Varying Probability sampling

In simple random sampling each and every unit of the population has an equal chance of being selected in the sample.

Whenever units vary in size, simple random sampling is not an appropriate method as no importance is given to the size of the units. Such ancillary information about the size of the units can be utilized in selecting the sample so as to get more efficient estimator of the population parameter. One such method is to assign unequal probabilities of selection to different units in the population depending on their sizes.

e.g. Ten orchards were having 125, 793, 970, 830, 1502, 864, 503, 106, 970, 312 fruit trees respectively. With orchards having varying numbers of fruit trees, it may be desirable to provide a sampling scheme in which orchards are selected with probabilities proportional to the number of trees in the orchards. When units vary in their sizes and the variate under study is highly correlated with the size of unit, the probability of selection may be assigned in proportion to the size of the unit. This type of the sampling procedure where the probability of selection is proportional to the size of the unit is known as probability proportional to size sampling, abbreviated as pps sampling.

There is a basic difference between simple random sampling and PPS sampling procedures. In simple random sampling unit at any given draw is the same, while in PPS sampling it differ from draw to draw.

Procedures of selecting a sample

There are two methods of selection

(i) Cumulative Total Method
(ii) Lahiri's Method

Cumulative Total Method

In this method sampling frame is available in the form of a list of units or in the form of map of units with values of auxiliary variable X.

Let the size of the units u_i be X_i where i = 1, 2, ...N

Let $P_i = P(u_i)$ = probability of selecting unit u_i at any draw.

π_i = Probability of selecting u_i at 1^{st} draw or 2^{nd} draw or.....or n^{th} draw

$= P_i + P_i + ...+ P_i = n P_i$

T_i = cumulative total for unit u_i

Algorithm

(1) Find cumulative total for unit u_i where i = 1, 2,.....N
i.e $T_i = X_1 + X_2 +.....+ X_i$

(2) Select a number ' R ' at random from 1 to X where X = $\sum_{i=1}^{N} Xi = T_N$

(3) Select unit u_i in the sample if $T_{i-1} < R \leq T_i$

Repeat the above procedure till the desired number of sample units are selected.

Note:- Consider $P_i = P(u_i)$ = probability of selecting unit u_i at any draw

$$= \frac{Ti - Ti-1}{TN} = \frac{Xi}{X}$$

i.e $P_i = \frac{Xi}{X}$

Thus probability of selecting unit u_i is proportional to seze X_i

Example :-

Draw a pps sample with replacement of size 2 from population of size 5

	u_1	u_2	u_3	u_4	u_5
Size (X_i)	10	25	35	46	20
Cumulative total (T_i)	10	35	70	116	136
$T_{i-1} < R \leq T_i$	1 – 10	11 – 35	36 – 70	71 – 116	117 - 136

Now select two numbers successively with replacement at random from random number table or other media. If selected number lies between 1 and 136, select the unit in which this number falls otherwise go to next number. Repeat the procedure till all required sample size is selected.

Let the first number from random table is 452 which is equivalent to 44 (452 -136 -136 -136 =44). This number lies in unit u_3 therefore u_3 is selected.

Let the next number is 33 which lies in unit u_2 therefore u_2 is selected.

Therefore u_2 and u_3 is selected in the sample

Lahiri's Method

Lahiri suggested an alternative procedure in which cumulations are avoided completely.

Step : 1 Select a number at random between 1 and N (say i is selected)

i.e $1 \leq i \leq N$, we select u_i provisionally

Step : 2 Select a number at random between 1 and M where M = Maxi. of size X_i

i.e $1 \leq R \leq M$

Step : 3 Select unit u_i finally if $R \leq X_i$ otherwise reject it.

Repeat the procedure till required number of units are selected.

Example :-

Select a pps sample with replacement of size 2 from population of size 5

Unit	u_1	u_2	u_3	u_4	u_5
Size (X_i)	10	23	15	22	80

Step : 1 Select a number at random say i such that $1 \leq i \leq 5$
Let a drawn number is i = 4 therefore unit u_4 is selected provisionally

Step : 2 Select a number R from 1 to M i.e from 1 to 80. Let the selected number is R = 80 but R is not less than or equal to u_4 = 22 therefore u_4 is not selected finally.

Repeat the procedure

Select i =2 => u_2 is selected provisionally.

Let another number between 1 and 80 is selected. Let it be R = 12

Since $R \leq u_2 = 23$ therefore u_2 is selected finally.

Draw another random number between 1 and 5.

Let it be i = 5 => u_5 is selected provisionally.

Draw another random number between 1 and 80. Let it be R = 40 which is < 80

Therefore u_5 is selected finally.

Therefore u_2 and u_5 selected in the sample by Lahiri's method.

Theorem :- 1. The probability of selecting the i^{th} unit in the first effective draw is $\frac{Xi}{X}$ in Lahiri's method of pps sampling.

Proof :-$P_1(u_i)$ = probability that unit u_i is selected at 1st trial

= (prob. of selecting unit provisionally) x (prob. of selecting unit finally)

$= \frac{1}{N} \times \frac{Xi}{M} = \frac{Xi}{NM}$

P(r) = Probability that no unit is selected at any draw

$= \frac{1}{N}(1 - \frac{X_1}{M}) + \frac{1}{N}(1 - \frac{X_2}{M}) + \frac{1}{N}(1 - \frac{X_3}{M}) + - - - - - + \frac{1}{N}(1 - \frac{X_N}{M})$

$= \frac{1}{N}(\frac{M-X_1}{M}) + \frac{1}{N}(\frac{M-X_2}{M}) + \frac{1}{N}(\frac{M-X_3}{M}) + - - - - - - + \frac{1}{N}(\frac{M-X_N}{M})$

$= \frac{1}{NM}\sum_{i=1}^{N}(M - X_i) = \frac{1}{NM}(NM - \sum_{i=1}^{N} X_i)$

$= 1 - \frac{\sum_{i=1}^{N} X_i}{NM} = 1 - \frac{\bar{X}}{M} = q$

$P_2(u_i)$ = probability that unit u_i is selected at 2nd trials

= (prob. of not selecting on 1st trial) x (selecting at 2nd trial)

$= \left(1 - \frac{\bar{X}}{M}\right)\frac{X_i}{NM}$

Similarly $P_3(u_i) = \left(1 - \frac{\bar{X}}{M}\right)^2 \frac{X_i}{NM}$

Therefore probability of selecting unit u_i at any draw

$= \frac{X_i}{NM} + \left(1 - \frac{\bar{X}}{M}\right)\frac{X_i}{NM} + \left(1 - \frac{\bar{X}}{M}\right)^2 \frac{X_i}{NM} + \left(1 - \frac{\bar{X}}{M}\right)^3 \frac{X_i}{NM} +$ ---------

$= \frac{X_i}{NM}(1 + q + q^2 + q^3 +$ ---------) where $q = 1 - \frac{\bar{X}}{M}$

This is geometric series

$= \frac{X_i}{NM} \cdot \frac{1}{1-q} = \frac{X_i}{NM} \cdot \frac{1}{1-\left(1-\frac{\bar{X}}{M}\right)} = \frac{X_i}{NM} \cdot \frac{M}{\bar{X}} = \frac{X_i}{X}$

Therefore P(selecting unit u_i at any draw) = $\frac{X_i}{X}$

Theorem :- 2 Let Y_i be the Y value of the unit drawn in the i^{th} draw and P_i be the corresponding selection probability, i = 1, 2, ----n, then an unbiased estimator for population total is

$$\hat{Y}_{pps} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i} \text{ and its variance is}$$

$$V(\hat{Y}_{pps}) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{Y_i}{P_i} - Y\right)^2 P_i$$

Proof :- Note that the ratio $\frac{Y_i}{P_i}$ can take any one of the N values $\frac{Y_j}{P_j}$, j = 1, 2, --- N with respective probabilities P_j

Therefore $E\left(\frac{Y_i}{P_i}\right) = \sum_{j=1}^{N}\frac{Y_j}{P_j}P_j = Y$

Therefore $E(\hat{Y}_{pps}) = E\left(\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}\right) = \frac{1}{n}\sum_{i=1}^{n}E(\frac{Y_i}{P_i}) = \frac{1}{n}\sum_{i=1}^{n}Y = \frac{nY}{n} = Y$

Thereforee $\hat{Y}_{pps}$ is unbiased for the population total under ppswr

Since draws are independent

$$V(\hat{Y}_{pps}) = V\left(\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}\right) = \frac{1}{n^2}\sum_{i=1}^{n}V\left(\frac{Y_i}{P_i}\right) \quad \text{-----} \quad (1)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}E\left(\frac{Y_i}{P_i} - E(\frac{Y_i}{P_i})\right)^2$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}E\left(\frac{Y_i}{P_i} - Y\right)^2 \text{ ---------- (2)}$$

Note that $\left(\frac{Y_i}{P_i} - Y\right)^2$ can take any one of the N values $\left(\frac{Y_j}{P_j} - Y\right)^2$ with respective probabilities P_j where j = 1, 2, ------ N

Therefore $E\left(\frac{Y_i}{P_i} - Y\right)^2 = \sum_{j=1}^{N}\left(\frac{Y_j}{P_j} - Y\right)^2 P_j$ ------- (3)

Substitute (3) in (2) , we get

$$V(\hat{Y}_{pps}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{N}\left(\frac{Y_j}{P_j} - Y\right)^2 P_j$$

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

***Remark** : The estimator $_{pps}$ is called the Hansen-Hurtwiz (HH) estimator and the strategy

(ppswr, $_{pps}$),the HH strategy.

Theorem :- 3 An unbiased estimator of $V(\hat{Y}_{pps})$ is $\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{Y_i}{P_i}-Y\right)^2$

Proof :- By theorem 2 $V(\hat{Y}_{pps}) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{Y_i}{P_i}-Y\right)^2 P_i$

$$= \frac{1}{n}\left(\sum_{i=1}^{N}\frac{Y_i^2}{P_i}-Y^2\right) \text{ ---------- (1)}$$

Also $E(V(\hat{Y}_{pps})) = V(\hat{Y}_{pps})$ - - - - - - - - - - - - - - - - - (2)

(Because E(C) = C)

$$= E(\hat{Y}_{pps}^2) - Y^2$$

Hence $Y^2 = E(\hat{Y}_{pps}^2) - E(V(\hat{Y}_{pps}))$

$$= E\left(\hat{Y}_{pps}^2 - V(\hat{Y}pps)\right) \text{ - - - - - - - - - - - - - - - - - - - (3)}$$

Also $\sum_{i=1}^{N}\frac{Y_i^2}{P_i} = E\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i^2}{P_i^2}\right)$ - - - - - - - - - - - - - - - - (4)

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

$Explation\ of\ (4)$

Using (2), (3) and (4) in (1) we get

$$E(V(\hat{\boldsymbol{Y}}_{pps})) = \frac{1}{n}\left(\sum_{i=1}^{N}\frac{Y_i^2}{P_i}-\boldsymbol{Y^2}\right)$$

$$= \frac{1}{n}\left(\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i^2}{P_i^2}\right) - \mathbf{E}\left(\hat{\boldsymbol{Y}}_{pps}^2 - \boldsymbol{V}(\hat{\boldsymbol{Y}}\mathbf{pps})\right)\right)$$

$$= \frac{1}{n}\,E\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i^2}{P_i^2}\right) - \frac{1}{n}E(\hat{Y}_{pps}^2) + \frac{1}{n}V(\hat{Y}pps)$$

$$E(V(\hat{\boldsymbol{Y}}_{pps})) - \frac{1}{n}E(V(\hat{\boldsymbol{Y}}_{pps})) = \frac{1}{n^2}E\left(\sum_{i=1}^{n}\frac{Y_i^2}{P_i^2}\right) - \frac{1}{n}\mathbf{E}(\hat{\boldsymbol{Y}}_{pps}^2)$$

$$\frac{\mathbf{n-1}}{\mathbf{n}}\left(\mathbf{E}(\mathbf{V}(\hat{\boldsymbol{Y}}\mathbf{pps}))\right) = \frac{1}{\mathbf{n}}\left(\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i}{P_i}-\hat{\boldsymbol{Y}}\mathbf{pps}\right)^2\right)\right)$$

$\therefore$ E(V($\widehat{\boldsymbol{Y}}_{\text{pps}}$)) = $\mathrm{E}\left(\frac{\mathbf{1}}{\boldsymbol{n(n-1)}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\left(\frac{\boldsymbol{Y_i}}{\boldsymbol{P_i}}-\widehat{\boldsymbol{Y}}\mathbf{pps}\right)^{\mathbf{2}}\right)$

Since E($\widehat{\boldsymbol{Y}}_{\text{pps}}$) = Y

$\therefore$ E(V($\widehat{\boldsymbol{Y}}_{\text{pps}}$)) = $\mathrm{E}\left(\frac{\mathbf{1}}{\boldsymbol{n(n-1)}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\left(\frac{\boldsymbol{Y_i}}{\boldsymbol{P_i}}-\boldsymbol{Y}\right)^{\mathbf{2}}\right)$

$\therefore$ $\mathbf{V(\widehat{Y}_{pps})} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{Y_i}{P_i}-Y\right)^2$

Theorem ;- 4 An unbiased estimator of the gain due to ppswr sampling as compared to srswr is $\frac{1}{\mathrm{n}^2}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\frac{\mathrm{Y_i^2}}{\mathrm{P_i}}\right)\left(\mathrm{N}-\frac{1}{\mathrm{P_i}}\right)$

Proof :- We know that under srswr.

$\mathrm{V}(\hat{Y})\text{srswr} = \mathrm{N}^2\frac{\sigma^2}{n}$

$= \mathrm{N}^2\frac{(N-1)}{Nn}\mathrm{S}^2 \qquad (Because\ \ (N-1)S^2 = N\sigma^2)$

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

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

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

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

Note that under ppswr, unbiased estimator of the quantities $\sum_{i=1}^{N}Y_i^2$ and Y^2 are $\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{\mathrm{Y_i^2}}{\mathrm{P_i}}$ and $\left(\hat{Y}_{pps}^2 - V(\hat{Y}\text{pps})\right)$ respectively. Put this in (1)

$\mathrm{V_{pps}}(\hat{Y})\text{srswr} = \frac{1}{\mathrm{n}}\left(\frac{N}{n}\sum_{i=1}^{n}\frac{Y_i^2}{P_i} - \left(\hat{Y}_{pps}^2 - V(\hat{Y}\text{pps})\right)\right)$

$= \frac{1}{n^2}\left(N\sum_{i=1}^{n}\frac{Y_i^2}{P_i} - n\hat{Y}_{pps}^2 + nV(\hat{Y}\text{pps})\right)$

$= \frac{1}{n^2}\left(N\sum_{i=1}^{n}\frac{Y_i^2}{P_i} - n\hat{\mathrm{Y}}_{\text{pps}}^2\right) + \frac{1}{\mathrm{n}}V(\hat{Y}\text{pps})$ - - - - - - - (2)

Already we have seen in Theorem 3, an unbiased estimator of $V(\hat{Y}\text{pps})$ is

$= \frac{1}{\mathrm{n(n-1)}}\sum_{i=1}^{n}\left(\frac{Y_i}{P_i} - \hat{Y}\text{pps}\right)^2$ - - - - - - - - - - - - - - - (3)

Subtract (3) from (2) , we get

$$\frac{1}{n^2}\left(N\sum_{i=1}^{n}\frac{Y_i^2}{P_i}-n\hat{Y}_{pps}^2\right)+\frac{1}{\mathrm{n}}V(\hat{Y}\mathrm{pps})-\frac{1}{\mathrm{n(n-1)}}\sum_{i=1}^{n}\left(\frac{Y_i}{P_i}-\hat{Y}\mathrm{pps}\right)^2$$

$$\frac{1}{n^2}\left(N\sum_{i=1}^{n}\frac{Y_i^2}{P_i}-n\hat{Y}_{pps}^2\right)+\frac{1}{\mathrm{n}}V(\hat{Y}\mathrm{pps})-\frac{1}{\mathrm{n(n-1)}}\sum_{i=1}^{n}\frac{Y_i^2}{P_i^2}+\frac{1}{(\mathrm{n-1})}\hat{\mathrm{Y}}_{\mathrm{pps}}^2$$

$$\left(\frac{\mathrm{N}}{\mathrm{n}^2}\sum_{\mathrm{i=1}}^{\mathrm{n}}\frac{\mathrm{Y_i^2}}{\mathrm{P_i}}-\frac{1}{\mathrm{n(n-1)}}\sum_{i=1}^{n}\frac{Y_i^2}{P_i^2}\right)-\frac{1}{\mathrm{n}}\cancel{\hat{\mathrm{Y}}_{\mathrm{pps}}^2}+\frac{1}{(\mathrm{n-1})}\cancel{\hat{\mathrm{Y}}_{\mathrm{pps}}^2}+\frac{1}{\mathrm{n}}V(\hat{Y}\mathrm{pps})$$

(for large n $\cong$ n-1)

$$\frac{1}{n^2}\left(N\sum_{\mathrm{i=1}}^{\mathrm{n}}\frac{\mathrm{Y_i^2}}{\mathrm{P_i}}-\frac{\cancel{n}}{(\mathrm{n-1})}\sum_{i=1}^{n}\frac{Y_i^2}{P_iP_i}\right)+\frac{1}{\mathrm{n}}V(\hat{Y}\mathrm{pps})$$

$$\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{Y_i^2}{P_i}\left(N-\frac{1}{\mathrm{P_i}}\right)\right)+\frac{1}{\mathrm{n}}V(\hat{Y}\mathrm{pps})$$

$\therefore$ Gain due to pps as

$$\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{Y_i^2}{P_i}\left(N-\frac{1}{\mathrm{P_i}}\right)\right)$$

PPS Sampling Without Replacement

Since the effective sample size in case of without replacement sampling is expected to be larger compared to sampling with replacement. The sampling without replacement can provide a more efficient estimator than sampling with replacement. A lot of work in the field of sampling with varying probabilities without replacement has been done, but most of the procedures are complex and not easily applicable in large scale surveys. If the sampling fraction is small in large scale survey, the efficiency of sampling with or without replacement will differ insignificantly. However, if the sampling fraction is larger, it is expected that the gain in efficiency due to pps sampling wor will be substantial.

Procedure of selection of a pps sample without replacement

There are several procedures for selecting samples with unequal probabilities wor.

1. General Selection Procedure

Suppose 'n' units are selected one by one with probability proportional to size measure x, at each draw without replacing the units selected in the previous draws. The probability of selection at the first draw for the j^{th} unit is given by

$P_j = \frac{X_j}{X}$, j = 1, 2, - - - - - - N where $X = \sum_{j=1}^{N} X_j$

Similarly the probability that the i^{th} unit is selected at the second draw when j^{th} unit has been selected at the first draw is given by

$P_{i|j} = \frac{P_i}{1-P_j}$, i = j and so on

This set up of sampling gives an ordered set of sample values $y_1, y_2, - - - y_n$ with probalities $P_1 + P_2 + - - - - - - - p_n$

2. **Sen – Midzuno Method.**

Under this system of selection probabilities due to Midzuno (1952), which consists in selecting the first unit with pps and the remaining (n-1) units from N-1 units of the population by simple random sampling wor.

Under this scheme

Π_i = Probability that i^{th} unit is included in the sample.

= P_i + Probability that i^{th} unit is not selected at the ith draw and is selected at any subsequent draw (n-1) draws.

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

$= \left(\frac{N-n}{N-1}\right)P_i + \frac{n-1}{N-1}$ for i = 1, 2, ----------N

π_{ij} = Probability that both y_i and y_j are included in the sample.

= Probability that i^{th} unit selected at 1^{st} draw and j^{th} at any (n-1) draws + Probability that j^{th} unit is selected at 1^{st} draw and i^{th} unit at any (n-1) draws + Probability that i and j^{th} unit not selected at first draw and both selected at subsequent (n-1) draws.

$= P_i \frac{n-1}{N-1} + P_j \frac{n-1}{N-1} + (1- P_i - P_j)\frac{n-1}{N-1}\frac{n-2}{N-2}$

$= \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(P_i + P_j) + \frac{n-2}{N-2}\right]$

By extension of the above argument, we can have y_i, y_j ---------y_q a sample of

$\pi_{ij\_\_\_\_q} = \frac{1}{\left(\binom{N-1}{n-1}\right)}(P_i + P_j + \_\_\_\_\_ P_q)$

Thus we have derived the first and second order inclusion probabilities under Midzuno sampling scheme. These expressions can be used in the Horvitz-Thompson estimator to estimate the population total unbiasedly and derive the variance of the estimator.

3. **Desraj Ordered Estimator**

$t_1 = \frac{Y_1}{P_1}$

$t_2 = Y_1 + \frac{Y_2}{P_2}(1-P_1)$

$t_3 = Y_1 + Y_2 + \frac{Y_3}{P_3}(1-P_1-P_2)$

.

.

$t_n = Y_1 + Y_2 + - - - Y_{n-1} + \frac{Y_n}{P_n}(1-P_1-P_2- \ldots .P_{n-1})$

The Desraj ordered estimator for the population total is defined as

$\hat{Y}_{DR} = \frac{1}{n}\sum_{i=1}^{n} t_i$

Theorem:- Under ppswor, $\hat{Y}_{DR}$ is unbiased for the population total and unbiased estimator of $V(\hat{Y}_{DR})$ is $\frac{1}{n(n-1)}\sum_{i=1}^{n}(t_i - \bar{t})^2$

Proof:- Note that the ratio $\frac{Y_1}{P_1}$ can take any one of N values $\frac{Y_j}{P_j}$ with respective probabilities P_j , j = 1, 2, ----N

Therefore $E\left(\frac{Y_1}{P_1}\right) = \sum_{j=1}^{N}\frac{Y_j}{P_j}P_j = \sum_{j=1}^{N} Y_j = Y$

i.e $E(t_1) = Y$

Hence t_1 is unbiased for the population Y

Let $t_2 = Y_1 + \frac{Y_2}{P_2}(1-P_1)$

$E_2\left(\frac{Y_2}{P_2}(1-P1)\middle|Y1\right) = \sum_{j=2}^{N}\frac{Y_j}{P_j}(1-P_1)\frac{P_j}{1-P_1} = \sum_{j=2}^{N} Y_j = Y - Y_1$

$\therefore E(t_2) = E_1E_2(t_2|t_1) = E_1\left(Y_1 + E_2\left(\frac{Y_2}{P_2}(1-P_1)\middle|Y_1\right)\right) = Y_1 + Y - Y_1 = Y$

$\therefore E(\hat{Y}_{DR}) = \frac{1}{2}\sum_{i=1}^{2} t_i = \frac{2Y}{2} = Y$

Similarly

$E(t_r) = E_1\left(Y_1 + Y_2 + - - - + Y_{r-1} + E_2\left(\frac{Y_r}{P_r}(1 - P_1 - - - P_{r-1})\middle|Y_1, Y_2, - - Y_{r-1}\right)\right)$

Where

$E_2\left(\frac{Y_r}{P_r}(1 - P_1 - - - P_{r-1})\middle|Y_1, Y_2, - - Y_{r-1}\right) = \sum_{j=r}^{N}\frac{Y_j}{P_j}(1 - P_1 - - - P_{r-1})\frac{P_j}{1-P_1 - - P_{r-1}}$

$= \sum_{j=r}^{N} Y_j = Y - Y_1 - - - - Y_{r-1}$

$E(t_r) = Y_1 + Y_2 + - - - + Y_{r-1} + Y - Y_1 - Y_2 - - - - Y_{r-1} = Y$

$\therefore E(\hat{Y}_{DR}) = \frac{1}{n}\sum_{i=1}^{n} E(t_r) = \frac{nY}{n} = Y$

Hence $\hat{Y}_{DR}$ is unbiased for the population total Y

We know that

$V(\hat{Y}_{DR}) = E(\hat{Y}_{DR}^2) - \left(E(\hat{Y}_{DR})\right)^2$

$= E(\hat{Y}_{DR}^2) - Y^2$ - - - - - - - - (1)

Consider $E(t_r t_s) = E_1E_2(t_r t_s | i_1, i_2, - - i_{s-1})$ (assuming that r < s)

$= E_1[t_r E_2(t_s | i_1, i_2, - - i_{s-1})]$

$= E_1[t_r Y] = YE[t_r] = Y.Y = Y^2$

Therefore $E\left(\frac{1}{n(n-1)}\sum_{r \neq s}^{n} t_r t_s\right) = Y^2$ - - - - - - - - - - - - - (2)

Substituting (2) in (1), we get

$V(\hat{Y}_{DR}) = E\left(\frac{1}{n}\sum_{i=1}^{n} t_i\right)^2 - E\left(\frac{1}{n(n-1)}\sum_{r \neq s}^{n} t_r t_s\right)$

$= E\left(\frac{1}{n^2}\sum_{r=1}^{n} t_r^2 + \frac{1}{n^2}\sum_{i=1}^{n} t_r t_s\right) - E\left(\frac{1}{n(n-1)}\sum_{r \neq s}^{n} t_r t_s\right)$

$= E\left(\left(\frac{1}{n^2}\sum_{r=1}^{n} t_r^2\right) + \left(\left(\frac{1}{n^2} - \frac{1}{n(n-1)}\right)\sum_{r \neq s}^{n} t_r t_s\right)\right)$

$= E\left(\left(\frac{1}{n^2}\sum_{r=1}^{n} t_r^2\right) - \left(\frac{1}{n^2(n-1)}\left(\left(\sum_{r=1}^{n} t_r\right)^2 - \sum_{r=1}^{n} t_r^2\right)\right)\right)$

$= E\left(\frac{1}{n(n-1)}\sum_{r=1}^{n} t_r^2 - \frac{1}{n-1}\bar{t}^2\right)$ where $\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$

$= \frac{1}{n-1}E\left(\frac{1}{n}\sum_{r=1}^{n}(t_r - \bar{t})^2\right)$

$= E\left(\frac{1}{n(n-1)}\sum_{r=1}^{n}(t_r - \bar{t})^2\right)$

$\therefore \quad V(\hat{Y}_{DR}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(t_i - \bar{t})^2$

Unordered Estimators

In the previous section we have considered the estimators which depends on the order of the units drawn in the sample. (eg. Des raj order estimator) Now we shall discuss such estimators which do not depend on the order in which the units are drawn within the sample. These estimators are generally called unordered estimators. Horvitz and Thompson (1952) and Murthy (1957) have shown that these estimators are more efficient than ordered estimators.

Horvitz – Thompson Estimator

Horvitz and Thompson suggested an estimator which is an unbiased estimator of the population total. Let us suppose that the initial probability of selection of the units U_i is P_i where $P_i = \frac{X_i}{X}$ for i = 1,2,-------N. The probability that units U_i is included in the sample would be given by

$\Pi_i = P_i + \sum_{i \neq j} \frac{P_j P_i}{1-P_j} = P_i[1 + \sum_{i \neq j} \frac{P_j}{1-P_j}]$

Further the probability that both the units U_i and U_j are included in the sample is

$\Pi_{ij} = P_i \frac{P_j}{1-P_i} + P_j \frac{P_i}{1-P_j} = P_i P_j [\frac{1}{1-P_i} + \frac{1}{1-P_j}]$

Suppose that Y_i be the value of the ith unit with Π_i the probability of inclusion in the sample. The Horvitz Thompson estimator is defined by

$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\Pi_i}$

Theorm:- In pps sampling wor, $\hat{Y}_{HT}$ is unbiased and its sampling variance is given by

$V_{HT}(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{(1-\Pi i)Y_i^2}{\Pi_i} + \sum_{i}^{N} \sum_{i \neq j}^{N} \frac{(\Pi ij - \Pi i \Pi j)}{\Pi i \Pi j} Y_i Y_j$ where Π_{ij} is the probability of inclusion of both the ith and jth unit in the sample.

Proof:- Let $\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\Pi_i}$

$\hat{Y}_{HT} = \sum_{i=1}^{N} a_i \frac{Y_i}{\Pi_i}$ where $a_i = \begin{cases} 1 & if\ ith\ unit\ is\ drawn \\ 0 & otherwise \end{cases}$

Obviously a_i is a random variable follows a binomial distribution with probability P_i .

Hence $E(a_i) = \Pi_i$

$V(a_i) = \Pi_i(1 - \Pi_i)$

$Cov(a_i, a_j) = \Pi ij - \Pi i \Pi j$

Now $E(\hat{Y}_{HT}) = E[\sum_{i=1}^{n} \frac{y_i}{\Pi_i}] = E[\sum_{i=1}^{N} a_i \frac{Y_i}{\Pi_i}]$

where $a_i = \begin{cases} 1 & if\ ith\ unit\ is\ drawn \\ 0 & otherwise \end{cases}$

$= \sum_{i=1}^{N} E(a_i)\frac{Y_i}{\Pi_i}$ where $E(a_i) = \sum a_i P(a_i) = 1.P(a_i) + 0.P(a_i) = \Pi_i$

$= \sum_{i=1}^{N} \Pi_i \frac{Y_i}{\Pi_i} = \sum_i^N Y_i = Y$ = Population total.

Hence $\hat{Y}_{HT}$ is an unbiased estimator of population total.

Further sampling variance is given by

$V_{HT}(\hat{Y}_{HT}) = V_{HT}(\sum_{i=1}^{n} \frac{y_i}{\Pi_i}) = V_{HT}(\sum_{i=1}^{N} a_i \frac{Y_i}{\Pi_i})$

$= \sum_i^N \frac{Y_i^2}{\Pi_i^2} V(a_i) + \sum_i^N \sum_{i\neq j}^N \frac{Y_i}{\Pi_i}\frac{Y_j}{\Pi_j}.Cov(a_i, a_j)$

$= \sum_i^N \Pi_i(1-\Pi_i).\frac{Y_i^2}{\Pi_i^2} + \sum_i^N \sum_{i\neq j}^N \frac{Y_i}{\Pi_i}\frac{Y_j}{\Pi_j}.\ \Pi ij - \Pi i \Pi j$

$= \sum_{i=1}^{N} \frac{(1-\Pi i)Y_i^2}{\Pi_i} + \sum_i^N \sum_{i\neq j}^N \frac{(\Pi ij - \Pi i \Pi j)}{\Pi i \Pi j} Y_i Y_j$

Murthy's unordered estimator

The Desraj ordered estimator depends upon the order in which the units are drawn. Murthy (1957) obtained the unordered estimator corresponding to Desraj ordered estimator. For the sake of simplicity, we shall restrict to sample of size 2 only. Suppose y_1 and y_2 are the values of the units selected in the first and second draws and p_1 and p_2 the corresponding initial selection probabilities.

The ordered estimator is

$\hat{y}_{DR}(1, 2) = \frac{1}{2}[\frac{y_1}{p_1} + y_1 + \frac{y_2}{p_2}(1-p_1)]$

$= \frac{1}{2}[\frac{y_1}{p_1}(1+p_1) + \frac{y_2}{p_2}(1-p_1)]$

On the other hand, if the same two units are drawn in the other order, the corresponding ordered estimator is given by

$\hat{y}_{DR}(2, 1) = \frac{1}{2}[\frac{y_2}{p_2}(1+p_2) + \frac{y_1}{p_1}(1-p_2)]$

Their corresponding selection probabilities are $p(1, 2) = p_1\frac{p_2}{1-p_1}$

and $p(2, 1) = p_2\frac{p_1}{1-p_2}$

The unordered estimator based on the ordered estimators $\hat{y}_{DR}(1, 2)$ and $\hat{y}_{DR}(2, 1)$ is given by

$\hat{y}_M = \frac{[\hat{y}_{DR}(1,2)p(1,2) + \hat{y}_{DR}(2,1)p(2,1)]}{[p(1,2)+p(2,1)]}$

$= \frac{[\frac{y_1}{p_1}(1-p_2) + \frac{y_2}{p_2}(1-p_1)]}{2-p_1-p_2}$

An unbiased estimator of $V(\hat{y}_M)$ is

$\frac{[(1-p_1-p_2)(1-p_1)(1-p_2)}{(2-p_1-p_2)^2}\ [\frac{y_1}{p_1} - \frac{y_2}{p_2}]^2$

Corollary:- Midzuno sampling design is one in which the Yates-Grundy estimator of variance is non-negative.

Proof:- Now we know that Yates-Grundy estimator of variance of Horvitz Thompson estimator is

$$\widehat{V}_{YG}(\widehat{Y}_{HT}) = \sum_i^n \sum_{j>i}^n \left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2$$

Where $\pi_i\pi_j - \pi_{ij} = \left(\frac{N-n}{N-1}P_i + \frac{n-1}{N-1}\right)\left(\frac{N-n}{N-1}P_j + \frac{n-1}{N-1}\right) - \frac{n-1}{N-1}\left(\frac{N-n}{N-2}(P_i + P_j) + \frac{n-2}{N-2}\right)$

$$= \left(\frac{N-n}{N-1}\right)^2 P_iP_j + \left(\frac{N-n}{N-1}\right)\left(\frac{n-1}{N-1}\right)P_i + \left(\frac{n-1}{N-1}\right)\left(\frac{N-n}{N-1}\right)P_j + \left(\frac{n-1}{N-1}\right)^2 - \left(\frac{n-1}{N-1}\right)\left(\frac{N-n}{N-2}\right)(P_i + P_j) - \left(\frac{n-1}{N-1}\right)\left(\frac{n-2}{N-2}\right)$$

$$= \left(\frac{N-n}{N-1}\right)^2 P_iP_j + \left(\frac{n-1}{N-1}\right)\left(\frac{N-n}{N-1}\right)(P_i + P_j) - \left(\frac{n-1}{N-1}\right)\left(\frac{N-n}{N-2}\right)(P_i + P_j) + \left(\frac{n-1}{N-1}\right)^2 - \left(\frac{n-1}{N-1}\right)\left(\frac{n-2}{N-2}\right)$$

$$= \left(\frac{n-1}{N-1}\right)\left[\left(\left(\frac{N-n}{N-1}\right) - \left(\frac{N-n}{N-2}\right)\right)(P_i + P_j)\right] + \left(\frac{N-n}{N-1}\right)^2 P_iP_j + \left(\frac{n-1}{N-1}\right)\left[\left(\frac{n-1}{N-1}\right) - \left(\frac{n-2}{N-2}\right)\right]$$

$$= \left(\frac{n-1}{N-1}\right)\left[\frac{(N-n)(-1)}{(N-1)(N-2)}\right](P_i + P_j) + \left(\frac{N-n}{N-1}\right)^2 P_iP_j + \left(\frac{n-1}{N-1}\right)\left[\frac{(N-n)}{(N-1)(N-2)}\right]$$

$$= \frac{(N-n)(n-1)}{(N-1)^2(N-2)}(1 - P_i - P_j) + \frac{(N-n)^2}{(N-1)^2}P_iP_j$$

$$\therefore \widehat{V}_{YG}(\widehat{Y}_{HT}) = \frac{(N-n)}{(N-1)^2}\left[\frac{\sum_i^n \sum_{i<j}^n (N-n)P_iP_j + \frac{(n-1)}{(N-2)}(1 - P_i - P_j)}{\pi_{ij}}\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2\right]$$

$$\textit{Which is} > 0 \ \textit{except when} \ \frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} = 0$$

RANDOM GROUP METHOD:

The **RANDOM GROUP METHOD** is due to *Rao, Hartley and Cochran* (1962).This method makes use of the size information and always yields sample containing distinct units. In this method, the population is randomly divided into 'n' mutually exclusive & exhaustive groups of sizes N_1, ….., N_n and one unit is drawn from each group with probability proportional to size of the units in that group. Here the group sizes N_1, ….., N_n are predetermined constants.

An unbiased estimator of the population total is $\hat{Y}_{RHC} = \sum_{i=1}^{n} \frac{Y_i}{P_i'}$;

Where Y_i is the Y-value drawn from the i[th] random group, and
P_i' is the selection probability of the unit drawn from the i[th] random group.

Let Y_{ij} and X_{ij} be the Y and X value of the j[th] unit in the i[th] random group w.r.t a given partition. Then Y_i can take any one of the N_i values Y_{ij}. j=1,2,…N_i ; and P_i' can take any one of the N_i values $\frac{X_{ij}}{\sum_{j=1}^{N_i} X_{ij}}$.

j=1,…N_i & i=1,…n.

Thm: The estimator $\hat{Y}_{RHC} = \sum_{i=1}^{n} \frac{Y_i}{P_i'}$ is unbiased for the population total **Y**.

Proof: $E[\hat{Y}_{RHC}] = E_1E_2[\hat{Y}_{RHC} | G_1, \dots, G_n]$ ----------(1)
where E_2 is the conditional expectation taken w.r.t. a given partitioning of the population and E_1 is the overall expectation.

Note that $E_2[\hat{Y}_{RHC} | G_1, \dots, G_n] = E_2\left[\sum_{i=1}^{n} \frac{Y_i}{P_i'} | G_1, \dots, G_n\right]$

$$= \sum_{i=1}^{n} E_2\left[\frac{Y_i}{P_i'} | G_i\right] \quad \text{----------(2)}$$

Since the ratio $\frac{Y_i}{P_i'}$ can take any one of the N_i values $Y_{ij} \Big/ \left[\frac{X_{ij}}{\sum_{j=1}^{N_i} X_{ij}}\right]$; j=1, … N_i.

With resp. probabilities $\frac{X_{ij}}{\sum_{j=1}^{N_i} X_{ij}}$.

Thus, $E_2\left[\frac{Y_i}{P_i'} | G_i\right] = \sum_{j=1}^{N_i} \frac{Y_{ij}}{\left(\frac{X_{ij}}{\sum_{j=1}^{N_i} X_{ij}}\right)} \cdot \left(\frac{X_{ij}}{\sum_{j=1}^{N_i} X_{ij}}\right) = \sum_{j=1}^{N_i} Y_{ij}$ ----------(3)

Substituting (3) in (2), we get;
$E_2[\hat{Y}_{RHC} | G_1, \dots, G_n] = \sum_{i=1}^{n} \sum_{j=1}^{N_i} Y_{ij} = Y$
Therefore, by (1); $\hat{Y}_{RHC}$ is unbiased for the population total under Random Group Method.

RESULT: $\sum_{i=1}^{N} \left[\frac{Y_i}{P_i} - Y\right]^2 . P_i = \sum_{i<j}^{N} \left[\frac{Y_i}{P_i} - \frac{Y_j}{P_j}\right]^2 P_iP_j$

PROOF: Note that,

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \left[\frac{Y_i}{P_i} - \frac{Y_j}{P_j}\right]^2 P_iP_j = \sum_{i=j}^{N} \left[\frac{Y_i}{P_i} - \frac{Y_j}{P_j}\right]^2 . {P_i}^2 + 2\sum_{i<j}^{N} \left[\frac{Y_i}{P_i} - \frac{Y_j}{P_j}\right]^2 P_iP_j \quad \text{------(1)}$$

$$\left[\begin{array}{c} This\ follows\ from\ the\ identity \\ \sum_{i=n}^{n} \sum_{j=1}^{n} a_{ij} = \sum_{i=1}^{n} a_{ii} + 2\sum_{i<j}^{n} a_{ij}\ ;\quad if\ a_{ij} = a_{ji} \end{array}\right]$$

Now (1) can be written as,

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{{Y_i}^2}{P_i}.P_j+\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{{Y_j}^2}{P_j}.P_i-2\sum_{i=1}^{N}\sum_{j=1}^{N}Y_iY_j=2\sum_{i<j}^{N}\left[\frac{Y_i}{P_i}-\frac{Y_j}{P_j}\right]^2P_iP_j$$

$$\left(\because\quad\sum_{i=1}^{N}\left[\frac{Y_i}{P_i}-\frac{Y_j}{P_j}\right]^2{P_i}^2=0\right)$$

- LHS:

$$=\sum_{i=1}^{N}\frac{{Y_i}^2}{P_i}\sum_{j=1}^{N}P_j+\sum_{j=1}^{N}\frac{{Y_j}^2}{P_j}\sum_{i=1}^{N}P_i-2(Y_1+Y_2+\ \ldots+Y_N)^2$$

$$\left[\because\sum_{i=1}^{2}\sum_{j=1}^{2}Y_iY_j=\ {Y_1}^2+{Y_2}^2+Y_1Y_2+Y_2Y_1\right.$$
$$\left.=(Y_1+Y_2)^2=\left(\sum_{i=1}^{2}Y_i\right)^2=Y^2\right]$$

$$=\sum_{i=1}^{N}\frac{{Y_i}^2}{P_i}+\ \sum_{j=1}^{N}\frac{{Y_j}^2}{P_j}-2\,Y^2$$

$$=2\ \sum_{i=1}^{N}\frac{{Y_i}^2}{P_i}-2\,Y^2$$

$$\therefore\ 2\ \sum_{i=1}^{N}\frac{{Y_i}^2}{P_i}-2\,Y^2\ =2\sum_{i<j}^{N}\left[\frac{Y_i}{P_i}-\frac{Y_j}{P_j}\right]^2P_iP_j$$

$$\therefore\ \sum_{i=1}^{N}\left[\frac{Y_i}{P_i}-Y\right]^2.P_i=\ \sum_{i<j}^{N}\left[\frac{Y_i}{P_i}-\frac{Y_j}{P_j}\right]^2P_iP_j$$

Hence the proof.

Thm: The variance of the estimator $\hat{Y}_{RHC}$ is $\left\{\sum_{i=1}^{n}\frac{N_i(N_i-1)}{N(N-1)}\right\}\left\{\sum_{j=1}^{N}\left[\frac{Y_j}{P_j}-Y\right]^2.P_j\right\}$

Proof: $V[\hat{Y}_{RHC}]= E_1V_2\,[\hat{Y}_{RHC}\,|G_1,\ldots,G_n]+ V_1E_2\,[\hat{Y}_{RHC}\,|G_1,\ldots,G_n]$

We have seen in the prev. thm

$$E_2\,[\hat{Y}_{RHC}\,|G_1,\ldots,G_n]= Y \qquad \text{-----(1)}$$

Thus,

$$V_1E_2\,[\hat{Y}_{RHC}\,|G_1,\ldots,G_n]= V_1(Y)=0 \qquad \text{----(2)}$$

Since one unit is drawn from each group independently,

$$V_2[\hat{Y}_{RHC}]= V_2\,[\sum_{i=1}^{n}\frac{Y_i}{P_i'}\,|G_1,\ldots,G_n]=\sum_{i=1}^{n}V_2\left[\left(\frac{Y_i}{P_i'}\right)\,|G_i\right]$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{N_i}\left[\frac{Y_{ij}}{P_{ij}}-E\left(\frac{Y_{ij}}{P_{ij}}\right)\right]^2.P_{ij} \qquad \text{----(3)}$$

Where, $P_{ij}=\frac{X_{ij}}{\sum_{k=1}^{N_i}X_{ik}}$

The RHS of above expression is obtained by taking n=1

Thus, $\sum_{j=1}^{N}\left[\frac{Y_j}{P_j}-E\left(\frac{Y_j}{P_j}\right)\right]^2.P_j=\sum_{j=1}^{N}\left[\frac{Y_j}{P_j}-Y\right]^2.P_j\ \begin{pmatrix}from\ above\ thm,\\ E\left[\frac{Y_j}{P_j}\right]=Y\end{pmatrix}$

$$=\sum_{j<k}^{N}\left[\frac{Y_j}{P_j}-\frac{Y_k}{P_k}\right]^2P_jP_k\ \begin{pmatrix}which\ has\ been\ proved\\ in\ prev.result\end{pmatrix}$$

Thus, $$V_2[\hat{Y}_{RHC}]=\sum_{i=1}^{n}\sum_{j<k}^{N}\left[\frac{Y_j}{P_j}-\frac{Y_k}{P_k}\right]^2P_jP_k. \qquad \text{-----(4)}$$

Since $\frac{N_i(N_i-1)}{N(N-1)}$ is the probability in a random split that, a pair of observations fall into the i^{th} group, we have

$$E_1V_2\,[\hat{Y}_{RHC}]=\sum_{i=1}^{n}\frac{N_i(N_i-1)}{N(N-1)}\sum_{j<k}^{N}\left[\frac{Y_j}{P_j}-\frac{Y_k}{P_k}\right]^2P_jP_k$$

$$\therefore V(\hat{Y}_{RHC}) = \sum_{i=1}^{n} \frac{N_i(N_i-1)}{N(N-1)} \sum_{j=1}^{N} \left[\frac{Y_j}{P_j} - Y\right]^2 P_j \quad \text{------(5)}$$

REMARK: When the groups are all of the same size, then $N_1=N_2=$ ---- $=N_n=N/n$

Thus, $\sum_{i=1}^{n} N_i(N_i-1) = \frac{N(N-n)}{n}$

Substituting this in (5), we get

$V(\hat{Y}_{RHC}) = \left[\frac{N(N-n)}{nN(N-1)}\right] \sum_{j=1}^{N} \left[\frac{Y_j}{P_j} - Y\right]^2 P_j = \frac{N-n}{n(N-1)} . V(\widehat{Y_{pps}})$

i.e. $V(\hat{Y}_{RHC}) < V(\widehat{Y_{pps}})$

From this we infer that "Random Group Method" is better than "probability Proportional to Size with replacement" whenever the groups are of same size.

PPS SYSTEMATIC SAMPLING:

In sampling of 'n' units, the cumulative total $\{T_i\}$, i=1,2,---,N; are determined and the unit corresponding to the number {r+jk}, j=0,1,2,---,(n-1); is selected where, k= T/n = X/n.

R: the random no. from 1 to k.

This is known as PPS Systematic Sampling.

In PPS Systematic Sampling, unit u_i is included in the sample if

$T_{i-1} < r+jk \le T_i$; for some value of j.

J=0,1,2,---,(n-1).

Since random no. is selected from 1 to k and X_i is favorable to inclusion of the i^{th} unit in the sample.

So the prob. $\prod_i$ of inclusion of u_i is $X_i/k = n X_i/X$; provided $k>X_i$.

If X/n is not integer, the sampling interval k can be taken as the integer nearest to X/n and in this case actual sample size differs from sample to sample.

The sampling scheme is termed as PPS Linear Systematic Sampling (ppslss) when the random start is taken from 1 to k. And the sampling scheme is termed as PPS Circular Systematic Sampling (ppscss) when the random start is taken from 1 to X and sampling is done in circular manner. However if, X/n is an integer, these two procedures will be equivalent.

Let us assume that X/n is an integer. An unbiased estimator of the populatn total Y, derived by Hartley and Rao, is given by

$$\hat{Y}_{H-R} = \sum_{i=1}^{n} \frac{Y_i}{\prod_i} = k \sum_{i=1}^{n} \frac{Y_i}{x_i} \qquad \because \prod_i = \frac{x_i}{k}$$

$= \frac{X}{n} \sum_{i=1}^{n} \frac{Y_i}{x_i} = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i}{P_i}$

(For fairly large values of 'N' and for values of 'n' relatively small compared to 'N' the approx. sampling variance can be written as $V(\hat{Y}_{H-R}) \approx \frac{1}{n} \sum_{i=1}^{N} \left(\frac{Y_i}{P_i} - Y\right)^2 . P_i\,[1-(n-1)P_i]$

Further this variance can be estimated by

$V(\hat{Y}_{H-R}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{i<j}^{n} \left[1 - n(P_i + P_j) + n \sum_{j=1}^{N} P_j^2\right] \left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j}\right)^2$

Ex: Take sample of size 2 factories from the populatn of 10 factories given in the following table:

Factory	No. of workers	Number associated	Cum. total
1	58	1-58	58
2	908	59-966	966
3	418	967-1384	1384
4	442	1385-1826	1826
5	615	1827-2441	2441
6	1972	2442-4413	4413
7	613	4414-5026	5026
8	734	5027-5760	5760
9	514	5761-6274	6274
10	846	6275-7120	7120

n=2

Total size X=7120. So sampling interval k=X/n=7120/2=3560.

Now select random no. from 1 to 3560.

Let it be r=2142. So selected sample of size 2 is r,r+k ie. 2142, 2142+3560

ie. 2142,5702.

So 5th and 8th units are selected to form the sample.

STRATIFIED PPS SAMPLING:

Let $U_1,U_2,---,U_N$ be the populatn of N units divided into 'k' strata with i^{th} stratum having N_i unit so that $\sum_{i=1}^{k} N_i = N$.

Suppose a sample of size n_i unit is selected from N_i unit of the i^{th} stratum with ppswr size being X.

Y_{ij}= value of j^{th} unit in the i^{th} stratum in the populatn. ; where i=1,2,---,k.

J=1,2,---,N_i.

And $P_{ij} = \frac{X_{ij}}{X_i}$ = The prob. of selecting j^{th} unit in the i^{th} stratum in population.

y_{ij}= value of j^{th} unit in i^{th} stratum in sample.

P_{ij}= $\frac{x_{ij}}{x_i}$

An unbiased estimator of Y is given by $\hat{Y} = \sum_{i=1}^{k} \hat{Y}_i = \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{Y_{ij}}{P_{ij}}$

And $V(\hat{Y}) = \sum_{i=1}^{k} V(\hat{Y}_i) = \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{N_i} \left(\frac{Y_{ij}}{P_{ij}} - Y_i\right)^2 . P_{ij}$

$= \sum_{i=1}^{k} \frac{1}{n_i} \left(\sum_{j=1}^{N_i} \frac{{Y_{ij}}^2}{P_{ij}} - {Y_i}^2\right)$

Estimate of $V(\hat{Y})$ is, $\hat{V}(\hat{Y}) = \sum_{i=1}^{k} V(\hat{Y}_i) = \sum_{i=1}^{k} \frac{1}{n_i(n_i-1)} \sum_{j=1}^{n_i} \left(\frac{Y_{ij}}{P_{ij}} - \hat{Y}_i\right)^2$

For effective stratification, strata should be formed in such a way that units within each strata are homogeneous w.r.t. the variable $\frac{Y_{ij}}{P_{ij}}$ and not w.r.t. the variable y or x taken separately.

PROPORTIONAL ALLOCATION:

The sampling variance for proportional allocation can be obtained by allocating 'n' proportional to strata totals of the size measure.

i.e. $n_i \propto X_i$

$$n_i = c.X_i \qquad \text{----(1)}$$

$$\Sigma n_i = c.\Sigma X_i$$

$$n = cX$$

$$\therefore c = \frac{n}{X} \qquad \text{----(2)}$$

Put (2) in (1), we get;

$$n_i = \frac{nX_i}{X}$$

$$\hat{Y} = \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{Y_{ij}}{P_{ij}}$$

$$= \sum_{i=1}^{k} \frac{X}{n X_i} \sum_{j=1}^{n_i} \frac{Y_{ij}}{P_{ij}}$$

$$= \frac{X}{n} \sum_{i=1}^{k} \frac{1}{X_i} \sum_{j=1}^{n_i} \frac{Y_{ij}}{P_{ij}}$$

$$V(\hat{Y}) = \sum_{i=1}^{k} \frac{1}{n_i} \left(\sum_{j=1}^{N_i} \frac{{Y_{ij}}^2}{P_{ij}} - {Y_i}^2 \right)$$

$$= \sum_{i=1}^{k} \frac{X}{n X_i} \sum_{j=1}^{N_i} \left(\frac{Y_{ij}}{P_{ij}} - Y_i \right)^2 . P_{ij}$$

$$= \frac{X}{n} \sum_{i=1}^{k} \frac{1}{X_i} \sum_{j=1}^{N_i} \left(\frac{Y_{ij}}{P_{ij}} - Y_i \right)^2 . P_{ij}$$

Now, $P_{ij} = \frac{X_{ij}}{X_i}$

$${P_{ij}}^* = \frac{X_{ij}}{X} = \frac{X_{ij}}{X_i} . \frac{X_i}{X}$$

$$\therefore {P_{ij}}^* = P_{ij} . {P_i}^* \qquad \text{where, } {P_i}^* = \frac{X_i}{X}$$

$$V_{prop}(\hat{Y}_{st}) = \frac{1}{n} \sum_{i=1}^{k} \frac{1}{{P_i}^*} \sum_{j=1}^{N_i} \left(\frac{Y_{ij}}{P_{ij}} - Y_i \right)^2 . P_{ij} \qquad \text{---(A)}$$

$$= \frac{1}{n} \sum_{i=1}^{k} \frac{1}{{P_i}^*} \sum_{j=1}^{N_i} \left(\frac{Y_{ij}}{P_{ij}} - Y_i \right)^2 . \frac{{P_{ij}}^*}{{P_i}^*}$$

$$= \frac{1}{n} \sum_{i=1}^{k} \frac{1}{({P_i}^*)^2} \sum_{j=1}^{N_i} \left(\frac{Y_{ij}}{P_{ij}} - Y_i \right)^2 . {P_{ij}}^*$$

Consider,

$$V_{unst}(\hat{Y}_{st}) = \frac{1}{n} \sum_{j=1}^{N} \left(\frac{Y_j}{P_j} - Y \right)^2 . P_j$$

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

$$= \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{N_i} \left(\frac{Y_{ij}}{P_{ij} {P_i}^*} - Y \right)^2 . P_{ij} . {P_i}^*$$

$$= \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{N_i} \frac{1}{{P_i}^*} \left(\frac{Y_{ij}}{P_{ij}} - Y {P_i}^* \right)^2 . P_{ij}$$

$$= \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{N_i} \frac{1}{{P_i}^*} \left[\frac{Y_{ij}}{P_{ij}} - Y_i + Y_i - Y {P_i}^* \right]^2 . P_{ij}$$

$$= \frac{1}{n} \sum_{i=1}^{k} \frac{1}{{P_i}^*} \sum_{j=1}^{N_i} \left(\frac{Y_{ij}}{P_{ij}} - Y_i \right)^2 P_{ij} + \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{N_i} (Y_i - Y {P_i}^*)^2 P_{ij} + \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{N_i} 2 \left(\frac{Y_{ij}}{P_{ij}} - Y_i \right) (Y_i - Y {P_i}^*) . P_{ij}$$

Since covariance part is zero ($\because$ between strata it is indpt)

$$\therefore V_{pps}(\hat{Y}_{unst}) = V_{prop}(\hat{Y}_{st}) + \frac{1}{n}\sum_{i=1}^{k}\sum_{i=1}^{N_i}(Y_i - YP_i{}^*)^2\, P_{ij}$$

The last term of RHS is ≥ 0.

$$\therefore V_{pps}(\hat{Y}_{unst}) \geq V_{prop}(\hat{Y}_{st})$$

The last term is zero if $Y_i = Y.P_i{}^*$

$$Y_i = Y\frac{X_i}{X}$$

$$\therefore Y_i \propto X_i$$

<u>OPTIMUM ALLOCATION</u>:

$$n_i \propto \sqrt{V_i{}^*}$$

$$n_i = c.\sqrt{V_i{}^*} \quad \text{---(1)}$$

$$\therefore \sum_{i=1}^{k} n_i = c.\sum_{i=1}^{k}\sqrt{V_i{}^*}$$

$$n = c.\sum_{i=1}^{k}\sqrt{V_i{}^*}$$

$$\therefore n = c\sum_{i=1}^{k}\sqrt{V_i{}^*}$$

$$\therefore c = \frac{n}{\sum_{i=1}^{k}\sqrt{V_i{}^*}} \quad \text{---(2)}$$

Put (2) in (1), we get;

$$n_i = \frac{n\sqrt{V_i{}^*}}{\sum_{i=1}^{k}\sqrt{V_i{}^*}} \quad \text{---(3)}$$

Where, $V_i{}^* = \sum_{j=1}^{N_i}\left(\frac{Y_{ij}}{P_{ij}} - Y_i\right)^2.P_{ij}$

$$V_{opt}(\hat{Y}_{st}) = \sum_{i=1}^{k}\frac{1}{n_i}.\sum_{j=1}^{N_i}\left(\frac{Y_{ij}}{P_{ij}} - Y_i\right)^2.P_{ij}$$

$$= \sum_{i=1}^{k}\frac{1}{n_i}V_i{}^*$$

Put value of n_i from (3),

$$= \frac{1}{n}\sum_{i=1}^{k}\frac{\sum_{i=1}^{k}\sqrt{V_i{}^*}}{\sqrt{V_i{}^*}}.V_i{}^*$$

$$= \frac{1}{n}\left(\sum_{i=1}^{k}\sqrt{V_i{}^*}\right)^2 \quad \text{---(B)}$$

Comparing (A) and (B) we get;

$$V_{prop}(\hat{Y}_{st}) = \frac{1}{n}\sum_{i=1}^{k}\frac{V_i{}^*}{P_i{}^*} \quad (\because V_i{}^* = \sum_{j=1}^{N_i}\left(\frac{Y_{ij}}{P_{ij}} - Y_i\right)^2.P_{ij})$$

$$V_{opt}(\hat{Y}_{st}) = \frac{1}{n}\left(\sum_{i=1}^{k}\sqrt{V_i{}^*}\right)^2$$

$$\therefore V_{prop}(\hat{Y}_{st}) - V_{opt}(\hat{Y}_{st}) = \frac{1}{n}\sum_{i=1}^{k}\frac{V_i{}^*}{P_i{}^*} - \frac{1}{n}\left(\sum_{i=1}^{k}\sqrt{V_i{}^*}\right)^2$$

$$= \frac{1}{n}\left[\sum_{i=1}^{k}\frac{V_i{}^*}{P_i{}^*} - \left(\sum_{i=1}^{k}\sqrt{V_i{}^*}\right)^2\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{k}\sqrt{\frac{V_i{}^*}{P_i{}^*}}^2 - \left(\sum_{i=1}^{k}\sqrt{V_i{}^*}\right)^2\right]$$

Since $\sum_{i=1}^{k}\frac{V_i{}^*}{P_i{}^*} > \sum_{i=1}^{k}V_i{}^*$ $\quad (\because P_i{}^* \; is \; between \; 0 \; \& \; 1)$

$\therefore$ RHS ≥ 0.

$$\Rightarrow V_{prop}(\hat{Y}_{st}) - V_{opt}(\hat{Y}_{st}) \geq 0$$

$$\Rightarrow V_{prop}(\hat{Y}_{st}) \geq V_{opt}(\hat{Y}_{st})$$

RATIO ESTIMATION:

Let $\hat{Y}$ and $\hat{X}$ be unbiased for the population totals Y and X of the study and auxiliary variable resp. The ratio estimator of the populatn total is defined as $\hat{Y}_R = \left[\frac{\hat{Y}}{\hat{X}}\right].X$

For eg. if Y is the no. of teak trees in a geographical region and X is its area in acres, the ratio $\frac{\hat{Y}}{\hat{X}}$ is an estimator of the no. of teak trees per acre of a region in the population. The product $\frac{\hat{Y}}{\hat{X}}$ with X, the total area in acres; would provide an estimator of Y, the total no. of teak trees in the populatn.

The estimator proposed above is meant for any sampling design yielding unbiased estimators for the populatn totals Y & X.

Let P(s) be any such sampling design. It may be noted that,

$E(\hat{Y}_R) = \sum_{S\in\Omega} \hat{Y}_R . P(s)$

$= X\sum_{S\in\Omega} \frac{\hat{Y}}{\hat{X}} . P(s)$

Since the RHS of the above expression is, in general, not equal to Y; the ratio estimator is, in general, biased for Y under the given sampling design.

THM 1: The approximate bias and mean square error of the ratio estimator are $B(\hat{Y}_R) = Y\left\{\left[\frac{V(\hat{X})}{X^2}\right] - \left[\frac{Cov(\hat{X},\hat{Y})}{XY}\right]\right\}$ and $MSE[\hat{Y}_R] = Y^2\left\{\left[\frac{V(\hat{Y})}{Y^2}\right] + \left[\frac{V(\hat{X})}{X^2}\right] - 2\left[\frac{Cov(\hat{X},\hat{Y})}{XY}\right]\right\}$.

PROOF: Let $e_0 = \frac{\hat{Y}-Y}{Y}$ and $e_1 = \frac{\hat{X}-X}{X}$

It may be noted that e_0 & e_1 satisfy the following;

$E(e_0) = \frac{E(\hat{Y}-Y)}{Y} = 0$ ---(1)

$E(e_1) = \frac{E(\hat{X}-X)}{X} = 0$ ---(2)

$E(e_0^2) = \frac{1}{Y^2} E(\hat{Y}-Y)^2 = V(\hat{Y})/Y^2$ ---(3)

$E(e_1^2) = \frac{1}{X^2} E(\hat{X}-X)^2 = V(\hat{X})/X^2$ ---(4)

$E[e_0 e_1] = \frac{1}{XY} E[(\hat{X}-X)(\hat{Y}-Y)$

$= \frac{Cov(\hat{X},\hat{Y})}{XY}$ ---(5)

Assume that the sample size is large enough so that $|e_0|<1$ and $|e_1|<1$.

[This is equivalent to assuming that for all possible samples $0<\hat{X}<2X$ and $0<\hat{Y}<2Y$].

Since $\hat{Y}=Y(1+e_0)$ and $\hat{X}= X(1+e_1)$, the estimator $\hat{Y}_R$ can be written as,

$\hat{Y}_R = \left[\frac{\hat{Y}}{\hat{X}}\right].X = \frac{Y(1+e_0)}{X(1+e_1)} . X$

$=Y(1+e_0)(1+e_1)^{-1}$

$=Y(1+e_0)(1- e_1+ e_1^2- \text{---})$

$=Y[1+e_0- e_1+ e_1^2- e_0 e_1 \text{------}]$

$Y + Y[e_0- e_1+ e_1^2- e_0 e_1 \text{------}]$

$\hat{Y}_R - Y = Y[e_0- e_1+ e_1^2- e_0 e_1 \text{------}]$

$\therefore E(\hat{Y}_R - Y) = E[Y(e_1^2 - e_0 e_1)]$ (Using (1) & (2) and ignoring terms of degree greater than two)

$=Y[E(e_1^2) - E(e_0 e_1)]$

$= Y\left\{\left[\frac{V(\hat{X})}{X^2}\right] - \left[\frac{Cov(\hat{X},\hat{Y})}{XY}\right]\right\}$ (Using (4) & (5).)

$E(\hat{Y}_R - Y)^2 = Y^2 E[e_0^2 + e_1^2 - 2e_0 e_1]$

$=Y^2\left\{\left[\frac{V(\hat{Y})}{Y^2}\right] + \left[\frac{V(\hat{X})}{X^2}\right] - 2\left[\frac{Cov(\hat{X},\hat{Y})}{XY}\right]\right\}$

Hence, the proof.

COROLLARY: Under Simple Random Sampling:

(i) $\hat{Y}_R = \frac{\sum Y_i}{\sum X_i} . X$

(ii) $B(\hat{Y}_R) = \left[\frac{N^2(N-n)}{Nn}\right] Y\left\{\left[\frac{S_x^2}{X^2}\right] - \left[\frac{S_{xy}}{XY}\right]\right\}$

(iii) $MSE(\hat{Y}_R) = \left[\frac{N^2(N-n)}{Nn}\right] Y^2\left\{\left[\frac{S_y^2}{Y^2}\right] + \left[\frac{S_x^2}{X^2}\right] - 2\left[\frac{S_{xy}}{XY}\right]\right\}$

PROOF: We know that under SRS;

$$V(\hat{Y})= \left[\frac{N^2(N-n)}{Nn}\right] {S_y}^2$$

$$V(\hat{X})= \left[\frac{N^2(N-n)}{Nn}\right] {S_x}^2$$

$$Cov(\hat{X},\hat{Y})= \left[\frac{N^2(N-n)}{Nn}\right] S_{xy}$$

Substituting these expressions in the results available in Thm:1; we get the required expressions.

THM 2: The ratio estimator $\hat{Y}_R$ is more efficient than the expansion estimator $\hat{Y}$, if

$$\rho > \frac{1}{2}\left[\frac{C_x}{C_y}\right]$$

Where, $C_y = \frac{S_y}{\bar{Y}}$, $C_x = \frac{S_x}{\bar{X}}$ and ρ is the coefficient of correlation.

PROOF: $V(\hat{Y}) > MSE[\hat{Y}_R]$

$$\therefore \left[\frac{N^2(N-n)}{Nn}\right] {S_y}^2 > \left[\frac{N^2(N-n)}{Nn}\right] Y^2 \left\{\left[\frac{{S_y}^2}{Y^2}\right] + \left[\frac{{S_x}^2}{X^2}\right] - 2\left[\frac{S_{xy}}{XY}\right]\right\}$$

$$\therefore {S_y}^2 > {S_y}^2 + \left[\frac{Y^2}{X^2}\right] {S_x}^2 - 2\left[\frac{Y}{X}\right] S_{xy}$$

$$\therefore 2\, S_{xy} > \left[\frac{Y}{X}\right] {S_x}^2$$

$$\therefore S_{xy} > \frac{1}{2}\left[\frac{Y}{X}\right] {S_x}^2$$

$$\therefore \frac{S_{xy}}{S_x\, S_y} > \frac{1}{2}\left[\frac{Y}{X}\right]\frac{S_x}{S_y} \quad \Rightarrow \rho > \frac{1}{2}\left[\frac{S_x/X}{S_{y/Y}}\right]$$

$$\rho > \frac{1}{2}\left[\frac{C_x}{C_y}\right]$$

Hence, the proof.

ESTIMATED MEAN SQUARE ERROR UNDER SIMPLE RANDOM SAMPLING:

Note that,

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

$$= \sum_{i=1}^{N}[Y_i - \bar{Y} + R\bar{X} - RX_i]^2 (\because R=\bar{Y}/\bar{X})$$

$$= \sum_{i=1}^{N}(Y_i - \bar{Y})^2 + R^2 \sum_{i=1}^{N}(X_i - \bar{X})^2 - 2\,R\, \sum_{i=1}^{N}(Y_i - \bar{Y})\,(X_i - \bar{X})$$

$$(N-1)^{-1} \sum_{i=1}^{N}[Y_i - RX_i]^2 = {S_y}^2 + R^2 {S_x}^2 - 2\,RS_{xy} \qquad \text{---(1)}$$

Substituting this in the expression for the mean square error, we get an equivalent expression;

$$MSE(\hat{Y}_R) = \left[\frac{N^2(N-n)}{Nn}\right]\left\{{S_y}^2 + R^2 {S_x}^2 - 2\,RS_{xy}\right\}$$

$$= \frac{N^2(N-n)}{Nn(N-1)} \sum_{i=1}^{N}[Y_i - RX_i]^2$$

Where, $\hat{R} = \frac{\sum Y_i}{\sum X_i}$

The ratio estimator considered in this section is not unbiased for the population total (mean).

UNBIASED RATIO TYPE ESTIMATORS:

We have seen that under simple random sampling, the ratio estimator takes the form; $\left(\frac{\sum Y_i}{\sum X_i}\right) X$.

The unbiased ratio type estimator is the average of ratios $\frac{Y_i}{X_i}$. ie $\frac{1}{n}\sum_i \frac{Y_i}{X_i}$

$$\hat{Y}_R = \frac{X}{n}\sum_i \frac{Y_i}{X_i}$$

It is reasonable to take $\hat{Y}_{R_0} = \frac{N}{n}\sum_i \left(\frac{Y_i}{X_i}\right)\bar{X}$

Like the ratio estimator, the above estimator is also biased for the population total.

Thm: The bias of the estimator $\hat{Y}_{R_0}$ is B$[\hat{Y}_{R_0}]$= $-[N-1]S_{zx}$

Where S_{zx} = $\frac{1}{N-1}\sum_i [Z_i - \bar{Z}][X_i - \bar{X}]$, $Z_i = \frac{Y_i}{X_i}$

PROOF: Taking $Z_i = \frac{Y_i}{X_i}$, i=1,2,----,N; the estimator $\hat{Y}_{R_0}$ can be written as

$$\hat{Y}_{R_0} = \frac{N}{n}\sum_i Z_i \bar{X} = \hat{Z}\ \bar{X} \qquad \text{where, } Z = \sum_{i=1}^{N} Z_i \qquad \text{---(1)}$$

We know that, $Cov(\hat{Z},\hat{X}) = \left[\frac{N^2(N-n)}{Nn}\right] \frac{1}{N-1}\sum_{i=1}^{N}[Z_i - \bar{Z}][X_i - \bar{X}]$

$$= \frac{N^2(N-n)}{Nn}\ \frac{1}{N-1}\ \sum_{i=1}^{N}[Z_i X_i - N\,\bar{Z}\,\bar{X}]$$

$$= \frac{N^2(N-n)}{Nn(N-1)}\ [N\bar{Y} - N\bar{Z}\,\bar{X}] \qquad \text{Using } Z_i = \frac{Y_i}{X_i}$$

$$= \frac{-N^2(N-n)}{Nn(N-1)}[\hat{Z}\bar{X} - Y]$$

$$= \frac{-N^2(N-n)}{Nn(N-1)}\ \mathrm{B}[\hat{Y}_{R_0}] \qquad \text{Using (1).}$$

$\therefore$B$[\hat{Y}_{R_0}]$= $\frac{-Nn(N-1)}{N^2(N-n)}\,.\,Cov(\hat{Z},\hat{X})$ ---(2)

We know that under SRS; $Cov(\hat{X},\hat{Y})$= $\left[\frac{N^2(N-n)}{Nn}\right]S_{xy}$.

Making use of this in (2), we get

$$\mathrm{B}\left[\hat{Y}_{R_0}\right] = -(\mathrm{N}-1)\,S_{zx}$$

NOTE: This thm helps to get an unbiased estimator for the population total.

- Find unbiased estimator of the population total.

It is known that, $S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \hat{\bar{X}}\right)\left(Y_i - \hat{\bar{Y}}\right)$

is unbiased for $S_{XY} = \frac{1}{N-1}\ \sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$ under SRS.

Therefore, an unbiased estimator or the bias given in the above thm is

$$\mathrm{B}\left[\hat{Y}_{R_0}\right] = -(\mathrm{N}-1)\,S_{zx}$$

$$= -\frac{(N-1)}{(n-1)}\sum_i^n\left(Z_i - \hat{\bar{Z}}\right)\left(X_i - \hat{\bar{X}}\right)$$

$$= -\frac{(N-1)}{(n-1)}\sum_i^n(Z_iX_i) - n\hat{\bar{Z}}\hat{\bar{X}}$$

$$= -\frac{(N-1)}{(n-1)}\sum_i^n Y_i - n\hat{\bar{Z}}\hat{\bar{X}}$$

$$= -\frac{n(N-1)}{(n-1)}\left[\hat{\bar{Y}} - \hat{\bar{Z}}\hat{\bar{X}}\right]$$

It may be observed that, if 'b' is an unbiased estimator of the bias of the estimator 'T' (which is meant for estimating the parameter Θ) then '-b' is unbiased for the parameter Θ.

Therefore, $\hat{Y}_{R_0} - B\left[\hat{Y}_{R_0}\right]$ is an unbiased estimator of the population total.

ie $\hat{Y}_{R_0} + \frac{n(N-1)}{(n-1)}\left[\hat{\bar{Y}} - \hat{\bar{Z}}\hat{\bar{X}}\right]$ is unbiased for *Y*.

Thus we have obtained an exactly unbiased ratio-type estimator by considering the 'mean of the ratios' of Y_i to X_i(rather than the 'ratio' of Y_i to X_i, as in case of ratio estimator) to form the estimator and correcting for the bias. The above estimator is due to *HARTLEY & ROSS(1954).*

ALMOST UNBIASED RATIO ESTIMATOR:

Suppose a sample of size 'n' is drawn in the form of 'm' independent sub-samples Y_i the same size, selected according to the same sampling design and $\hat{Y}_\iota$ and $\hat{X}_\iota$, i= 1,2,---,m; are unbiased estimators of the population totals Y and X based on the 'm' subsamples. The following two estimates can be considered for Y:

$$\widehat{Y_1} = \left[\frac{\hat{Y}}{\hat{X}}\right]X \text{ , where} \hat{Y} = \frac{1}{m}\sum_{i=1}^m Y_i \text{ and } \hat{X} = \frac{1}{m}\sum_{i=1}^m X_i \qquad \text{---(1)}$$

$$\text{And } \widehat{Y_m} = \frac{1}{m}\sum_{i=1}^m\left[\frac{\hat{Y}_\iota}{\hat{X}_\iota}\right]X = \frac{1}{m}\sum_{i=1}^m r_i\ X \text{ where } r_i = \left[\frac{\hat{Y}_\iota}{\hat{X}_\iota}\right] \qquad \text{---(2)}$$

Under the usual assumptions, the bias of the estimator Y_1 is

$$B_1 = Y[R.V(\hat{X}) - Cov(\hat{X}, \hat{Y})]$$

$$= Y[R.V\left(\frac{1}{m}\sum_{i=1}^m \hat{X}_\iota\right) - Cov(\frac{1}{m}\sum_{i=1}^m \hat{X}_\iota\ , \frac{1}{m}\sum_{i=1}^m \hat{Y}_\iota)\ ]$$

$$= \frac{1}{m^2}Y\sum_{i=1}^m[R.V(\hat{X}_\iota) - Cov(\hat{X}_\iota, \hat{Y}_\iota)]$$

$$= \frac{1}{m^2}\sum_{i=1}^m B(r_i) \qquad \text{---(3)}$$

Where $B(r_i) = Y[R.V(\hat{X}_\iota) - Cov(\hat{X}_\iota, \hat{Y}_\iota)]$

And the bias of the estimator Y_m is $B_m = B(\widehat{Y_m}) = \frac{1}{m}\sum B(r_i)$ ---(4)

Comparing (3) & (4), we get;

$$mB_1 = B_m$$

This shows that the bias of the estimator Y_m is m times that of $\widehat{Y_1}$

Note that $B_m - B_1 = E\left[\widehat{Y_m} - Y\right] - E[\widehat{Y_1} - Y]$

$$= E\left[\widehat{Y_m} - \widehat{Y_1}\right]$$

Therefore $E\left[\widehat{Y_m} - \widehat{Y_1}\right] = (m-1)B_1$

Hence $\frac{\widehat{Y_m} - \widehat{Y_1}}{m-1}$ is an unbiased estimator of B_1.

Thus after correcting the estimator $\widehat{Y_1}$ for its bias we get an unbiased estimator for the population total.

$$\widehat{Y_{AV}} = \widehat{Y_1} - \left[\frac{\widehat{Y_m} - \widehat{Y_1}}{m-1}\right] = \frac{m\widehat{Y_1} - \widehat{Y_m}}{m-1}$$

Since the estimator given above is obtained by correcting only the approximate bias (not the exact bias), it is known as 'ALMOST UNBIASED RATIO-TYPE ESTIMATOR'.

BOUND FOR BIAS:

We know that the bias of the ratio estimator $\widehat{Y_R}$ is $B(\widehat{Y_R}) = E[\widehat{Y_R}] - Y$

We know that, $Cov(\widehat{Y_R}, \hat{X}) = E[\widehat{Y_R}\,\hat{X}] - E[\widehat{Y_R}]E[\hat{X}]$

$$= E\left[\frac{\hat{Y}}{\hat{X}}\, X\, \hat{X}\right] - E[\widehat{Y_R}]X$$

$$= XE[\hat{Y}] - XE[\widehat{Y_R}]$$

$$= X\,Y - X\,E[\widehat{Y_R}]$$

$$= X[Y - E(\widehat{Y_R})]$$

$$= -X\,B(\widehat{Y_R})$$

Therefore $Cor(\widehat{Y_R}, \hat{X}) = \frac{Cov(\widehat{Y_R}, \hat{X})}{SD(\widehat{Y_R}).SD(\hat{X})} = \frac{-X\,B(\widehat{Y_R})}{SD(\widehat{Y_R}).SD(\hat{X})}$

Since $\left|Cor(\widehat{Y_R}, \hat{X})\right| \leq 1$

$$\left|\frac{-X\,B(\widehat{Y_R})}{SD(\widehat{Y_R}).SD(\hat{X})}\right| \leq 1$$

$$\frac{X|B(\widehat{Y_R})|}{SD(\widehat{Y_R}).SD(\hat{X})} \leq 1$$

$$\frac{|B(\widehat{Y_R})|}{SD(\widehat{Y_R})} \leq \frac{SD(\hat{X})}{X}$$

The above is due to *HARTLEY & ROSS(1954).*

PRODUCT ESTIMATION:

We have proved under SRS, the ratio estimator is more precise than the expansion estimator, when the variables X and Y have high positive correlation. In fact it is not difficult to see under any sampling design, $\widehat{Y_R}$ is more efficient than $\hat{Y}$ if $\rho(\hat{X}, \hat{Y}) > \frac{1}{2}\frac{C(\hat{X})}{C(\hat{Y})}$

Where $C(\hat{X}) = \frac{SD(\hat{X})}{X}$ and $C(\hat{Y}) = \frac{SD(\hat{Y})}{Y}$

This shows that if correlation between X & Y is negative, the ratio estimator will not be more precise than the conventional estimator.

For such situations, *MURTHY(1964)* suggested another method of estimation, which is expected to be more efficient than $\hat{Y}$ in situations where $\widehat{Y_R}$ is less efficient than $\hat{Y}$.

In this method termed **'PRODUCT METHOD OF ESTIMATION'** the population total is estimated by using the estimator $\widehat{Y_p} = \left[\frac{\hat{Y}}{X}\right]\hat{X}$

Since the estimator uses the product $\hat{Y}\hat{X}$ rather than the ratio $\frac{\hat{Y}}{X}$ it is known as PRODUCT ESTIMATOR.

Thm: The exact bias and the approximate mean square error of the product estimator are given by

$$B(\widehat{Y_p}) = \frac{Cov(\hat{X},\hat{Y})}{X} \quad \text{and } MSE(\widehat{Y_p}) = V(\hat{Y}) + 2\,R\,Cov(\hat{X},\hat{Y}) + R^2V(\hat{X})$$

PROOF: Let $e_0=\frac{\hat{Y}-Y}{Y}$ and $e_1=\frac{\hat{X}-X}{X}$

It may be noted that e_0 and e_1 satisfy the following $E(e_0)= E(e_1)=0$.

$E(e_0^2)=\frac{V(\hat{Y})}{Y^2}$, $E(e_1^2)=\frac{V(\hat{X})}{X^2}$, $E(e_0e_1)=\frac{Cov(\hat{X},\hat{Y})}{XY}$

Since $\hat{Y}=Y(1+e_0)$ and $\hat{X}= X(1+e_1)$, the estimator $\hat{Y}_p$ can be written as,

$\hat{Y}_p= Y(1+e_0)\,(1+e_1)$

$=Y(1+ e_0+ e_1+ e_0\, e_1)$

$=Y + Y(e_0+ e_1+ e_0\, e_1)$

$\hat{Y}_p-Y= Y(e_0+ e_1+ e_0\, e_1)$ ----(1)

Taking expectation on both sides of (1);

$$E(\hat{Y}_p - Y) = YE(e_0\, e_1)= Y\frac{Cov(\hat{X},\hat{Y})}{XY}= \frac{Cov(\hat{X},\hat{Y})}{X}= B(\widehat{Y_p})$$

Squaring & taking expectation on both sides we get of (1) and ignoring terms of degree greater than two, we get the approximate mean sq. error

$$\begin{aligned} MSE(\hat{Y}_p) = E(\hat{Y}_p - Y)^2 &= Y^2E[e_0^2 + e_1^2 + 2e_0e_1] \\ &= Y^2\left[\frac{V(\hat{Y})}{Y^2} + \frac{V(\hat{X})}{X^2} + 2\frac{Cov(\hat{X},\hat{Y})}{XY}\right] \\ &= V(\hat{Y}) + R^2V(\hat{X}) + 2\,R\,Cov(\hat{X},\hat{Y}) \end{aligned}$$

Hence the proof.

Thm: The product estimator $\hat{Y}_p$ is more efficient than $\hat{Y}$ if $\rho(\hat{X},\hat{Y}) < -\frac{1}{2}\frac{C(X)}{C(Y)}$

PROOF: $V(\hat{Y}) > MSE(\hat{Y}_p)$

$$\left[\frac{N^2(N-n)}{Nn}\right]S_y^2 > \frac{N^2(N-n)}{Nn}\left[S_y^2 + \frac{Y^2}{X^2}S_x^2 + 2\frac{Y}{X}S_{xy}\right]$$

$$\begin{aligned} \therefore S_y^2 &> S_y^2 + \frac{Y^2}{X^2}S_x^2 + 2\frac{Y}{X}S_{xy} \\ 0 &> \frac{Y}{X}\left[\frac{Y}{X}S_x^2 + 2S_{xy}\right] \\ 2S_{xy} &< -\frac{Y}{X}S_x^2 \\ S_{xy} &< -\frac{1}{2}\frac{Y}{X}S_x^2 \end{aligned}$$

Divide both sides by S_xS_y

$$\frac{S_{xy}}{S_xS_y} < -\frac{1}{2}\frac{Y}{X}\frac{S_x}{S_y}$$

$$\therefore \rho_{xy} < -\frac{1}{2}\frac{C(X)}{C(Y)} \qquad \text{where } C(X) = \frac{S_x}{X}, C(Y) = \frac{S_y}{Y}$$

$\therefore$ Product estimator $\hat{Y}_p$, is more efficient than $\hat{Y}$ if $\rho_{xy} < -\frac{1}{2}\frac{C(X)}{C(Y)}$

Thm: Under SRS $\widehat{Y_p} + \frac{N^2(N-n)}{Nn}\frac{S_{xy}}{\bar{X}}$ is unbiased for the population total.

Proof: We know that under SRS $Cov(\hat{X},\hat{Y}) = \left[\frac{N^2(N-n)}{Nn}\right]S_{xy}$

Thus, the true bias of the product estimator under SRS is,

$$-\frac{N^2(N-n)}{Nn}\frac{S_{xy}}{\bar{X}}$$

Since s_{xy} is unbiased for S_{xy} under SRS an unbiased estimator of the bias of the product estimator is

$$-\frac{N^2(N-n)}{Nn}\frac{s_{xy}}{\bar{X}}$$

Therefore $\widehat{Y_p} + \frac{N^2(N-n)}{Nn}\frac{s_{xy}}{\bar{X}}$

The above estimator is obtained by correcting only the approximate bias. Thus, it is approximately unbiased.

TWO PHASE SAMPLING:

The ratio and product estimators introduced in this chapter assume the knowledge of the population total **X** of the auxiliary variable . However there are some situations where the population total of the auxiliary variable will not be known in advance. In such cases, 'two phase sampling' can be used for getting ratio or product estimator.

In two phase sampling, a sample of size 'n' is selected initially by using a suitable sampling design and the population total 'X' is estimated; and then a sample of size 'n' is selected to estimate the population totals of the study as well as auxiliary variables. The second phase sample can be either a subsample of the first phase sample or it can be directly drawn from the given population. The sampling designs used in the first & second phases need not be the same. Depending on the situation, different sampling designs can also be used. Generally 'two phase sampling' is recommended only when the cost of conducting 'first phase' survey is more economical than compared to that of the 'second phase'.

Let $\hat{X}_d$ be an unbiased estimator of 'X' based on the first phase, and $\hat{X}, \hat{Y}$ be the unbiased estimators of X and Y based on the second phase sample. Then the ratio & product estimators based on 'two-phase sampling' are:

$\hat{Y}_{RD} = \left(\frac{\hat{Y}}{\hat{X}}\right) \hat{X}_d$ → Ratio estimator

$\hat{Y}_{PD} = \left(\frac{\hat{Y}\hat{X}}{\hat{X}_d}\right)$ → Product estimator

The following thms. give the approximate bias and MSE (mean square error) of the ratio & product estimator under different cases of 'two-phase sampling'.

THM 1: (i) When the samples are drawn independently in the two phases of sampling, the approximate bias of the ratio estimator is:

$$B(\hat{Y}_{RD}) = Y\left[\frac{V(\hat{X})}{X^2} - \frac{Cov\,(\hat{X},\hat{Y})}{X\,Y}\right]$$

(ii) When the second phase sample is a subsample of the first phase sample, the approximate bias of the ratio estimator is

$$B(\hat{Y}_{RD}) = Y\left[\frac{V(\hat{X})}{X^2} - \frac{Cov\,(\hat{X},\hat{Y})}{X\,Y} - \frac{Cov\,(\hat{X},\hat{X}_d)}{X^2} + \frac{Cov\,(\hat{Y},\hat{X}_d)}{Y\,X}\right]$$

PROOF: (i) It should be noted that when the samples are drawn independently in the two phases of sampling $Cov\,(\hat{X},\hat{X}_d) = 0$ and $Cov\,(\hat{Y},\hat{X}_d) = 0$ -----(1)

Define $e_o = \frac{\hat{Y}-Y}{Y}$ and $e_1 = \frac{\hat{X}-X}{X}$ and $e_d = \frac{\hat{X}_d - X}{X}$

It may be noted that e_0, e_1 & e_d satisfy the following:

$$\left.\begin{aligned} E(e_0) = E(e_1) = E(e_d) = 0 \\ E({e_0}^2) = \frac{V(\hat{Y})}{Y^2}, E({e_1}^2) = \frac{V(\hat{X})}{X^2}, E({e_d}^2) = \frac{V(\hat{X}_d)}{X^2} \\ E(e_0 e_1) = \frac{Cov(\hat{Y}, \hat{X})}{Y\,X}, E(e_0 e_d) = \frac{Cov(\hat{Y}, \hat{X}_d)}{Y\,X} = 0 \\ and\ E(e_1 e_d) = \frac{Cov(\hat{X}, \hat{X}_d)}{X^2} = 0 \end{aligned}\right\} \quad ---(2)$$

The ratio estimator can be expressed in terms of e_o, e_1 and e_d as follows:

$$\hat{Y}_{RD} = [Y(1+e_0)][X(1+e_1)]^{-1}[X(1+e_d)]$$

$$= Y(1+e_0)(1+e_d)(1+e_1)^{-1}$$

$$= Y(1+e_0+e_d+e_0e_d)(1-e_1+{e_1}^2-{e_1}^3+\quad ---)$$

$$= Y(1-e_1+{e_1}^2+e_0-e_0e_1+e_d-e_1e_d+e_0e_d)$$

(ignoring terms of degree > 2)

$$\hat{Y}_{RD} - Y = Y(e_0 - e_1 + e_d + {e_1}^2 - e_0e_1 - e_1e_d + e_0e_d) \qquad ---(3)$$

Taking expectation on both sides of (3) and using (1) & (2); we get

$$B(\hat{Y}_{RD}) = E[\hat{Y}_{RD} - Y]$$

$$= Y[E({e_1}^2) - E(e_0e_1)]$$

$$= Y\left[\frac{V(\hat{X})}{X^2} - \frac{Cov(\hat{Y}, \hat{X})}{Y\,X}\right]$$

(ii) When the 'second phase sample is a subsample of the first phase sample, then $E(e_0e_d) = \frac{Cov(\hat{Y}, \hat{X}_d)}{Y\,X}$ and $E(e_1e_d) = \frac{Cov(\hat{X}, \hat{X}_d)}{X^2}$.

∴By using (3); $B(\hat{Y}_{RD}) = E[\hat{Y}_{RD} - Y] = Y[E({e_1}^2) - E(e_0e_1) - E(e_1e_d) + E(e_0e_d)]$

$$= Y\left[\frac{V(\hat{X})}{X^2} - \frac{Cov(\hat{X}, \hat{Y})}{X\,Y} - \frac{Cov(\hat{X}, \hat{X}_d)}{X^2} + \frac{Cov(\hat{Y}, \hat{X}_d)}{Y\,X}\right]$$

Hence the proof.

<u>Thm 2</u>: (i) When the samples are drawn independently in the two phases of sampling, the approximate mean square error of the ratio estimator is

$$MSE(\hat{Y}_{RD}) = Y^2\left[\frac{V(\hat{Y})}{Y^2} + \frac{V(\hat{X})}{X^2} + \frac{V(\hat{X}_d)}{X^2} - 2\,\frac{Cov(\hat{X},\hat{Y})}{XY}\right]$$

(ii) When the second phase sample is a subsample of the first phase sample, the approximate mean square error is

$$MSE(\hat{Y}_{RD}) = Y^2\left[\frac{V(\hat{Y})}{Y^2} + \frac{V(\hat{X})}{X^2} + \frac{V(\hat{X}_d)}{X^2} - 2\,\frac{Cov(\hat{X},\hat{Y})}{XY} - 2\,\frac{Cov(\hat{X},\hat{X}_d)}{X^2} + 2\,\frac{Cov(\hat{Y},\hat{X}_d)}{YX}\right]$$

PROOF: From the theorem 1;

$$\hat{Y}_{RD} - Y = Y(e_0 - e_1 + e_d + {e_1}^2 - e_0e_1 - e_1e_d + e_0e_d)$$

$$(\hat{Y}_{RD} - Y)^2 = Y^2(e_0 - e_1 + e_d)^2 \qquad \text{(ignoring higher order degree >2)}$$

$$= Y^2({e_0}^2 + {e_1}^2 + {e_d}^2 - 2e_0e_1 - 2e_1e_d + 2e_0e_d)$$

(i) $MSE(\hat{Y}_{RD}) = E(\hat{Y}_{RD} - Y)^2$

$= Y^2[E({e_0}^2) + E({e_1}^2) + E({e_d}^2) - 2E(e_0e_1) - 2E(e_1e_d) + 2E(e_0e_d)]$

$= Y^2\left[\frac{V(\hat{Y})}{Y^2} + \frac{V(\hat{X})}{X^2} + \frac{V(\hat{X}_d)}{X^2} - 2\,\frac{Cov(\hat{X},\hat{Y})}{XY}\right]$

$\{\because\ E(e_1e_d) = 0 = E(e_0e_d)\}$

(ii) When the second phase sample is a subsample of the first phase sample;

$E(e_1e_d) \neq 0, E(e_0e_d) \neq 0$

$$MSE(\hat{Y}_{RD}) = E(\hat{Y}_{RD} - Y)^2$$

$$MSE(\hat{Y}_{RD}) = Y^2\left[\frac{V(\hat{Y})}{Y^2} + \frac{V(\hat{X})}{X^2} + \frac{V(\hat{X}_d)}{X^2} - 2\,\frac{Cov(\hat{X},\hat{Y})}{XY} - 2\,\frac{Cov(\hat{X},\hat{X}_d)}{X^2} + 2\,\frac{Cov(\hat{Y},\hat{X}_d)}{YX}\right]$$

Hence the proof.

THM 3: (i) When the sample are drawn independently the approximate bias of the product estimator in two phase sampling is

$$B(\hat{Y}_{PD}) = Y\left[\frac{V(\hat{X}_d)}{X^2} + \frac{Cov(\hat{Y},\hat{X})}{YX}\right]$$

(ii) When the second phase sample is a subsample, the approximate bias of the product estimator is

$$B(\hat{Y}_{PD}) = Y\left[\frac{V(\hat{X}_d)}{X^2} + \frac{Cov(\hat{Y},\hat{X})}{YX} - \frac{Cov(\hat{Y},\hat{X}_d)}{YX} - \frac{Cov(\hat{X},\hat{X}_d)}{X^2}\right]$$

PROOF: It should be noted that when the samples are drawn independently in the two phases of sampling

$Cov\left(\hat{X},\hat{X}_d\right)=0$ and $Cov\left(\hat{Y},\hat{X}_d\right)=0$ ---(1)

Define $e_0=\frac{\hat{Y}-Y}{Y}$ and $e_1=\frac{\hat{X}-X}{X}$ and $e_d=\frac{\hat{X}_d-X}{X}$

$$\left.\begin{aligned} E(e_0)=E(e_1)=E(e_d)=0 \\ E({e_0}^2)=\frac{V(\hat{Y})}{Y^2}, E({e_1}^2)=\frac{V(\hat{X})}{X^2}, E({e_d}^2)=\frac{V(\hat{X}_d)}{X^2} \\ E(e_0e_1)=\frac{Cov\left(\hat{Y},\hat{X}\right)}{Y\,X}, E(e_0e_d)=\frac{Cov\left(\hat{Y},\hat{X}_d\right)}{Y\,X} \\ and\ E(e_1e_d)=\frac{Cov\left(\hat{X},\hat{X}_d\right)}{X^2} \end{aligned}\right\} \quad ---(2)$$

The product estimator can be expressed in terms of e_0, e_1 and e_d as follows:

$$\hat{Y}_{PD}=\frac{\hat{Y}\,\hat{X}}{\hat{X}_d}$$

$$=\frac{[Y(1+e_0)][X(1+e_1)]}{X(1+e_d)}$$

$$=Y(1+e_0)(1+e_1)(1+e_d)^{-1}$$

$$=Y(1+e_0+e_1+e_0e_1)(1-e_d+{e_d}^2-{e_d}^3+\quad ---)$$

$$=Y(1-e_d+{e_d}^2+e_0-e_0e_d+e_1-e_1e_d+e_0e_1)$$

$$\hat{Y}_{PD}-Y=Y(e_0-e_d+e_1+{e_d}^2-e_0e_d-e_1e_d+e_0e_1) \quad ---(3)$$

(i) Taking expectation on both sides of (3) & using (1) & (2);

$$B\left(\hat{Y}_{PD}\right)=E\left[\hat{Y}_{PD}-Y\right]=Y[E({e_d}^2)+E(e_0e_1)]$$

$$=Y\left[\frac{V(\hat{X}_d)}{X^2}+\frac{Cov\left(\hat{Y},\hat{X}\right)}{Y\,X}\right]$$

$(\because E(e_0e_d)=E(e_1e_d)=0)$

Hence the proof

(ii) When the second phase sample is a sub sample, the approximate bias of the product estimator is obtain by taking expectation on both sides of (3)

$$B\left(\hat{Y}_{PD}\right)=E\left[\hat{Y}_{PD}-Y\right]=Y[E({e_d}^2)-E(e_0e_d)-E(e_1e_d)+E(e_0e_1)]$$

$$B\left(\hat{Y}_{PD}\right)=Y\left[\frac{V(\hat{X}_d)}{X^2}+\frac{Cov\left(\hat{Y},\hat{X}\right)}{Y\,X}-\frac{Cov\left(\hat{Y},\hat{X}_d\right)}{Y\,X}-\frac{Cov\left(\hat{X},\hat{X}_d\right)}{X^2}\right]$$

Hence proof

THM 4: (i) When the samples are independently drawn, the approximate mean square error of the product estimator is

$$MSE(\hat{Y}_{PD}) = Y^2\left[\frac{V(\hat{Y})}{Y^2} + \frac{V(\hat{X})}{X^2} + \frac{V(\hat{X}_d)}{X^2} + 2\,\frac{Cov(\hat{Y},\hat{X})}{XY}\right]$$

(ii) When the second phase sample is a subsample, the approximate MSE of the product estimator is

$$MSE(\hat{Y}_{PD}) = Y^2\left[\frac{V(\hat{Y})}{Y^2} + \frac{V(\hat{X})}{X^2} + \frac{V(\hat{X}_d)}{X^2} + 2\,\frac{Cov(\hat{Y},\hat{X})}{XY} - 2\,\frac{Cov(\hat{Y},\hat{X}_d)}{YX} - 2\frac{Cov(\hat{X},\hat{X}_d)}{X^2}\right]$$

PROOF: From them. 3

$$\hat{Y}_{PD} - Y = Y(e_0 - e_d + e_1 + {e_d}^2 - e_0e_d - e_1e_d + e_0e_1) \text{ ----------(1)}$$

Taking square on both sides and ignoring degrees greater than 2,

$$(\hat{Y}_{PD} - Y)^2 = Y^2(e_0 - e_d + e_1)^2 \quad \text{(ignoring higher order degree >2)}$$
$$= Y^2({e_0}^2 + {e_1}^2 + {e_d}^2 - 2e_0e_d - 2e_1e_d + 2e_0e_1) \text{ ------- (2)}$$

Taking expectation on both sides i.e $E(\hat{Y}_{PD} - Y)^2$

$$MSE(\hat{Y}_{PD}) = Y^2[E({e_0}^2) + E({e_1}^2) + E({e_d}^2) - 2E(e_0e_d) - 2E(e_1e_d) + 2E(e_0e_1)] \text{ -----(3)}$$

Case (i)

Since in case (i) samples are drawn independently $Cov(\hat{Y},\hat{X}_d) = 0$ ie $E(e_0e_d) = 0$

And $Cov(\hat{X},\hat{X}_d) = 0$ ie $E(e_1e_d) = 0$

$\therefore\ MSE(\hat{Y}_{PD}) = E[\hat{Y}_{PD} - Y]^2$

$$= Y^2\left[\frac{V(\hat{Y})}{Y^2} + \frac{V(\hat{X})}{X^2} + \frac{V(\hat{X}_d)}{X^2} + 2\,\frac{Cov(\hat{Y},\hat{X})}{XY}\right]$$

Hence proof.

(ii) When the second phase sample is a subsample of the first phase sample, then

$Cov(\hat{Y},\hat{X}_d) \neq 0$ and $Cov(\hat{X},\hat{X}_d) \neq 0$

Therefore by using (3) we get

$\therefore\ MSE(\hat{Y}_{PD}) = E[\hat{Y}_{PD} - Y]^2$

$$= Y^2\left[\frac{V(\hat{Y})}{Y^2} + \frac{V(\hat{X})}{X^2} + \frac{V(\hat{X}_d)}{X^2} + 2\,\frac{Cov(\hat{Y},\hat{X})}{XY} - 2\,\frac{Cov(\hat{Y},\hat{X}_d)}{YX} - 2\frac{Cov(\hat{X},\hat{X}_d)}{X^2}\right]$$

Hence the proof of case (ii).

APPROXIMATE BIAS & MEAN SQUARE ERROR OF THE RATIO ESTIMATOR OF TWO-PHASE SAMPLING WHEN SIMPLE RANDOM SAMPLING IS USED:

Let S_1 & S_2 be the samples obtained in the two phases of sampling, where, in the first phase a large sample of size n_1 is drawn to estimate the population total X and in the second phase a sample of size 'n' is drawn to estimate the populatn totals 'X' and 'Y', both by using SRS. Here 'n' is assumed to be small when compared to n_1.

Let $\hat{X}_d = \frac{N}{n_1}\sum_{i\in S_1} X_i$, $\hat{X} = \frac{N}{n_2}\sum_{i\in S_2} X_i$ and $\hat{Y} = \sum_{i\in S_2} Y_i$.

Note that when S_2 is a subsample of S_1

(i) $E(\hat{X}_d) = \sum_{i=1}^{N} X_i$ = X ---(1)

(ii) $E(\hat{X}) = E_I E_{II}(\hat{X}|S_1) = E_I\left[E_{II}\left(\frac{N}{n_1}\sum_{i\in S_1} X_i \,\middle|\, S_1\right)\right]$

$= E_I\left[\frac{N}{n_1}\sum_{i\in S_1} X_i\left(\frac{n_1}{n}\right)\right]$

$= X$ ---(2)

(iii) $V(\hat{X}_d) = \frac{N^2(N-n_1)}{N\,n_1} {S_X}^2$ ---(3)

(iv) $V(\hat{X}) = E_I V_{II}(\hat{X}|S_1) + V_I E_{II}(\hat{X}|S_1)$

$= E_I\left[\frac{N^2(n_1-n)}{n_1 n}\,.{S_X}^2\right] + V_I(\hat{X}_d)$

$= \frac{N^2(n_1-n)}{n_1 n}\,.{S_X}^2 + \frac{N^2(N-n_1)}{N\,n_1}{S_X}^2$

$= \frac{N^2(N-n)}{N\,n}\,{S_X}^2$ ---(4)

(v) $V(\hat{Y}) = \frac{N^2(N-n)}{N\,n}\,{S_Y}^2$ ---(5)

(vi) $Cov(\hat{X}, \hat{X}_d) = E(\hat{X}\hat{X}_d) - E(\hat{X})E(\hat{X}_d)$

$= E_I E_{II}(\hat{X}\hat{X}_d\,|S_1) - X^2$

$= E_I[\hat{X}_d^2] - X^2$

$= V(\hat{X}_d) = \frac{N^2(N-n_1)}{N\,n_1}{S_X}^2$ ---(6)

(vii) $Cov\left(\hat{Y},\hat{X}_d\right) = \frac{N^2(N-n_1)}{N\,n_1} S_{XY}$ ---(7)

(viii) $Cov\left(\hat{Y},\hat{X}\right) = E\left(\hat{X}\hat{Y}\right) - XY$

$= E_I E_{II}\left(\hat{X}\hat{Y}\middle|S_1\right) - XY$

$= E_I\left[\frac{N^2(n_1-n)}{n_1 n}\,.S_{1(XY)} + \hat{Y}_d\,\hat{X}_d\right] - XY$

$= \frac{N^2(n_1-n)}{n_1 n} S_{XY} + E_I\left(\hat{X}_d\,\hat{Y}_d\right) - XY$

$= \frac{N^2(n_1-n)}{n_1 n} S_{XY} + Cov\left(\hat{X}_d\,,\hat{Y}_d\right)$

$= \frac{N^2(n_1-n)}{n_1 n} S_{XY} + \frac{N^2(N-n_1)}{N n_1} S_{XY}$

$= \frac{N^2(N-n)}{N\,n} S_{XY}$ ---(8)

Here $\hat{Y}_d$ and $S_{1(XY)}$ are their analogues of Y and S_{xy} resp. based on the sample S_1.

When the samples drawn independently, the results derived in the SRS can be used directly without any difficulty.

THM 5: When SRS is used in both phases of sampling and samples are drawn independently

$$B\left(\hat{Y}_{RD}\right) = \frac{N^2(N-n)}{N\,n}\,Y[C_{XX} - C_{XY}]$$

And $MSE\left(\hat{Y}_{RD}\right) = \frac{N^2(N-n)}{N\,n}\,Y^2[C_{YY} + C_{XX} - 2C_{XY}] + \frac{N^2(N-n_1)}{N\,n_1} Y^2 C_{XX}$

Where $C_{XX} = \frac{{S_X}^2}{X^2}, C_{YY} = \frac{{S_Y}^2}{Y^2}, C_{XY} = \frac{S_{XY}}{XY}$

PROOF: We know from the thm 1 , when the samples are drawn independently in the two phases of sampling,

$$B\left(\hat{Y}_{RD}\right) = Y\left[\frac{V\left(\hat{X}\right)}{X^2} - \frac{Cov\left(\hat{Y},\hat{X}\right)}{Y\,X}\right]$$

By using (4) & (8), substitute value of $V\left(\hat{X}\right)$ and $Cov\left(\hat{Y},\hat{X}\right)$., we get

$$B\left(\hat{Y}_{RD}\right) = Y\left[\frac{N^2(N-n)}{N\,n}\frac{{S_X}^2}{X^2} - \frac{N^2(N-n)}{N\,n}\frac{S_{XY}}{XY}\right]$$

$$= \frac{N^2(N-n)}{N\,n} Y\left[\frac{{S_X}^2}{X^2} - \frac{S_{XY}}{XY}\right]$$

$$= \frac{N^2(N-n)}{N\,n} Y[C_{XX} - C_{XY}]$$

And $MSE(\hat{Y}_{RD})$ by using thm 2;

$$MSE(\hat{Y}_{RD}) = Y^2 \left[\frac{V(\hat{Y})}{Y^2} + \frac{V(\hat{X})}{X^2} + \frac{V(\hat{X}_d)}{X^2} - 2\,\frac{Cov(\hat{X},\hat{Y})}{X\,Y}\right]$$

Substitute values of $V(\hat{Y}), V(\hat{X}), V(\hat{X}_d)$ &$Cov(\hat{X},\hat{Y})$ by using (3),(4),(5) & (8), we get

$$MSE(\hat{Y}_{RD}) = Y^2 \left[\frac{N^2(N-n)}{N\,n}\frac{{S_Y}^2}{Y^2} + \frac{N^2(N-n)}{N\,n}\frac{{S_X}^2}{X^2} + \frac{N^2(N-n_1)}{N\,n_1}\frac{{S_X}^2}{X^2} - 2\frac{N^2(N-n)}{N\,n}\frac{S_{XY}}{XY}\right]$$

$$= \frac{N^2(N-n)}{N\,n} Y^2[C_{YY} + C_{XX} - 2C_{XY}] + \frac{N^2(N-n_1)}{N\,n_1} Y^2\,C_{XX}$$

<u>THM 6</u>: When SRS is used in both the phases of sampling and the second phase sample is a subsample of the first phase sample then

$$B(\hat{Y}_{RD}) = \frac{N^2(n_1-n)}{n_1\,n} Y\,[C_{XX} - C_{XY}]$$

And $MSE(\hat{Y}_{RD}) = \frac{N^2(N-n)}{N\,n}\,Y^2[C_{YY} + C_{XX} - 2C_{XY}] + \frac{N^2(N-n_1)}{N\,n_1} Y^2[C_{XY} - C_{XX}]$

<u>PROOF</u>: When the second phase sample is a subsample of the first phase sample,

$$B(\hat{Y}_{RD}) = Y\left[\frac{V(\hat{X})}{X^2} - \frac{Cov(\hat{Y},\hat{X})}{Y\,X} - \frac{Cov(\hat{X},\hat{X}_d)}{X^2} + \frac{Cov(\hat{Y},\hat{X}_d)}{YX}\right]$$

$$= Y\left[\frac{N^2(N-n)}{N\,n}\frac{{S_X}^2}{X^2} - \frac{N^2(N-n)}{N\,n}\frac{S_{XY}}{XY} - \frac{N^2(N-n_1)}{N\,n_1}\frac{{S_X}^2}{X^2} + \frac{N^2(N-n_1)}{N\,n_1}\frac{S_{XY}}{XY}\right]$$

$$= Y\left[\frac{N^2(N-n)}{N\,n}\frac{{S_X}^2}{X^2} - \frac{N^2(N-n_1)}{N\,n_1}\frac{{S_X}^2}{X^2}\right] - Y\left[\frac{N^2(N-n)}{N\,n}\frac{S_{XY}}{XY} - \frac{N^2(N-n_1)}{N\,n_1}\frac{S_{XY}}{XY}\right]$$

$$= \frac{N^2(n_1-n)}{n_1\,n}\,Y\,\frac{{S_X}^2}{X^2} - \frac{N^2(n_1-n)}{n_1\,n}\,Y\frac{S_{XY}}{XY}$$

$$= \frac{N^2(n_1-n)}{n_1\,n}\,Y[C_{XX} - C_{XY}]$$

And $MSE(\hat{Y}_{RD}) = Y^2\left[\frac{V(\hat{Y})}{Y^2} + \frac{V(\hat{X})}{X^2} + \frac{V(\hat{X}_d)}{X^2} - 2\,\frac{Cov(\hat{X},\hat{Y})}{X\,Y} - 2\frac{Cov(\hat{X},\hat{X}_d)}{X^2} - 2\frac{Cov(\hat{Y},\hat{X}_d)}{Y\,X}\right]$

Substituting all values of (3), (4), (5), (6), (7) & (8), we get;

$$MSE(\hat{Y}_{RD}) = Y^2\left[\frac{N^2(N-n)}{N\,n}\frac{{S_Y}^2}{Y^2} + \frac{N^2(N-n)}{N\,n}\frac{{S_X}^2}{X^2} + \frac{N^2(N-n_1)}{N\,n_1}\frac{{S_X}^2}{X^2} - 2\frac{N^2(N-n)}{N\,n}\frac{S_{XY}}{XY}\right.$$
$$\left. - 2\frac{N^2(N-n_1)}{N\,n_1}\frac{{S_X}^2}{X^2} + 2\frac{N^2(N-n_1)}{N\,n_1}\frac{S_{XY}}{XY}\right]$$

$$= Y^2\left[\frac{N^2(N-n)}{N\,n}\frac{{S_Y}^2}{Y^2} + \frac{N^2(N-n)}{N\,n}\frac{{S_X}^2}{X^2} - \frac{N^2(N-n_1)}{N\,n_1}\frac{{S_X}^2}{X^2} - 2\frac{N^2(N-n)}{N\,n}\frac{S_{XY}}{XY}\right.$$
$$\left. + \ 2\frac{N^2(N-n_1)}{N\,n_1}\frac{S_{XY}}{XY}\right]$$

$$= \frac{N^2(N-n)}{N\,n}\,.Y^2\left[\frac{{S_Y}^2}{Y^2} + \frac{{S_X}^2}{X^2} - 2\frac{S_{XY}}{XY}\right] + \frac{N^2(N-n_1)}{N\,n_1}\,.Y^2\left[\frac{S_{XY}}{XY} - \frac{{S_X}^2}{X^2}\right]$$

$$= \frac{N^2(N-n)}{N\,n}\,.Y^2[C_{YY} + C_{XX} - 2C_{XY}] + \frac{N^2(N-n_1)}{N\,n_1}Y^2[C_{XY} - C_{XX}]$$

Hence the proof.

USE OF MULTI_AUXILIARY INFORMATION:

There are many situations in which in addition to the study variable, information on several related auxiliary variables will not be available. In such situations, the ratio estimator can be extended in several ways. In this section one straight forward extension due to ***OLKIN(1958)*** is considered.

Let $\hat{X}_i$ be unbiased for X_i (i=1,2,---,k), the population total of the i[th] auxiliary variable and $\hat{Y}$ be unbiased for Y, the population total of the study variable. ***OLKIN*** suggested a composite estimator of the form $\hat{Y}_{RK} = \sum_{i=1}^{k} w_i \left[\frac{\hat{Y}}{\hat{X}_i}\right] X_i$ ---(1)

Where $w_1, w_2, - - -, w_k$ are predetermined constants satisfying $\sum_{i=1}^{k} w_i = 1$.

When k=2, the above estimator reduces to

$$\hat{Y}_{R2} = w_1\left[\frac{\hat{Y}}{\hat{X}_1}\right]X_1 + w_2\left[\frac{\hat{Y}}{\hat{X}_2}\right]X_2 \qquad \text{---(2)}$$

Where $w_1 + w_2 = 1$.

In order to make the expressions compact, the following notations are used.

$$V_0 = \frac{V(\hat{Y})}{Y^2}, V_1 = \frac{V(\hat{X}_1)}{{X_1}^2}, V_2 = \frac{V(\hat{X}_2)}{{X_2}^2}, C_{01} = \frac{Cov\,(\hat{Y}, \hat{X}_1)}{Y\,X_1}, C_{02} = \frac{Cov\,(\hat{Y}, \hat{X}_2)}{Y\,X_2},$$

$$C_{12} = \frac{Cov\,(\hat{X}_1, \hat{X}_2)}{X_1 X_2}$$

THM 1: The approximate bias and MSE of $\hat{Y}_{R2}$ are

$$B(\hat{Y}_{R2}) = Y[V_2 - C_{02} + w_1(C_{02} - C_{01} - V_2)]$$

And $MSE(\hat{Y}_{R2})= Y^2[V_0 + V_2 - 2C_{02} + {w_1}^2(V_2 + V_1 - 2C_{12}) - 2w_1(C_{01} + V_2 - C_{02} - C_{12})]$

PROOF: Let $e_0 = \frac{\hat{Y}-Y}{Y}$, $e_1=\frac{\hat{X}_1-X_1}{X_1}$, $e_2=\frac{\hat{X}_2-X_2}{X_2}$

The estimator $\hat{Y}_{R2}$ can be written as

$$\hat{Y}_{R2} = w_1\, Y(1+e_0)(1+e_1)^{-1} + w_2\, Y(1+e_0)(1+e_2)^{-1}$$

$$= Y[w_1(1+e_0)(1-e_1+{e_1}^2- \ \dots) + w_2(1+e_0)(1-e_2+{e_2}^2- \ \dots)$$

$$= Y[w_1(1-e_1+{e_1}^2+e_0-e_0e_1 \ \dots) + (1-w_1)(1-e_2+{e_2}^2+e_0-e_0e_2 \dots)]$$

$$= Y[(1-e_2+{e_2}^2+e_0-e_0e_2 \dots) + w_1(1-e_1+{e_1}^2+e_0-e_0e_1 \dots \ -1+e_2-{e_2}^2-e_0+e_0e_2 \dots)]$$

$$\hat{Y}_{R2} - Y = Y[(e_0 - e_2 - e_0e_2 + {e_2}^2 \dots) + w_1(e_2 - e_1 - e_0e_1 - {e_2}^2 + e_0e_2 \dots)] \qquad \text{---(1)}$$

Taking expectation on both sides, after ignoring terms with degree >2, we get the approximate bias;

$$B(\hat{Y}_{R2}) = E(\hat{Y}_{R2} - Y) = Y\left[\left(E({e_2}^2) - E(e_0e_2)\right) + w_1\left(E(e_0e_2) - E(e_0e_1) - E({e_2}^2)\right)\right]$$

$$= Y[(V_2 - C_{02}) + w_1(C_{02} - C_{01} - V_2)] \qquad \text{---(2)}$$

Taking expectation on both sides of (1), after squaring & ignoring terms with degree >2, we get the approximate MSE;

$$MSE(\hat{Y}_{R2}) = E(\hat{Y}_{R2} - Y)^2$$

$$= E[Y^2\{(e_0 - e_2)^2 + {w_1}^2\,(e_2 - e_1)^2 + 2w_1(e_0e_2 - e_0e_1 - \ {e_2}^2 + e_1e_2)\}]$$

$$= Y^2\left[E({e_0}^2) + \ E({e_2}^2) - 2E(e_0e_2) + {w_1}^2\left(E({e_2}^2) + E({e_1}^2) - 2E(e_1e_2)\right) + 2w_1\left(E(e_0e_2) - E(e_0e_1) - E({e_2}^2) + E(e_1e_2)\right)\right]$$

$$= Y^2[V_0 + V_2 - 2C_{02} + {w_1}^2(V_2 + V_1 - 2C_{12}) - 2w_1(C_{01} + V_2 - C_{02} - C_{12})] \qquad \text{---(3)}$$

REMARK: Note that the MSE given in (3) attains minimum if,

$$w_1 = \frac{C_{01}+V_2-C_{02}-C_{12}}{V_2+V_1-2C_{12}} \qquad \text{---(4)}$$

The minimum MSE of the estimator $\widehat{Y_{R2}}$ obtained by substituting (4) in (3) is,

$$Y^2\left[(V_0 + V_2 - 2C_{02}) - \frac{(V_2+C_{01}-C_{02}-C_{12})^2}{V_2+V_1-2C_{12}}\right] \qquad \text{---(5)}$$

It is to note that the denominator of the above expression is nothing but the variance of the difference e_1-e_2. Thus, the minimum MSE given in (5) is always less than or equal to

$Y^2(V_0 + V_2 - 2C_{02})$. Which is nothing but the approximate MSE of the ratio estimator based on the auxiliary variable X_2 . Thus, we infer that- by using the additional auxiliary variable, the efficiency of the ratio estimator can be increased.

RATIO ESTIMATION IN STRATIFIED SAMPLING:

When the sample is selected in the form of a stratified sample, the ratio estimator can be constructed in two different ways.

Let $\hat{Y}_h$and $\hat{X}_h$ (h=1,2,---,L) be unbiased for the population totals Y_h and X_h, the h[th] stratum totals of the study and auxiliary variables resp. Using these estimates, the population total can be estimated by using one of the following estimates

$\hat{Y}_{RS} = \sum_{h=1}^{L} \left[\frac{\hat{Y}_h}{\hat{X}_h}\right] X_h$ ---(1)

$\hat{Y}_{RC} = \left[\frac{\sum_{h=1}^{L} \hat{Y}_h}{\sum_{h=1}^{L} \hat{X}_h}\right] X$ ---(2)

The estimators $\widehat{Y_{RS}}$ and $\widehat{Y_{RC}}$ are known as separate ratio estimator and combined ratio estimator resp. The separate ratio estimator can be used to estimate the population total only when the true stratum total $\widehat{X_h}$ of the auxiliary variable is known in all strata.

THM 1: The approximate bias and MSE of the separate ratio estimator are

$$B(\hat{Y}_{RS}) = \sum_{h=1}^{L} Y_h \left\{\left[\frac{V(\hat{X}_h)}{{X_h}^2}\right] - \frac{Cov\,(\hat{X}_h, \hat{Y}_h)}{X_h Y_h}\right\}$$

And $MSE(\hat{Y}_{RS}) = \sum_{h=1}^{L} {Y_h}^2 \left[\frac{V(\hat{Y}_h)}{{Y_h}^2} + \frac{V(\hat{X}_h)}{{X_h}^2} - 2\frac{Cov\,(\hat{X}_h, \hat{Y}_h)}{X_h Y_h}\right]$

PROOF: Note that $E(\hat{Y}_{RS} - Y) = \sum_{h=1}^{L} E[\{\left[\frac{\hat{Y}_h}{\hat{X}_h}\right] X_h\} - Y]$

$= \sum_{h=1}^{L} Y_h \left[\frac{V(\hat{X}_h)}{{X_h}^2} - \frac{Cov\,(\hat{X}_h, \hat{Y}_h)}{X_h Y_h}\right]$ ---(1) $\left(using\ thm\ B(\hat{Y}_R) = Y\left[\frac{V(\hat{X})}{X^2} - \frac{Cov\,(\hat{Y}, \hat{X})}{Y\,X}\right]\right)$

Thus,

$B(\hat{Y}_{RS}) = \sum_{h=1}^{L} Y_h \left\{\left[\frac{V(\hat{X}_h)}{{X_h}^2}\right] - \frac{Cov\,(\hat{X}_h, \hat{Y}_h)}{X_h Y_h}\right\}$ ---(2)

The MSE of the separate ratio estimator is

$$MSE(\hat{Y}_{RS}) = E(\hat{Y}_{RS} - Y)^2 = E\sum_{h=1}^{L} \left\{\left[\frac{\hat{Y}_h}{\hat{X}_h}\right] X_h - Y_h\right\}^2$$

$$= \sum_{h=1}^{L} {Y_h}^2 \left[\frac{V(\hat{Y}_h)}{{Y_h}^2} + \frac{V(\hat{X}_h)}{{X_h}^2} - 2\frac{Cov\,(\hat{X}_h,\hat{Y}_h)}{X_h Y_h}\right] \qquad \text{---(3)}$$

$$\left(Using\ thm\ MSE\left(\widehat{Y_R}\right) = Y^2\left[\frac{V(\hat{Y})}{Y^2} + \frac{V(\hat{X})}{X^2} - 2\,\frac{Cov\,(\hat{X},\hat{Y})}{X\,Y}\right]\right)$$

THM 2: The approximate bias and MSE are $B\left(\hat{Y}_{RC}\right) = Y\left[\frac{V(\sum_{h=1}^{L}\hat{X}_h)}{X^2} - \frac{Cov(\sum_{h=1}^{L}\hat{X}_h,\sum_{h=1}^{L}\hat{Y}_h)}{XY}\right]$

$$= Y\left[\frac{\sum_{h=1}^{L} V(\hat{X}_h)}{X^2} - \frac{\sum_{h=1}^{L} Cov\,(\hat{X}_h,\hat{Y}_h)}{XY}\right] \qquad \text{---(1)}$$

And $MSE(\hat{Y}_{RS}) = Y^2\left[\frac{V(\sum_{h=1}^{L}\hat{Y}_h)}{Y^2} + \frac{V(\sum_{h=1}^{L}\hat{X}_h)}{X^2} - 2\frac{Cov(\sum_{h=1}^{L}\hat{X}_h,\sum_{h=1}^{L}\hat{Y}_h)}{XY}\right]$

$$= Y^2\left[\sum_{h=1}^{L}\left\{\frac{V(\hat{Y}_h)}{Y^2} + \frac{V(\hat{X}_h)}{X^2} - 2\frac{Cov\,(\hat{X}_h,\hat{Y}_h)}{XY}\right\}\right] \qquad \text{---(2)}$$

PROOF: H.W.

REGRESSION ESTIMATION:

Like Ratio Estimation, regression estimation is another method of estimation of a finite population total using the knowledge of an auxiliary variable 'X' which is closely related to the study variable 'Y'.

We know that, when the variables 'X' & 'Y' are linearly related, the least square estimates of the slope & intercept are respectively $\hat{\beta} = \frac{S_{XY}}{S_X{}^2}$ and $\hat{\alpha} = \hat{\bar{Y}} - \hat{\beta}\hat{\bar{X}}$.

The population total 'Y' can be written as

$Y = \sum_{i\in S} Y_i + \sum_{i\in S'} Y_i$, where S'=S-s. ---(0)

Once the sample is observed, the first term in the RHS becomes fully known. Using the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$, each observed Y_i can be estimated by

$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ for i∈S' ---(1)

$= \hat{\bar{Y}} - \hat{\beta}\hat{\bar{X}} + \hat{\beta}X_i$

Summing both the sides over S', we get

$$\sum_{i\in S'} \hat{Y}_i = (N-n)\left(\hat{\bar{Y}} - \hat{\beta}\hat{\bar{X}}\right) + \hat{\beta}\sum_{i\in S'} X_i$$

$$= (N-n)\left(\hat{\bar{Y}} - \hat{\beta}\hat{\bar{X}}\right) + \hat{\beta}\left(X - \sum_{i\in S} X_i\right)$$

$$= (N-n)\left(\hat{\bar{Y}} - \hat{\beta}\hat{\bar{X}}\right) + \hat{\beta}\left(N\bar{X} - n\hat{\bar{X}}\right)$$

Substituting these estimated values for the unobserved Y_i values in (0), we get an estimator for the population total Y as

$$\hat{Y} = \sum_{i\in S} Y_i + (N-n)\left(\hat{\bar{Y}} - \hat{\beta}\hat{\bar{X}}\right) + \hat{\beta}\left(N\bar{X} - n\hat{\bar{X}}\right)$$

$$= n\hat{\bar{Y}} + (N-n)\left(\hat{\bar{Y}} - \hat{\beta}\hat{\bar{X}}\right) + \hat{\beta}\left(N\bar{X} - n\hat{\bar{X}}\right)$$

$$= N\hat{\bar{Y}} + N\hat{\beta}\left(\bar{X} - \hat{\bar{X}}\right)$$

The above estimator is known as the **Linear Regression Estimator** of the population total Y.

It should be noted that the above estimator is not unbiased for the population total under SRS.

THM 1: The approximate Mean Square Error of the regression estimator under SRS is

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

PROOF: Define $e_0 = \frac{\hat{Y}-Y}{Y}$, $e_1 = \frac{\hat{X}-X}{X}$, $e_2 = \frac{s_{XY}-S_{XY}}{S_{XY}}$ and $e_3 = \frac{{s_X}^2-{S_X}^2}{{S_X}^2}$

It is noted that $E(e_i) = 0$, for i=1,2,3.

And $E(e_0^2)=\frac{V(\hat{Y})}{Y^2}=\frac{N^2\left[\frac{N-n}{Nn}\right]{S_Y}^2}{Y^2}$ similarly $E(e_1^2)$, $E(e_2^2)$ and $E(e_3^2)$

The regression estimator can be written as

$$\hat{Y}_{LR} = Y(1+e_0) + \left[\frac{S_{XY}(1+e_2)}{{S_X}^2(1+e_3)}\right][X - X(1+e_1)]$$

$$= Y(1+e_0) - \left[\frac{XS_{XY}}{{S_X}^2}\right](1+e_2)(1+e_3)^{-1}e_1$$

Assuming $|e_i| < 1$, the linear regression estimator can be rewritten as

$$\hat{Y}_{LR} - Y = Ye_0 - BXe_1(1+e_2)(1-e_3+{e_3}^2- \ \ldots)$$

Where $B = \frac{S_{XY}}{{S_X}^2}$

$$= Ye_0 - BX[e_1 - e_1e_3 + e_1e_2] \qquad \text{(ignoring terms of degree >2)}$$

Squaring both the sides and taking expectations we get,

$$E\left[\hat{Y}_{LR} - Y\right]^2 = Y^2E({e_0}^2) + X^2B^2E({e_1}^2) - 2XYB\ E(e_0e_1)$$

$$= N^2\left[\frac{N-n}{Nn}\right][{S_Y}^2 + B^2{S_X}^2 - 2BS_{XY}]$$

$$= N^2\left[\frac{N-n}{Nn}\right]\left[{S_Y}^2 + \left(\frac{{S_{XY}}^2}{{S_X}^4}\right){S_X}^2 - 2\left(\frac{S_{XY}}{{S_X}^2}\right)S_{XY}\right]$$

$$= N^2\left[\frac{N-n}{Nn}\right]{S_Y}^2\left[1 - \left(\frac{{S_{XY}}^2}{{S_X}^2{S_Y}^2}\right)\right]$$

$$= N^2\left[\frac{N-n}{Nn}\right]{S_Y}^2(1-\rho^2)$$

THM 2: Under SRS $V(\hat{Y}_{SRS}) > MSE(\hat{Y}_{LR})$ and $MSE(\hat{Y}_R) > MSE(\hat{Y}_{LR})$

PROOF: Since -1<ρ<1, we have $(1-\rho^2) < 1$.

Therefore, $N^2\left[\frac{N-n}{Nn}\right]S_Y{}^2(1-\rho^2) < N^2\left[\frac{N-n}{Nn}\right]S_Y{}^2$

Hence $V(\hat{Y}_{SRS}) > MSE(\hat{Y}_{LR})$

Consider the difference,

$MSE(\hat{Y}_R) - MSE(\hat{Y}_{LR})$ where $\hat{Y}_R$ is ratio estimator.

$$MSE(\hat{Y}_R) - MSE(\hat{Y}_{LR}) = N^2\left[\frac{N-n}{Nn}\right][S_Y{}^2 + R^2S_X{}^2 - 2RS_{XY} - S_Y{}^2 + S_Y{}^2\rho^2]$$

$$= N^2\left[\frac{N-n}{Nn}\right]\left[S_Y{}^2\rho^2 + R^2S_X{}^2 - 2RS_{XY}\right]$$

$$= N^2\left[\frac{N-n}{Nn}\right]\left[\frac{S_Y{}^2S_{XY}{}^2}{S_X{}^2S_Y{}^2} + R^2S_X{}^2 - 2RS_{XY}\right]$$

$$= N^2\left[\frac{N-n}{Nn}\right]\frac{\left[S_{XY} - RS_X{}^2\right]^2}{S_X{}^2}$$

Since the RHS of the above expression is always non negative, the result follows.

DIFFERENCE ESTIMATION:

The ratio estimator which is obtained by multiplying the conventional estimator $\hat{Y}$ by the factor $X/\hat{X}$ is an alternative to the estimator $\hat{Y}$.Here we shall examine the possibility of improving upon $\hat{Y}$ by considering the estimator obtained by adding $\hat{Y}$ with constant times the difference $X - \hat{X}$ whose expected value is zero.

That is, as an estimator of Y, we suggest $\hat{Y}_{DR} = \hat{Y} + \lambda\left[X - \hat{X}\right]$ ---(1)

Where λ is a predetermined value such that the above estimator depends on the difference $X - \hat{X}$ rather than the ratio $X/\hat{X}$. It is termed as **DIFFERENCE ESTIMATOR.**

The difference estimator is evidently unbiased for the population total Y and its variance is $MSE(\hat{Y}_{DR}) = E\left[\hat{Y}_{DR} - Y\right]^2$

$$= E\left[(\hat{Y} - Y) + \lambda(X - \hat{X})\right]^2 = E\left[(\hat{Y} - Y) - \lambda(\hat{X} - X)\right]^2$$

$$= E\left[\hat{Y} - Y\right]^2 + \lambda^2E\left[\hat{X} - X\right]^2 - 2\lambda E\left[(\hat{X} - X)(\hat{Y} - Y)\right]$$

$$= V(\hat{Y}) + \lambda^2V(\hat{X}) - 2\lambda Cov(\hat{X}, \hat{Y})$$

The above expression for variance is applicable for any sampling design yielding unbiased estimators for Y and X.

It can be seen that the above variance is minimum if $\lambda = \frac{Cov(\hat{X},\hat{Y})}{V(\hat{X})}$

And the resulting min variance is $V(\hat{Y})(1-{\rho_{XY}}^2)$ where ρ_{XY} is the coefficient of correlation between X and Y.

It is interesting to note that, when SRS is used, the optimum value of $\lambda = \frac{S_{XY}}{{S_X}^2}$ and the min variance happens to be $N^2\left[\frac{N-n}{Nn}\right]{S_Y}^2(1-\rho^2)$ which is nothing but the approx mean square error of the linear regression estimator. It is to note that the optimum value of λ depends on S_{XY} which in general will not be known. Normally in such situations, survey practitioners use unbiased estimators for unknown quantities. The value derived from the optimal choice happens to be the least squares estimate. Thus, the estimator $\hat{Y}_{DR}$ reduces to the linear regression estimator.It is also to be noted that the difference estimator reduces to the ratio estimator when $\lambda = \frac{\hat{Y}}{\hat{X}}$

DOUBLE SAMPLING IN DIFFERENCE ESTIMATOR:

As in the case of ratio estimation, here one can employ double sampling method to estimate the population total Y whenever the population total X of the auxiliary variable is not known.The difference estimator for population total under double sampling is defined as

$$\hat{Y}_{DD} = \hat{Y} + \lambda(\hat{X}_D - \hat{X})$$

Where $\hat{X}_D$ is the unbiased estimator of the population total X based on the first phase sample. Evidently the difference estimator is unbiased for the population total in both cases of double sampling.

Note that $V(\hat{Y}_{DD}) = E[\hat{Y}_{DD} - Y]^2$

$$= E[\hat{Y} + \lambda(\hat{X}_d - \hat{X}) - Y]^2$$

$$= E[\hat{Y} - Y + \lambda\{(\hat{X}_d - X) - (\hat{X} - X)\}]^2$$

$$= E\left[(\hat{Y}-Y)^2 + \lambda^2\{(\hat{X}_d - X) - (\hat{X} - X)\}^2 + 2\lambda(\hat{Y}-Y)\{(\hat{X}_d - X) - (\hat{X} - X)\}\right]$$

$$= E[(\hat{Y}-Y)^2 + \lambda^2\left\{(\hat{X}_d - X)^2 + (\hat{X} - X)^2 - 2(\hat{X}_d - X)(\hat{X} - X)\right\}$$

$$+2\lambda\{(\hat{Y}-Y)(\hat{X}_d - X) - (\hat{Y}-Y)(\hat{X} - X)\}]$$

$$= E(\hat{Y}-Y)^2 + \lambda^2\left[E(\hat{X}_d - X)^2 + E(\hat{X} - X)^2 - 2E(\hat{X}_d - X)(\hat{X} - X)\right]$$

$$+2\lambda\left[E\left(\hat{Y}-Y\right)\left(\hat{X}_d-X\right)-E\left(\hat{Y}-Y\right)\left(\hat{X}-X\right)\right]$$

$$=V\left(\hat{Y}\right)+\lambda^2\left[V\left(\hat{X}_d\right)+V\left(\hat{X}\right)-2Cov\left(\hat{X},\hat{X}_d\right)\right]+2\lambda\left[Cov\left(\hat{Y},\hat{X}_d\right)-Cov\left(\hat{Y},\hat{X}\right)\right]$$

-------(1)

When the samples are drawn independently, the above variance reduces to

$V\left(\hat{Y}_{DD}\right)=V\left(\hat{Y}\right)+\lambda^2\left[V\left(\hat{X}_d\right)+V\left(\hat{X}\right)\right]-2\,\lambda Cov\left(\hat{Y},\hat{X}\right)$ -------(2)

THM 1: When the samples are drawn independently in two phases of sampling with the help of SRS the variance of the difference estimator is

$$V\left(\hat{Y}_{DD}\right)=N^2\left[f{S_Y}^2+\lambda^2(f+f')S_X{}^2-2\lambda fS_{XY}\right]$$

Where $f=\frac{N-n}{Nn}\ and\ f'=\frac{N-n\prime}{Nn\prime}$

Where n' and n are sample sizes corresponding to the first and second phases of sampling.

Further the min variance of difference estimator in this case is $N^2f{S_Y}^2\left[1-\left(\frac{f}{f+f'}\right)\rho^2\right]$

Where ρ is correlation coefficient between X and Y.

PROOF: when the samples are drawn independently

$$V\left(\hat{Y}_{DD}\right)=V\left(\hat{Y}\right)+\lambda^2\left[V\left(\hat{X}_d\right)+V\left(\hat{X}\right)\right]-2\,\lambda Cov\left(\hat{Y},\hat{X}\right)$$

When the samples are drawn with SRS,

$$V\left(\hat{Y}\right)=N^2\left[\frac{N-n}{Nn}\right]{S_Y}^2, V\left(\hat{X}_d\right)=N^2\left[\frac{N-n'}{Nn'}\right]S_X{}^2$$

$$V\left(\hat{X}\right)=N^2\left[\frac{N-n}{Nn}\right]S_X{}^2\,, Cov\left(\hat{Y},\hat{X}\right)=N^2\left[\frac{N-n}{Nn}\right]S_{XY}$$

$$\therefore V\left(\hat{Y}_{DD}\right)=N^2\left[\frac{N-n}{Nn}\right]{S_Y}^2+\lambda^2\left[N^2\left[\frac{N-n'}{Nn'}\right]S_X{}^2+N^2\left[\frac{N-n}{Nn}\right]S_X{}^2\right]-2\,\lambda N^2\left[\frac{N-n}{Nn}\right]S_{XY}$$

$$=N^2f{S_Y}^2+\lambda^2\left[N^2f'S_X{}^2+N^2f{S_X}^2\right]-2\,\lambda N^2fS_{XY}$$

$=N^2\left[f{S_Y}^2+\lambda^2(f+f')S_X{}^2-2\,\lambda fS_{XY}\right]$ ----(3)

Where $f=\frac{N-n}{Nn}\ and\ f'=\frac{N-n\prime}{Nn\prime}$

Differentiating the above variance partially w.r.t. λ and equating the derivative to zero, we get $\lambda = \left(\frac{f}{f+f'}\right)\frac{S_{XY}}{{S_X}^2}$

Substituting this value in (3), we get the minimum variance

$$N^2\left[f{S_Y}^2 + \frac{f^2}{(f+f')^2}(f+f')\frac{{S_{XY}}^2}{{S_X}^4}{S_X}^2 - 2\frac{f}{f+f'}\frac{{S_{XY}}^2}{{S_X}^2}fS_{XY}\right]$$

$$= N^2 f{S_Y}^2\left[1 + \frac{f}{f+f'}\frac{{S_{XY}}^2}{{S_X}^2{S_Y}^2} - 2\frac{f}{f+f'}\frac{{S_{XY}}^2}{{S_X}^2{S_Y}^2}\right]$$

$$= N^2 f{S_Y}^2\left[1 - \left(\frac{f}{f+f'}\right)\rho^2\right]$$

It is to be noted that the 2nd order derivative is always positive.

THM 2: When the second phase sample is a subsample of the first phase sample & SRS is used in both phases of sampling the variance of the difference estimator is

$$V\left(\hat{Y}_{DD}\right) = N^2\left[f{S_Y}^2 + \lambda^2(f+f'){S_X}^2 - 2\,\lambda(f-f')S_{XY}\right]$$

The minimum variance of the difference estimator in this case is $N^2{S_Y}^2[f'\rho^2 + f(1-\rho^2)]$

Where f and f' are defined as $f = \frac{N-n}{Nn}\ and\ f' = \frac{N-n'}{Nn'}$

PROOF: When the 2nd sample is a subsample of the 1st sample

$$V\left(\hat{Y}_{DD}\right) = V\left(\hat{Y}\right) + \lambda^2\left[V\left(\hat{X}_d\right) + V\left(\hat{X}\right) - 2Cov\left(\hat{X},\hat{X}_d\right)\right] + 2\lambda\left[Cov\left(\hat{Y},\hat{X}_d\right) - Cov\left(\hat{Y},\hat{X}\right)\right]$$

When the samples are drawn from SRS,

$$V\left(\hat{Y}_{DD}\right) = N^2\left[\frac{N-n}{Nn}\right]{S_Y}^2 + \lambda^2\left[N^2\left[\frac{N-n'}{Nn'}\right]{S_X}^2 + N^2\left[\frac{N-n}{Nn}\right]{S_X}^2 - 2N^2\left[\frac{N-n'}{Nn'}\right]{S_X}^2\right]$$
$$+ 2\,\lambda\left[N^2\left[\frac{N-n'}{Nn'}\right]S_{XY} - N^2\left[\frac{N-n}{Nn}\right]S_{XY}\right]$$

$$= N^2 f{S_Y}^2 + \lambda^2\left[N^2 f{S_X}^2 - N^2 f'{S_X}^2\right] + 2\,\lambda[N^2 f' S_{XY} - N^2 f S_{XY}]$$

$$= N^2\left[f{S_Y}^2 + \lambda^2(f-f'){S_X}^2 + 2\,\lambda(f'-f)S_{XY}\right] \qquad \text{-----(1)}$$

For minimum variance, differentiating w.r.t. λ and equate to zero.

$$2\,\lambda(f'-f){S_X}^2 + 2(f'-f)S_{XY} = 0$$

$\therefore \lambda(f'-f){S_X}^2 = (f'-f)S_{XY}$

$\therefore \lambda = \frac{S_{XY}}{{S_X}^2}$

Substitute the value of λ in (1), we get minimum variance

$$= N^2\left[f{S_Y}^2 + \frac{{S_{XY}}^2}{{S_X}^4}(f-f'){S_X}^2 + 2\frac{S_{XY}}{{S_X}^2}(f'-f)S_{XY}\right]$$

$$= N^2{S_Y}^2\left[f + \frac{{S_{XY}}^2}{{S_X}^2{S_Y}^2}(f-f') - 2\frac{{S_{XY}}^2}{{S_X}^2{S_Y}^2}(f-f')\right]$$

$$= N^2{S_Y}^2[f-\rho^2(f-f')]$$

$$= N^2{S_Y}^2[f'\rho^2 + f(1-\rho^2)]$$

MULTIVARIATE DIFFERENCE ESTIMATION:

When information about more than one auxiliary variable is known, the difference estimator defined can be extended in a straightforward manner.

Let $\hat{Y}, \hat{X}_1$ and $\hat{X}_2$ be unbiased for population totals Y, X_1 and X_2 of the study variable Y, the auxiliary variables X_1 and X_2 respectively. The difference estimator of the population total Y is defined as

$\hat{Y}_{D2} = \hat{Y} + \beta_1(X_1 - \hat{X}_1) + \beta_2(X_2 - \hat{X}_2)$ -----(1)

Where the constants β_1 and β_2 are predetermined.

The estimator $\hat{Y}_{D2}$ is unbiased for the population total and its variance is

$$V(\hat{Y}_{D2}) = E[\hat{Y}_{D2} - Y]^2 = E[(\hat{Y}-Y) - \beta_1(X_1 - \hat{X}_1) - \beta_2(X_2 - \hat{X}_2)]^2$$

$$= V(\hat{Y}) + {\beta_1}^2V(\hat{X}_1) + {\beta_2}^2V(\hat{X}_2) - 2\beta_1 Cov(\hat{Y},\hat{X}_1) - 2\beta_2 Cov(\hat{Y},\hat{X}_2) + 2\beta_1\beta_2 Cov(\hat{X}_1,\hat{X}_2)$$

Denote $V_0 = V(\hat{Y}), V_1 = V(\hat{X}_1), V_2 = V(\hat{X}_2), C_{01} = Cov(\hat{Y},\hat{X}_1), C_{02} = Cov(\hat{Y},\hat{X}_2)\ and$

$$C_{12} = Cov(\hat{X}_1,\hat{X}_2)$$

Differentiating the variance expression partially w.r.t. β_1 and β_2 and equating derivatives to zero, we get

$V_1\beta_1 + C_{12}\beta_2 = C_{01}$ -----(2)

$C_{12}\beta_1 + V_2\beta_2 = C_{02}$ ------(3)

Solving these two equations we get

$\beta_1 = \frac{C_{01}V_2 - C_{12}C_{02}}{V_1V_2 - {C_{12}}^2}$ -----(4)

$\beta_2 = \frac{C_{02}V_1 - C_{01}C_{12}}{V_1V_2 - {C_{12}}^2}$ ----(5)

Substituting these values in variance expression, we get

$$V(\hat{Y}_{D2}) = V_0 + \left(\frac{C_{01}V_2 - C_{12}C_{02}}{V_1V_2 - {C_{12}}^2}\right)^2 V_1 + \left(\frac{C_{02}V_1 - C_{01}C_{12}}{V_1V_2 - {C_{12}}^2}\right)^2 V_2 - 2\left(\frac{C_{01}V_2 - C_{12}C_{02}}{V_1V_2 - {C_{12}}^2}\right)C_{01} - 2\left(\frac{C_{02}V_1 - C_{01}C_{12}}{V_1V_2 - {C_{12}}^2}\right)C_{02} + 2\left(\frac{C_{01}V_2 - C_{12}C_{02}}{V_1V_2 - {C_{12}}^2}\right)\left(\frac{C_{02}V_1 - C_{01}C_{12}}{V_1V_2 - {C_{12}}^2}\right)C_{12}$$

After simplification, we get $V(\hat{Y})\left[1 - {R_{Y.X_1X_2}}^2\right]$ ---(6)

Where $R_{Y.X_1X_2}$ is the multiple correlation between $\hat{Y}$and $\hat{X}_1$and $\hat{X}_2$

Since the multiple correlation between $\hat{Y}$and $\hat{X}_1$ and $\hat{X}_2$ is always greater than the correlation between $\hat{Y}$and $\hat{X}_1$and that of $\hat{Y}$and $\hat{X}_2$we infer that the use of additional auxiliary information will always increase the efficiency of the estimator. However it should be noted that the values of β_1 and β_2 given in (4) & (5) depend on C_{01} and C_{02} which in general will not be known.

The following theorem proves that whenever b_1 and b_2 are used in place of β_1 and β_2, the resulting estimator will have mean square error that is approximately equal to minimum variance given in (6), where

$b_1 = \frac{c_{01}V_2 - c_{12}c_{02}}{V_1V_2 - {c_{12}}^2}$ ----(7)

$b_2 = \frac{c_{02}V_1 - c_{01}c_{12}}{V_1V_2 - {c_{12}}^2}$ ---(8)

Here c_{01} and c_{02} are unbiased for C_{01} and C_{02} resp.

THM 3: The approximate mean square error of the estimator

${\hat{Y}_{D2}}^* = \hat{Y} + b_1(X_1 - \hat{X}_1) + b_2(X_2 - \hat{X}_2)$ is same as that of the difference estimator defined in (1), where b_1 and b_2 are defined as in (7) & (8) resp.

PROOF: Let $e_0 = \frac{\hat{Y}-Y}{Y}, e_1 = \frac{\hat{X}_1 - X_1}{X_1}, e_2 = \frac{\hat{X}_2 - X_2}{X_2}, e_{01} = \frac{c_{01}-C_{01}}{C_{01}}, e_{02} = \frac{c_{02}-C_{02}}{C_{02}}$

It can be seen that $b_1 = \frac{[c_{01}(1+e_{01})\,V_2 - c_{12}c_{02}(1+e_{02})]}{V_1V_2 - {c_{12}}^2} = \frac{c_{01}V_2 - c_{12}c_{02}}{V_1V_2 - {c_{12}}^2} + \frac{c_{01}V_2e_{01} - c_{12}c_{02}e_{02}}{V_1V_2 - {c_{12}}^2}$

$= \beta_1 + \frac{c_{01}V_2e_{01} - c_{12}c_{02}e_{02}}{V_1V_2 - {c_{12}}^2}$ ---(9)

Similarly $b_2 = \beta_2 + \frac{c_{02}V_1e_{02} - c_{01}c_{12}e_{01}}{V_1V_2 - {c_{12}}^2}$ ----(10)

Using (9) and (10), the estimator ${\hat{Y}_{D2}}^*$ can be written as

$$\hat{Y}^*_{D2} = \hat{Y} + \left\{\beta_1 + \frac{c_{01}V_2e_{01} - c_{12}c_{02}e_{02}}{V_1V_2 - {c_{12}}^2}\right\}(X_1 - \hat{X}_1) + \left\{\beta_2 + \frac{c_{02}V_1e_{02} - c_{01}c_{12}e_{01}}{V_1V_2 - {c_{12}}^2}\right\}(X_2 - \hat{X}_2) \quad \text{---(11)}$$

Replacing $\hat{Y}, \hat{X}_1$ and $\widehat{X_2}$ by $Y(1+e_0), X_1(1+e_1)$ & $X_2(1+e_2)$ in (11) we get

$$\hat{Y}_{D2}^{*} - Y = Ye_0 + \left\{\beta_1 + \frac{c_{01}V_2 e_{01} - c_{12}c_{02}e_{02}}{V_1V_2 - {c_{12}}^2}\right\}(-X_1e_1) + \left\{\beta_2 + \frac{c_{02}V_1 e_{02} - c_{01}c_{12}e_{01}}{V_1V_2 - {c_{12}}^2}\right\}(-X_2e_2)$$

Squaring both sides and ignoring terms of degree greater than 2,

$$V(\hat{Y}_{D2}^{*}) = E\left[\hat{Y}_{D2}^{*} - Y\right]^2$$

$$= Y^2E[{e_0}^2] + {\beta_1}^2{X_1}^2E[{e_1}^2] + {\beta_2}^2{X_2}^2E[{e_2}^2] - 2\beta_1YX_1E(e_0e_1) - 2\beta_2YX_2E(e_0e_2)$$

$$= V(\hat{Y}) + {\beta_1}^2V(\hat{X}_1) + {\beta_2}^2V(\hat{X}_2) - 2\beta_1Cov(\hat{Y}, \hat{X}_1) - 2\beta_2Cov(\hat{Y}, \hat{X}_2)$$

Substituting the values available in (4) & (5), in above equation we get approximate mean square error of $\hat{Y}_{D2}^{*}$ as $V(\hat{Y})\left[1 - {R_{Y.x_1x_2}}^2\right]$

STUDY OF NON SAMPLING ERRORS:

We have seen I the theory of sampling that the true value of each unit in the population can be obtained and tabulated without any error. Accordingly one would expect that a complete enumeration of all the units in the population would give rise to data free from errors. This is not usually the case in practice. For instance, its difficult to completely avoid errors of observations. So also in the processing of data, tabulated errors may be committed affecting the final results. Errors arising in this manner are termed as non sampling errors. Thus in large scale census & survey work, occurrence of non sampling errors is not only possible but its also unavoidable. Thus data obtained in a census by complete enumeration, although free from sampling error would still be subject to non sampling error. Whereas the result of a sampling survey would be subject to sampling error as well as non sampling errors.

SOURCES OF NON SAMPLING ERRORS:

Non sampling errors may be due to following factors:

(i) Data specification being inadequate & inconsistent w.r.t the objectives of the survey.
(ii) Omission or duplication of units due to imprecise definition, incomplete or wrong identification of units, faulty method of enumerations.
(iii) Inaccurate method of interview.
(iv) Lack of trained and experienced investigators.
(v) Lack of adequate inspection and supervision of primary staff.
(vi) Inadequate scrutiny of the basic data.

(vii) Errors in data processing operations such as coding, punching, verification, tabulation etc.

(viii) Errors committed during presentation and printing of tabulated results etc.

These sources are not exhaustive, but are given to indicate some of the possible sources of erros.

NON RESPONSE ERROR:

One of the sources of error in censuses and surveys mentioned earlier is incomplete coverage in respect of units. This may occur due to refusal by respondents to give information or their being not at home, sample units being inaccessible and so on. The error in this case would arise because the set of units getting excluded may have characteristics so different from the set of units actually surveyed as to make the results biased. This type of error is termed as non response error. The non response error is not important if the characteristics of the non responding units are similar to those of responding units. But such similarity of characteristics between the two types of units is not always obtained in practice.