** VARYING PROBABILITY SAMPALING **

Auxiliary Information:-

(i) Planning stage – e.g. Stratified Sampling
(ii) Specified Stage
(iii) Estimation Stage – e.g. Ratio, Regression, Product method of estimation.

Unlike SRS, here the units gets different probability of selection with the help of the use auxiliary variable i.e. We associate to vector Y' = $(y_1,\ldots\ldots y_n)$, a vector P'= $(p_1,p_2\ldots\ldots..p_n)$ such that $\sum_{i=1}^{N} p_i = 1$, $p_i > 0$.The rational behind the use of varying probability of selection is that large units contribute more towards the population total and hence they should have large probability of selection in the sample. If X_i denotes the size of i^{th} unit. Then $p_i \propto x_i$

$$\therefore \sum_{i=1}^{N} p_i = k\sum_{i=1}^{N} x_i \Rightarrow k = \frac{1}{X} \quad where \quad X = \sum_{i=1}^{N} x_i$$

$$\therefore p_i = \frac{x_i}{X} \qquad \text{where i = 1,2,……..N.}$$

This sampling is also known as PPS (Probability Proportional to Size)

There are predominately two methods by which we select the PPS sample.

(i) Cumulative total method.
(ii) Laheri`s method.

(1) Cumulative Total Method:-

Here we make the cumulative total of the units of the population given as bellow,

Unit No.	Size	Cumulative total
1	x_1	$T_1 = x_1$
2	x_2	$T_2= x_1+x_2$
3	x_3	$T_3= x_1+x_2+x_3$
.	.	……..
.	.	……
I	x_i	$T_i = x_1+\ldots..+x_i$
.	.	…
N	x_N	$T_N = x_1+\ldots..+x_N$

Now select a random number between 1 to X. Where X = total of X_i's in population. If r^{th} random number is selected and $T_{i-1} < r \leq T_i$. Then i^{th} unit is selected. This ensure that $p_i = x_i / X$. If we allow repetition of units in the sample. It is known as PPSWR sampling procedure and if we do not allow repetition then resultant will be PPSWOR sampling procedure.

One of the major defect of this sampling procedure is that for large N. Preparation of cumulative total table requires large cost and time, so Laheri suggested the following procedure.

(2) Laheri`s method of selection of a PPS sample:-

Laheri`s Scheme consists in 1st deciding a number in which it is larger than the sizes of units. Then select a pair of random number say (i, j), if $x_i \le x_j$ retain i^{th} unit in the sample otherwise reject the pair. Repeat it till we get the sample size n. Here also if retain all units even with repetition, it results an PPSWR. If we retain only distinct units in sample, then it is PPSWOR.

** Estimation Of population mean or total.

Let a sample of $(y_1,\ldots,y_n)$ is selected with PPSWR in which probability of selecting i^{th} unit of y is p_i (i = 1,2,...,N) and we are interested in population mean.

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

let $Z_i = \frac{Y_i}{NP_i}$ i = 1,2,….,N.

then $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{NP_i}$

is an unbiased estimator of population mean, that is,

$$E(\bar{z}) = \frac{1}{n}\sum_{i=1}^{n} E(z_i) = \frac{1}{n}\sum_{i=1}^{n}\left[\Sigma Z_i\, p_i\right] = \frac{1}{n}\sum_{i=1}^{n}\Sigma \frac{Y_i}{NP_i} p_i$$

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

$$E(z_i) = \sum_{i=1}^{N} P_i Z_i = \sum_{i=1}^{N} P_i \frac{Y_i}{NP_i} = \sum_{i=1}^{N} Y_i = \bar{Y}$$

$$V(z_i) = E(z_i - \bar{Y})^2 = \sum_{i=1}^{N} P_i (Z_i - \bar{Y})^2 = \sum_{i=1}^{N} P_i \left(\frac{Y_i}{NP_i} - \bar{Y}\right)^2 = \sigma_z^2 \text{ , say}$$

$$V(\bar{z}) = E(\bar{z}^2) - \left[E(\bar{z})\right]^2 \qquad (1)$$

$$E(\bar{z}^2) = E\left[\frac{1}{n}\sum_{i=1}^{n} z_i\right]^2$$

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

$$= \frac{1}{n^2} E\left[\sum_i z_i^2 + \sum_{i\ne j}^{n} z_i z_j\right]$$

$$= \frac{1}{n^2}\left[\sum_i E(z_i^2) + \sum_{i\ne j}^{n} E(z_i)E(z_j)\right] \qquad \text{(ii)}$$

as z_i and z_j are independent.

$$E(z_i^2) = \sum_i z_i^{\,2} p_i$$
$$E(z_i) = \bar{z} \qquad \text{(iii)}$$

On putting this value in (ii), we have

$$E(\bar{Z}^2) = \frac{1}{n^2}\left[\sum_i \sum_i z_i^{\,2} p_i + \sum_{i \neq j}^{n} z^{-2}\right]$$
$$= \frac{1}{n^2}\left[n \sum_i z_i^{\,2} p_i + n(n-1) z^{-2}\right]$$
$$= \frac{1}{n}\left[\sum_i z_i^{\,2} p_i + (n-1) z^{-2}\right]$$
$$= \frac{1}{n}\left[\sum_i z_i^{\,2} p_i + nz^{-2} - z^{-2}\right]$$

On putting this value in (i), we have

$$\text{Var}(\bar{Z}) = \frac{1}{n}\left[\sum_i z_i^{\,2} p_i + nz^{-2} - z^{-2}\right] - z^{-2}$$
$$= \frac{1}{n}\left[\sum_i z_i^{\,2} p_i + nz^{-2} - z^{-2} - nz^{-2}\right]$$
$$= \frac{1}{n}\left[\sum_i z_i^{\,2} p_i - z^{-2}\right]$$
$$= \frac{1}{n}\sum_i p_i (z_i - \bar{z})^2$$
$$= \frac{\sigma^2 z}{n} \qquad \text{where } \sigma^2 z = \sum_i p_i (z_i - \bar{z})^2$$

** Unbiased Estimator V($\bar{Z}$) :-

Let us consider $s_z^{\,2} = \frac{1}{n-1}\sum_1^n \left[(z_i - \bar{z})^2\right]$

Then, $E(s_z^{\,2}) = \frac{1}{n-1} E\left[\sum_1^n (z_i - \bar{z})^2\right]$

$$(n-1)\, E(s^2{}_z) = E\left[\sum_1^n (z_i^{\,2} - nz^{-2}\right]$$
$$= \sum_1^n E(z_i^{\,2}) - nE(z^{-2})$$
$$= n\sum_1 z_i^{\,2} p_i - n\left[Var(\bar{z}) + \left[E(\bar{z})\right]^2\right]$$
$$= n\sum_1 z_i^{\,2} p_i - n\left[\frac{\sigma_z^{\,2}}{n} \bar{z}^2\right]$$

$$= n\sum_1 z_i^{\,2} p_i - \sigma_z^{\,2} - n\,\bar{Z}^2$$

$$= n\left[\sum_1 z_i^{\,2} p_i - \bar{z}^{2}\right] - \sigma_z^{\,2}$$

$$= n\,\sigma_z^{\,2} - \sigma_z^{\,2}$$

$$= (n-1)\,\sigma_z^{\,2}$$

$$E(s_z^{\,2}) = \sigma_z^{\,2}$$

$$\therefore \quad V(\bar{Z}) = \sigma_z^{\,2}/n = E(s_z^{\,2})/n$$

Unbiased estimator of $V(\bar{Z}) = s_z^{\,2}/n$

** PPSWOR :-

This was first considered Horvits & Thompson (1952). The use of PPWOR for estimating population mean. The suggested that

$$\bar{Y}_{HT} = \sum_{i=1}^{N} \frac{y_i}{N\pi_i}$$

where π_i = inclusion probability of i[th] unit being included in the sample. This is different from initial probability of selection which is given by p_i. Then,

$$= \sum_{s>i} p(s) \qquad \ldots\ldots\ldots\ldots(1)$$

If us define an indicator fn. δ_i.

$$\delta_i = \begin{cases} 1, \; if \;\; i \in s \\ 0, \; o.w. \end{cases}$$

Then $\bar{Y}_{HT} = \sum_{i=1}^{N} \frac{Y|\delta|}{N\pi_i}$

$$E(\bar{Y}_{HT}) = \sum_{i=1}^{N} \frac{E(\delta_i) y_i}{N\pi_i}$$

but $E(\delta_i) = \sum_{s>i} 1\,p(s) + \sum_{s>i} 0 \;\; p(s) = \pi_i$ (from (i))

$$\therefore E(\bar{Y}_{HT}) = \sum_{i=1}^{N} \frac{\pi_i y_i}{N\pi_i} = \sum_{i=1}^{N} \frac{y_i}{N} = Y_N$$

Further $v(\delta_i) = E(\delta_i^{\,2}) - [E(\delta_i)]^2 \qquad (\because \delta_i^{\,2} = \delta_i)$

$$= \pi_i - \pi_i^{\,2}$$

$$= \pi_i(1-\pi_i)$$

and $\operatorname{cov}(\delta_i, \delta_j) = E(\delta_i \delta_j) - E(\delta_i)\;E(\delta_j)$

$$= \pi_{ij} - \pi_i \pi_j$$

Where π_{ij} is the probability of including units i and j in the sample.

$$\pi_{ij} = \sum_{s>(i,j)} p(s)$$

$$v(\overline{Y_{HT}}) = v\left[\sum_{i=1}^{N} \frac{Y_i s_i}{N\pi_i}\right]$$

$$= \frac{1}{N^2}\left[\sum_{i=1}^{N} \frac{Y_i^2 v(s_i)}{\pi_i^2} + \sum_{i\neq j}^{N} \frac{Y_i Y_j \ \mathrm{cov}(\delta_i, \delta_j)}{\pi_i \pi_j}\right]$$

$$= \frac{1}{N^2}\left[\sum_{i=1}^{N} \frac{Y_i^2 \pi_i (1-\pi_i)}{\pi_i^2} + \sum_{i\neq j}^{N} \frac{Y_i Y_j \ (\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j}\right]$$

$$= \frac{1}{N^2}\left[\sum_{i=1}^{N} \frac{Y_i^2 (1-\pi_i)}{\pi_i} + \sum_{i\neq j}^{N} \frac{Y_i Y_j \ (\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j}\right]$$

** Estimation of $v(\overline{Y_{HT}})$:-

$$v(\overline{Y_{HT}}) = \frac{1}{N^2}\left[\sum_{i=1}^{N} \frac{\pi_i \ Y_i^2 (1-\pi_i)}{\pi_i^2} + \sum_{i\neq j}^{N} \frac{Y_i Y_j \ (\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j}\right]$$

$$= \frac{1}{N^2}\left[\sum_{i=1}^{N} \frac{\pi_i (1-\pi_i) Y_i^2}{\pi_i^2} + \sum_{i\neq j}^{N} \frac{Y_i Y_j Y_i \ (\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j}\right]$$

* Evaluation of inclusion probability n =2.

$$\pi_i = \sum_{r=1}^{n} p_{ir}$$

$$= \sum_{r=1}^{2} p_{ir}$$

$$= p_{i1} + p_{i2}$$

$$= p_i + \sum_{j(\neq i)=1}^{N} \frac{p_j p_i}{1-p_j}$$

$$= p_i\left(1 + s - \frac{p_i}{1-p_i}\right) \quad \textit{where} \quad s = \sum_{i=1}^{N} \frac{p_i}{1-p_i}$$

$$\pi_{ij} = p_{i1\,j2} + p_{j1i2}$$

$$= \frac{p_i \ p_j}{1-p_i} + \frac{p_j p_i}{1-p_j}$$

$$= p_i \ p_j\left[\frac{1}{1-p_i} + \frac{1}{1-p_j}\right]$$

** Yates and Grondy proved the following relationship.

An elegent expression for the variance of the H-T estimator is given by Yates and Grondy. Since,δ_I= 1 exactly n units in the probability and zero for the rest , we have,

$$\sum_{i=1}^{N} \delta_i = 1 \qquad \text{……..(1)}$$

Take expectation of both sides, we have

$$\sum_{i=1}^{N} E(\delta_i) = E(n)$$

$$\sum_{i=1}^{N} \pi_i = n$$

Squaring (1) and taking the expectation, we have

$$\sum_{i \neq j}^{N} \pi_{ij} = n(n-1)$$

By definition; $\pi_{ij} = \sum_{i=1}^{n} p_{ir}$

$$\pi_{ij} = \sum_{i=1} p_{ir}$$

$$\sum_{i} \pi_i = \sum_{i}^{n} \sum_{r}^{n} p_{ir} = \sum_{r}^{n} \sum_{i}^{n} p_{ir} = \sum_{r}^{n} 1 = n$$

We have, $\sum_{j(\neq i)=1}^{N} \pi_{ij} = \sum_{j(\neq i)}^{N} \sum_{s(\neq i)=1}^{N} (p_{ij}\, js)$

$$= \sum_{j(\neq i)}^{N} \sum_{r=1}^{N} \sum_{s(\neq r)=1}^{N} p_{ir}\, p_{js/ir} \qquad [\because p(AB) = p(A)\ (B/A)]$$

$$= \sum_{r=1}^{N} p_{ir} \sum_{s(\neq r)=1}^{N} \sum_{j(\neq i)}^{N} p_{js/ir}$$

$$= \sum_{r=1}^{N} p_{ir} \sum_{s(\neq r)=1}^{N} 1$$

$$= \pi_i (n-1)$$

$$\sum_{i=1}^{N} \sum_{j(\neq i)=1}^{N} \pi_{ij} = n(n-1)$$

$$= \sum_{i=1}^{N} \pi_i\, (n-1)$$

$$= \text{n (n = 1)}$$

$$v(\overline{Y_{HT}}) = \sum_{i=1}^{N} \frac{Y_i^2 (1-\pi_i)\pi_i}{N^2 \pi_i^{\,2}} + \sum_{j \neq i}^{N} \frac{Y_i Y_j (\pi_{ij} - \pi_i \pi_j)\pi_i}{N^2 \pi_i \pi_j}$$

$$N^2\, v(\overline{Y_{HT}}) = \sum_{i=1}^{N} \frac{Y_i^2 (1-\pi_i)}{\pi_i} + \sum_{j \neq i}^{N} \frac{Y_i Y_j (\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} \qquad \text{……..(1)}$$

but $$\sum_{j\neq i}^{N}(\pi_{ij}-\pi_i\pi_j)=\sum_{j\neq i}^{N}\pi_{ij}-\sum_{j\neq i}^{N}\pi_i\,\pi_j$$

$$=\pi_i(n-1)-\pi_i(n-\pi_i)$$

$$=\pi_i(n-1-n-1\,\pi_i)$$

$$=\pi_i(1-\pi_i) \qquad \text{……..(2)}$$

This value put in (1)

$$N^2\,v(\overline{Y_{HT}})=\sum_{i=1}^{N}\frac{Y_i^2}{\pi_i^{\,2}}\sum_{j\neq i}^{N}(\pi_i\pi_j-\pi_{ij})-\sum_{j\neq i}^{N}\frac{Y_iY_j(\pi_i\pi_j-\pi_{ij})}{\pi_i\pi_j}$$

$$=\sum_{j\neq i}^{N}(\pi_i\pi_j-\pi_{ij})\left(\sum_{i=1}^{N}\frac{Y_i^2}{\pi_i^{\,2}}-\frac{Y_iY_j}{\pi_i\pi_j}\right)$$

$$=\sum_{j\neq i}^{N}(\pi_i\pi_j-\pi_{ij})\left(\frac{Y_i}{\pi_i}-\frac{Y_j}{\pi_j}\right)^2$$

$$V_{YG}(\overline{Y_{HT}})=\frac{1}{N^2}\sum_{j\neq i}^{N}\left(\frac{Y_i}{\pi_i}\cdot\frac{Y_j}{\pi_j}\right)^2(\pi_i\pi_j-\pi_{ij})$$

$$V_{YG}(\overline{Y_{HT}})=\frac{1}{N^2}\sum_{j\neq i}^{N}\delta_i\delta_j\,\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}$$

$$E\left(V_{yG}(\overline{Y_{HT}})\right)=\frac{1}{N^2}\sum_{j\neq i}^{N}E\left(\delta_i\delta_j\right)\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}$$

$$=\frac{1}{N^2}\sum_{j\neq i}^{N}\pi_{ij}\,\frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}}$$

$$=\frac{1}{N^2}\sum_{j\neq i}^{N}(\pi_i\pi_j-\pi_{ij})$$

** The Major Draw Back of H.T Estimator are:-

(1) Estimate of variance can assume negative value and have no inference can be

(2) Calculation of π_i, π_{ij}, are difficult. The author have suggested that the population many be deeply stratified that a sample of 2 or 3 is sufficient to represent the strata. Yotes and Grandy claimed that the estimator proposed by them is always positive, Then, illustrated it by an exp. Of 5 popes for n = 2, Then claim is however not true for n > 2.

So it is also taken negative value but less frequently then that of H.T.

** To show that $\hat{V}_{yG}(\overline{Y_{HT}})$ is always positive for n = 2.

We know that; $\pi_i=p_i\left(1+s-\frac{p_i}{1-p_i}\right)$ *and* $\pi_{ij}=p_ip_j\left(\frac{1}{1-p_i}+\frac{1}{1-p_j}\right)$

Now $\pi_i=p_i\left(1+\frac{p_i}{1-p_i}+A\right)$ where $A=\sum_{k(\neq ij)}^{N}\frac{p_k}{1-p_k}$ $s=A+\frac{p_i}{1-p_i}+\frac{p_j}{1-p_j}$

$$\pi_j = p_j\left(1+\frac{p_i}{1-p_i}+A\right)$$

$$\pi_i\,\pi_j = p_i p_j\left(1+\frac{p_j}{1-p_j}+A\right)\left(1+\frac{p_i}{1-p_i}+A\right)$$

$$= p_i p_j\left(1+\frac{p_i}{1-p_i}+A+\frac{p_j}{1-p_j}+\frac{p_i p_j}{(1-p_i)(1-p_j)}+A\frac{p_j}{1-p_j}+A+\frac{p_i A}{1-p_i}+A^2\right)$$

$$= p_i p_j\left(1+\frac{p_i}{1-p_i}+\frac{p_j}{1-p_j}+2A+\frac{p_i p_j}{(1-p_i)(1-p_j)}+\left(\frac{p_i}{1-p_i}+\frac{p_j}{1-p_j}\right)A+A^2\right)$$

$$\pi_i\pi_j - \pi_{ij} = p_i p_j\left(1+\frac{p_i}{1-p_i}+\frac{p_j}{1-p_j}+2A+\frac{p_i p_j}{(1-p_i)(1-p_j)}+\left(\frac{p_i}{1-p_i}+\frac{p_j}{1-p_j}\right)A+A^2\right)$$

$$- p_i p_j\left(\frac{1}{1-p_i}+\frac{1}{1-p_j}\right)$$

$$= p_i p_j\left(1+\frac{p_i}{1-p_i}-\frac{1}{1-p_i}+\frac{p_j}{1-p_j}-\frac{1}{1-p_j}+2A+\frac{p_i p_j}{(1-p_i)(1-p_j)}+\left(\frac{p_i}{1-p_i}+\frac{p_j}{1-p_j}\right)A+A^2\right)$$

$$= p_i p_j\left(1-\frac{1-p_i}{1-p_i}-\frac{1-p_j}{1-p_j}+2A+\frac{p_i p_j}{(1-p_i)(1-p_j)}+\left(\frac{p_i}{1-p_i}+\frac{p_j}{1-p_j}\right)A+A^2\right)$$

$$= p_i p_j\left(1-1-1+2A+\frac{p_i p_j}{(1-p_i)(1-p_j)}+\left(\frac{p_i}{1-p_i}+\frac{p_j}{1-p_j}\right)A+A^2\right)$$

but, min $A = \dfrac{N-2}{N-1}$ $\qquad\left(\because \dfrac{(N-2)1/N}{1-1/N}\right)$

Consider, $\quad = \dfrac{2(N-2)}{N-1}-1$

$$= \frac{2N-4-N+1)}{N-1} = \frac{N-3}{N-2} \geq 0 \qquad \text{if N < 3.}$$

** Midzuno Scheme of Sampling:-

Select the first unit of the sample with PPS and the remaining (n-1) units with SRSWOR.

$\pi_i(n) =$ Probability of selecting i^{th} unit at the 1^{st} draw + probability of selecting at any of (n-1) draw

$$= p_i + (1-p_i)\frac{n-1}{N-1}$$

$$= \frac{N-n}{N-1} p_i + \frac{\mu - 1}{N-1}$$

$\pi_{ij}(n) =$ Probability of selecting i[th] first draw and j[th] at any of the (n-1) draws + probability j[th] is selecting the first draw p i[th] included at any of (n-1) draws + neither i[th] nor j[th] unit is selected at the first draw but are included at remaining (n-1) draw

$$= p_i \frac{n-1}{N-1} + p_j \frac{n-1}{N-1} + (1 - p_i - p_j) \frac{n-1}{N-1} \frac{n-2}{N-2}$$

$$- \left(\frac{n-1}{N-1}\right)(p_i + p_j) + \left(\frac{n-1}{N-1}\right)\left(\frac{n-2}{N-2}\right) - (p_i + p_j)\left(\frac{n-1}{N-1}\right)\left(\frac{n-2}{N-2}\right)$$

$$= (p_i + p_j)\left(\frac{n-1}{N-1}\right)\left(\frac{N-n}{N-2}\right) + \left(\frac{n-1}{N-1}\right)\left(\frac{n-2}{N-2}\right)$$

$$\pi_{ijk}(n) = \frac{N-n}{N-3} \frac{n-2}{N-2} \frac{n-1}{N-1}(p_i + p_j + p_k) + \left(\frac{n-1}{N-1}\right)\left(\frac{n-2}{N-2}\right)\left(\frac{n-3}{N-3}\right)$$

$$\pi_{i1,i2,\ldots\ldots ir}(n) = \frac{N-n}{N-r} \frac{n-r+1}{N-r+1} \ldots\ldots \frac{n-1}{N-1}(p_{i1} + p_{i2} + \ldots + p_{ir}) + \frac{(n-1)\ldots\ldots\ldots\ldots(n-r)}{(N-1)\ldots\ldots\ldots\ldots(N-r)}$$

Thus in this sampling scheme the probability of selection of units in the sample can always be made propertinoned to the sum of their sizes.

The another property of this sampling scheme is that Yates-Grandy is estimator or Horwity-Thompson.

Estimator's variance is always positive. They can be proved so follows,

We know that Yot's estimators of $V(\overline{Y_{HT}})$ will always be positive if $\pi_i \pi_j - \pi_{ij} > 0, \; k_{ij}$

$$\pi_i \pi_j - \pi_{ij} = \left(\frac{N-n}{N-1} p_i + \frac{n-1}{N-1}\right)\left(\frac{N-n}{N-1} p_j + \frac{n-1}{N-1}\right) - \frac{N-n}{N-2}, \frac{n-1}{N-1}(p_i + p_j)$$

$$- \frac{n-1}{N-1}, \frac{n-2}{N-2}$$

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

$$- \left(\frac{N-n}{N-2}\right)\left(\frac{n-1}{N-1}\right)(p_i + p_j) - \left(\frac{n-1}{N-1}\right)\left(\frac{n-2}{N-2}\right)$$

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

$$+ \left(\frac{n-1}{N-1}\right)\left(\frac{n-1}{N-1} - \frac{n-2}{N-2}\right)$$

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

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

$$= \left(\frac{N-n}{N-1}\right)^2 p_i p_j - \left(\frac{n-1}{N-1}\right)\left[(p_i + p_j)\ \frac{N-n}{(N-1)(N-2)} - \frac{N-n}{(N-1)(N-2)}\right]$$

$$= \frac{N-n}{(N-1)^2}\left[(N-n)\, p_i p_j - \frac{n-1}{N-2}(p_i + p_j) - \frac{n-1}{N-2}\right] > 0$$

$$\therefore \quad (N-n)\, p_i p_j - \frac{n-1}{N-2}(p_i + p_j) - \left(\frac{n-1}{N-2}\right) > 0$$

$$\therefore \quad \pi_i \pi_j > \pi_{ij}$$

** RATIO ESTIMATION **

In sample random sampling we considered estimators using observed values of characteristic under study. Many a time the characteristic Y under study is closely related to a auxiliary characteristic X and data on X are either readily available or can be easily collected for all the unit in the population. In such situations, it is customary to consider estimators of $\overline{Y}$ that are the data on X and are more efficient then the estimators which use data on the characteristic Y alone fact that, the data on the auxiliary variable can be used data at a later stage after selecting the sample, encourages such processors two type of these commonly used methods are

(1) The ratio type methods.

(2) The regression methods.

** NOTATION **

Y_i = the value of Y, the characteristic under study for the i^{th} unit in the population. i = 1, 2, ….., N

X_i = the value of X, the auxiliary characteristic for the i^{th} unit in the population.

$\hat{R} = \overline{Y} / \overline{X}$: The ratio of population means.

$\hat{R} = \overline{y} / \overline{x}$: The ratio of sample means.

The ratio estimators of the population mean $\overline{Y}$ is that defined as $\overline{y_r} = \hat{R}\overline{X}$

Example:-Let Y = number of bullocks on a holding.

X = total area in acres

The ratio $\hat{R}$ is an estimator of the number of bullocks for a car for holding in the population, the product of $\hat{R}$ with $\overline{X}$. The coverage size of a holding in acres would prosed in estimator of $\overline{Y} - 1$ the coverage number of bullocks per holding in the population.

y = population of Rajkot in 1992.

x = population of Rajkot in 1987.

X = total population of Rajkot in 1987.

Let us have a SRSWOR of n households for 1987 and hence get the sample total y. Then the ratio estimator of the total population of Rajkot in 1992 us given by,

$$\hat{y}_R = \frac{y}{x} X$$

** BIAS OF RATIO ESTIMATORS:-

The ratio estimator of the population mean $\overline{Y}$ is given by,

$$\hat{y}_R = \frac{\overline{y}}{\overline{x}} \overline{X} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\text{(i)}$$

Let us define $e_0 = \frac{\overline{y} - \overline{Y}}{\overline{Y}}$ ……(ii)

$$e_1 = \frac{\overline{x} - \overline{X}}{\overline{X}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\text{(iii)}$$

$$E(e_0) = E\left[\frac{\overline{y} - \overline{Y}}{\overline{Y}}\right] = \frac{E(\overline{y}) - \overline{Y}}{\overline{Y}} = \frac{\overline{Y} - \overline{Y}}{\overline{Y}} = 0$$

$$E(e_1) = E\left[\frac{\overline{x} - \overline{X}}{\overline{X}}\right] = \frac{E(\overline{x}) - \overline{X}}{\overline{X}} = \frac{\overline{X} - \overline{X}}{\overline{X}} = 0$$

$$E(e_0^{\,2}) = E\left[\frac{\overline{y} - \overline{Y}}{\overline{Y}}\right]^2 = \frac{\text{var}(\overline{y})}{\overline{Y}^2} = \frac{1}{\overline{Y}^2} \; \frac{N-n}{Nn} \; S_y^{\,2}$$

$$E(e_1^{\,2}) = E\left[\frac{\overline{x} - \overline{X}}{\overline{X}}\right]^2 = \frac{\text{var}(\overline{x})}{\overline{X}^2} = \frac{1}{\overline{X}^2} \; \frac{N-n}{Nn} \; S_x^{\,2}$$

$$E(e_0e_1) = E\left[\frac{\bar{y}-\bar{Y}}{\bar{Y}}\ \frac{\bar{x}-\bar{X}}{\bar{X}}\right] = \frac{\mathrm{cov}\,(\bar{y},\bar{x})-\bar{X}}{\bar{Y}\bar{X}} = \frac{1}{\bar{Y}\bar{X}}\ \frac{N-n}{Nn}\ S_{yx}^2$$

where $S_y^2 = \frac{1}{N-1}\sum (y_i - \bar{Y})^2$

$$S_x^2 = \frac{1}{N-1}\sum (x_i - \bar{X})^2$$

$$S_{yx}^2 = \frac{1}{N-1}\sum (x_i - \bar{X})(y_i - \bar{Y})$$

From (ii) $\bar{y} = \bar{Y}(1+e_0)$

From (iii) $\bar{x} = \bar{X}(1+e_1)$ put this value in equation (i)

$$\bar{y}_R = \frac{\bar{Y}(1+e_0)}{\bar{X}(1+e_1)}\bar{X} = \bar{y}\left(\frac{1+e_0}{1+e_1}\right) = \bar{y}\ (1+e_0)(1+e_1)^{-1}$$

$$= \bar{y}\ (1+e_0)(1-e_1+e_1^{\,2}) \quad \left[\begin{array}{l}\text{In the exparition s terms having}\\ \text{than power 3 are negleted}\end{array}\right.$$

$$= \bar{y}\ (1+e_0-e_1-e_0e_1+e_1^{\,2})$$

Subtract $\bar{Y}$ from both the side in above expression we get,

$$\bar{Y}_R - \bar{Y} = \bar{Y}\ (e_0-e_1-e_0e_1+e_1^{\,2})$$

Bias of the ratio estimator to the first degree of approximation is given by,

$$\text{Bias}\ (\bar{y}_R) = E\left[(\bar{y}_R)-\bar{y}\right] = E(\bar{Y}_R)-\bar{Y}$$

$$= \bar{Y}\,E\,(e_0-e_1-e_0e_1+e_0^{\,2})$$

$$= \bar{Y}\,E\,(e_1^{\,2}-e_0e_1) \qquad [\because E(e_0)=E(e_1)=0]$$

$$= \bar{Y}\left[E\,(e_1^{\,2})-E(e_0e_1)\right]$$

$$= \bar{Y}\left[\frac{1}{\bar{X}^2}\frac{N-n}{Nn}\ S_x^2 - \frac{1}{\bar{X}\bar{Y}}\frac{N-n}{Nn}\ S_{yx}^2\right]$$

$$= \bar{Y}\left[\frac{N-n}{Nn}\left(\frac{S_x^2}{\bar{X}^2}\ \frac{S_{yx}^2}{\bar{Y}\bar{X}}\right)\right]$$

$$= \bar{Y}\ \frac{1-f}{n}(CX^2 - \tau C_yC_x) \qquad \left[\because \frac{n}{N} = f,\ S_{yx} = \tau S_yS_x,\quad \tau = S_{yx}/S_yS_x\right.$$

where, τ = correlation coefficient between y and x.

$C_x = S_x/\bar{X}$ = coefficient of variation of x.

$C_y = S_y/\bar{Y}$ = coefficient of variation of y.

f = n/N sample fraction.

REMARK:-

Ratio estimator is unbiased if

bias $(\overline{Y_R}) = 0$. i.e. if $\overline{Y}\,\frac{1-f}{n}\left[\frac{S_x^2}{\overline{X}^2} - \tau\frac{S_y S_x}{\overline{Y}\,\overline{X}}\right] = 0$

i.e. if $\frac{S_x^2}{\overline{X}^2} - \frac{S_y S_x}{\overline{Y}\,\overline{X}} = 0$

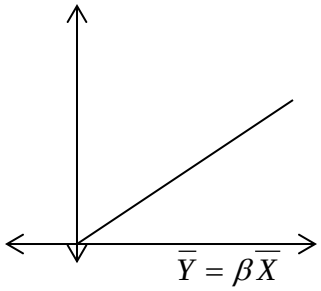
i.e. if $\frac{S_x^2}{\overline{X}^2} = \frac{S_{yx}}{\overline{Y}\,\overline{X}}$

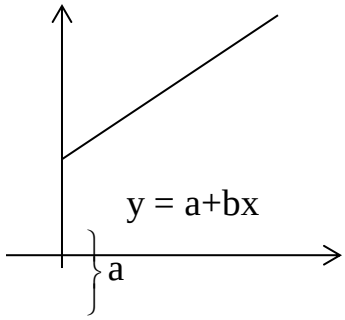
i.e. if $\overline{Y} = \frac{S_{yx}}{S_x^2}\,\overline{X}$

i.e. if $\overline{Y} = \beta\,\overline{X}, \quad \beta = \frac{S_{yx}}{S_x^2} =$ regression coefficient of y on x

i.e. if the line of regression of y on x is a straight line passing through origin.

i.e. if





** MEAN SQUARE ERROR OF RATIO ESTIMATOR :-

The mean square error of ratio estimator of the populoation mean is given by,

$$MSE(\overline{Y_R}) = E\left[\overline{Y_R} - \overline{Y}\right]^2$$

$$= E\left[\overline{Y}(e_0 - e_1)\right]^2$$

$$= \overline{Y}^2 E(e_0^{\,2} - 2e_0 e_1 + e_1^{\,2})$$

$$= \overline{Y}^2\left[E(e_0^{\,2}) - 2E(e_0 e_1) + E(e_1^{\,2})\right]$$

$$= \overline{Y}^2\,\frac{N-n}{Nn}\left[\frac{S_y^2}{\overline{Y}^2} - \frac{2S_{yx}}{\overline{Y}\,\overline{X}} + \frac{S_x^2}{\overline{X}^2}\right]$$

$$= \frac{N-n}{Nn}\left[S_y^2 - 2S_{yx}\frac{\overline{Y}}{\overline{X}} + S_x^2\frac{\overline{Y}^2}{\overline{X}^2}\right]$$

$$= \frac{N-n}{Nn}\left[S_y^2 - 2RS_{yx} + R^2 S_x^2\right] \qquad \left[\because R = \frac{\bar{Y}}{\bar{X}}\right]$$

OR $$MSE(\bar{y}_R) = \bar{Y}^2 \frac{1-f}{n}\left[C_y^2 - 2\rho C_y C_x + C_x^2\right]$$

** COMPARISON OF RATIO ESTIMATOR WITH SRSWOR:-

In SRSWOR, we have

$$V(\bar{y}_n)_{SRS} = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 = \frac{1-f}{n} S_y^2 \qquad (\because f = n/N)$$

For ratio estimator we have,

$$MSE(\bar{y}_R) = \frac{1-f}{n}\left[S_y^2 - 2RS_{yx} + R^2 S_x^2\right]$$

Ratio estimator is better then SRSWOR if

$$V(\bar{y}_n)_{SRS} - MSE(\bar{y}_R) > 0$$

i.e. if $$\frac{1-f}{n}\left[S_y^2\right] - \frac{1-f}{n}\left[S_y^2 - 2RS_{yx} + R^2 S_x^2\right] > 0$$

$$\frac{1-f}{n}\left[2RS_{yx} + R^2 S_x^2\right] > 0$$

i.e. if $$2RS_{yx} > R^2 S_x^2$$

$$2R\tau S_y S_x > R^2 S_x^2$$

i.e. if $$\rho > \frac{1}{2}\frac{S_x}{S_y} \; R > \frac{1}{2}\frac{S_x}{S_y}\frac{\bar{Y}}{\bar{X}} \qquad (\because -1 < \rho < 1)$$

i.e. if $$\rho > \frac{1}{2}\frac{C_x}{C_y}$$

i.e. if $$\rho > \frac{1}{2} \quad when \quad C_x = C_y$$

** ESTIMATION OF VARIANCE OF RATIO ESTIMATOR:-

The estimate of the variance of ratio estimator is given by,

$$EST(\bar{y}_R) = \frac{1-f}{n}\left[s_y^2 - 2rs_{yx} + r^2 s_x^2\right]$$

where $r = \bar{y}/\bar{x}$

$$s_y^2 = 1/n-1 \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$s_x^2 = 1/n-1 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$s_{yx}^2 = 1/n-1 \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

Now,

$$Est\ V(\bar{y}_R) = \frac{1-f}{n}\frac{1}{n-1}\left[\sum_{i=1}^{n}(y_i-\bar{y})^2 - 2r\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x}) + r^2\sum_{i=1}^{n}(x_i-\bar{x})^2\right]$$

$$= \frac{1-f}{n}\frac{1}{n-1}\left[\sum_{i=1}^{n}\{(y_i-\bar{y}) - r(x_i-\bar{x})\}^2\right]$$

$$= \frac{1-f}{n}\frac{1}{n-1}\left[\sum_{i=1}^{n}\{y_i - rx_i\}^2\right]$$

$$= \frac{1-f}{n}\frac{1}{n-1}\left[\sum_{i=1}^{n}y_i^{\,2} - 2r\sum_{i=1}^{n}y_i x_i + r^2\sum_{i=1}^{n}x_i^{\,2}\right]$$

** UNBIASED RATIO TYPE ESTIMATOR:-

An Unbiased Ratio Type Estimator of the population mean $\bar{Y}$ is defined as,

$$\bar{Y}_{GH} = \bar{r}\,\bar{x} + \frac{n(N-1)}{N(n-1)}(\bar{y} - \bar{r}\,\bar{x}) \qquad \text{.....(i)}$$

where $\bar{r} = 1/n \sum_{i=1}^{n} r_i$ $\qquad \bar{x} = 1/n \sum_{i=1}^{n} x_i$

$\bar{R} = 1/N \sum_{i=1}^{N} r_i$ $\qquad \bar{y} = 1/n \sum_{i=1}^{n} y_i$

$r_i = y_i / x_i$

$\bar{X} = 1/N \sum_{i=1}^{N} x_i$ $\qquad \bar{Y} = 1/N \sum_{i=1}^{N} y_i$

Now, taking expectation on both sides,

$$E\left[\bar{Y}_{GH}\right] = E\left[\bar{r}\,\bar{x}\right] + \frac{n(N-1)}{N(n-1)}\ E(\bar{y} - \bar{r}\,\bar{x})$$

$$= \bar{R}\,\bar{X} + \frac{N-1}{N}\ E\left[\frac{n\bar{y} - n\bar{r}\,\bar{x}}{n-1}\right]$$

$$= \bar{R}\,\bar{X} + \frac{N-1}{N}\ E\left[\frac{\sum_{i=1}^{n} y_i - n\bar{r}\,\bar{x}}{n-1}\right] \qquad \left[\because \sum y_i/n = \bar{Y}\right]$$

$$= \bar{R}\,\bar{X} + \frac{N-1}{N}\ E\left[\frac{\sum_{i=1}^{n} r_i x_i - n\bar{r}\,\bar{x}}{n-1}\right] \qquad \left[\because \sum y_i/x_i = r_i\right]$$

$$= \bar{R}\,\bar{X} + \frac{N-1}{N}\; E[s_{rx}]$$

$$= \bar{R}\,\bar{X} + \frac{N-1}{N}\; S_{rx} \qquad [\because (Es_{rx}) \cong S_{rx}\;]$$

$$= \bar{R}\,\bar{X} + \frac{N-1}{N}\left[\frac{\sum_{i=1}^{N} r_i x_i \;-\; N\,\bar{R}\,\bar{X}}{N-1}\right]$$

$$= \bar{R}\,\bar{X} + \frac{1}{N}\sum_{i=1}^{N} r_i x_i \;-\; \bar{R}\,\bar{X}$$

$$= \frac{1}{N}\sum_{i=1}^{N} r_i x_i$$

$$= \frac{1}{N}\sum_{i=1}^{N} y_i \qquad [\because\; r_i = y_i / x_i]$$

$$= \bar{Y}$$

$$E(\bar{y}_{GH}) = \bar{Y}$$

EXAMPLE :- Complete the variance of ratio estimators of the population mean :

When N = 100, n = 25, ρ= ½ $C_x = C_y$ complete it with SRSWOR :-

$$MSE\,(\overline{y_R}) = \bar{Y}^2\;\; \frac{1-f}{n}\left[C_y^2 - 2\rho C_y C_x + C_x^2\right]$$

$$= \bar{Y}^2\;\; \frac{1-f}{n}\left[C_y^2 - C_y{}^2 2\rho + C_y^2\right]$$

$$= \bar{Y}^2\;\; \frac{1-f}{n}\left[2C_y^2 - 2\rho C_y{}^2\right]$$

$$= \bar{Y}^2\;\; \frac{1-f}{n}\; 2C_y^2\;[1-\rho]$$

$$= \bar{Y}^2\;\; \frac{1-f}{n}\;\frac{2S_y^2}{\bar{Y}^2}\;[1-\rho] \qquad [\because C_y^2 = S_y^2/\bar{Y}^2\;]$$

$$f = \frac{n}{N} = \frac{25}{100} = \frac{1}{4} = 0.25$$

$$= \frac{0.75}{25} \times 2.\; S_y^2\left(1 - \frac{1}{2}\right)$$

$$= \frac{0.75}{25}\;\; S_y^2$$

$$= \frac{75}{25}\;\frac{S_y^2}{100}$$

$$MSE\;(\bar{y}_R) = 0.03\;\; S_y^2$$

For SRSWOR $V(\bar{y}_n) = \frac{N-n}{Nn} S_y^2$

$= \frac{100-25}{2500} S_y^2$

$= 0.03\, S_y^2$

** REGRESSION METHOD OF ESTIMATOR:-

The difference estimator of the population mean $\bar{Y}$ is defined as,

$\bar{Y}_d = \bar{y} + d(\bar{X} - \bar{x})$ ……..(*)

where d is some constant Now,

$E(\bar{Y}_d) = E(\bar{y}) + d\, E(\bar{X} - \bar{x})$

$= \bar{y} + d\,(\bar{X} - \bar{X})$

$= \bar{y}$

i.e. $\bar{y}_d$ is an unbiased estimator of $\bar{y}$

next,

$V(\bar{y}_d) = E\left[\bar{y}_d - E(\bar{y}_d)\right]^2$

$= E\left[\bar{y}_d - \bar{y}\right]^2$

$= E\left[\bar{y} + d\,(\bar{X} - \bar{x}) - \bar{y}\right]^2$ ……..[from (*)]

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

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

$= E(\bar{y} - \bar{Y})^2 - 2dE(\bar{y} - \bar{Y})(\bar{x} - \bar{X}) + d^2 E(\bar{x} - \bar{X})^2$

$= V(\bar{y}) - 2d\,\mathrm{cov}(\bar{y}, \bar{x}) + d^2 V(\bar{x})$

To determine constant d we minimize $V(\bar{y}_d)$ w.r.t. d.

$\frac{\partial V(\bar{Y}_d)}{\partial d} = 0 \Rightarrow \frac{\partial}{\partial d}\left[V(\bar{y}) - 2d\,\mathrm{cov}(\bar{y}, \bar{x}) + d^2 V(\bar{x})\right] = 0$

$\Rightarrow -2\,\mathrm{cov}(\bar{y}, \bar{x}) + 2\,d\,V(\bar{x}) = 0$

$\Rightarrow 2\,\mathrm{cov}(\bar{y}, \bar{x}) = 2\,d\,V(\bar{x})$

$\Rightarrow \mathrm{cov}(\bar{y}, \bar{x}) = d\,V(\bar{x})$

$\Rightarrow \quad d = \frac{\mathrm{cov}(\bar{y}, \bar{x})}{V(\bar{x})}$

$= \frac{\frac{1-f}{n} S_{yx}}{\frac{1-f}{n} S_x^{\,2}} = \frac{S_{yx}}{S_x^{\,2}} = \beta$

β is Regression coefficient of y over x.

$\therefore$ the difference estimators becomes

$$\bar{y}_{\beta} = \bar{y} + \beta(\bar{X} - \bar{x})$$

Where β the population regression coefficient of y on x is unknown it will be replaced by simple regression coefficient of y on x, b-we get the linear regression estimator of $\bar{Y}$.

i.e. $\bar{Y}_{lr} = \bar{y} + b(\bar{X} - \bar{x})$ where b = s_{yx}/s_x^2, lr = linear regression.

** BIAS OF REGRESSION ESTIMATOR:-

Let us define, (1) $e_0 = \dfrac{\bar{y} - \bar{Y}}{\bar{Y}}$ $E(e_0) = 0$

(2) $e_1 = \dfrac{\bar{x} - \bar{X}}{\bar{X}}$ $E(e_1) = 0$

(3) $e_2 = \dfrac{b - \beta}{\beta}$ $E(e_2) = 0$

$$E(e_0^{\,2}) = \frac{V(\bar{y})}{\bar{Y}^2} = \frac{1}{\bar{Y}^2}\frac{N-n}{Nn} \quad S_y^2 = \frac{1-f}{n}S_y^2/\bar{Y}^2$$

$$E(e_1^{\,2}) = \frac{1-f}{n}S_x^2/\bar{X}^2$$

$$E(e_0 e_1) = \frac{1-f}{n}E\left[\frac{(\bar{y}-\bar{Y})(\bar{x}-\bar{X})}{\bar{Y}\,\bar{X}}\right] = \frac{\text{cov}\,(\bar{y},\bar{x})}{\bar{Y}\,\bar{X}} = \frac{1-f}{n}\frac{S_{yx}}{\bar{x}\,\bar{y}}$$

$$E(e_1 e_2) = \frac{\text{cov}\,(\bar{x},b)}{\bar{X}\,\beta}$$

There fore the regression estimator becomes.

$$\begin{aligned}\bar{Y}_{lr} &= \bar{y} + b(\bar{X} - \bar{x}) \\ &= \bar{y}(1+e_0) + (e_2+1)\beta(-e_1\bar{X}) \\ &= \bar{y}(1+e_0) - \beta\bar{X}\,e_1(1+e_2)\end{aligned}$$

i.e. $\bar{Y}_{lr} - \bar{Y} = \bar{y}e_0 - \beta\bar{X}\,e_1(1+e_2)$

The bias of regression estimator is given by:

$$\begin{aligned}Bias\,(\bar{Y}_{lr}) &= E\left[\bar{Y}_{lr} - \bar{Y}\right] \\ &= E\left[\bar{Y}_{e_0} - \beta\bar{X}\;e_1(1+e_2)\right] \\ &= E\left[\bar{Y}_{e_0} - \beta\bar{X}\;(e_1 + e_1e_2)\right] \\ &= \bar{Y}\,E(e_0) - \beta\bar{X}\,E(e_1 + e_1e_2) \\ &= \bar{Y}\,0 - \beta\bar{X}\,E(e_1) - \beta\bar{X}\,E(e_1e_2) \\ &= 0 - 0 - \beta\bar{X}\,E(e_1e_2) \\ &= -\beta\bar{X}\,E(e_1e_2) \\ &= -\beta\bar{X}\,\frac{\text{cov}\,(\bar{x},b)}{\beta\,\bar{X}}\end{aligned}$$

$$Bias\ (\bar{Y}_{lr}) = -\operatorname{cov}\ (\bar{x}, b)$$

** MEAN SQUARE ERROR OF REGRESSION ESTIMATOR:-

Mean square error of the regression estimator of the population mean to the first degree of approximation is given by:

$$\begin{aligned} MSE\ (\bar{Y}_{lr}) &= E\left[\bar{Y}_{lr} - \bar{Y}\right]^2 \\ &= E\left[\bar{Y}e_0 - \beta\bar{X}\ e_1(1+e_2)\right]^2 \\ &= E\left[\bar{Y}^2 e_0^{\ 2} - 2\beta\bar{X}\bar{Y}\ e_1 e_0(1+e_2) + \beta^2\bar{X}^2\ e_1^{\ 2}(1+e_2)^2\ \right] \end{aligned}$$

$$= \bar{Y}^2 E(e_0^{\ 2}) - 2\beta\bar{X}\bar{Y}E\ (e_1 e_0)\frac{b}{\beta} + \beta^2\bar{X}^2\ e_1^{\ 2}\frac{b^2}{\beta^2} \qquad \left[\because \quad 1+e_2 = \frac{b}{\beta}\ \right]$$

$$= \bar{Y}^2 E(e_0^{\ 2}) - 2\bar{X}\bar{Y}bE\ (e_1 e_0) + b^2\bar{X}^2\ E(e_1^{\ 2})$$

$$= \bar{Y}^2\frac{1-f}{n}\frac{S_y^2}{\bar{Y}^2} - 2\bar{X}\bar{Y}b\frac{1-f}{n}\frac{S_{yx}}{\bar{Y}\ \bar{X}} + b^2\bar{X}^2\ \frac{1-f}{n}\frac{S_x^2}{\bar{X}}$$

$$= \frac{1-f}{n}\left[S_y^2 - 2bS_{yx} + b^2\ S_x^2\right]$$

$$= \frac{1-f}{n}\left[S_y^2 - 2b\rho S_y\ S_x + b^2\ S_x^2\right] \qquad \left[\because S_{yx} = \rho S_y S_x\right]$$

$$= \frac{1-f}{n}\left[S_y^2 - 2\frac{\rho S_y}{S_x}\ S_y\ S_x\rho + b^2\ S_x^2\right] \qquad \left[b = \frac{S_{yx}}{S_x^2} = \frac{\rho S_y S_x}{S_x^2}\right]$$

$$= \frac{1-f}{n}\left[S_y^2 - 2\rho^2\ S_y^{\ 2} + \rho^2\ S_y^2\right]$$

$$= \frac{1-f}{n}\left[S_y^2 - \rho^2\ S_y^{\ 2}\right] = \frac{1-f}{n}S_y^2(1-\rho^2)$$

Which is mean square error of regression estimator.

** COMPARISON OF REGRESSION ESTIMATOR WITH SRSWOR AND RATIO ESTIMATOR:-

In SRSWOR we have $V(\bar{Y}_n) = \frac{1-f}{n}S_y^2$

For regression estimation we have $MSE\ (\hat{\bar{Y}}_{lr}) = \frac{1-f}{n}S_y^2(1-\rho^2)$

now, $V(\bar{Y}_n) - MSE\ (\hat{\bar{Y}}_{lr}) = \frac{1-f}{n}S_y^2 - \frac{1-f}{n}S_y^2(1-\rho^2)$

$$= \frac{1-f}{n}S_y^2\ \rho^2 > 0$$

i.e. $V(\bar{Y}_n) > MSE(\hat{Y}_{lr})$

i.e. Regression estimator is superior to SRSWOR for ratio estimator, we have,

$$MSE\ (\hat{\bar{Y}}_R) = \frac{1-f}{n}\left[S_y^2 - 2R\rho S_y S_x + R^2 S_x{}^2\right]$$

now

$$MSE\ (\hat{\bar{Y}}_R) - MSE\ (\bar{Y}_{lr}) = \frac{1-f}{n}\left[S_y^2 - 2R\rho S_y S_x + R^2 S_x{}^2\right] - \frac{1-f}{n}\left[S_y^2(1-\rho)^2\right]$$

$$= \frac{1-f}{n}\left[S_y^2 - 2R\rho S_y S_x + R^2 S_x{}^2 - S_y^2 + \rho^2 S_y^2\right]$$

$$= \frac{1-f}{n}\left[R^2 S_x{}^2 - 2R\rho S_y S_x + \rho^2 S_y^2\right]$$

$$= \frac{1-f}{n}\left[RS_x - \rho S_y\right]^2 \geq 0$$

i.e. $MSE\ (\hat{\bar{Y}}_R) \geq MSE\ (\bar{Y}_{lr})$

i.e. regression estimation is superior to ratio estimation and equally efficient.

If $RS_x - \rho S_y = 0$

i.e. If $RS_x = \rho S_y$

i.e. if $\frac{\bar{Y}}{\bar{X}} S_x = \rho S_y$

i.e. if $\bar{Y} = \rho \frac{S_y}{S_x} \bar{X}$ $\quad \left| \because \beta = \frac{S_{yx}}{S_x{}^2} = \frac{\rho S_y S_x}{S_x{}^2} = \frac{S_y}{S_x} \right.$

i.e. if $\bar{Y} = \beta \bar{X}$

i.e. if the line of regression of Y on X is a straight line passing through origin.

NOTE:- When $\bar{Y}_{lr}$ *and* $\hat{\bar{Y}}_R$ are equally efficient there use $\hat{\bar{Y}}_R$ because it dose not require the compi
tation of b_i .

** APPLICATION:-

When the measurement of one variable (study variable) is complicative costly and time consuming where as the measurement of another variable (auxiliary variable) is simple , cheaper and quicker then the regression estimator is used.

Suppose, We wish to estimate the average area of a leaf on a certain plants the weight of a leaf. Here the measurement of y is complicative, costly

and time consuming where as measurement of x is simple, cheaper and quicker.

We can weight all the leaves together and get the total $\sum X = N\overline{X}$. Let us have a SRSWOR of n leaves and hence the value of y and x we compute $\overline{Y}, \overline{X}$ and b for the sample then the regression estimator of the average of a leaf is given by,

$$\overline{Y}_{lr} = \overline{y} + b(\overline{X} - \overline{x})$$

** SYSTEMETIC SAMPLING **

In systematic sampling the 1st sample is selected according to given pattern. Suppose, there are N units in population. Let n be a sample size. Assume that N = nk, We list the units of population. We draw a random number term 1 to k. Suppose it is i (i =1,2,…,k.) then the i^{th} unit is selected in the sample the remaining (n-1) units are selected as the units having no.

i+k, i+2k, i+3k,…….i+(n-1)k.

** SYSTEMETIC SAMPLING **

In systematic sampling the 1st sample is selected according to given pattern. Suppose, there are N units in population. Let n be a sample size. Assume that N = nk, We list the units of population. We draw a random number term 1 to k. Suppose it is i (i =1,2,…,k.) then the i^{th} unit is selected in the sample the remaining (n-1) units are selected as the units having no.

i+k, i+2k, i+3k,…….i+(n-1)k.

thus the systematic sample of size n the units with no consist of i , i+k, i+2k, i+3k,…….i+(n-1)k. Where k is called the interval of systematic sample. This is called the linear systematic sampling.

e.g. N = 100, n = 10, k = N/n = 10.

The 1st unit is selected by drawing a random no. from 1 to 10. Suppose it is 7. Then 7^{th} unit is selected in the sample. Thus, systematic sample of size 10 contains the units as 7,17, 27, 37, 47, 57, 67, 77, 87, 97.

** SELECTION OF ALL SYSTEMATIC SAMPLE:-

$$\begin{array}{ccccccc} 1 & 2 & \cdots & i & \cdots & k \\ 1+k & 2+k & \cdots & i+k & \cdots & 2k \\ 1+2k & 2+2k & \cdots & i+2k & \cdots & 3k \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1+(n-1)k & 2+(n-1)k & \cdots & i+(n-1)k & \cdots & (n-1)k \end{array}$$

1) There k systematic sample each of size n
2) Each systematic sample has probability 1/k of being selected.

** RELATION WITH STRATIFIED SAMPLING:-

Systematic sampling is similar to Stratified sampling in which N units are grouped into n strata each of size k and one unit per stratum is selected. But the difference between them lies in fact that the unit occupy the same relative position in different strata where as in stratified sampling units are selected at random in different strata.

* VALUE OF VARIABLE Y IN THE POPULATION:-

1^{st}	2^{nd}	3^{rd}	$\cdots$	i^{th}	$\cdots$	n^{th}	*mean*
Y_{11}	Y_{21}	Y_{31}	$\cdots$	Y_{i1}	$\cdots$	Y_{n1}	$\bar{Y}1$
Y_{12}	Y_{22}	Y_{32}	$\cdots$	Y_{i2}	$\cdots$	Y_{n2}	$\bar{Y}2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_{1j}	Y_{2j}	Y_{3j}	$\cdots$	Y_{ij}	$\cdots$	Y_{nj}	$\bar{Y}\ j$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_{1n}	Y_{2n}	Y_{3n}	$\cdots$	Y_{in}	$\cdots$	Y_{nn}	$\bar{Y}n$
mean $\bar{Y}1$	$\bar{Y}2$	$\bar{Y}3$	$\cdots$	$\bar{Y}i$	$\cdots$	$\bar{Y}n$	$\bar{\bar{Y}}$

* NOTATIONS:-

Let, Y_{ij} = value of j^{th} unit in the i^{th} systematic sample.
Whrer i =1,2,…k. and k = 1,2,…..n.

$\bar{Y}_i = \sum_{j=1}^{n} y_{ij}/k$ = mean of the i^{th} systematic sample.

$\bar{Y}_j = \sum_{i=1}^{k} y_{ij}/k$ = mean of the j^{th} unit.

$$\bar{Y} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} Y_{ij}}{nk} = \frac{1}{k}\sum_{j=1}^{n}\bar{Y}_i$$

= mean of the population.

= mean of the systematic sample mean.

**** EXPECTED VALUE OF SYSTEMATIC SAMPLE MEAN:-**

$$E(\bar{Y}_{sys}) = \sum_{i=1}^{k}\frac{1}{k}\,\bar{Y}_i$$

$$= \frac{1}{k}\sum_{i}\bar{Y}_i$$

$$= \frac{1}{nk}\sum_{i}\sum_{j} Y_{ij} \qquad \left| \because \bar{Y}_i = \sum_{j} Y_{ij} \Big/ n \right.$$

$$= \frac{1}{N}\sum_{i}\sum_{j} Y_{ij}$$

$$= \bar{Y} = \text{Population mean}$$

$$E(\bar{Y}_{sys}) = \bar{Y} = \bar{Y}_N$$

Systematic sample mean $(\bar{Y}_{sys})$ is an unbiased estimator of population mean $\bar{Y}_N$.

*** Variance of Systematic Mean:-**

$$V(\bar{Y}_{sys}) = E\left[\bar{Y}_{sys} - E(\bar{Y}_{sys})\right]^2$$

$$= E\left[\bar{Y}_{sys} - \bar{Y}..\right]^2$$

$$= E\left[\frac{\sum_{i}\sum_{j} Y_{ij}}{nk} - \bar{Y}..\right]^2$$

$$= E\left[\sum_{i}\frac{1}{k}\bar{Y}_i - \bar{Y}..\right]^2$$

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

(I) ALTERNATIVE EXPRESSION OF VARIANCE :-

$$V(\bar{Y}_{sys}) = \frac{N-1}{N} S_y^2 - \frac{N-k}{N} S_{wsys}^2$$

Where $S_y^2 = \frac{1}{N-1}\sum_i \sum_j (Y_{ij} - \bar{Y}..)^2$ ……(1)

= mean square among the units in the whole population.

$$S_{wsys}^2 = \frac{1}{k}\sum_i S_i^2 = \frac{1}{k(n-1)}\sum_i \sum_j (Y_{ij} - Y_i)^2$$ ……..(2)

=mean square among the units within the systematic sample.

Proof:- we have,

$$\sum_i \sum_j (Y_{ij} - \bar{Y}..)^2 = \sum_i \sum_j (Y_{ij} - \bar{Y}_i + \bar{Y}_i - \bar{Y}..)^2$$

$$= \sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2 + \sum_i \sum_j (\bar{Y}_i - \bar{Y}..)^2 + PT$$

i.e. $(N-1)S_y^2 = k(n-1)S_{wsys}^2 + n\sum (Y_i - \bar{Y}..)^2$ from (1) and (2)

i.e. $(N-1)S_y^2 = (N-k)S_{wsys}^2 + nk\ V(\bar{Y}_{sys})$

i.e. $V(\bar{Y}_{sys}) = \frac{N-1}{N} S_y^2 - \frac{N-k}{N} S_{wsys}^2$

* COMPARISION WITH SRSWOR:-

In SRSWOR we have,

$$V(\bar{Y}_n) = \frac{N-n}{Nn} S_y^2$$

In systematic sampling we have,

$$V(\bar{Y}_{sys}) = \frac{N-1}{N} S_y^2 - \frac{N-K}{N} S_{wsys}^2$$

Systematic sample is better then SRSWOR

i.e If $V(\bar{Y}_n) > V(\bar{Y}_{sys})$

i.e If $\frac{N-n}{Nn} S_y^2 > \frac{N-1}{N} S_y^2 - \frac{N-k}{N} S_{wsys}^2$

i.e. If $\left(\frac{N-1}{N} - \frac{N-n}{Nn}\right) S_y^2 < \frac{N-k}{N} S_{wsys}^2$

i.e. If $\left(\frac{n-1}{n}\right) S_y^2 < \frac{k(n-1)}{nk} S_{wsys}^2$

i.e. If $S_y^2 < S_{wsys}^2$

i.e. If mean square among the units in the whole population < mean square among the units within the systematic sample.

(II) $V(\bar{Y}_{sys}) = \frac{N-1}{nN} S_y^2 [1+(n-1)\rho]$

where, $\rho = \dfrac{\sum_{j\neq i}^{n}\sum_{i=1}^{k}(y_{ij}-\bar{Y})(y_{ij}-\bar{Y}..) \Big/ kn(n-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{Y}..)^2 \Big/ nk}$

$$(1) \leftarrow \quad = \frac{\sum_{i=1}^{k}\sum_{j\neq i=1}^{n}(y_{ij}-\bar{Y}..)(y_{ij}-\bar{Y}..)}{(n-1)\ (N-1)\ S_y^2}$$

= intraclass correlation coefficient between pair of units within the systematic sample.

Proof:- we have,

$$V(\bar{Y}_{sys}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{Y}_i - \bar{Y}..)^2$$

$$n^2 V(\bar{Y}_{sys}) = \frac{n^2}{k}\sum_{i=1}^{k}(\bar{Y}_i - \bar{Y}..)^2 \qquad \text{[multiplying by } n^2$$

$$n^2 k V(\bar{Y}_{sys}) = \sum_i (n\bar{Y}_i - n\bar{Y}..)^2$$

$$n^2 k V(\bar{Y}_{sys}) = \sum_i \left(\sum_j Y_{ij} - n\bar{Y}..\right)^2 \qquad \left| \frac{\sum_j Y_{ij}}{n} = \bar{Y}_i \right.$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}..)^2 + \sum_i \sum_{j\neq j=1}(Y_{ij}-\bar{Y}..)(Y_{ij}-\bar{Y}..)$$

$$= (N-1)S_y^2 + \rho(n-1)(N-1)S_y^2 \qquad \left| \because \textit{ from (1) \& equation of } S_y^2 \right.$$

$$n^2 k V(\bar{Y}_{sys}) = (N-1)S_y^2[1+(n-1)\rho]$$

i.e. $$V(\bar{Y}_{sys}) = \frac{N-1}{nN} S_y^2 [1+(n-1)\rho]$$

* Comparison With SRSWOR:-

In SRSWOR,

$$V(\bar{Y}_n) = \frac{N-n}{Nn} S_y^2$$

$$V(\bar{Y}_{sys}) = \frac{N-1}{nN} S_y^2 [1+(n-1)\rho]$$

Systematic sampling is better than SRSWOR

i.e If $V(\bar{Y}_{sys}) < V(\bar{Y}_n)$

i.e. if $\frac{N-1}{nN} S_y^2 [1+(n-1)\rho] < \frac{N-n}{Nn} S_y^2$

i.e. if $\left(\frac{N-1}{nN} - \frac{N-n}{Nn}\right) S_y^2 < -\frac{N-1}{nN} S_y^2 (n-1)\rho$

i.e. if $\frac{N-1}{nN} S_y^2 < -\frac{N-1}{nN} S_y^2 (n-1)\rho$

i.e. if $1 < -(N-1)\rho$

i.e. if $1 > (N-1)\rho$

i.e. if $\rho < \frac{1}{N-1}$

(III) $V(\bar{Y}_{sys}) = \frac{S_{wst}^2}{n}\left[\frac{N-n}{N} + (n-1)\ \rho_{wst}\right]$

(i) where $S_{wst}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_j)^2 \big/ n(k-1)$

$$\rho_{wst} = \frac{\sum_{i=1}^{k}\sum_{j\neq i=1}^{n}(Y_{ij} - \bar{Y}_{\cdot j})(Y_{ij} - \bar{Y}_{\cdot j}) \big/ kn(n-1)}{\sum_{i}\sum_{j}(Y_{ij} - \bar{Y}_{\cdot j})^2 \big/ n(k-1)}$$

(2) $$= \frac{\sum_{i=1}^{k}\sum_{j\neq i=1}^{n}(Y_{ij} - \hat{Y}_{ij})(Y_{ij} - \bar{Y}_{\cdot j}) \big/ kn(n-1)}{S_{wst}^2}$$

= Correlation coefficient between pairs of units within the strata.

Proof:- $V(\bar{Y}_{sys}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{Y}_i - \bar{Y}..)^2$

$n^2 k V(\bar{Y}_{sys}) = \sum_{i=1}^{k}(n\bar{Y}_i - \bar{Y}..)^2$ $\quad \big| \because$ *multiplying by* n^2

$= \sum_{i=1}^{k}\left(\sum_{j=1}^{n} Y_{ij} - n\bar{Y}..\right)^2$ $\quad \left| \bar{Y}_i = \sum_j Y_{ij} \right/ n$

$= \sum_{i=1}^{k}\left(\sum_{j=1}^{n}(Y_{ij} - {}^{n}\bar{Y}..)\right)^2$

$= \sum_{i=1}^{k}\left(\sum_{j=1}^{n} Y_{ij} - \sum_i y_{\cdot j}\frac{n}{n}\right)^2$ $\quad \left| \bar{Y} = \frac{\sum_j y_{\cdot j}}{n} \right.$

$$= \sum_{i=1}^{k}\left(\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{\cdot j})\right)^2$$

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

$$= n(k-1)S_{wst}^2 + kn(n-1)\rho_{wst}S_{wst}^2 \qquad \because \text{ from (1) \& (2)}$$

$$= S_{wst}^2\left[n(k-1)+kn(n-1)\rho_{wst}\right]$$

$$V(\bar{Y}_{sys}) = \frac{S_{wst}^2}{n^2k}\left[n(k-1)+kn(n-1)\rho_{wst}\right]$$

$$= \frac{S_{wst}^2}{n^2k}\left[\frac{n(k-1)}{nk}+\frac{kn(n-1)\rho_{wst}}{nk}\right]$$

$$= \frac{S_{wst}^2}{n}\left[\frac{nk-n}{nk}+(n-1)\rho_{wst}\right]$$

$$= \frac{S_{wst}^2}{n}\left[\frac{N-n}{N}+(n-1)\rho_{wst}\right]$$

* Comparison With Stratified Sampling

There are N units divided into n strata each of size k and 1 unit per strata is selected.

$$V(\bar{Y}_{st}) = \sum_{j=1}^{n}p_j^2\left(\frac{1}{n_j}-\frac{1}{N_j}\right)S_j^2$$

Where $p_j = \frac{N_j}{N} = \frac{k}{nk} = \frac{1}{n} = \frac{n_j}{n}$ $\qquad \therefore n_j = 1, \quad N_j = k$

$$S_j^2 = \frac{1}{k-1}\sum_{i=1}^{k}(Y_{ij}-\bar{Y}_j)^2$$

$$V(\bar{Y}_{st}) = \sum_{j=1}^{n}\frac{1}{n^2}\left(\frac{1}{1}-\frac{1}{k}\right)\frac{1}{k-1}\sum_{i=1}^{k}(Y_{ij}-\bar{Y}_{\cdot j})^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n}\frac{1}{n^2(k-1)}\left(\frac{k-1}{k}\right)(Y_{ij}-\bar{Y}_{\cdot j})^2$$

$$= \frac{1}{n^2k}\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{\cdot j})^2$$

$$= \frac{1}{nN}\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{\cdot j})^2$$

$$n(k-1)V(\bar{Y}_{st}) = \frac{n(k-1)}{nN}\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{\cdot j})^2 \qquad \Big| \because \text{ multiply by } \frac{1}{n(k-1)}$$

$$= S_{wst}^2 \quad \frac{N-n}{Nn}$$

On comparison we note that Systematic Sampling has the same precision as corresponding equivalent stratified sampling if $\rho_{wst} = 0$.

** SPECIAL POPULATION:-

If we have special population we have these type of situations.

	1	2	$\cdots$	i	$\cdots$	k
1	$\mu+h$	$\mu+2h$	$\cdots$	$\mu+ih$	$\cdots$	$\mu+kh$
2	$\mu+(k+1)h$	$\mu+(k+2)h$	$\cdots$	$\mu+(k+i)h$	$\cdots$	$\mu+2kh$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$\mu+\{(n-1)(k+1)\}h$	$\mu+\{(n-1)(k+2)\}h$	$\cdots$	$\mu+\{(n-1)(k+i)\}h$	$\cdots$	$\mu+nkh$

The value of population increase accordingly to linear law i.e. differ by constant amount h(say).

Example:-Suppose t values of X are μ+h, μ+2h,….., μ+lh,…., μ+th

Mean of this series is,

$$\text{Mean} = \frac{1}{t}\sum_{i=1}^{t}(\mu+ih)$$

$$= \frac{1}{t}\left[\mu t + h\frac{t(t+1)}{2}\right] \qquad \left|\sum_{i=1}^{t} i = \frac{t(t+1)}{2}\right.$$

$$\text{Mean} = \mu + h\frac{(t+1)}{2}$$

$$\text{Mean Square} = \frac{1}{N-1}\sum_{i}(y_i - \bar{Y})^2$$

$$= \frac{1}{t-1}\left[\sum_{i=1}^{t}\left\{\mu+ih-\mu-h\frac{(t+1)}{2}\right\}^2\right]$$

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

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

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

$$= \frac{h^2}{t-1}\frac{t(t+1)}{12}\left[2(2t+1)-(t+1)6+3(t+1)\right]$$
$$= \frac{h^2}{t-1}\frac{t(t+1)}{12}\left[4t+2-6t-6+3t+3)\right]$$
$$= \frac{h^2}{12}\frac{t(t+1)}{t-1}\left[t-1\right]$$
$$= \frac{h^2}{12}t\left[t-1\right]$$

****Theorem:-**For special population with usual notation prove that $V_{st}:V_{sy}:V_{ran}$: :1/n:1:n if k is large and hence,

$E_{st} : E_{sys} : E_{ran}$: : n^2 : n : 1

Proof:-(1) In SRSWOR, we have,

$$V_{ran} = \frac{N-n}{nN}S_y^2$$
$$= \frac{N-n}{nN}\frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{Y}..)^2$$

There are nk = N values in the population differing by constant amount h.

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

(2) In systematic sampling, we have,

$$V_{sys} = \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_i-\bar{Y}..)^2$$
$$= \frac{k-1}{k}\frac{1}{k-1}\sum_{i}(\bar{y}_i-\bar{Y}..)^2$$

$\bar{Y}_{i.}$ differs by a constant amount h and there are such k values.

$$V_{sys} = \frac{k-1}{k}\ \frac{h^2k(k-1)}{12}$$
$$= \frac{h^2(k^2-1)}{12}$$

(3) In stratified sampling, we have.

$$V_{st} = \sum_{j=1}^{n}p_j^2\left(\frac{1}{n_j}-\frac{1}{N_j}\right)S_j^2$$
$$= \sum_{j=1}^{n}p_j^2\left(\frac{1}{n_j}-\frac{1}{N_j}\right)\frac{1}{k-1}\sum_{i=1}^{k}(Y_{ij}-Y._j)^2$$

There are k values each is differing by a constant h.

$$= \frac{\sum_{i=1}^{k} (Y_{ij} - Y_{.j})^2}{k-1} = \frac{h^2\ k\,(k+1)}{12}$$

There are N = nk values divided in n strata each of size k and one unit per stratum to selected.

$$V_{st} = \sum_{j=1}^{n} \frac{1}{n^2}\left(\frac{1}{1} - \frac{1}{k}\right) \frac{h^2 k(k-1)}{12}$$

$$= \frac{h^2(k^2-1)}{12n}$$

There fore

$$V_{st} : V_{sy} : V_{ran} :: \frac{h^2(k^2-1)}{12n} : \frac{h^2(k^2-1)}{12} : \frac{h^2(k-1)(nk+1)}{12n}$$

$$:: \frac{k+1}{n} : (k+1) : nk+1$$

$$: \frac{1}{n} : 1 : n \qquad \text{[coordinate of k is taken]}$$

if k is large.

Next, $\frac{V_{ran}}{V_{st}} : \frac{V_{ran}}{V_{sy}} : \frac{V_{ran}}{V_{ran}} :: n^2 : n : 1$ if n is large,

i.e. $E_{st} : E_{sy} : E_{ran} : : n^2 : n : 1.$

** CIRCULAR SYSTEM SAMPLING:-

The case N = nk is called linear systematic sampling. Let as examine what happens when N ≠ nk.

e.g. N = 11, n = 3, k = N/n = 11/3 ≅ 4

Random strata	Systematic Sample i, i+k, i+2k,……i+(n-1)k			Probability
1	Y_1	Y_5	Y_9	¼
2	Y_2	Y_6	Y_{10}	¼
3	Y_3	Y_7	Y_{11}	¼
4	Y_4	Y_8		¼

$$E(\bar{Y}_{sy}) = \frac{1}{k}\sum_{j=1}^{n} Y_i$$

$$= \frac{1}{4}\left[\bar{Y}_1. + \bar{Y}_2. + \bar{Y}_3. + \bar{Y}_4.\right]$$

$$= \frac{1}{4}\left[\frac{Y_1+Y_5+Y_9}{3}+\frac{Y_2+Y_6+Y_{10}}{3}+\frac{Y_3+Y_7+Y_{11}}{3}+\frac{Y_4+Y_8}{2}+\frac{Y_4+Y_8}{3}-\frac{Y_4+Y_8}{3}\right]$$

$$= \frac{1}{4}\left[\sum_{i=1}^{11}\frac{Y_1}{3}+\{Y_4+Y_8\}\left\{\frac{1}{2}-\frac{1}{3}\right\}\right]$$

$$= \frac{\sum_i Y_i}{12}+\frac{Y_4+Y_8}{24}$$

$$= \frac{11\bar{Y}..}{12}+\frac{Y_4+Y_8}{24} \qquad \left| \because Y.. = \frac{\sum Y_i}{11}\right.$$

$$= \left(1-\frac{1}{12}\right)\bar{Y}..+\frac{Y_4+Y_8}{24}$$

$$= \bar{Y}..+\left(\frac{Y_4+Y_8}{24}-\frac{\bar{Y}..}{12}\right)$$

$$E(\bar{Y}_{sys}) \neq \bar{Y}..$$

** DRAW BACK OF LINEAR SYS SAMPLING (L.S.S.)

(1) All the sys sampling are not at the same size.
(2) Sys sampling mean is not an unbiased estimator of the population mean.

To remove these draw back a rular Sys. Sampling is suggested. In C.S.S. the 1st unit is selected by drawing a random no. from 1 to N and then select every k^{th} unit in a cyclic manner till a sample of size n is obtain where,

K = the nearest integer N/n

i.e. $k \cong N/n$

e.g. N = 11, n = 3, $k = N/n \cong 4$

Suppose the first random no. is 4 then C.S.S. of size 3. Contains the unit as Y_4, Y_8 , Y_i .

** TWO STAGE SAMPLING **

In cluster sampling the whole population is divided in to clusters and a sample of clusters is selected and the elements of the selected clusters are enumerated. If the elements of the clusters are more or less homogeneous then it is uneconomical to enumerate all the elements of the selected dusters. In this case two-stage or sub-sampling is suggested.

The sampling procedure of first selecting a sample of clusters and then selecting a sample of clusters and then selecting a sample of element from each selected clusters is known as two-stage or sub-sampling. The clusters which form sampling units at the first stage are called first-stage units and the elements within the cluster which form sampling units at the second stage are called second stage units.

For an example, while sampling of fields in a taluka, first a sample of villages is selected and from each selected village a sample of fields is selected.

This procedure can be extended and we have three stage or multi-stage sampling.

**** Notations :-**

N = number of first stage units in the population.

M = number of second stage units in the population.

Y_{ij} = Value of the j^{th} second stage unit in the i^{th} first stage unit.

i = 1,2,….,N.

j = 1,2,…….,M.

$\bar{Y}_i = \frac{1}{M}\sum_{j=1}^{M} Y_{ij}$ = Mean per second stage unit in i^{th} first stage unit in the population.

$$\bar{Y}.. = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} Y_{ij}$$

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

= Population mean = (mean per second stage unit in the whole population.)

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

= Mean square among the second stage units in the i^{th} first age units in the population.

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

= Mean square among the second stage units within the first age units in the population.

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

= Mean square among the first age units in the population.

We assume that a SRSWOR of n first age unit is known from the population and from each First stage unit a SRSWOR of m second stage unit is taken, thus we have a sample of nm second stage units.

$$\bar{Y}_{im} = \frac{1}{m}\sum_{j=1}^{m} Y_{ij}$$

= Mean per second stage unit in the i[th] first age unit in the sample.

$$\bar{Y}_{nm} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij} = \frac{1}{n}\sum_{i=1}^{n} \bar{Y}_{nm}$$

= Two stage sample mean. or mean per second stage unit in the i[th] first stage unit in the sample.

$$s_i^2 = \frac{1}{m-1}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{im})^2$$

= Mean square among the second stage units in the i[th] first age units in the sample.

$$\bar{S}_w^{\,2} = \frac{1}{n}\sum_{i=1}^{n} s_i^{\,2} = \frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{im})^2$$

= Mean square among the second stage units within the i[th] first age units in the sample.

$$s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_{im} - \bar{y}_{nm})^2$$

= Mean square among the first stage units in the sample.

We know that $V(X) = EV(X/Y) + V\,E(X/Y)$

$E(X) = E_1\ E_2(X/Y)$

*** Theorem:-** Prove that,

$$E(\bar{Y}_{nm}) = \bar{Y}..$$

$$V(\bar{Y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2$$

*** Proof:-**

$$E(\bar{Y}_{nm}) = E\left[\frac{1}{n}\sum_{i=1}^{n}\bar{Y}_{im}\right]$$

$$= E_1E_2\left[\frac{1}{n}\sum_{i=1}^{n}\bar{Y}_{im}/i\right]$$

$$= E_1\left[\frac{1}{n}\sum_{i=1}^{n}E_2(\bar{Y}_{im}/i)\right]$$

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

$$= \frac{1}{n}\sum_{i=1}^{n}E_1[\bar{Y}_{i}.]$$

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

$$= \bar{Y}.. \qquad (1)$$

$$V(\bar{Y}_{nm}) = V_1E_2\left[\frac{1}{n}\sum_{i=1}^{n}\bar{Y}_{im}/i\right] + E_1V_2\left[\frac{1}{n}\sum_{i=1}^{n}\bar{Y}_{im}/i\right]$$

$$= V_1\left[\frac{1}{n}\sum_{i=1}^{n}E_2(\bar{Y}_{im}/i)\right] + E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}V_2(\bar{Y}_{im}/i)\right]$$

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

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

$$= \left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{N}\sum_{i=1}^{N}S_i^2$$

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

So from (i) and (ii)

$$E(\bar{Y}_{nm}) = \bar{Y}..$$

$$V(\bar{Y}_{nm}) = \left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2$$

Estimation of $V(\bar{Y}_{nm})$

Theorem:- An unbiased estimator of the variance of two-stage sample mean is given by,

$$Est\ V(\bar{Y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2 + \frac{1}{N}\left(\frac{1}{m} - \frac{1}{M}\right) \bar{s}_w^2$$

where, $s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n} (\bar{y}_{im} - \bar{y}_{nm})^2$

$$s_i^2 = \frac{1}{m-1}\sum_{j=1}^{m} (y_{ij} - \bar{y}_{im})^2$$

$$\bar{s}_w^2 = \frac{1}{n}\sum_{i=1}^{n} s_i^2$$

Proof:-

$$V(\bar{Y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right) \bar{s}_w^2$$

i.e. $$Est\ V(\bar{Y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) Est\, s_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right) Est\, \bar{s}_w^2 \qquad \text{(iii)}$$

Now, $$E(\bar{s}_w^2) = E\left[\frac{1}{n}\sum_{i=1}^{n} s_i^2\right]$$

$$= E_1 E_2\left[\frac{1}{n}\sum_{i=1}^{n} s_i^2 / i\right]$$

$$= E_1\left[\frac{1}{n}\sum_{i=1}^{n} E_2(s_i^2 / i)\right]$$

$$= E_1\left[\frac{1}{n}\sum_{i=1}^{n} S_i^2\right]$$

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

i.e. $$Est\, \bar{s}_w^2 = \bar{S}_w^2 \qquad \text{(iv)}$$

and $$s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n} (\bar{y}_{im} - \bar{y}_{nm})^2$$

i.e. $$(n-1)\ E(s_b^2) = E\left[\sum_{i=1}^{n} (\bar{y}_{im})^2\right] - nE\left[\sum_{i=1}^{n} (\bar{y}_{nm})^2\right]$$

$$= E_1 E_2\left[\sum_{i=1}^{n} \bar{y}_{im}^2 / i\right] - n\left[V(\bar{y}_{nm}) + \bar{Y}^2..\right]$$

[Since $V(\bar{Y}_{nm}) = E(\bar{Y}_{nm}^2) - \left[E(\bar{Y}_{nm})\right]^2 = E(\bar{Y}_{nm}^2) - \bar{Y}..^2$

So, $E(\bar{Y}_{nm}^2) = V(\bar{Y}_{nm}) + \bar{Y}..^2$]

$$= E_1\left[\sum_{i=1}^{n} E_2(\bar{y}_{im}^2 / i)\right] - n\left[\left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2\right] - n\bar{Y}..^2$$

$$= E_1\left[\sum_{i=1}^{n}\left\{V_2(\bar{y}_{im}^2/i) + \left(E_2(\bar{y}_{im}/i)\right)^2\right\}\right] - n\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 - \left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 - n\bar{Y}_{..}^{\,2}$$

$$= E_1\left[\sum_{i=1}^{n}\left\{\left(\frac{1}{m}-\frac{1}{M}\right)S_i^2 + \bar{Y}_{i\cdot}^{\,2}\right\}\right] - n\bar{Y}_{..}^{\,2} - n\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 - \left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2$$

$$= n\left(\frac{1}{m}-\frac{1}{M}\right)E_1\left[\frac{1}{n}\sum_{i=1}^{n}S_i^2\right] + nE_1\left[\frac{1}{n}\sum_{i}\bar{Y}_{i\cdot}^{\,2}\right] - n\bar{Y}_{..}^{\,2} - n\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 - \left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2$$

$$= n\left(\frac{1}{m}-\frac{1}{M}\right)\left[\frac{1}{N}\sum_{i=1}^{N}S_i^2\right] + nE_1\left[\frac{1}{n}\sum_{i}\bar{Y}_{i\cdot}^{\,2}\right] - n\bar{Y}_{..}^{\,2} - n\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 - \left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2$$

$$= n\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 - n\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 - \left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + n\left(\frac{1}{N}\sum_{i}^{N}\bar{Y}_{i\cdot}^{\,2}\right) - n\bar{Y}_{..}^{\,2}$$

$$= \left(\frac{1}{m}-\frac{1}{M}\right)(n-1)\bar{S}_w^2 - n\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \frac{n}{N}\left(\sum_{i=1}^{N}\bar{Y}_{i\cdot}^{\,2} - N\bar{Y}_{..}^{\,2}\right)$$

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

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

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

$$(n-1)E(s_b^2) = (n-1)\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 - (n-1)S_b^2$$

i.e. $$E(s_b^2) = \left(\frac{1}{m}-\frac{1}{M}\right)E(\bar{s}_w^2) - S_b^2 \qquad \text{(v)}$$

**** Efficiency of Two-Stage Sampling and SRSWOR:-**

In SRSWOR we have,

$$V(\bar{Y}_{nm})\ SRSWOR = \left(\frac{1}{nm}-\frac{1}{NM}\right)S^2$$

$$\cong \frac{1}{nm}S^2 \quad \left(\frac{1}{M}\ \text{is negligible}\right) \qquad \text{or (M is large.)}$$

In two-stage sampling, we have,

$$V(\bar{Y}_{nm})\ two-stage = \left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2$$

$$\cong \left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \frac{1}{nm}\bar{S}_w^2 \quad \left(\frac{1}{M}\ is\ negligible\right)$$

now we have $M^2 (N-1) S_b^2 = (NM-1) S^2 [1+(M-1)\rho]$

i.e. $S_b^2 = \frac{NM-1}{M(N-1)} S^2\left[\frac{1+(M-1)\rho}{M}\right]$

$= \frac{N}{N-1} S^2\rho \qquad (\frac{1}{M}\ is\ \ negligible)$

Next we have,

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

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

$= \frac{\left(N-\frac{1}{M}\right)S^2 - (N-1)S_b^2}{N\left(1-\frac{1}{M}\right)}$

$= \frac{NS^2-(N-1)S_b^2}{N}$

$= S^2 - \left(\frac{N-1}{N}\right)S_b^2 \qquad (\because \frac{1}{M}\text{ is negligible })$

$= S^2 - \frac{N-1}{N}\ \frac{N}{N-1} S^2\ \rho$

$= S^2 - S^2\ \rho$

$= S^2(1-\rho)$

There fore $V(\bar{Y}_{nm})$ two-stage becomes.

$V(\bar{Y}_{nm})\ two\ stage \cong \frac{N-n}{nN}\ \frac{N}{N-1} S^2\ \rho\ + \frac{S^2}{nm}(1-\rho)$

$\cong \frac{N-n}{n(N-1)}\ S^2\ \rho + \frac{S^2}{nm}(1-\rho)$

$\cong \frac{S^2}{nm}\left[\frac{m(N-n)}{N-1}\ \rho + 1 - \rho\right]$

$\cong \frac{S^2}{nm}\left[1+\left\{\frac{m(N-n)}{N-1}\ -1\right\}\rho\right]$

Efficiency of two-stage sampling over SRSWOR is given by,

$\text{Efficiency} \cong \frac{V(\bar{Y}_{nm})\ SRSWOR}{V(\bar{Y}_{nm})\ two-stage}$

$\cong \frac{S^2/nm}{\frac{S^2}{nm}\left[1+\left\{\frac{m(N-n)}{N-1}-1\right\}\rho\right]}$

$$\text{Eff.} \quad \cong \frac{1}{1+\left\{\frac{m(N-n)}{N-1}-1\right\}\rho}$$

Obtain the efficiency of two stage sampling compared to SRSWOR in forms of intra class correlation to efficient.

Optimum Value of n and m :- (For Two-Stage Sampling)

$$V(\bar{Y}_{nm}) = \left(\frac{1}{n}-\frac{1}{N}\right) s_b^2 + \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right) \bar{s}_w^2$$

Consider the cost function.

$c = c_1 n + c_2 nm.$

Where c = total cost

c_1 = cost for collecting information per first stage unit.

c_2 = cost for collecting information per second stage unit.

Determine the optimum value of n and m such that variance.

i.e $V(\bar{Y}_{nm})$ is minimized for fined cost w, say,

$$\phi = \left(\frac{1}{n}-\frac{1}{N}\right) s_b^2 + \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right) \bar{s}_w^2 + \lambda[c_1 n + c_2 nm] = 0$$

$$\frac{\partial \phi}{\partial n} = -\frac{1}{n^2}\bar{s}_b^2 + (\frac{1}{m}-\frac{1}{M})\bar{s}_w^2 + \lambda(c_1 + c_2 m) = 0$$

$$\Rightarrow n^2 = \frac{s_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\bar{s}_w^2}{\lambda[c_1 + c_2 m]} \qquad \text{(i)}$$

$$\frac{\partial \phi}{\partial m} = -\frac{1}{nm^2}\bar{s}_w^2 + \lambda c_2 n = 0$$

$$\Rightarrow -\lambda c_2 n = -\frac{\bar{s}_w^2}{nm^2}$$

$$\Rightarrow n^2 = \frac{\bar{s}_w^2}{\lambda c_2 m^2} \qquad \text{(ii)}$$

From (i) and (ii) we have,

$$\frac{s_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\bar{s}_w^2}{\lambda[c_1 + c_2 mn]} = \frac{\bar{s}_w^2}{\lambda c_2 m^2}$$

i.e $$c_2 m^2\left[s_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\bar{s}_w^2\right] = [c_1 + c_2 m]\ \bar{s}_w^2$$

i.e $$c_2 m^2\left[s_b^2 - \left(\frac{1}{M}\bar{s}_w^2\right)\right] = c_1\ \bar{s}_w^2$$

i.e $c_2 m^2 D = c_1 \bar{s}_w^2$, *where* $D = s_b^2 - \frac{1}{M}\bar{s}_w^2$

$$m^2 = \frac{c_1 \bar{s}_w^2}{c_2 D}$$

If D > 0

$$\hat{m} = \sqrt{\frac{c_1 \bar{s}_w^2}{c_2 D}}$$

Now, $c_0 = c_1 n + c_2 n \hat{m}$

i.e. $\hat{n} = \frac{c_0}{c_1 + c_2 \hat{m}}$

If D < 0 , the above procedure fails in this case we proceed as follows.

$$V(\bar{Y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right) \bar{s}_w^2$$

$$= \frac{1}{n}\left[s_b^2 - \frac{1}{M}\bar{s}_w^2\right] + \frac{1}{nm}\bar{s}_w^2 - \frac{1}{N}s_b^2$$

$$= \frac{1}{n} D + \frac{1}{nm}\bar{s}_w^2 - \frac{1}{N}s_b^2$$

now, $c_0 = c_1 n + c_2 nm$.

i.e. $n = \frac{c_0}{c_1 + c_2 m}$ …….(iii)

Substituting the above , value of n from (iii) becomes,

$$V(\bar{Y}_{nm}) = \frac{c_1 + c_2 m}{c_0} D + \frac{c_1 + c_2 m}{c_0 m} \bar{S}_w^2 - \frac{1}{N} S_b^2$$

As m increases, $V(\bar{Y}_{nm})$ decreases and $V(\bar{Y}_{nm})$ is minimum when m is the largest positive integer.

i.e. $\hat{m} = M$. Then , n is obtained from the following.

$$\hat{n} = \frac{c_0}{c_1 + c_2 \hat{m}}$$

$$= \frac{c_0}{c_1 + c_2 M}$$

So $\hat{n} = \frac{c_0}{c_1 + c_2 \hat{m}}$ and $\hat{m} = M$ optimum value of n and m .

** CLUSTER SAMPLING**

A sample procedure pre-suppose the division of the population in to a finik no. of distinct and identifying units that called sampling unit. The smallest units in to which the population can be divided are called the element of population and the group of element are called clusters. Some times instead of taking an element as the sampling unit cluster is taken as the sampling unit. This is useful when the list of element is not available while the list of cluster is available.

Thus, in cluster sampling, A population is divided into cluster and sample random sample of clusters is selected and information for all elements of the selected cluster is collected.

EXAMPLE:-

While sampling of population is a certain city, the list of person residing in that city is not available, but the list of households is readily available.

Value of study variable y in the population:-

	1	2	…..	i	…..	N
1	Y_{11}	Y_{21}	….	Y_{i1}	…..	Y_{N1}
2	Y_{12}	Y_{22}	….	Y_{i2}	…..	Y_{N2}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
j	Y_{1j}	Y_{2j}	….	Y_{ij}	…..	Y_{Nj}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
M	Y_{1M}	Y_{2M}	….	Y_{iM}	…..	Y_{NM}

NM = Number of elements in the population.

N = Number of clusters in the population.

M = Number of elements in each clusters.

Y_{ij} = Value of j^{th} elements in the i^{th} cluster (i= 1,2,…,N. j = 1,2,…..,M.)

$\bar{Y}_i. = \frac{1}{M}\sum_{j=1}^{M} Y_{ij}$ = Mean of the i^{th} cluster.

$\bar{Y}_{NM} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} Y_{ij} = \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i .$

= population mean.

$S_i^2 = \frac{1}{M-1}\sum_{j=1}^{M}(Y_{ij} - \bar{Y}_{NM})^2$

= mean square between elements in i[th] cluster.

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

= mean square between elements within the clusters in the population.

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

= Mean square between elements in the whole population.

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

= Mean square between clusters in the whole population.

Suppose a simple random sample of n cluster is drawn from N elements and all the elements of selected clusters are enumerated. Thus, we have a cluster sample of NM elements.

$$\bar{Y}_{nM} = \frac{1}{nM}\sum_i\sum_j y_{ij}$$

$$= \frac{1}{n}\sum_i \bar{Y}_i \quad = \text{Sample mean.}$$

$$\bar{s}_w^2 = \frac{1}{n}\sum_i s_i^2$$

$$= \frac{1}{n(M-1)}\sum_i\sum(Y_{ij} - \bar{Y}..)^2$$

= Mean square between elements within the cluster in the sample.

$$s_b^2 = \frac{1}{n-1}\sum_i (y_{ij} - \bar{Y}_{nM})^2$$

= Mean square between clusters in the sample.

$$s^2 = \frac{1}{nM-1}\sum_i\sum_j (y_{ij} - \bar{Y}_{nM})^2$$

= Mean square between elements in the whole sample.

*** Theorem:**-Prove that-

1) $E(\bar{Y}_{nM}) = \bar{Y}_{NM}$
2) $E(\bar{Y}_{nM}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2$

*** Proof:**-Let Q_i be the quantity associated with the i[th] unit of the population (i= 1,2,…,N)

Let $\overline{Q} = \frac{1}{N}\sum_{i=1}^{N} Q_i$ = population mean.

$$S_Q^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Q_i - \overline{Q}_N)^2$$

Suppose a simple random sample of size n is taken from the population of size N.

Let $\overline{Q}_n = \frac{1}{n}\sum_{i=1}^{n} Q_i$

$$S_Q^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Q_i - \overline{Q}_n)^2$$

For **SRSWOR** we have,

i) $E(\overline{Q}_n) = \overline{Q}_N$

ii) $V(\overline{Q}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) S_Q^2$

iii) $E(s_Q^2) = S_Q^2$

For cluster sampling, let us define $Q_i = Y_i$, $\overline{Q}_n = \overline{Y}_{nM}$, $\overline{Q}_N = \overline{Y}_{NM}$, $S_Q^2 = S_b^2$

Appling (i) and (ii) we get,

$$E(\overline{Y}_{nM}) = \overline{Y}_{NM}$$

$$V(\overline{Y}_{nM}) = \left(\frac{1}{n} - \frac{1}{M}\right) S_b^2, \quad \text{Hence proved.}$$

**** Efficiency of Cluster Sampling Compared To SRSWOR:-**

In **SRSWOR,** we have

$$V(\overline{Y}_{nM}) = \left(\frac{1}{nm} - \frac{1}{NM}\right) S^2$$

For cluster sampling, we have

$$V(\overline{Y}_{nM}) = \left(\frac{1}{n} - \frac{1}{M}\right) S_b^2$$

Efficiency of cluster sampling over SRSWOR is given by:

$$E = \frac{V(\overline{Y}_{nM})}{V_{ci}(\overline{Y}_{nM})} = \frac{1/M\left(\frac{1}{n} - \frac{1}{N}\right)S^2}{\left(\frac{1}{n} - \frac{1}{N}\right)S_b^2} = \frac{S^2}{MS_b^2}$$

* ALTERNATIVE FORMS OF EFFICIENCY:-

(1) We have,

$$(NM-1)s^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{ij}-\bar{Y}_{NM})^2$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{ij}-\bar{Y}_i+\bar{Y}_i-\bar{Y}_{NM})^2$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{ij}-\bar{Y}_i)^2+\sum_i\sum_j(\bar{Y}_i-\bar{Y}_{NM})^2+ \textit{ product term}$$

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

where $\rho = \dfrac{\sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{ij}-\bar{Y}_i)(Y_{ij}-\bar{Y}_i)\big/NM(M-1)}{\sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{ij}-\bar{Y}_{NM}{}^2)\ \big/NM}$

$$= \frac{\sum_i\sum_j(Y_{ij}-\bar{Y}_{NM})(Y_{ij}{}'-\bar{Y}_{NM})}{(M-1)\ (NM-1)S^2}$$

= Intra class (intra cluster) correlation co efficient between pairs of elements within clusters.

*** Proof:-** $\left[\sum_{j=1}^{M}(y_{ij}-\bar{Y}_{NM})\right]^2$

$$= \sum_{j=1}^{M}(y_{ij}-\bar{Y}_{NM})^2 + \sum_{j\neq j'=1}^{M}(y_{ij}-\bar{Y}_{NM})(y_{ij}-\bar{Y}_{NM})$$

i.e. $\sum_{i=1}^{N}\left[\sum_{j=1}^{M}(y_{ij}-\bar{Y}_{NM})\right]^2$

$$= \sum_i\sum_j(y_{ij}-\bar{Y}_{NM})^2+\sum_i\sum_{j\neq j'}(y_{ij}-\bar{Y}_{NM})(y_{ij'}-\bar{Y}_{NM})$$

$$= (NM-1)S^2+(NM-1)(M-1)\rho S^2$$

$$= (NM-1)S^2[1+(M-1)\rho] \qquad \ldots\ldots\ldots\ldots\ldots\ldots(\text{i})$$

$$\left[\sum_{j=1}^{M}(y_{ij}-\bar{Y}_{NM})\right]^2 = \left[M\bar{Y}_i.-M\bar{Y}_{NM}\right]^2$$

$$= M^2\left[\bar{Y}_i.-\bar{Y}_{NM}\right]^2$$

$$= M^2(N-1)S_b^2 \qquad \ldots\ldots\ldots\ldots\ldots\ldots(\text{ii})$$

From (i) and (ii) we get

$(NM-1)\ S^2\ [1+(M-1)\ \rho] = M^2\ (N-1)\ S_b^2$

$$\therefore \quad E = \frac{S^2}{MS_b^2} = \frac{M(N-1)}{(NM-1)[1+(M-1)\rho]}$$

if N is large

$$E \cong 1/1+(M-1)\rho$$

i.e. $$S^2 = \frac{N(M-1)}{NM-1} S_w^{-2} + \frac{M(N-1)}{NM-1} S_b^2$$

we have ,

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

$$= \frac{\frac{N(M-1)}{NM-1} S_w^{-2} + \frac{M(N-1)}{NM-1} S_b^2}{MS_b^2}$$

$$= \frac{N(M-1)}{M(NM-1)} \frac{S_w^{-2}}{S_b^2} + \frac{N-1}{NM-1}$$

$$E \cong \frac{M-1}{M} \frac{S_w^{-2}}{S_b^2} + \frac{1}{M}; \qquad \text{if N is large .}$$

*** ANOVA TABLE ***

Source of variation	Sum of squares	Degree of fraction	M.S.S.	E
Between cluster	$\sum_i \sum_j (Y_i. - \bar{Y}_{NM})^2$	N-1	MS_b^2	S^2/MS_b^2
Within cluster	$\sum_i \sum_j (Y_{ij} - \bar{Y}_i.)^2$	N(M-1)	S_w^{-2}	
Total	$\sum_i \sum_j (Y_{ij} - \bar{Y}_{NM})^2$	NM-1	S^2	

(2) Efficiency of cluster sampling over SRSWOR in forms of Intra-class correlation co-efficient is given by:

$$E \cong \frac{1}{1+(M+1)\rho}$$

**** ESTIMATION OF EFFICIENCY:-**

A estimate of the efficiency of cluster sampling over SRSWOR is given by:

$$\hat{E} = \frac{M-1}{M} \frac{s_w^{-2}}{MS_b^2} + \frac{1}{M} \qquad \text{if N is large.}$$

Proof:- We have to show that

$$E(s_w^{-2}) = S_w^{-2} \quad and \quad E(s_b^2) = S_b^2$$

Let Q_i be the quantity associated with the i^{th} unit of the population i = 1,2,…,N.

Let $\overline{Q}_N = \frac{1}{N}\sum_{i=1}^{N} Q_i$

$$S_Q^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Q_i - \overline{Q}_n)^2$$

Suppose a simple random sample of n unit is drawn from N unit.

Let $\overline{Q}_n = \frac{1}{n}\sum_{i=1}^{n} Q_i$

$$s_Q^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Q_i - \overline{Q}_n)^2$$

For SRSWOR we have,

i) $E(\overline{Q}_n) = \overline{Q}_N$

ii) $V(\overline{Q}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) S_Q^2$

iii) $E(s_Q^2) = S_Q^2$

For cluster sampling let us define.

$$Q_i = S_i^2, \quad \overline{Q}_n = \frac{1}{n}\sum_{i=1}^{n} S_i^2 = s_w^{-2}, \quad \overline{Q}_N = S_w^2,$$

Appling (i) we get,

$$E(s_w^{-2}) = S_w^{-2}$$

Again for cluster sampling let us define

$$Q_i = \overline{Y}_i, \ \overline{Q}_n = \overline{Y}_{nm}, \quad \overline{Q}_N = \overline{Y}_{NM}, \quad s_Q^2 = s_b^2, \quad S_Q^2 = S_b^2$$

Appling (i) we have

$$E(s_b^2) = S_b^2$$

** OPTIMUM VALUE OF n AND M:-

Faifield Smith (1938).

$S_b^2 = \frac{S^2}{M}$ is some constant less than 1.

Mahalnibis (1940) and Jessen (1942)

$S_w^{-2} = a.M^b$ a and b are constant greater then zero.

We have $\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \overline{Y}_{NM})^2 = \sum_i\sum_j(y_{ij} - \overline{Y}_i.)^2 + \sum_i\sum_j(\overline{Y}_i. - \overline{Y}_{NM})^2$

i.e. $(NM-1)\,S^2 = N(M-1)\,S_w^{-2} + M(N-1)S_b^2$

$$S_b^2 = \frac{(NM-1)S^2 N(M-1)S_w^{-2}}{M(N-1)}$$

$$\cong S^2 - S_w^{-2}\frac{(M-1)}{M} \qquad \text{if N is large}$$

$$\cong S^2 - a(M-1)(M)^{b-1}\ ; \qquad S_w^{-2} = a.M^b$$

$$V(\bar{Y}_{nM}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2$$

$$\cong \frac{1}{n} S_b^2 \qquad \text{for large N}$$

$$\cong \frac{1}{n}\left[S^2 - a(M-1)M^{b-1}\right] \cong V$$

let the cost function bc

$c = c_1 nM + c_2\sqrt{n}$

where, c_1 = Cost for colleting information per element

c_2 = Cost proportional to unit distance between cluster.

Our problem is to determine the optimum value of n and M, so that variance is minimized for fixed cost c_0 say.

We minimize,

$$\phi = \frac{1}{n}\left[S^2 - a(M-1)M^{b-1}\right] + \mu\left[c_1 nM + c_2\sqrt{n}\right]$$

$$\frac{\partial\phi}{\partial n} = 0 \Rightarrow \frac{-1}{n^2}\left[S^2 - a(M-1)M^{b-1}\right] + \mu\left[c_1 M + \frac{c_2}{2\sqrt{n}}\right] = 0$$

$$\Rightarrow \frac{-V}{n} + \mu\left[c_1 M + \frac{c_2}{2\sqrt{n}}\right] = 0$$

$$\Rightarrow \mu = V/\ n\left[c_1 M + c_2/2\sqrt{n}\right] \qquad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

$$\frac{\partial\phi}{\partial M} = 0 \Rightarrow \frac{\partial V}{\partial M} + \mu a n = 0$$

$$\Rightarrow \frac{\partial V}{\partial M} = -\mu a n$$

$$= -\frac{V\ c_1\ n}{n\left[c_1 M + c_2/2\sqrt{n}\right]} \qquad \text{(using (1))}$$

$$\Rightarrow \frac{1}{V}\frac{\partial V}{\partial M} = -\frac{-c_1}{c_1 M + c_2/2\sqrt{n}} \qquad \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

Solving $c_1 nM + c_2\sqrt{n} - c_0 = 0$ for $\sqrt{n}$

$$\sqrt{n} = \frac{-c_2 \pm \sqrt{\Delta}}{2a\,M}\ , \qquad \Delta = c_2^2 + 4c_1 c_0 M$$

$$= \frac{-c_2 + \sqrt{\Delta}}{2a\,M} \qquad (\because \sqrt{n} \text{ being positive})$$

now, $$c_1 M + \frac{c_2}{2\sqrt{n}} = c_1 M + \frac{c_2\, 2c_1 M}{2(-c_2 + \sqrt{\Delta})}$$

$$= c_1 M + \frac{c_2 c_1 M}{-c_2 + \sqrt{\Delta}}$$

$$= \frac{c_1 M \sqrt{\Delta}}{-c_2 + \sqrt{\Delta}}$$

(2) reduce to i : $$\frac{1}{V}\frac{\partial V}{\partial M} = \frac{-c_1[-c_2 + \sqrt{\Delta}]}{aM\sqrt{\Delta}}$$

$$\Rightarrow \frac{M}{V}\frac{\partial V}{\partial M} = \frac{-c_2 + \sqrt{\Delta}}{\sqrt{\Delta}}$$

$$\Rightarrow \frac{M}{V}\frac{\partial V}{\partial M} = \frac{c_2 - 1}{\sqrt{\Delta}}$$

$$\Rightarrow \frac{M}{V}\frac{\partial V}{\partial M} = \frac{c_2}{\sqrt{c_2^2 + 4c_1 c_0 M}} - 1$$

$$\Rightarrow \frac{M}{V}\frac{\partial V}{\partial M} = \frac{c_2}{\sqrt{c_2^2\left(\frac{1 + 4c_1 c_0 M}{c_2^2}\right)}} - 1$$

$$= \frac{1}{\sqrt{\frac{1 + 4c_1 c_0 M}{c_2^2}}} - 1$$

$$\therefore \frac{M}{V}\frac{\partial V}{\partial M} = -1 + \left[\frac{1 + 4c_1 c_0 M}{c_2^2}\right]^{-1/2} \qquad \ldots\ldots\ldots\ldots\ldots\ldots(3)$$

$$\frac{\partial V}{\partial M} = \frac{\partial}{\partial M}\left[\frac{1}{n}\left\{S^2 - a(M-1)M^{b-1}\right\}\right]$$

$$= \frac{1}{n}\left[0 - \left\{aM^{b-1} + a(M-1)(b-1)M^{b-2}\right\}\right]$$

$$= \frac{1}{n}\left[-aM^{b-1} - a(M-1)(b-1)M^{b-2}\right]$$

$$\Rightarrow M\frac{\partial V}{\partial M} = \frac{1}{n}\left[-aM^{b} - a(M-1)(b-1)M^{b-1}\right]$$

$$\Rightarrow \frac{M}{v}\frac{\partial V}{\partial M} = \frac{1/n\left[-aM^{b} - a(M-1)(b-1)M^{b-1}\right]}{1/n\left[S^2 - a(M-1)M^{b-1}\right]}$$

$$= \frac{\left[-aM^{b} - a(M-1)(b-1)M^{b-1}\right]}{\left[S^{2} - a(M-1)M^{b-1}\right]} \quad \ldots\ldots\ldots\ldots\ldots\ldots(4)$$

From (3) and (4) we find

$$-1+\left[\frac{1+4c_1c_0M}{c_2^2}\right]^{-1/2} = \frac{-aM^{b} - a(M-1)(b-1)M^{b-1}}{S^{2} - a(M-1)M^{b-1}}$$

This equation is to be solved by mal and error method. It will give the optimum value of M. Optimum value M is called $\hat{M}$ say.

Then the optimum value of n, $\hat{n}$ say is obtained by,

$$\sqrt{n} = \frac{-c_2 + \sqrt{c_2^2 + 4c_1c_0\hat{M}}}{2a\hat{M}}$$

For cluster sampling obtain for minimize variable cost is fixed.

** DOUBLE SAMPLING **
OR
** SUB SAMPLING **

Question:-Why double sampling is used or necessary ?

Answer:-When the information auxiliary variate is available, we have seen how it could be utilized to obtain the more efficient estimator.

e.g. Ratio and regression methods estimators.

If the information on auxiliary variable x is not available, we select an initial sample from the population and information on x is collected. Then we consider a second sample from the initial sample and collect the information on y. This procedure of first selecting an initial sample and then a second sample from the initial sample is known as double sampling or sub-sampling.

* Double Sampling for PPS Sampling. :-

Suppose we wish to estimate the population mean $\bar{Y}$ using probability proportioned to size with replacement (PPSWR) scheme. Where the size being the value of some auxiliary variate x. Suppose the information on auxiliary variate x is not available. Then, we select an initial sample of size n' using SRSWOR and collect the information on x. Then, we select a second sample of size n from the initial sample using probability proportional to size with replacement scheme.

Then, $\hat{\bar{Y}} = \frac{1}{n}\sum z_i, \qquad z_i = \frac{Y_i}{np_i}, \qquad p_i = \frac{x_i}{x'}$

Is an unbiased estimators of $\bar{Y}$.. with variance is given that,

$$Var\ (\hat{\bar{Y}}) = \frac{1}{N(N-1)}\ V_p(Y)\ \frac{n'-1}{nn'} + \frac{N-n'}{n'N} S_y^2$$

where $V_P(Y) = \sum_{i=1}^{N} \frac{x_i}{X}\left(\frac{Y_i}{x_i/X} - Y\right)^2$

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

$$x' = \sum_{i=1}^{n'} x_i$$

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

$$X = \sum_{i=1}^{N} X_i$$

Now, $E(\hat{\bar{Y}}) = E\left[\frac{1}{n}\sum_{i=1}^{n} z_i\right]$

$$= E_1E_2\left[\frac{1}{n}\sum_{i=1}^{n} z_i\right]$$

$E_2 \Rightarrow$ second sample.

$E_1 \Rightarrow$ initial (first) sample.

When E_2 denotes the conditional expectation over second sample when initial sample is fixed.

$$E(\hat{\bar{Y}}) = E_1(\hat{\bar{y}}) \qquad E_1E_2\left[\frac{1}{n}\sum_{i=1}^{n} z_i\right]$$

$$= \bar{Y} \qquad E_1\left[E_2\ z_i\right] = E_1[\bar{y}'] = \bar{Y}$$

Next $Var\ (\hat{\bar{Y}}) = E_1 V_2(\hat{\bar{y}}) + V_1E_2(\hat{\bar{y}})$

Consider $V_1E_2[\hat{\bar{y}}] = V_1[\bar{Y}']$

$$= \frac{N-n'}{n'N} S_y^2 \qquad \text{(i)}$$

Further the conditional variance of $\hat{\bar{Y}}$ when the initial sample fixed is given by,

$$V_z(\hat{\bar{Y}}) = \frac{1}{n}\sum_{i=1}^{n'} \frac{x_i}{x'}\left(\frac{Y_i}{n'x_i/x'} - Y^{-1}\right)^2$$

$$= \frac{1}{nn'^2}\sum_{i=1}^{n'}\frac{x_i}{x'}\left(\frac{Y_i x'}{x_i} - Y'\right)^2 \qquad (\because n'Y^{-1} = Y')$$

$$= \frac{1}{nn'^2}\sum_{i=1}^{n'} x_i x'\left(\frac{Y_i}{x_i} - \frac{Y'}{x'}\right)^2$$

$$= \frac{1}{nn'^2}\sum_{i=1}^{n'}\sum_{i<j}^{n'} x_i x_j\left(\frac{Y_i}{x_i} - \frac{Y_j}{x_j}\right)^2$$

$$= \frac{1}{nn'^2}\sum_{i=1}^{N}\sum_{i<j}^{N} a_i a_j\ x_i x_j\left(\frac{Y_i}{x_i} - \frac{Y_j}{x_j}\right)^2$$

Where $a_j = 1$ if the i^{th} unit is included in the sample.

$a_i = 0$ if the i^{th} unit is not included in the sample.

$$E(a_i) = \frac{n'}{N}$$

$$E(a_i, a_j) = \frac{n'(n'-1)}{N(N-1)}$$

for a SRSWOR of size n'

$$E_1V_2[\hat{\bar{Y}}] = 1\left[\frac{1}{nn'^2}\sum_{i=1}^{N}\sum_{i\neq j}^{N} a_i a_j\ x_i x_j\left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2\right]$$

$$= \frac{1}{nn'^2}\frac{n'(n'-1)}{N(N-1)}\sum_{i=1}^{N}\sum_{i\neq j}^{N} x_i x_j\left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2$$

$$= \frac{1}{N(N-1)}\frac{n'-1}{nn'}\sum_{i=1}^{N} x_i X\left(\frac{y_i}{x_i} - \frac{Y}{X}\right)^2$$

$$= \frac{1}{N(N-1)}\frac{n'-1}{nn'}\sum_{i=1}^{N}\frac{x_i}{X}\left(\frac{y_i X}{x_i} - Y\right)^2$$

$$= \frac{1}{N(N-1)}\frac{n'-1}{nn'}\sum_{i=1}^{N}\frac{x_i}{X}\left(\frac{y_i}{x_i/X} - Y\right)^2$$

$$= \frac{1}{N(N-1)}\frac{n'-1}{nn'}V_p(Y) \qquad \text{(ii)}$$

Substitution (i) and (ii)

$Var\ (\hat{\bar{Y}})$ becomes;

$$Var\ (\hat{\bar{Y}}) = \frac{1}{N(N-1)}\frac{n'-1}{nn'}V_p(Y) + \frac{N-n'}{n'N}S_y^2$$

** DOUBLE SAMPLING FOR UNBIASED RATIO ESTIMATOR:-

Suppose we wish to estimate the population mean $\bar{Y}$ using ratio method of estimation. Suppose the information on auxiliary variable x is not available.

Select a SRSWOR of size n' from the population and collect the information on y and x

Consider the estimator as $\hat{\bar{Y}} = \frac{\bar{Y}}{\bar{x}}\, x'$

Where $\bar{Y}(\bar{x})$ = mean of the second sample of size n of variate Y(x)

x' = mean of the initial sample of size n' of variate x.

now, $E(\hat{\bar{Y}}) = E\left[\frac{\bar{Y}}{\bar{x}}\, \bar{x}'\right]$

$= E_1 E_2\left[\frac{\bar{Y}}{\bar{x}}\, \bar{x}'\right]$

Where E_2 denotes the conditional expectation over second sample when the initial sample is fixed.

$= E_1\left[\sum^{(n'_n)} prob.\frac{\bar{Y}}{\bar{x}}\, \bar{x}'\right]$

The probability of selecting the second sample for the fixed initial sample fixed is given by $\frac{\bar{x}}{\binom{n'}{n}\bar{x}'}$

$p_i = \frac{x_i}{X} = \frac{x_i}{\Sigma x_i}$ probability of choosing n from n' (n'_n)

so. $x^{-1} = \frac{1}{(n'_n)}\sum_I^{(n'_n)} \bar{x}$

$$E[\hat{\bar{Y}}] = E_1\left[\sum_I^{\binom{n'}{n}} \frac{\bar{x}}{\binom{n'}{n}(x^{-1})}\, \frac{\bar{y}}{x} x^{-1}\right]$$

$$= E_1\left[\frac{I}{\binom{n'}{n}}\sum_I^{\binom{n'}{n}} \bar{y}\right]$$

$$= E_1\left[Y^{-1}\right]$$

$$= \bar{Y}$$

Next $Var\ (\hat{\bar{Y}}) = E\left[\hat{\bar{Y}}^2\right] - \left(E\left[\hat{\bar{Y}}\right]\right)^2$

$$E\left[\hat{\bar{Y}}^2\right] = E\left[\frac{\bar{Y}^2}{\bar{x}^2}\ \bar{x}'^2\right]$$

$$= E_1 E_2\left[\frac{\bar{Y}^2}{\bar{x}^2}\ \bar{x}'^2\right]$$

$$= E_1\left[\sum_I^{\binom{n'}{n}} prob.\frac{\bar{Y}^2}{\bar{x}^2}\ \bar{x}'^2\right]$$

$$= E_1\left[\sum_I^{\binom{n'}{n}} \frac{\bar{x}}{\binom{n'}{n}x^{-1}}\frac{\bar{Y}^2}{\bar{x}^2}\ \bar{x}'^2\right]$$

$$= \frac{1}{\binom{n'}{n}} E_1\left[x^{-1}\sum_I^{\binom{n'}{n}}\frac{\bar{Y}^2}{\bar{x}^2}\right]$$

$$= \frac{1}{\binom{n'}{n}}\left[\sum_I^{\binom{N}{n'}}\frac{1}{\binom{N}{n'}}\left\{x^{-1}\sum_I^{\binom{n'}{n}}\frac{\bar{Y}^2}{\bar{x}^2}\right\}\right]$$

$$= \frac{1}{\binom{N}{n'}\binom{n'}{n}}\left[\sum_I^{\binom{N}{n'}} x^{-1}\left\{\sum_I^{\binom{n'}{n}}\frac{\bar{Y}^2}{\bar{x}^2}\right\}\right]$$

There fore variance becomes,

$$Var[\hat{\bar{Y}}] = \frac{1}{\binom{N}{n'}\binom{n'}{n}}\left[\sum^{\binom{N}{n'}} x^{-1}\left\{\sum_I^{\binom{n'}{n}}\frac{\bar{Y}^2}{\bar{x}^2}\right\}\right] - \bar{Y}^2$$

** Non-Sampling Errors:-

The theory of sampling scheme is assume that,

(i) Some probability sampling scheme is used.

(ii) The observation on the i^{th} unit of the population i.e. Y_i is the correct value.

Then the error in the estimate is surely due to random sampling this error is known as sampling error. In general, these are other type of errors, these errors are due to measurement feabalation, editing, etc. These errors are known as non-sampling errors. The main sources of non-sampling errors are,

(i) Lake of precision in reporting observation.

(ii) Incomplete coverage of the sample.

(iii) Faulty method of estimation is used.

When a complete count is made there is no sampling error but there will be non-sampling errors.

As the sample size increases the sampling error will tend to decrease and non-sampling errors will tend to increase.

When a sample is selected both the type of errors will remain present.