bar{x}$$

$$\varepsilon_3 = \frac{s_{xy} - s_{xy}}{s_{xy}} \Rightarrow s_{xy} = (1+\varepsilon_3) s_{xy}$$

$$\varepsilon_4 = \frac{s_x^2 - s_x^2}{s_x^2} \Rightarrow s_x^2 = (1+\varepsilon_4) s_x^2$$

$E(\varepsilon_1) = 0,\ E(\varepsilon_2) = 0,\ E(\varepsilon_3) = 0,\ E(\varepsilon_4) = 0.$

Define,

$$\mu_{21} = E\left[(\bar{x} - \bar{X})^2 (y - \bar{Y})\right]$$

$$\mu_{30} = E(\bar{x} - \bar{X})^3.$$

Estimation error of $\bar{\bar{Y}}_{Rd}$.

$$\bar{\bar{Y}}_{Rd} = \frac{(1+\varepsilon_0)}{(1+\varepsilon_1)}(1+\varepsilon_2)\frac{\bar{Y}}{\bar{X}}\bar{x}$$

$$= \bar{Y}(1+\varepsilon_0)(1+\varepsilon_2)(1+\varepsilon_1)^{-1}$$

$$= \bar{Y}(1+\varepsilon_0)(1+\varepsilon_2)(1-\varepsilon_1+\varepsilon_1^2-\dots)$$

$$= \bar{Y}(1+\varepsilon_0+\varepsilon_2+\varepsilon_0\varepsilon_2-\varepsilon_1-\varepsilon_0\varepsilon_1-\varepsilon_1\varepsilon_2+\varepsilon_1^2)$$

Bias of $\bar{Y}_{Rd}$:

$$E(\bar{\bar{Y}}_{Rd}) = \bar{Y}\left[1+0+0+E(\varepsilon_0\varepsilon_2)-0-E(\varepsilon_0\varepsilon_1)-E(\varepsilon_1\varepsilon_2)+E(\varepsilon_1^2)\right]$$

$$\text{Bias}(\bar{\bar{Y}}_{Rd}) = E(\bar{\bar{Y}}_{Rd}) - \bar{Y}$$

$$= \bar{Y}\left[E(\varepsilon_0\varepsilon_2) - E(\varepsilon_0\varepsilon_1) - E(\varepsilon_1\varepsilon_2) + E(\varepsilon_1^2)\right]$$

$$= \bar{Y}\left[\left(\frac{1}{n'}-\frac{1}{N}\right)\rho C_x C_y - \left(\frac{1}{n}-\frac{1}{N}\right)\rho C_x C_y - \left(\frac{1}{n'}-\frac{1}{N}\right)C_x^2 + \left(\frac{1}{n}-\frac{1}{N}\right)C_x^2\right]$$

$$= \bar{Y}\left(\frac{1}{n} - \frac{1}{n'}\right)\left(C_x^2 - \rho C_x C_y\right)$$

$$= \bar{Y}\left(\frac{1}{n} - \frac{1}{n'}\right) C_x\left(C_x - \rho C_y\right).$$

MSE of $\hat{\bar{Y}}_{Rd}$:

$$MSE\left(\hat{\bar{Y}}_{Rd}\right) = E\left(\hat{\bar{Y}}_{Rd} - \bar{Y}\right)^2$$

$$= \bar{Y}^2 E\left(\varepsilon_0 + \varepsilon_2 - \varepsilon_1\right)^2$$

$$= \bar{Y}^2 E\left[\varepsilon_0^2 + \varepsilon_1^2 + \varepsilon_2^2 + 2\varepsilon_0\varepsilon_2 - 2\varepsilon_0\varepsilon_1 - 2\varepsilon_1\varepsilon_2\right]$$

$$= \bar{Y}^2 E\left[\varepsilon_0^2 + \varepsilon_1^2 + \varepsilon_2^2 + 2\varepsilon_0\varepsilon_2 - 2\varepsilon_0\varepsilon_1 - 2\varepsilon_2^2\right]$$

$$= \bar{Y}^2\left[\left(\frac{1}{n} - \frac{1}{N}\right) C_y^2 + \left(\frac{1}{n} - \frac{1}{N}\right) C_x^2 - \left(\frac{1}{n'} - \frac{1}{N}\right) C_x^2 + 2\left(\frac{1}{n'} - \frac{1}{N}\right)\right.$$

$$\rho C_x C_y \; 2\left(\frac{1}{n} - \frac{1}{N}\right)\rho C_x C_y$$

$$= \bar{Y}^2\left(\frac{1}{n} - \frac{1}{N}\right)\left(C_{oc}^2 + C_y^2 - 2\rho C_x C_y\right) + \bar{Y}^2\left(\frac{1}{n'} - \frac{1}{n}\right) C_x +$$

$$\left(2\rho C_y - C_x\right).$$

$$= MSE(\text{ratio estimator}) + \bar{Y}^2\left(\frac{1}{n'} - \frac{1}{n}\right)\left(2\rho C_x C_y - C_x^2\right.$$

$$2P.C_x C_y - C_x^2 > 0$$

Or

$$P > \frac{1}{2} \frac{C_x}{C_y}$$

Estimation Error :-

Then,

$$\hat{\bar{Y}}_{reg} = \bar{y} + \hat{\beta}(\bar{x}' - \bar{x})$$

$$= \bar{y} + \frac{S_{xy}(1+\varepsilon_3)}{S_x^2(1+\varepsilon_4)} (\varepsilon_2 - \varepsilon_1)\bar{x}$$

$$= \bar{y} + \bar{x} \frac{S_{xy}}{S_x^2} (1+\varepsilon_3)(\varepsilon_2 - \varepsilon_1)(1+\varepsilon_4)^{-1}$$

$$= \bar{y} + \bar{x} \beta (1+\varepsilon_3)(\varepsilon_2 - \varepsilon_1)(1 - \varepsilon_4 + \varepsilon_4^2 - \ldots)$$

$$\hat{\bar{Y}}_{reg} = \bar{y} + \bar{x}\beta(\varepsilon_2 + \varepsilon_2\varepsilon_3 - \varepsilon_2\varepsilon_4 - \varepsilon_1 - \varepsilon_1\varepsilon_3 + \varepsilon_1\varepsilon_4)$$

Bias :-

$$E(\hat{\bar{Y}}_{reg}) = \bar{Y} + \bar{X}\beta[E(\varepsilon_2\varepsilon_3) - E(\varepsilon_2\varepsilon_4) - E[\varepsilon_1\varepsilon_3] E(\varepsilon_1\varepsilon_4)$$

$$Bias(\hat{\bar{X}}_{reg}) = E(\hat{\bar{Y}}_{reg}) - \bar{X}$$

$$= \bar{Y}\beta\left[\left(\frac{1}{n'}-\frac{1}{N}\right)\frac{1}{N}\sum\frac{(\bar{x}'-\bar{x})(S_{xy}-S_{xy})}{\bar{x}S_{xy}}\right]$$

$$-\left(\frac{1}{n'}-\frac{1}{N}\right)\frac{1}{N}\left(\sum\frac{(\bar{x}'-\bar{x})\,(S_x^2-S_x^2)}{\bar{x}S_x^2}\right)$$

$$-\left(\frac{1}{n'}-\frac{1}{N}\right)\frac{1}{N}\sum\left[\frac{(\bar{x}-\bar{x})\,(S_{xy}-S_{xy})}{\bar{x}S_{xy}}\right]$$

$$+\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{N}\left[\sum\frac{(\bar{x}-\bar{x})(S_x^2-S_x^2)}{\bar{x}S_x^2}\right]$$

$$= \bar{Y}\beta\left[\left(\frac{1}{n'}-\frac{1}{N}\right)\frac{\mu_{21}}{\bar{x}S_{xy}}-\frac{1}{n'}-\frac{1}{N}\right)\frac{\mu_{30}}{\bar{x}S_x^2}$$

$$-\left(\frac{1}{n}-\frac{1}{N}\right)\frac{\mu_{21}}{\bar{x}S_{xy}}+\left(\frac{1}{n}-\frac{1}{N}\right)\frac{\mu_{30}}{\bar{x}S_x^2}\Bigg]$$

$$= -\beta\left(\frac{1}{n}-\frac{1}{n'}\right)\left(\frac{\mu_{21}}{S_{xy}}-\frac{\mu_{30}}{S_x^2}\right)$$

Double Sampling for Stratification :-

The method of post-stratificat is useful only if the relative proportion of each stratum in the population $w_h = \frac{N_h}{N}$ is known for each stratum h. If these Proportion are not known, double sampling may be used, with an initial (large) sample used to classify the units into strata and then a stratified sample selected from the initial sample.

Step 1 :- n' initial simple random sample are selected from a population of N units. these units are classified into strata. with n'_h observed to be in stratum h. The population proportion $W_h = \frac{N_h}{N}$ is estimated by the sample proportion :

$$w_h = \frac{n'_h}{n'} \quad h = 1 \ldots L.$$

Step 2 : A second sample is then selected by stratified random sampling from the first sample. These units are classified into strata with n_h units selected from the n'_h sample units in stratum h. Measurement of Y_{hi} is recorded for each unit in the second sample.

We denote the sample mean is Stratum h in the second sample as :

$$\bar{y}_h = \sum_{i=1}^{n_h} Y_{hi}/n_h$$