Unit: 3

Ratio Estimators.

Definition and Notation:

y_i - the value of the characteristics under study for the i^{th} unit of the population.

x_i - the value of the auxiliary characteristics on the same unit.

Y - the total of y characteristics of the population

X - the total of x characteristics of the population.

$R = Y/X = \bar{Y}/\bar{X}$ - the ratio of the population totals or mean of character ~~the~~ y and x.

ρ = the correlation coefficient between x and y in the population

The ratio estimators of the population ratio $Y/X = R$, the total Y, and the mean $\bar{Y}$,

$$\hat{R} = y/x = \bar{y}/\bar{x}$$

$$\hat{Y}_R = \frac{y}{x} X = \bar{y}/\bar{x} \, X = \hat{R} X.$$

$$\hat{\bar{Y}}_R = y/x \, \bar{X} = \hat{R} \bar{X}$$

Bias of Ratio estimators:

Biased Estimators are to be used when these are comparatively more efficient. There can be many illustrations where one would like to use such estimators in order to obtain higher precision without changing the cost

Theorem: 1

In Simple random sampling the bias of the ratio estimator $\hat{R}$ is given by

$$B(\hat{R}) = -\frac{Cov(\hat{R}, \bar{x})}{\bar{X}}.$$

Proof:

$$Cov(u,v) = E(uv) - E(u)E(v)$$

$$Cov(\hat{R}, \bar{x}) = E(\hat{R}\bar{x}) - E(\bar{x})E(\hat{R})$$

$$= E\left(\frac{\bar{y}}{\bar{x}} \cdot \bar{x}\right) - E(\bar{x})E(\hat{R})$$

$$= E(\bar{y}) - E(\bar{x})E(\hat{R}).$$

$$= \bar{Y} - \bar{X}E(\hat{R})$$

$$Cov(\hat{R}, \bar{x}) = \bar{x}\left(\frac{\bar{Y}}{\bar{x}} - E(\hat{R})\right)$$

$$\frac{Cov(\hat{R}, \bar{x})}{\bar{x}} = \left(\frac{\bar{Y}}{\bar{x}} - E(\hat{R})\right)$$

$$= R - E(\hat{R}) = -B(\hat{R}).$$

Corollary: 1

$$\frac{|B(\hat{R})|}{\sigma_{\hat{R}}} \leq CV(\bar{x}).$$

CV - ~~correlation~~ coefficient of Variation.

Corollary: 2

$$B(\hat{Y}_R) = \bar{X}B(\hat{R}) \text{ and } B(\hat{\bar{Y}}_R) = \bar{X}B(\hat{R}).$$

Theorem:- 2

S.T the first approximation to the relative bias of the ratio estimator in simple random sampling, wor, is given by.

$$\frac{B(\hat{R})}{R} \cong \frac{1-f}{n\bar{x}\bar{y}} (RS_x^2 - \rho S_x S_y)$$

$$\cong \frac{1-f}{n} (C_x^2 - \rho C_x C_y).$$

Where $C_x = S_x/\bar{x}$; $C_y = S_y/\bar{y}$ are the ~~correlation~~ coefficient of variation of x and y respectively

Proof:

$$\hat{R} - R = \frac{\bar{y}}{\bar{x}} - R = \frac{\bar{y} - R\bar{x}}{\bar{x}} = \frac{\bar{y} - R\bar{x}}{\bar{X} + (\bar{x} - \bar{X})}$$

$$= \frac{1}{\bar{X}} (\bar{y} - R\bar{x}) \left(1 + \frac{\bar{x} - \bar{X}}{\bar{X}}\right)^{-1}$$

Expanding by a Taylor's series we get

$$\hat{R} - R \cong \frac{1}{\bar{X}} (\bar{y} - R\bar{x}) \left(1 - \frac{\bar{x} - \bar{X}}{\bar{X}} + \dots\right)$$

Ignoring the second and higher orders.

$$\hat{R} - R \cong \frac{1}{\bar{X}} \left[(\bar{y} - R\bar{x}) - \frac{1}{\bar{X}} E(\bar{y} - R\bar{x})(\bar{x} - \bar{X})\right]$$

$E(\bar{y} - R\bar{x}) = \bar{Y} - R\bar{X} = 0$ respectively.

$$E(\bar{y} - R\bar{x})(\bar{x} - \bar{X}) = E[\bar{y}\cdot(\bar{x} - \bar{X}) - RE\{\bar{x}\cdot(\bar{x} - \bar{X})\}$$

$$= E[(\bar{y} - \bar{Y})(\bar{x} - \bar{X}) - RE(\bar{x} - \bar{X})^2$$

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

$$B(\hat{R}) = \left\{\frac{1-f}{n\bar{x}^2}\right\}(R s_x^2 - \rho s_x s_y)$$

$$= \frac{1-f}{n\bar{x}\bar{y}}(R^2 s_x^2 - \rho s_x s_y)$$

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

$$= \frac{\Sigma(\bar{y}-\bar{Y})(\bar{x}-\bar{X})}{n} - R\,\frac{\Sigma(\bar{x}-\bar{X})^2}{n}$$

$$\left[\rho = \frac{\Sigma(x_i-\bar{X})(y_i-\bar{Y})}{n\,s_x s_y}\right.$$

$$\left.\rho\, s_x s_y = \frac{\Sigma(x_i-\bar{X})\,\Sigma(y_i-\bar{Y})}{n}\right]$$

$$B(\hat{R}) = \frac{1-f}{n\bar{x}\bar{y}}\left[\rho s_x s_y - R s_x^2\right]$$

$$B(\hat{R}) = \frac{(1-f)}{n\bar{x}\bar{y}}\left[R s_x^2 - \rho s_x s_y\right] \quad \left[\because R = \frac{\bar{y}}{\bar{x}}\right]$$

$$= \frac{1-f}{n}\left[\frac{s_x^2}{\frac{\bar{x}}{\bar{y}} \times (\bar{x}\bar{y})} - \rho\frac{s_x s_y}{\bar{x}\bar{y}}\right]$$

$$= \frac{1-f}{n}\left[\frac{s_x^2}{\bar{x}^2} - \frac{\rho s_x s_y}{\bar{x}\bar{y}}\right]$$

$$= \frac{1-f}{n}\left[C_x^2 - \rho C_x C_y\right] //.$$

Approximate Variance of Ratio Estimator

Theorem: 3

In SRSWOR for large n $V(\hat{R}) = \frac{1-f}{n\bar{x}^2} \frac{\sum_{i=1}^{N} (y_i - Rx_i)^2}{N-1}$

Pf:

$$V(\hat{R}) = E\left[\hat{R} - E(\hat{R})\right]^2$$

$$= E\left(\frac{\bar{y}}{\bar{x}} - R\right)^2$$

$$= E\left[\frac{\bar{y} - R\bar{x}}{\bar{x}}\right]^2$$

$$= \frac{E(\bar{y} - R\bar{x})^2}{E(\bar{x})^2} \qquad E(\bar{x})^2 = \bar{x}^2$$

$$V(\hat{R}) = \frac{E(\bar{y} - R\bar{x})^2}{\bar{x}^2}$$

let $u_i = y_i - Rx_i$

$\bar{y} = \bar{y} - R\bar{x}$ for sample

$\bar{v} = \bar{y} - R\bar{x}$ for population

$\bar{v} = \bar{y} - \frac{\bar{y}}{\bar{x}}\bar{x} = 0.$

$$V(\hat{R}) = \frac{E(\bar{u} - \bar{v})^2}{\bar{x}^2}$$

In SRSWOR, $V(\bar{y}) = (1-f)\, S^2/n$

Here $S^2 = \sum \frac{(y_i - \bar{y})^2}{N-1}$

$$V(\hat{R}) = \frac{1}{\bar{x}^2} \frac{(1-f)}{n} \sum_{i=1}^{N} \frac{(\bar{y}_i - \bar{y})^2}{N-1}$$

$$= \frac{1-f}{n\bar{x}^2} \sum_{i=1}^{N} \frac{(y_i - Rx_i)^2}{N-1}$$

COROLLARY:

$$V(\hat{\bar{y}}) = V(\hat{R}\bar{x})$$

$$= \bar{x}^2 V(\hat{R})$$

$$= \bar{x}^2 \frac{1-f}{n\bar{x}^2} \sum_{i=1}^{N} \frac{(y_i - Rx_i)^2}{N-1}$$

$$V(\hat{\bar{y}}) = \frac{1-f}{n} \sum_{i=1}^{N} \frac{(y_i - Rx_i)^2}{N-1}$$

Theorem. 4

Show that 1st order of approximate variance of $\hat{R}$ can be expressed.

$$V(\hat{R}) = \frac{1-f}{n\bar{x}^2} [S_y^2 + R^2 S_x^2 - 2\rho R S_x S_y]$$

$$= \frac{(1-f)R^2}{n} [C_y^2 + C_x^2 - 2\rho C_x C_y].$$

Pf:

Since $\bar{y} = R\bar{x}$

$$y_i = y_i - Rx_i = (y_i - \bar{y}) - R(x_i - \bar{x})$$

Then,

$$V(\hat{R}) = \frac{1-f}{n\bar{x}^2(N-1)} \sum_{i=1}^{N} [(y_i - \bar{y}) - R(x_i - \bar{x})]^2$$

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

$$S_y^2 = \frac{\Sigma (y_i - \bar{y})^2}{N-1} \quad S_x^2 = \frac{\Sigma (x_i - \bar{x})^2}{N-1}$$

$$V(\hat{R}) = \frac{1-f}{n\bar{x}} \left[S_y^2 + R^2 S_x^2 - 2C S_x S_y \right]$$

$$= \frac{(1-f)R^2}{n} \left[C_y^2 + C_x^2 - 2\rho C_x C_y \right]$$

Then,

$$\bar{Y} = R\bar{X}$$

$$V(\hat{R}) \simeq \frac{(1-f)}{n\bar{x}^2} \frac{\sum (y_i - Rx_i)^2}{N-1}$$

$$= \frac{(1-f)}{n\bar{x}^2(N-1)} \sum (y_i - Rx_i)^2$$

$$\Rightarrow \boxed{u_i = y_i - Rx_i = (y_i - \bar{Y}) - R(x_i - \bar{X})}$$

Now $V(\hat{R}) = \frac{(1-f)}{n\bar{x}^2(N-1)} \sum [(y_i - \bar{Y}) - R(x_i - \bar{X})]^2$

$$\frac{1-f}{n\bar{x}^2(N-1)} \left[\sum (y_i - \bar{y})^2 + R^2 \sum (x_i - \bar{x})^2 - 2R \sum (x_i - \bar{x})(y_i - \bar{y}) \right]$$

$$= \frac{1-f}{n\bar{x}^2} \left[\frac{\sum (y_i - \bar{y})^2}{N-1} + R^2 \frac{\sum (x_i - \bar{x})^2}{N-1} - 2R \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1} \right]$$

Cone coff.

$$\rho = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{(N-1) S_x S_y}$$

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

$$= \frac{(1-f)}{n\bar{x}^2} R^2 \left[S^2y/R^2 + S^2x - 2\rho \frac{S_x S_y}{R} \right]$$

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

$$= \frac{(1-f)}{n} R^2 \left[\frac{S^2y}{\bar{x}^2\bar{y}^2}\bar{x}^2 + \frac{S^2x}{\bar{x}^2} - 2\rho \frac{S_x S_y}{\bar{y}\cdot\bar{x}^2}\bar{x} \right]$$

$$= \frac{1-f}{n} R^2 \left[S^2y/\bar{y}^2 + \frac{S^2x}{\bar{x}^2} - 2\rho \frac{S_x \cdot S_y}{\bar{x}\cdot\bar{Y}} \right]$$

$$= \frac{1-f}{n} R^2 \left[C^2y + C^2x - 2\rho C_x C_y \right].$$

Corollary:

$$V(\bar{y}_R) = \frac{1-f}{n} \left[S^2y + R^2 S^2x - 2R\rho S_x S_y \right]$$

$$= \frac{1-f}{n} \left[C_y^2 \bar{y}^2 + R^2 C_x^2 \bar{x}^2 - 2R\rho C_x \bar{x} C_y \bar{y} \right]$$

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

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

Corollary:

Ratio Estimators in Stratified Sampling:

Separate Ratio estimator:

If y_m, x_m are the sample totals in the m^{th} stratum and X_m is the m^{th} stratum total. We may define the estimator $\hat{Y}_{rs}$ (S for separate) as,

$$\hat{Y}_{RS} = \sum_m \frac{y_m}{x_m} \cdot X_m = \sum_m \bar{y}_m / \bar{x}_m \cdot X_m.$$

Theorem: 5 If the sample size n_m is moderate in all strata & SRS, wor, is done independently within each stratatum, show that $\hat{Y}_R$ is a biased estimator with negligible bias and its sampling variance,.

$$V(\hat{R}_{YS}) = \sum_m N_m^2 \frac{(1-f_m)}{n_m} (S_{ym}^2 + R_m^2 S_{xm}^2 - 2R_m \rho_m S_{ym} S_{xm})$$

where $R_m = Y_m / X_m$ and ρ_m are true ratio and coefficient of correlation, resp. in the m^{th} stratum.

Proof:

$$\hat{Y}_{RM} = \bar{y}_m / \bar{x}_m \cdot X_m$$

$$\hat{Y}_{RS} = \sum_m \hat{Y}_{RM}$$

$$\frac{B(\hat{Y})}{Y} = \frac{1-f}{n} (C_x^2 - \rho C_x C_y) \quad [\text{To refer Theorem: 2}]$$

$$= \sum_m \frac{(1-f)}{n_m} (C_{xm}^2 - \rho C_{xm} C_{ym}).$$

$$V(\hat{Y}_{RS}) = V\left(\sum_m \hat{Y}_{RM}\right)$$

$$= \sum_m V(\hat{Y}_{Rm})$$

$$= \sum_m N_m^2 \frac{(1-f)}{n_m} (S_{ym}^2 + R_m^2 S_{xm}^2$$

$$= \sum_m N_m^2 \frac{(1-f_m)}{n_m} \left(S^2_{y_m} + R_m^2 S^2_{x_m} - 2R_m f_m S_{y_m} S_{x_m} \right).$$

COROLLARY

In stratified random sampling, usor, almost unbiased estimator of $V(\hat{\bar{y}}_{RS})$ is given by

$$v(\hat{\bar{y}}_{RS}) = \sum_m N_m^2 \frac{(1-f_m)}{n_m} \left(s^2_{y_m} + R_m^2 s^2_{x_m} - 2R_m f_m s_{xy_m} \right).$$

Combined Ratio estimator:

It was assumed, in case of separate estimator, that nm's were large in each stratum. However, it may not hold good always in practice. To overcome this difficulty, Hansen, Hurwitz and Gruney (1946) suggested a combined ratio estimator $\hat{\bar{y}}_{RC}$ (c for combined) for a sample from a stratified population as.

$$\hat{\bar{y}}_{RC} = \frac{\hat{\bar{y}}_{st}}{\hat{\bar{x}}_{st}} \cdot X = \frac{\bar{y}_{st}}{\bar{x}_{st}} \cdot X$$

where $\bar{y}_{st} = \frac{\hat{\bar{y}}_{st}}{N} = \sum_m \frac{N_m}{N} \bar{y}_m$

$$\bar{x}_{st} = \frac{\hat{\bar{x}}_{st}}{N} = \sum_m \frac{N_m}{N} \bar{x}_m.$$

are estimated population means from a stratified sample and X is overall total of x.

Theorem: 6

If the total sample size n is large and simple random sampling, wor, is done in each stratum independently, then $\hat{Y}_{RC}$ is a consistent estimator and its sampling variance is given by

$$V(\hat{Y}_{RC}) = \sum_m N_m^2 \frac{(1-f_m)}{n_m} (S_{ym}^2 + R^2 S_{xm}^2 - 2R\rho_m S_{ym} S_{xm})$$

Proof:

Since R is a consistent estimator, $\hat{Y}_{RC}$ is a ratio estimator and therefore, also a consistent estimator of the population total, to derive its bias.

$$\hat{Y}_{RC} - Y = \frac{N\bar{x}}{\bar{x}_{st}} (\bar{y}_{st} - R\bar{x}_{st}) \cong N(\bar{y}_{st} - R\bar{x}_{st})$$

Let us define a variate $U_{mj} = y_{mj} - Rx_{mj}$

$$\bar{u}_{st} = \bar{y}_{st} - R\bar{x}_{st} \quad \text{and.}$$

$$U = \bar{Y} - R\bar{X} = 0.$$

$$E(\bar{y} - R\bar{x})(\bar{x} - \bar{X}) = E[\bar{y} \cdot (\bar{x} - \bar{X})] - R E[\bar{x} \cdot (\bar{x} - \bar{X})]$$

$$= E(\bar{y} - \bar{Y})(\bar{x} - \bar{X}) - R E(\bar{x} - \bar{X})(\bar{x} - \bar{X})$$

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

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

$$B(\hat{Y}_{RC}) = \sum_m \frac{(1-f)}{n\bar{X}^2} \left[\frac{R S_x^2}{\bar{X}} - \rho S_x S_y \right]$$

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

$$\sum_m \frac{1-f}{n} N \left[\frac{S_x^2}{\bar{X}^2} - \frac{\rho S_x S_y}{\frac{\bar{Y}}{\bar{X}} \cdot \bar{X}^2} \right]$$

$$B(\hat{Y}_{RC}) = \sum_{i}^{m} \frac{1-f_m}{n_m} W_i^2 y_i \left(\frac{S_x^2}{\bar{x}} - \rho \frac{S_x S_y}{\bar{x}\,\bar{y}} \right)$$

$$\frac{B(\hat{Y}_{RC})}{\bar{Y}} = \sum_{m} \frac{(1-f_m)}{n_m} W_m^2 \left(\frac{S_{xm}^2}{\bar{x}^2} - \frac{\rho_m S_{xm} S_{ym}}{\bar{x}\,\bar{y}} \right)$$

$$\therefore C_x = S_x/\bar{x} \; ; \; C_y = S_y/\bar{y}$$

$$B = \frac{1-f}{n} (C_x^2 - \rho C_x C_y)$$

$$V(\hat{Y}_{RC}) = V(N \bar{y}_{st}) = N^2 V(\bar{y}_{st})$$

$$= \sum_{m} N_m^2 \underbrace{\frac{(1-f_m)}{n_m} S_u^2}_{V(\hat{R})}$$

$$V(\hat{R}) = \frac{1-f}{n\bar{x}^2(N-1)} \sum_{i}^{N} \left((y_i - \bar{y}) - R(x_i - \bar{x}) \right)^2$$

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

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

$$\left[\rho = \sum_{i}^{N} \frac{(y_i - \bar{y})(x_i - \bar{x})}{(N-1) S_x S_y} \right]$$

$$\left[\rho (n-1) S_x S_y = \sum_{i}^{N} (y_i - \bar{y})(x_i - \bar{x}) \right]$$

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

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

$$= \frac{1-f}{n} R^2 \left[\frac{S_y^2}{\bar{Y}^2 \cdot \bar{x}^2} + \frac{S^2_x}{\bar{x}^2} - \frac{2\rho S_x S_y}{\bar{Y}\bar{x} \cdot \bar{x}^2} \right]$$

$$= \frac{1-f}{n} R^2 \left[\frac{S_y^2}{\bar{Y}^2} + \frac{S^2_x}{\bar{x}^2} - \frac{2\rho S_x S_y}{\bar{y}\bar{x}} \right]$$

$$V(\hat{Y}_{RC}) = \sum_m N_m^2 \frac{(1-f_m)}{n_m} \left[\frac{S^2_{ym}}{\bar{Y}_m^2} + \frac{S^2_{xm}}{\bar{x}_m^2} - \frac{2\rho S_{xm} S_{ym}}{\bar{x}_m \bar{Y}_m} \right]$$

(or)

$$V(\hat{Y}_{RC}) = \sum_m N_m^2 \frac{(1-f_m)}{n_m} \left[S^2_{ym} + R^2 S^2_{xm} - 2R\rho_m S_{xm} S_{ym} \right]$$

Thus proved.

Comparison of separate and combined Ratio estimators:-

$$V(\hat{Y}_{RC}) - V(\hat{Y}_{RS}) = \sum_m N_m^2 \frac{(1-f_m)}{n_m} \left[(R^2 - R_m^2) S^2_{xm} - 2(R - R_m) \rho_m S_{ym} S_{xm} \right]$$

$$= \sum_m N_m^2 \frac{(1-f_m)}{n_m} \left[(R - R_m)^2 S^2_{xm} + 2(R - R_m)(\rho_m S_{ym} S_{xm} - R_m S_{xm}) \right]$$

Since the last term is small, the right side is likely to be positive. Thus, the separate ratio estimator is expected to be more precise provided the sample in each stratum is large enough for the approximate variance formula to be applied. The combined estimator will have a large variance as compared to the separate estimator, but the bias in the former is expected to be smaller than in the latter. Unless the population ratio of y to x in different strata varies considerably, the combined estimator will usually have negligible bias and precision will be as high as the separate estimator. In cases when the line of regression of y on x passes through the origin within

each stratum, the separate ratio estimator will be more precise than the combined one. Hence the rules for choice between the two methods should be

⇒ 1. If the sample taken from each stratum is small the combined ratio estimator should be used unless there is a wide difference between the strata ratios R_m. When there are such difference and regrouping of the strata is possible so that each group will not differ much and has a large sample size, the separate ratio estimator should be used.

2. If the sample size in each stratum is large so that the approximate variance formula can be applied, then it is better to use the separate ratio estimator unless it involves extra calculation work. If it does, then one should make sure of the gain before actually using it.

3. If X_m are known independently from stratum to stratum the separate ratio estimator may be used as a method of estimation. When $X_m = N_m$, the estimator becomes ~~are known independently~~ equivalent to the stratified estimator and there is no problem of bias from these estimates.

4. If the sampling units are the elementary units and the denominator of the ratio is simply the number of elementary units in the sample, both the estimators are unbiased. In such cases, the estimator should be chosen by consideration of the variance and gain to be obtained by using it.

Regression Estimators:-

Linear regression estimators also make use of auxiliary information for increasing precision. It was been that the ratio estimator provides a precise estimate of the population mean if regression is linear and the line does not go through the origin, it is better to use estimator based on linear regression. In other words, if the study variate (y) is approximately a constant and a multiple of the auxiliary variate, it is more precise to estimate the population mean or total by fitting a linear regression. such an estimator is called a Regression Estimator.

Difference Estimator:-

Let y and x be correlated characteristics. to estimate $\bar{Y}$, if from a ~~SRS~~ SRS, we obtain the unbiased estimator $\bar{y}$ and $\bar{x}$ of $\bar{Y}$ and $\bar{X}$ respectively, then we can improve upon the estimator $\bar{y}$ by introducing a difference function. Thus a simple difference estimator is defined by,

$$\hat{y}_D = \bar{y} + (\bar{X} - \bar{x})$$

It is assumed that there is a unit change in y when a unit change a made in x. x and y having equal variances. This assumption not be valid if the relationship is of the type $y = cx + k$. where c & k are constants. In Difference Estimator is defined as,

$$\hat{y}_D = \bar{y} + c(\bar{X} - \bar{x}).$$

c & k $\Rightarrow$ are known Quantities.

Theorem: In SRS, wor, the sampling variance of $\bar{y}_D$ is obtained by.

$$V(\bar{y}_D) = \frac{1-f}{n}\left[S_y^2 + c^2 S_x^2 - 2\rho c S_x S_y\right)$$

where c is some specified quantity.

Proof:

Since c is a given constant and $E(\bar{x} - \bar{x}) = 0$ the unbiased of $\bar{y}_D$ is proved.

$$u_i = y_i - c(x_i - \bar{x})$$

$$\bar{u} = \bar{y} - c(\bar{x} - \bar{x}),$$

$$\bar{U} = \bar{Y} - c(\bar{X} - \bar{X}) = \bar{Y}.$$

$$V(\bar{y}_D) = V(\bar{u}) = \frac{1-f}{n} S_u^2. \qquad S_x^2 + c^2 S_x^2 - 2c\rho S_y S_x$$

where $S_u^2 = \dfrac{\sum_i^N (u_i - \bar{U})^2}{(N-1)}$

$$= \frac{\sum_i^N \left[(\bar{y} - c(x_i - \bar{x})) - \bar{Y}\right]^2}{N-1}$$

$$= \frac{\sum_i^N \left[(\bar{y} - \bar{Y}) - c(\bar{x} - \bar{X})\right]^2}{N-1}$$

$$= \frac{1}{N-1}\left[\sum_i^N (\bar{y} - \bar{Y})^2 + c^2 \sum_i^N (\bar{x} - \bar{X})^2 - 2c \sum_i^N (\bar{x} - \bar{X})(\bar{y} - \bar{Y})\right]$$

$$= \sum_{i}^{N} \frac{(y_i - \bar{Y})^2}{N-1} + c^2 \sum_{i}^{N} \frac{(x_i - \bar{X})^2}{N-1} - 2c\rho S_x S_y]$$

$$S_u^2 = S_y^2 + c^2 S_x^2 - 2c\rho S_x S_y]$$

$$V(\hat{\bar{Y}}_D) = \frac{1-f}{n} [S_y^2 + c^2 S_x^2 - 2c\rho S_x S_y].$$

Hence Proved.

Corollary:-

1. For the case $c = R (= \bar{Y}/\bar{X})$, the variance of the difference estimator $\bar{y}_D$ is exactly the same as the first order approximation of $V(\hat{\bar{Y}}_R)$

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

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

2. In SRS, WOR, an unbiased estimator of $V(\bar{y}_D)$ is obtained by,

$$v(\bar{y}_D) = v(\bar{y}) + c^2 v(\bar{x}) - 2c \, Cov(\bar{y}, \bar{x})$$

$$= \frac{1-f}{n} (S_y^2 + c^2 S_x^2 - 2c S_{yx})$$

Regression Estimator:

While discussing the difference estimator, it was seen that the optimum value to be given to c is β, where β is the regression coefficient of y and x. Generally, β is not known in the advance and its value in estimate from the sample. Suppose y_i and x_i are obtained for each unit at the sample, then a least square estimate of β is,

$$b = \frac{\sum_{i}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i}^{n} (x_i - \bar{x})^2}$$

Thus Linear regression estimator of the population mean $\bar{Y}$, are population mean Y are given by,

$$\bar{y}_1 = \bar{y} + b(\bar{X} - \bar{x})$$

$$\hat{Y}_1 = N[\bar{y} + b(\bar{X} - \bar{x})]$$

$$\hat{Y}_1 = N \hat{y}_1$$

Therefore, we shall present the theory of linear regression estimator $\bar{y}_i$ only. Since b is random variate exact expression of the mean and variance of the regression estimator are difficult to derive. large sample approximation to its bias and sampling variance will be given.

Bias of Regression Estimator:-

The regression estimator $\bar{y}_1$ is biased since,

i) β is generally estimated by taking the ratio of the estimate cov $(\bar{y}, \bar{x})$ to that $V(\bar{x})$

ii) It involves the product of 2 variables viz. $b\bar{x}$. The bias regression estimator will usually be trivial and will decrease as sample size increase.

Theorem: In SRS, the bias of $\hat{y}_i$ is a approximate by. $B(\bar{y}_i) \cong - \text{cov}(\bar{x}, b)$

Which will be negligible if the sample size is ~~less~~ large.

Proof:

Suppose $\bar{y} = \bar{Y}(1+e)$

$\bar{x} = \bar{X}(1+e_1)$

$b = \beta(1+e_2)$

where $E(e) = E(e_1) = E(e_2) = 0.$

$\bar{y}_1 = \bar{y} + b(\bar{x} - \bar{x})$

$$\bar{y}_1 = \bar{y}(1+e) + \beta(1+e_2)(\bar{x} - \bar{x}(1+e_1))$$

$$= \bar{y} + e\bar{y} + (\beta + \beta e_2)(\bar{x} - \bar{x} - e_1\bar{x})$$

$$= \bar{y} + e\bar{y} + (\bar{x}\beta - \bar{x}\beta - \beta e_1\bar{x} + \beta e_2\bar{x} - \beta e_2\bar{x} - \beta e_1 e_2\bar{x})$$

$$= \bar{y} + e\bar{y} - \beta e_1\bar{x} - \beta e_1 e_2\bar{x}$$

$$\bar{y}_1 = \bar{y} + (e\bar{y} - e_1\beta\bar{x}) - e_1 e_2 \beta\bar{x}$$

∴ The bias of ~~estimator~~ regression estimator is

$$B(\bar{y}_1) = \bar{y}E(e) - \beta\bar{x}E(e_1) - \beta\bar{x}E(e_1 e_2)$$

$$B(\bar{y}_1) = -\text{cov}(\bar{x}, b) \quad [\because E(e) = E(e_1) = 0.]$$

For large samples, usually $\text{cov}(\bar{x}, b)$ decreases. It becomes zero if the joint distribution of y and x is a bivariate normal.

Sampling Variance of Regression Estimator:-

Theorem: In SRS, the large sample ~~variable~~ variance of the regression estimator is given by

$$V(\bar{y}_1) = V(\bar{y}) + \beta^2 V(\bar{x}) - 2\beta\,\text{cov}(\bar{y}, \bar{x}).$$

PF:

$$\bar{y}_1 = \bar{y} + (e\bar{y} - e_1\beta\bar{x}) - e_1 e_2 \beta\bar{x}$$

$$\bar{y}_1 - \bar{y} = e\bar{y} - e_1\beta\bar{x} - e_1 e_2 \beta\bar{x}$$

If the term involving $e_1 e_2$ is ignoring, we have

$$y_1 - \bar{y} = e\bar{y} - e_1\beta\bar{x}$$

Therefore,

$$V(\bar{y}_i) = V(e\bar{Y} - e_i \beta \bar{X})$$

$$= V(e\bar{Y}) + \beta^2 V(e_i\bar{X}) - 2\beta \operatorname{cov}(e\bar{Y}, e\bar{X})$$

$$V(\bar{y}_i) = V(\bar{y}) + \beta^2 V(\bar{x}) - 2\beta \operatorname{cov}(\bar{y}, \bar{x}).$$

Combined with the mean per unit and Ratio estimators

For a large ~~numbers~~ sample size n, the variance for ~~minimal~~ estimators of the population mean $\bar{Y}$,

$$V(\bar{y}) = \frac{1-f}{n} S_y^2 \quad \text{(mean per unit)}$$

$$V(\bar{y}_R) = \frac{1-f}{n} [S_y^2 - 2R\rho S_x S_y + R^2 S_x^2] \quad \text{(ratio)}.$$

$$V(\bar{y}_i) = \frac{1-f}{n} (1-\rho^2) S_y^2. \quad \text{(regression)}.$$

Finally, comparing with the mean per unit estimator, we observe that the variance of the regression estimator always smaller unless $\rho = 0$ in case $\rho = 0$ the variation for both are equal. The reduction in variance is large when ρ is high and small when ρ is low.

Comparing next with the ratio estimator, it can be seen that the variances of the regression estimator is less than that of the ratio estimator if

$$(\rho S_y - R S_x)^2 > 0$$

$$(\beta - R)^2 > 0$$

which is always true unless $\beta = R$. In this situation, both estimators have the same variance and this occurs only when the regression of y on x is a straight line through the origin.

The comparision suggests that all these estimators belong to the class of estimators.

$$\bar{y}_R = \bar{y} + z(\bar{X} - \bar{x}).$$

where z is a random variate having some values in a finite range.

Some rules for a choice among alternatives are as follows:

1. When advance information on an approximate value of $\beta(=c)$ is available, then with simple computations good results can be obtained with such value of c with the difference estimator.

2. When the correlation coefficient between variates is nearly equal to the ratio of their S.D ie, $\rho \cong \frac{\sigma_x}{\sigma_y}$. the difference estimator with $c=1$ will attain equally precise results as the regression estimator.

3. When the correlation coefficient between variates is nearly equation to the ratio of their coefficients of variation ie, $\rho \cong C_x/C_y$, the ratio estimator will provide equally precise results as the regression estimator.

4. When ρ is different form σ_x/σ_y the regression estimator should be preferred. In this situation, the difference estimator with $c=1$ and not provide precise results.

5. When ρ is different form C_x/C_y the regression estimator should be preferred. In this situation, the variance of the ratio estimator will be longer than that of regression estimator.

6. When computations for the regression estimator are heavy, time consuming and expensive, its uses is recommended only if the gains from such computations are such more significant than the additional cases.

Regression Estimators in stratified Sampling:-

Nm units * SRSWOR like the ratio estimator, two regression are possible Separate regression Estimator.

* The Combined regression Estimator which is obtained by getting common regression coefficient $\forall$ the strata.

Separate Regression Estimator:-

Theorem:- If sampling is independent in different strata are sample size is large enough in each stratum, then $\bar{y}_{lr}$ is ~~biased~~. is an almost unbiased estimator. To the 1st order of approximation its variance is given by,

$$V(\bar{y}_{lr}) = \sum_{m}^{k} W_m^2 \frac{(1-f_m)}{n_m} \left[S_{ym}^2 - 2 b_m \rho_m S_{ym} S_{xm} + b_m^2 S_{xm}^2 \right]$$

Proof:

$$u_i = y_i - b(x_i - \bar{x}), \quad \bar{u} = \bar{y} - b(\bar{x} - \bar{x})$$

$$\bar{u} = \bar{y} - b(\bar{x} - \bar{x})$$

$$V(\bar{y}_{lr}) = V(\bar{u}) = \frac{1-f}{n} S_u^2$$

$$S_u^2 = \frac{\sum_i^N (u_i - \bar{u})^2}{N-1}$$

$$= \frac{\sum_i^N [(\bar{y} - \bar{y})^2 - b(\bar{x} - \bar{x})^2]^2}{N-1}$$

$$= \frac{1}{N-1} \left[\sum_i^N (\bar{y} - \bar{y})^2 + b^2 \sum_i^N (\bar{x} - \bar{x})^2 - 2b \sum (\bar{x} - \bar{x})(\bar{y} - \bar{y}) \right]$$

$$= \left[\sum_i^N \frac{(\bar{y} - \bar{y})^2}{N-1} + b^2 \sum_i^N \frac{(\bar{x} - \bar{x})^2}{N-1} - 2b \frac{\sum (\bar{x} - \bar{x})(\bar{y} - \bar{y})}{N-1} \right]$$

$$S_u^2 = \left[S_{y_{res}}^2 + b^2 S_x^2 - 2b\rho S_x S_y \right]$$

$$V(\bar{y}_{lr}) = \sum_m^k w_m^2 \frac{(1-f_m)}{n_m} \left[S_{ym}^2 + b_m^2 S_{xm}^2 - 2 b_m \rho_m S_{xm} S_{ym} \right].$$

Thus Proved.

Combined Regression Estimator:-

Theorem:- If Sampling is independent in different strata and sample size is large enough in each stratum, the variance of $\bar{y}_{lc}$ is given by,

$$V(\bar{y}_{lc}) = \sum_{m}^{k} W_m^2 \frac{(1-f_m)}{n_m} (S_{ym}^2 - 2b\rho_m S_{ym} S_{xm} + b^2 S_{xm}^2)$$

[The proof is previous theorem]

Comparision of Combined and Separate Estimators:-

Some caution is required when deciding whether the separate or combined regression estimator is to be used at any specific situation. The bias in the separate regression estimator is likely to be larger when sample size in each stratum is small. On the other hand, the defect of the combined estimator is that it variance is inflated, if the population regression coeff. differ from stratum to stratum. If the regression are approximately linear and the regression coeff. appear to be the same in all strata, use of the combined regression estimator is preferred. If the regression are linear but the coeff. differ from stratum to stratum, it is advisable to adopt the separate estimator. Though the separate regression estimator is likely to be more efficient than the combined regression estimator, the latter is likely to have smaller bias.