CHAPTER 7

Ratio and Product Methods of Estimation

7.1 NEED FOR RATIO ESTIMATION

In the preceding chapters, we have discussed some methods of using information on an auxiliary variable for improving the precision of the estimates of population mean/total. In chapter 4, the selection probabilities for the population units were determined from the measures of size provided by such supplementary information. Also, the use of information on the auxiliary variable for the purpose of stratification has been discussed in chapter 5. In this chapter, and also in the following chapter, we present some other estimators that make use of auxiliary information for achieving higher efficiency.

In socio-economic surveys, one may be interested in estimating ratios like per capita income or expenditure. Similarly, estimation of yield per unit area, or the use of fertilizer/pesticides per hectare for a particular crop, could be of importance in case of agricultural surveys. Estimation of percent relative fall in real estate prices or input-output ratio, are useful for industry and commerce.

Population ratio R is the ratio of two population parameters. Mostly, these parameters are population totals or means. For instance, the yield per hectare $R=Y/X$, where Y is the total production and X the corresponding total area under the crop. The ratio R is usually estimated by the ratio of unbiased estimators $\hat{Y}$and $\hat{X}$ of Y and X respectively. It could also be equivalently estimated by $\hat{R}=\bar{y}/\bar{x}$, where $\bar{y}$ and $\bar{x}$ unbiasedly estimate the population means $\bar{Y}$ and $\bar{X}$ respectively. Here x, the area under the crop, is treated as the auxiliary variable. The estimators of such population ratios are known as *ratio estimators.*

The estimators of population ratio R can also be used for building up the estimators of population mean/total for the study variable y. In situations, where the study variable y is highly correlated with the auxiliary variable x, and the two are also approximately proportional, the ratio of y to x is expected to be less variable than the y's themselves. In such a situation, it would, therefore, be better to estimate R from the sample and multiply the estimator of R with the known population mean/total of the auxiliary variable x, to obtain an estimator for the population mean/total of the study variable y. The estimator, so obtained, is also called ratio estimator of population mean $\bar{Y}$ or the total Y.

We first consider the estimator of population ratio R. The estimators for mean/ total shall be discussed subsequently.

7.2 ESTIMATION OF POPULATION RATIO

Suppose that a WOR simple random sample of n units is drawn from a population of N units to estimate the population ratio

$$R = \frac{Y}{X} \tag{7.1}$$

where, as mentioned in the preceding section, $Y = \Sigma\, Y_i$ and $X = \Sigma\, X_i$, $i = 1, 2, ..., N$, are the population totals for the estimation variable y and the auxiliary variable x respectively. We assume that the population mean $\bar{X} = X/N$ is known. All the sample units are then observed for the variables y and x. Let $(y_1, x_1), (y_2, x_2), ..., (y_n, x_n)$ denote the set of these observations, whereas $\bar{y}$ and $\bar{x}$ represent the corresponding sample means.

As defined earlier, the population mean squares and product are given by

$$\left.\begin{aligned} S_y^2 &= \frac{1}{N-1}\left(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right) \\ S_x^2 &= \frac{1}{N-1}\left(\sum_{i=1}^{N} X_i^2 - N\bar{X}^2\right) \\ S_{xy} &= \frac{1}{N-1}\left(\sum_{i=1}^{N} X_i Y_i - N\bar{X}\,\bar{Y}\right) \end{aligned}\right] \tag{7.2}$$

Also, their respective sample estimators are

$$\left.\begin{aligned} s_y^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) \\ s_x^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \\ s_{xy} &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}\right) \end{aligned}\right] \tag{7.3}$$

Further, let ρ and r be the population and sample correlation coefficients between the variables y and x respectively. We then have :

Estimator of population ratio R :

$$\hat{R} = \frac{\bar{y}}{\bar{x}} \tag{7.4}$$

Approximate bias of estimator $\hat{R}$:

$$B(\hat{R}) = \left(\frac{N-n}{Nn}\right)\left(\frac{1}{\bar{X}^2}\right)(RS_x^2 - S_{xy}) \tag{7.5}$$

Approximate mean square error of estimator $\hat{R}$:

$$\text{MSE}(\hat{R}) = \left(\frac{N-n}{Nn}\right)\left(\frac{1}{\overline{X}^2}\right)(S_y^2 + R^2 S_x^2 - 2RS_{xy}) \tag{7.6}$$

Estimator of MSE ($\hat{R}$) :

$$\text{mse}(\hat{R}) = \left(\frac{N-n}{Nn}\right)\left(\frac{1}{\overline{X}^2}\right)(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}\, s_{xy}) \tag{7.7}$$

In case $\overline{X}$ is not known, the sample mean $\overline{x}$ could be used in its place in (7.7) to find the value of mse ($\hat{R}$).

Example 7.1

A survey project was undertaken by a graduate student in a small town of USA consisting of N=1620 family households. The purpose of the survey was to estimate the proportion of monthly family income spent on purchasing milk. An SRS without replacement sample of n=30 households was selected for this purpose. The sample data, in respect of monthly income (x) and expenditure on milk (y), both in dollars, are listed in table 7.1.

Table 7.1 Monthly family income and the amount spent on milk

Household	Monthly income	Expenditure on milk	Household	Monthly income	Expenditure on milk
1	2000	60	16	1620	60
2	2900	85	17	1880	48
3	2400	80	18	2402	67
4	3200	106	19	2665	68
5	2400	60	20	1948	62
6	3260	84	21	1870	50
7	1600	45	22	3400	94
8	3080	106	23	1700	53
9	2445	75	24	1805	50
10	3600	102	25	2850	100
11	2960	75	26	3700	108
12	1750	60	27	2960	84
13	1980	60	28	2730	54
14	4260	108	29	2840	66
15	3370	91	30	2620	52

Taking $\overline{X}$ = \$2550, find both point and interval estimates of R.

Solution

First of all, we work out the following statistics :

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

$$= \frac{1}{30} (2000 + 2900 + ... + 2620)$$

$$= 2606.5$$

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

$$= \frac{1}{30} (60 + 85 + ... + 52)$$

$$= 73.77$$

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

$$= \frac{1}{30-1} [(2000)^2 + (2900)^2 + ... + (2620)^2 - 30\,(2606.5)^2]$$

$$= 489759.15$$

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

$$= \frac{1}{30-1} [(60)^2 + (85)^2 + ... + (52)^2 - 30(73.77)^2]$$

$$= 417.88$$

$$s_{xy} = \frac{1}{n-1} (\sum_{i=1}^{n} x_i\, y_i - n\bar{x}\,\bar{y})$$

$$= \frac{1}{30-1} [(2000)\,(60) + (2900)\,(85) + ... + (2620)\,(52)$$

$$- 30\,(2606.5)\,(73.77)]$$

$$= \frac{1}{29} [6125305 - 30\,(2606.5)\,(73.77)]$$

$$= 12305.51$$

We now compute estimate of ratio R from (7.4) as

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

$$= \frac{73.77}{2606.5}$$

$$= .02830$$

The estimate of mean square error for $\hat{R}$ is provided by (7.7). Using the sample values computed above, we have

$$\begin{aligned}\text{mse}\,(\hat{R}) &= \left(\frac{N-n}{Nn}\right)\left(\frac{1}{\overline{X}^2}\right)(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}\, s_{xy})\\ &= \left[\frac{1620-30}{(1620)\,(30)}\right]\left[\frac{1}{(2550)^2}\right][417.88 + (.02830)^2\ (489759.15)\\ &\quad - 2(.02830)\,(12305.51)]\\ &= 5.71713\times10^{-7}\end{aligned}$$

Now, we proceed to work out confidence interval for the population ratio R. It can be computed from

$$\begin{aligned}&\hat{R} \pm 2\sqrt{\text{mse}\,(\hat{R})}\\ &= .02830 \pm 2\sqrt{5.71713\times10^{-7}}\\ &= .02830 \pm .00151\\ &= .02679,\ .02981\end{aligned}$$

Thus the proportion of monthly income spent on milk is estimated to be 2.830%, and the confidence limits above indicate that the population proportion is most likely to be in the range of 2.679% to 2.981%. ■

7.3 RATIO ESTIMATOR FOR POPULATION MEAN/ TOTAL

As mentioned earlier, the estimator of population mean $\overline{Y}$ can be obtained by multiplying the estimator $\hat{R}$ by $\overline{X}$, the population mean for auxiliary variable x. The expressions for bias, mean square error, and the estimator of mean square error for the estimator of mean can also be obtained by multiplying the expression for bias $B(\hat{R})$ in (7.5) by $\overline{X}$, and the expressions for MSE $(\hat{R})$ and mse $(\hat{R})$ in (7.6) and (7.7) by $\overline{X}^2$. This yields the following results :

Ratio estimator of population mean $\overline{Y}$:

$$\bar{y}_r = \frac{\bar{y}\,\overline{X}}{\bar{x}} \qquad (7.8)$$

Approximate bias of estimator $\bar{y}_r$:

$$\left.\begin{aligned}B(\bar{y}_r) &= \left(\frac{N-n}{Nn}\right)\left(\frac{1}{\overline{X}}\right)(RS_x^2 - S_{xy})\\ &= \left(\frac{N-n}{Nn}\right)\overline{Y}\,(C_x^2 - \rho\, C_x\, C_y)\end{aligned}\right] \qquad (7.9)$$

Approximate mean square error of estimator $\bar{y}_r$:

$$\begin{aligned} MSE(\bar{y}_r) &= \frac{N-n}{Nn}\,(S_y^2 + R^2 S_x^2 - 2RS_{xy}) \\ &= \left(\frac{N-n}{Nn}\right)\bar{Y}^2\,(C_y^2 + C_x^2 - 2\rho\, C_x C_y) \end{aligned} \qquad (7.10)$$

Estimator of MSE $(\bar{y}_r)$:

$$mse\,(\bar{y}_r) = \frac{N-n}{Nn}\,(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}\, s_{xy}) \qquad (7.11)$$

where C_y and C_x are the population coefficients of variation for the variables y and x respectively.

The expressions relating to estimator of population total can be easily obtained from the expressions for the mean.

Ratio estimator of population total Y :

$$\hat{Y}_r = N\,\bar{y}_r = \frac{\bar{y}\,X}{\bar{x}} \qquad (7.12)$$

Bias of estimator $\hat{Y}_r$:

$$B(\hat{Y}_r) = N\,B(\bar{y}_r) \qquad (7.13)$$

Mean square error of estimator $\hat{Y}_r$:

$$MSE\,(\hat{Y}_r) = N^2\,MSE\,(\bar{y}_r) \qquad (7.14)$$

Estimator of MSE $(\hat{Y}_r)$:

$$mse\,(\hat{Y}_r) = N^2\,mse(\bar{y}_r) \qquad (7.15)$$

The MSE($\bar{y}_r$) and MSE($\hat{Y}_r$) given in (7.10) and (7.14), will be smaller than their respective counterparts V($\bar{y}$) and V($\hat{Y}$) in (3.9) and (3.12), for the usual estimators based on simple random sampling, when

$$\rho > \frac{C_x}{2C_y} \qquad (7.16)$$

If x is the same character as y but has been measured on an earlier occasion, the coefficients of variation C_x and C_y may be taken as equal. In that case, it pays to use the ratio method of estimation in place of simple mean estimator if ρ >.5. However, one should keep in mind that the inequality (7.16) is based on approximation. It should also be noted that the ratio estimate may not be as good as the simple average, even in the presence of perfect correlation between y and x, when the regression line of y on x passes through a point on y-axis that is far from origin, since the near proportionality between y and x does not exist in that situation. To summarize :

> The ratio estimators of population mean and total will be more efficient than the usual SRS based respective estimators if $\rho > C_x / 2C_y$ and the regression line passes through, or nearly through, the origin.

It would be a sound practice to examine the relationship between y and x on the basis of past surveys, and use this information during future studies.

Example 7.2

The data on study variable (y) and auxiliary variable (x) given below, are for a hypothetical population of 8 units :

y :	10	12	15	17	18	22	24	30
x :	4	6	9	10	13	14	16	20

Work out the efficiency of ratio estimator $\bar{y}_r$ in relation to the usual estimator $\bar{y}$ for WOR simple random samples of size 3.

Solution

For calculating the desired relative efficiency, we shall need the values of R, S_y^2, S_x^2, and S_{xy}. So, we first obtain these values from the population data. Thus,

$$\bar{Y} = \frac{1}{8}(10+12+\ldots+30)$$

$$= 18.5$$

$$\bar{X} = \frac{1}{8}(4+6+\ldots+20)$$

$$= 11.5$$

$$R = \frac{\bar{Y}}{\bar{X}} = \frac{18.5}{11.5} = 1.6087$$

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

$$= \frac{1}{8-1}\left[(10)^2 + (12)^2 + \ldots + (30)^2 - 8\,(18.5)^2\right]$$

$$= \frac{1}{7}\left[3042 - 8\,(18.5)^2\right]$$

$$= 43.4286$$

$$S_x^2 = \frac{1}{N-1} (\sum_{i=1}^{N} X_i^2 - N\overline{X}^2)$$

$$= \frac{1}{8-1} [(4)^2 + (6)^2 + ... + (20)^2 - 8(11.5)^2]$$

$$= \frac{1}{7} [1254 - 8(11.5)^2]$$

$$= 28.0000$$

$$S_{xy} = \frac{1}{N-1} (\sum_{i=1}^{N} X_i Y_i - N\overline{X}\,\overline{Y})$$

$$= \frac{1}{8-1} [(4)(10) + (6)(12) + ... + (20)(30) - 8(11.5)(18.5)]$$

$$= \frac{1}{7} [1943 - 8(11.5)(18.5)]$$

$$= 34.4286$$

The correlation coefficient between variables y and x is equal to

$$\rho = \frac{S_{xy}}{S_x S_y}$$

$$= \frac{34.4286}{\sqrt{(28.0000)(43.4286)}}$$

$$= .9873$$

On using above computed values in (3.9) and (7.10), we obtain variance $V(\overline{y})$ and mean square error $MSE(\overline{y}_r)$. Thus,

$$V(\overline{y}) = \frac{N-n}{Nn} S_y^2$$

$$= \frac{8-3}{(8)(3)} (43.4286)$$

$$= 9.0476$$

$$MSE(\overline{y}_r) = \frac{N-n}{Nn} (S_y^2 + R^2 S_x^2 - 2RS_{xy})$$

$$= \frac{8-3}{(8)(3)} [43.4286 + (1.6087)^2 (28.0000) - 2(1.6087)(34.4286)]$$

$$= 1.0666$$

The $MSE(\overline{y}_r)$ is seen to be less than $V(\overline{y})$, implying that, the ratio estimator of population mean is more efficient than the usual SRS based estimator $\overline{y}$. The percent relative efficiency of the estimator $\overline{y}_r$ with respect to $\overline{y}$ is obtained as

$$\begin{aligned}\text{RE} &= \frac{V(\bar{y})}{\text{MSE}\,(\bar{y}_r)}\,(100)\\ &= \frac{9.0476}{1.0666}\,(100)\\ &= 848.27\ \blacksquare\end{aligned}$$

Example 7.3

The agricultural wing of a district administration wishes to estimate the area under paddy harvested with combine. The district is comprised of 520 villages, including small towns, and the total area under paddy is 36,000 hectares. Keeping in view the budget at disposal, a WOR simple random sample of 26 villages was drawn for the purpose. The information on area in hectares, collected in respect of these villages, is given below in table 7.2.

Table 7.2 Area under paddy (x) and the area harvested with combine (y)

Village	x	y	Village	x	y
1	120	82	14	71	56
2	47	18	15	57	41
3	62	47	16	64	52
4	90	77	17	49	30
5	83	67	18	68	43
6	106	98	19	56	32
7	52	36	20	54	27
8	57	37	21	81	61
9	81	58	22	66	39
10	52	37	23	112	87
11	66	52	24	86	61
12	78	68	25	48	22
13	41	15	26	67	35

Estimate the total area under paddy in the district, harvested with combine, and obtain the confidence limits for this area.

Solution

We have X=36,000. As in example 7.1, we first work out the following sample estimates :

$$\begin{aligned}\bar{y} &= \frac{1}{26}\,(82 + 18 + \ldots + 35)\\ &= 49.15\end{aligned}$$

$$\begin{aligned}\bar{x} &= \frac{1}{26}\,(120 + 47 + \ldots + 67)\\ &= 69.77\end{aligned}$$

$$s_y^2 = \frac{1}{26-1}\,[(82)^2 + (18)^2 + ... + (35)^2 - 26(49.15)^2]$$

$$= 467.02$$

$$s_x^2 = \frac{1}{26-1}\,[(120)^2 + (47)^2 + ... + (67)^2 - 26(69.77)^2]$$

$$= 422.74$$

$$s_{xy} = \frac{1}{26-1}\,[(120)(82) + (47)(18) + ... + (67)(35) - 26(69.77)(49.15)]$$

$$= \frac{1}{25}\,[99624 - 26\,(69.77)\,(49.15)]$$

$$= 418.60$$

Using (7.12), we now estimate the total area under paddy harvested with combine. It gives

$$\hat{Y}_r = \frac{\bar{y}\,X}{\bar{x}}$$

$$= \frac{(49.15)\,(36000)}{69.77}$$

$$= 25360.47$$

Also,

$$\hat{R} = \frac{49.15}{69.77} = .7045$$

Estimate of mean square error can be obtained from (7.15) and (7.11). Thus,

$$\text{mse}\,(\hat{Y}_r) = N^2\,\text{mse}\,(\bar{y}_r)$$

$$= \frac{N\,(N-n)}{n}\,(s_y^2 + \hat{R}^2\,s_x^2 - 2\hat{R}\,s_{xy})$$

$$= \frac{520\,(520-26)}{26}\,[467.02 + (.7045)^2\,(422.74) - 2\,(.7045)\,(418.60)]$$

$$= 859826.98$$

Following (2.8), the confidence limits for population total are obtained as

$$\hat{Y}_r \pm 2\sqrt{\text{mse}\,(\hat{Y}_r)}$$

$$= 25360.47 \pm 1854.54$$

$$= 23505.93,\ 27215.01$$

It means, the total area under paddy, harvested with combine, in the district is likely to fall in the closed interval [23505.93, 27215.01] hectares with probability approximately equal to .95.

It must be pointed out here, that if the information on the auxiliary variable (area under paddy) was not used in the form of ratio estimator to estimate the total paddy area harvested with combine, the other alternative was to estimate it through the estimator $\hat{Y} = N\bar{y}$, where $\bar{y}$ is the usual mean based on WOR simple random sample. The estimator of the variance $V(\hat{Y})$ would then have been, from (3.10) and (3.13), as

$$\begin{aligned} v(\hat{Y}) &= \frac{N(N-n)}{n} s_y^2 \\ &= \frac{520(520-26)}{26} (467.02) \\ &= 4614157.60 \end{aligned}$$

which is much larger than $\text{mse}(\hat{Y}_r)$. Estimated percent relative efficiency of the ratio estimator, with respect to the SRS without replacement estimator $\hat{Y}$, is given by

$$\begin{aligned} RE &= \frac{v(\hat{Y})}{\text{mse}\,(\hat{Y}_r)} \cdot (100) \\ &= \frac{4614157.60}{859826.98} (100) \\ &= 536.64 \end{aligned}$$

Thus the ratio estimator $\hat{Y}_r$ for the total combine harvested area is over five times more efficient than the without replacement SRS estimator. This increase in efficiency is due to the use of the auxiliary information on the area under paddy. ■

7.4 DETERMINING THE SAMPLE SIZE FOR ESTIMATION OF RATIO, MEAN, AND TOTAL

Once the sampling design for the study has been chosen, the next question that the investigator faces is to decide about the number of units to be selected in the sample so as to get estimators with predetermined precision. In this section, we discuss the procedure to determine the required sample size for estimating population parameters R, $\bar{Y}$, and Y with B as the bound on error of estimation.

To determine the number of sample units required to estimate the population ratio R, defined in (7.1), with B as the bound on the estimation error, we need to solve the equation

$$2\sqrt{\text{mse}\,(\hat{R})} = B \tag{7.17}$$

for n. The expression for mse($\hat{R}$) has been given in (7.7). For evaluating the value of mse($\hat{R}$), we make use of the observations on the variables y and x for a preliminary sample of size n_l. Thus,

$$\text{mse}(\hat{R}) = \left(\frac{N-n}{Nn}\right)\frac{1}{\overline{X}^2}(s_{yl}^2 + \hat{R}_l^2 s_{xl}^2 - 2\hat{R}_l\, s_{xyl})$$

where the sample mean squares and product s_{yl}^2, s_{xl}^2, and s_{xyl} and the ratio $\hat{R}_l$ are obtained from the preliminary sample of size n_l. The solution of (7.17) for n, gives the following rule for choosing the required sample size.

Sample size to estimate R with a bound B on error of estimation :

$$n = \frac{Ns_r^2}{ND + s_r^2} \tag{7.18}$$

where

$$s_r^2 = s_{yl}^2 + \hat{R}_l^2 s_{xl}^2 - 2\hat{R}_l\, s_{xyl} \tag{7.19}$$

and

$$D = \frac{\overline{X}^2 B^2}{4}$$

If $n_l \geq n$, the preliminary sample of size n_l is enough. Include (n-n_l) additional units in the sample, otherwise.

Example 7.4

Assuming the sample of 30 houses drawn in example 7.1 as a preliminary sample, determine the sample size required for estimating ratio R of example 7.1 with a margin of error .001.

Solution

Since the sample of 30 houses of example 7.1 is now taken as the preliminary sample, the sample estimates worked out in example 7.1 will be treated as estimates obtained from the preliminary sample. This means,

$$n_l = 30,\ \overline{x} = \overline{x}_l = 2606.5,\ \hat{R} = \hat{R}_l = .02830,\ s_y^2 = s_{yl}^2 = 417.88,$$

$$s_x^2 = s_{xl}^2 = 489759.15, \text{ and } s_{xy} = s_{xyl} = 12305.51.$$

From (7.19), we then work out

$$\begin{aligned} s_r^2 &= s_{yl}^2 + \hat{R}_l^2 s_{xl}^2 - 2\hat{R}_l\, s_{xyl} \\ &= 417.88 + (.02830)^2\,(489759.15) - 2\,(.02830)\,(12305.51) \\ &= 113.6313 \end{aligned}$$

Also,

$$\begin{aligned} D &= \frac{\overline{X}^2 B^2}{4} \\ &= \frac{(2550)^2\,(.001)^2}{4} \\ &= 1.6256 \end{aligned}$$

From (7.18), the sample size needed to estimate population ratio R, with a bound on the error of estimation as .001, would be

$$\begin{aligned} n &= \frac{Ns_r^2}{ND + s_r^2} \\ &= \frac{(1620)\,(113.6313)}{(1620)\,(1.6256) + 113.6313} \\ &= 67.01 \\ &\approx 67 \end{aligned}$$

This means that the investigator needs to select $n\text{-}n_1 = 67\text{-}30 = 37$ more households to estimate the population ratio R with the desired precision. ■

Similarly, when ratio estimator is used for estimating the population mean/total, the required sample size is obtained by solving the following equations for n :

$$2\sqrt{\text{mse}\,(\overline{y}_r)} = B \quad \text{(when estimating mean)}$$

$$2\sqrt{\text{mse}\,(\hat{Y}_r)} = B \quad \text{(when estimating total)}$$

where, as in (7.17), the value of $(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy})$ involved in mse$(\overline{y}_r)$ and mse$(\hat{Y}_r)$, defined in (7.11) and (7.15), is calculated from the n_1 observations on variables y and x from the preliminary sample, and B is the magnitude of bound on the error of estimation. This yields the following rule :

Sample size to estimate the population mean / total with a permissible error B :

$$n = \frac{Ns_r^2}{ND + s_r^2} \tag{7.20}$$

where

$D = B^2/4$ (when estimating mean)

$D = B^2/4N^2$ (when estimating total)

and s_r^2 is as defined in (7.19). If $n_1 \geq n$, the sample size n_1 is sufficient, otherwise, one needs to select $(n\text{-}n_1)$ additional units in the sample.

Example 7.5

Treating the sample of 26 villages drawn in example 7.3 as the preliminary sample, determine the sample size required to estimate the total area under paddy, harvested with combine, with the bound on the error of estimation as 1500 hectares.

Solution

As the sample of 26 villages selected in example 7.3 is assumed as a preliminary sample, the sample estimates worked out there would be treated as estimates provided by the preliminary sample. We thus have

$$n_l = 26, \hat{R} = \hat{R}_l = .7045, s_y^2 = s_{yl}^2 = 467.02, s_x^2 = s_{xl}^2 = 422.74,$$

$$\text{and } s_{xy} = s_{xyl} = 418.60.$$

Then we compute

$$\begin{aligned} s_r^2 &= s_{yl}^2 + \hat{R}_l^2\, s_{xl}^2 - 2\,\hat{R}_l\, s_{xyl} \\ &= 467.02 + (.7045)^2\,(422.74) - 2(.7045)\,(418.60) \\ &= 87.0270 \end{aligned}$$

Further,

$$D = \frac{B^2}{4N^2} = \frac{(1500)^2}{4(520)^2} = 2.0803$$

The required sample size can then be computed by using (7.20). Thus,

$$\begin{aligned} n &= \frac{Ns_r^2}{ND + s_r^2} \\ &= \frac{(520)\,(87.0270)}{(520)\,(2.0803) + 87.0270} \\ &= 38.7 \\ &\approx 39 \end{aligned}$$

This means that the investigator will be required to select 39-26=13 more villages, if the population total under study has to be estimated with a maximum margin of error as 1500 hectares. ■

7.5 SEPARATE AND COMBINED RATIO ESTIMATORS

For the reasons mentioned in chapter 5, it is sometimes desirable to stratify the population and then use ratio estimators for estimating population mean or total. For the discussion in this section, we shall assume that the sample drawn from each stratum is large enough, so that, the mean square approximations work fairly well.

Let us assume that the population of N units is divided into L strata, such that, h-th stratum has N_h units. Thus $\Sigma N_h = N$, $h = 1,2,\ldots,L$. From the h-th stratum, a WOR simple random sample of n_h units is selected, so that, the total sample size over all the strata becomes n. Also, let $\bar{Y}_h$ and $\bar{X}_h$, h=1,2,...,L, denote the h-th stratum means for the variables y and x respectively, whereas $\bar{y}_h$ and $\bar{x}_h$ denote their sample counterparts. The h-th stratum mean squares and product for the two variables are defined as

$$\left.\begin{aligned} S^2_{hy} &= \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 \\ S^2_{hx} &= \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)^2 \\ S_{hxy} &= \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)(Y_{hi} - \bar{Y}_h) \end{aligned}\right] \qquad (7.21)$$

Similarly, s^2_{hy}, s^2_{hx}, and s_{hxy} are the sample mean squares and product which unbiasedly estimate S^2_{hy}, S^2_{hx}, and S_{hxy} respectively.

Now we discuss two different methods for constructing ratio estimators in stratified sampling. One is to build up the ratio estimators $(\bar{y}_h/\bar{x}_h)\bar{X}_h$, h = 1,2,...,L, within each stratum, and then form a weighted average of these separate estimators as a single estimator of population mean. The estimator, so obtained, is known as *separate ratio estimator*. Alternatively, one can estimate population mean $\bar{Y}$ by $\bar{y}_{st}$ and the mean $\bar{X}$ by $\bar{x}_{st}$ using stratified sampling. Then $(\bar{y}_{st}/\bar{x}_{st})\,\bar{X}$ can be used as an estimator of population mean $\bar{Y}$. This estimator is known as *combined ratio estimator*, and was proposed by Hansen *et al.* (1946).

Separate ratio estimator of population mean $\bar{Y}$:

$$\bar{y}_{sr} = \sum_{h=1}^{L} \frac{W_h \bar{y}_h}{\bar{x}_h} \bar{X}_h \qquad (7.22)$$

Approximate bias of the estimator $\bar{y}_{sr}$:

$$B(\bar{y}_{sr}) = \sum_{h=1}^{L} W_h \left(\frac{N_h - n_h}{N_h n_h}\right)\left(\frac{1}{\bar{X}_h}\right)(R_h S^2_{hx} - S_{hxy}) \qquad (7.23)$$

Approximate mean square error of estimator $\bar{y}_{sr}$:

$$\text{MSE}(\bar{y}_{sr}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right)(S^2_{hy} + R_h^2 S^2_{hx} - 2R_h S_{hxy}) \qquad (7.24)$$

Estimator of MSE($\bar{y}_{sr}$) :

$$\text{mse}(\bar{y}_{sr}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right)(s^2_{hy} + \hat{R}_h^2 s^2_{hx} - 2\hat{R}_h s_{hxy}) \qquad (7.25)$$

where $\hat{R}_h = \bar{y}_h/\bar{x}_h$.

Example 7.6

A farm owner wishes to conduct a survey to estimate per tree yield for orange variety Mandarins (*Citrus reticulata*) in his orchard. The orchard has 8 rows of 30 trees each. The trees in the first 6 rows were planted 10 years back, and are now fully developed and matured. The last 2 rows consist of younger trees that were planted 6 years back. Record of yield for the preceding year for all the trees is available. Per tree yield for the matured plants was 98 kg, and it was 53 kg for the younger plants. In order to estimate current average yield, a stratified WOR simple random sample of 24 trees was selected using proportional allocation. Out of 24 trees in the sample, n_1=18 trees were selected from the first six rows (stratum I), and n_2 = 6 trees came from the last two rows of younger trees (stratum II). All the selected trees were observed for yield. The yield (in kg) figures for the preceding year (x) and the current year (y) are given below :

Table 7.3 Orange yields for sample trees

Stratum I				Stratum II			
x	y	x	y	x	y	x	y
90.5	94.5	110.2	116.1	103.7	108.5	50.7	66.0
84.0	88.8	98.6	102.4	90.4	96.4	55.9	58.0
92.0	90.0	105.9	100.3	115.5	121.0	67.0	76.4
76.5	82.0	83.1	90.9	95.1	105.0	56.5	72.8
105.3	109.4	74.0	82.6	100.4	102.3	58.4	77.8
85.6	90.0	107.5	114.5	80.0	87.1	62.3	72.6

Estimate average yield per tree, using separate ratio estimator, and place confidence limits on it.

Solution

From the statement of the example, we have N = 240, N_1 = 180, N_2 = 60, n = 24, n_1 = 18, n_2 = 6, $\bar{X}_1$ = 98, and $\bar{X}_2$ = 53. The sample means, sample mean squares and products for the two strata have been computed, and are given below in table 7.4 along with certain other sample and population characteristics.

Table 7.4 Certain calculated strata values

Stratum I		Stratum II	
n_1 = 18	s^2_{1x} = 150.134	n_2 = 6	s^2_{2x} = 31.659
N_1 = 180	s^2_{1y} = 137.261	N_2 = 60	s^2_{2y} = 54.848
W_1 = .75	s_{1xy} = 136.985	W_2 = .25	s_{2xy} = 24.056
$\bar{X}_1$ = 98	r_1 = .954	$\bar{X}_2$ = 53	r_2 = .577
$\bar{x}_1$ = 94.350	$\hat{R}_1$ = 1.049	$\bar{x}_2$ = 58.467	$\hat{R}_2$ = 1.208
$\bar{y}_1$ = 98.989		$\bar{y}_2$ = 70.600	

The separate ratio estimate given by (7.22) yields

$$
\begin{aligned}
\bar{y}_{sr} &= \frac{W_1\bar{y}_1\bar{X}_1}{\bar{x}_1} + \frac{W_2\bar{y}_2\bar{X}_2}{\bar{x}_2} \\
&= W_1\hat{R}_1\bar{X}_1 + W_2\hat{R}_2\bar{X}_2 \\
&= (.75)\,(1.049)\,(98) + (.25)\,(1.208)\,(53) \\
&= 93.108
\end{aligned}
$$

We now work out mean square error of estimator $\bar{y}_{sr}$. Thus, from (7.25)

$$
\begin{aligned}
\text{mse}\,(\bar{y}_{sr}) &= \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_n\, n_h}\right) (s_{hy}^2 + \hat{R}_h^2\, s_{hx}^2 - 2\hat{R}_h\, s_{hxy}) \\
&= (.75)^2 \left[\frac{180-18}{(180)\,(18)}\right] [137.261 + (1.049)^2\,(150.134) \\
&\quad - 2(1.049)\,(136.985)] + (.25)^2 \left[\frac{60-6}{(60)\,(6)}\right] [54.848 \\
&\quad + (1.208)^2\,(31.659) - 2(1.208)\,(24.056)] \\
&= .4240 + .4024 \\
&= .8264
\end{aligned}
$$

The next step is to compute confidence limits. This we do through (2.8) as

$$
\begin{aligned}
&\bar{y}_{sr} \pm 2\sqrt{\text{mse}\,(\bar{y}_{sr})} \\
&= 93.108 \pm 1.818 \\
&= 91.290,\ 94.926
\end{aligned}
$$

Hence, the average orange yield per tree is estimated as 93.108 kg. Also, it is indicated that had all the trees in the orchard been observed for yield, the per tree yield would most probably have taken a value in the closed interval [91.290, 94.926]. ■

Unless the ratio R_h is constant from stratum to stratum, the separate ratio estimator is likely to be more precise than the combined estimator. For the separate ratio estimator, the sample size in all the strata should be sufficiently large, otherwise, it will have appreciable bias and the MSE($\bar{y}_{sr}$) approximations will not be good enough. It also needs the knowledge of $\bar{X}_h$ for each stratum. With only a small sample in each stratum, the combined estimator is to be recommended unless there is good empirical evidence to indicate wide differences in the strata ratios. Various expressions corresponding to the *combined ratio estimator* are given in the following box :

Combined ratio estimator of mean $\bar{Y}$:

$$\bar{y}_{cr} = \frac{\bar{y}_{st}}{\bar{x}_{st}}\bar{X} = \left[\frac{\sum_{h=1}^{L} N_h \bar{y}_h}{\sum_{h=1}^{L} N_h \bar{x}_h}\right]\bar{X} \qquad (7.26)$$

Approximate bias of the estimator $\bar{y}_{cr}$:

$$B(\bar{y}_{cr}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right)\left(\frac{1}{\bar{X}}\right)(RS_{hx}^2 - S_{hxy}) \qquad (7.27)$$

Approximate mean square error of estimator $\bar{y}_{cr}$:

$$\text{MSE}(\bar{y}_{cr}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_n n_h}\right)(S_{hy}^2 + R^2 S_{hx}^2 - 2RS_{hxy}) \qquad (7.28)$$

Estimator of MSE($\bar{y}_{cr}$) :

$$\text{mse}(\bar{y}_{cr}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_n n_h}\right)(s_{hy}^2 + \hat{R}^2 s_{hx}^2 - 2\hat{R}\, s_{hxy}) \qquad (7.29)$$

where $\hat{R} = \bar{y}_{st}/\bar{x}_{st}$.

Example 7.7

Since the ratios $\hat{R}_1$ and $\hat{R}_2$ computed in example 7.6 do not differ much, one can also use combined ratio estimator in place of separate ratio estimator. Thus, estimate the per tree yield using the estimator $\bar{y}_{cr}$, and also obtain the confidence interval for it.

Solution

Most of the intermediate values required for the purpose of estimation are already available in table 7.4. Hence, we have

$$\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h\bar{y}_h = \sum_{h=1}^{L} W_h\bar{y}_h$$

$$= (.75)(98.989) + (.25)(70.600)$$

$$= 91.892$$

$$\bar{x}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h\bar{x}_h = \sum_{h=1}^{L} W_h\bar{x}_h$$

$$= (.75)(94.350) + (.25)(58.467)$$

$$= 85.379$$

Also, the population mean $\overline{X}$ for all the 240 trees is calculated from the relation

$$\begin{aligned}\overline{X} &= \sum_{h=1}^{L} W_h\overline{X}_h \\ &= (.75)\,(98) + (.25)\,(53) \\ &= 86.75\end{aligned}$$

The combined ratio estimator of the average yield per tree, from (7.26), would be

$$\begin{aligned}\overline{y}_{cr} &= \frac{\overline{y}_{st}}{\overline{x}_{st}}\,\overline{X} \\ &= \frac{(91.892)\,(86.75)}{85.379} \\ &= 93.368\end{aligned}$$

For computing estimated mean square error of $\overline{y}_{cr}$, we are to use single pooled value $\hat{R}$ in place of different $\hat{R}_1$ and $\hat{R}_2$ values for the two strata. This pooled $\hat{R}$ value is obtained from combined estimators $\overline{y}_{st}$ and $\overline{x}_{st}$. Thus,

$$\begin{aligned}\hat{R} &= \frac{\overline{y}_{st}}{\overline{x}_{st}} \\ &= \frac{91.892}{85.379} \\ &= 1.076\end{aligned}$$

We now work out estimated mean square error of $\overline{y}_{cr}$. From (7.29),

$$\text{mse}(\overline{y}_{cr}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h\, n_h}\right) (s_{hy}^2 + \hat{R}^2 s_{hx}^2 - 2\hat{R}\, s_{hxy})$$

Using above computed $\hat{R}$ and other values from example 7.6, one gets

$$\begin{aligned}\text{mse}(\overline{y}_{cr}) &= (.75)^2 \left[\frac{180-18}{(180)\,(18)}\right] [137.261+(1.076)^2\,(150.134) \\ &\quad - 2\,(1.076)\,(136.985)] + (.25)^2 \left[\frac{60-6}{(60)\,(6)}\right] [54.848 \\ &\quad + (1.076)^2\,(31.659) - 2\,(1.076)\,(24.056)] \\ &= .4582 + .3725 \\ &= .8307\end{aligned}$$

To work out the confidence interval we again use (2.8). Thus, we have the interval limits as

$$\bar{y}_{cr} \pm 2\sqrt{\text{mse}\ (\bar{y}_{cr})}$$

$$= 93.368 \pm 2\sqrt{.8307}$$

$$= 93.368 \pm 1.823$$

$$= 91.545,\ 95.191$$

The use of combined ratio estimator indicates that the population mean yield, based on all the 240 units, would most probably take a value in the range of 91.545 to 95.191 kg. ■

7.6 SOME FURTHER REMARKS

7.1 Let a sample of n units be selected using SRS without replacement. Also, if $r_i = y_i/x_i$ is the ratio of y and x values for i-th sample unit, and $\bar{r} = (\Sigma r_i)/n$, i=1,2,..., n, then the estimator

$$\bar{y}_{hr} = \bar{r}\bar{X} + \frac{n}{N}\left(\frac{N-1}{n-1}\right)(\bar{y} - \bar{r}\ \bar{x}) \tag{7.30}$$

is unbiased for population mean $\bar{Y}$. The variance of this estimator, to the first order of approximation, is equal to the MSE($\bar{y}_r$). The estimator $\bar{y}_{hr}$ was proposed by Hartley and Ross (1954).

7.2 If the sample is selected using Sen-Midzuno's scheme (Sen, 1952; Midzuno, 1952), where the first unit in the sample is selected with probability proportional to the auxiliary variable value and the remaining (n-1) units are selected through simple random sampling WOR, the ratio estimator $\bar{y}_r$, defined in (7.8), becomes unbiased for the population mean.

7.3 Another technique used to obtain unbiased estimators is based on splitting at random the sample of size n into k groups, each of size m = n/k. Let

$$\bar{y}_r^{(j)} = \frac{\bar{y}'_j}{\bar{x}'_j}\bar{X} \tag{7.31}$$

where $\bar{y}'_j$ and $\bar{x}'_j$ are the sample means based on a sample of (n-m) units obtained by omitting the j-th group of m units. Then the estimator

$$\bar{y}_m = \frac{1}{k}\sum_{j=1}^{k}\bar{y}_m^{(j)} \tag{7.32}$$

where

$$\bar{y}_m^{(j)} = \bar{y}_r^{(j)} + \frac{(N-n+m)}{N}k\left(\bar{y} - \bar{y}_r^{(j)}\frac{\bar{x}}{\bar{X}}\right) \tag{7.33}$$

is unbiased for the population mean $\bar{Y}$. This estimator is due to Mickey (1959).

7.4 Several other workers including Murthy and Nanjamma (1959), Quenouille (1956), Beale (1962), and Tin (1965) have proposed procedures of obtaining almost unbiased ratio type estimators for the population mean $\bar{Y}$. Some generalized estimators of population mean have been proposed by Srivastava (1967) and Diana (1993). Swain (1964) provides theoretical details for ratio estimators based on systematic samples.

7.7 PRODUCT METHOD FOR ESTIMATING MEAN / TOTAL

In section 7.3, we have seen that the ratio method of estimation provides a more efficient estimator of population mean/total than the usual SRS estimators of mean/total provided the values for the variables y and x are nearly proportional, and the correlation between them is positive and high. This means that the ratio estimator $\bar{y}_r$ can not be used to improve upon the conventional estimator $\bar{y}$ in situations where ρ is negative. For such cases, we consider an estimator known as *product estimator.*

Product estimator of population mean $\bar{Y}$:

$$\bar{y}_p = \frac{\bar{y}\,\bar{x}}{\bar{X}} \tag{7.34}$$

Approximate bias of estimator $\bar{y}_p$:

$$\left.\begin{aligned} B(\bar{y}_p) &= \left(\frac{N-n}{Nn}\right)\frac{S_{xy}}{\bar{X}} \\ &= \left(\frac{N-n}{Nn}\right)\bar{Y}\rho\, C_x C_y \end{aligned}\right] \tag{7.35}$$

Approximate mean square error of estimator $\bar{y}_p$:

$$\left.\begin{aligned} \text{MSE}(\bar{y}_p) &= \frac{N-n}{Nn}\left(S_y^2 + R^2 S_x^2 + 2RS_{xy}\right) \\ &= \left(\frac{N-n}{Nn}\right)\bar{Y}^2\left(C_y^2 + C_x^2 + 2\rho C_x\, C_y\right) \end{aligned}\right] \tag{7.36}$$

Estimator of MSE ($\bar{y}_p$) :

$$\text{mse}(\bar{y}_p) = \frac{N-n}{Nn}\left(s_y^2 + \hat{R}^2 s_x^2 + 2\hat{R}s_{xy}\right) \tag{7.37}$$

where C_y and C_x are the population coefficients of variation for the variables y and x respectively.

As usual, the expressions for the estimator of population total, its bias, mean square error, and estimator of mean square error can be written from the expressions for the estimator $\bar{y}_p$.

Estimator of population total Y:

$$\hat{Y}_p = N\bar{y}_p = \frac{N\bar{y}\,\bar{x}}{\bar{X}} \qquad (7.38)$$

Bias of estimator $\hat{Y}_p$:

$$B(\hat{Y}_p) = N\,B(\bar{y}_p) \qquad (7.39)$$

Mean square error of estimator $\hat{Y}_p$:

$$MSE\,(\hat{Y}_p) = N^2\,MSE\,(\bar{y}_p) \qquad (7.40)$$

Estimator of MSE $(\hat{Y}_p)$:

$$mse\,(\hat{Y}_p) = N^2\,mse(\bar{y}_p) \qquad (7.41)$$

The MSE($\bar{y}_p$) and MSE($\hat{Y}_p$), given in (7.36) and (7.40), will be smaller than the corresponding variances V($\bar{y}$) and V($\hat{Y}$), in (3.9) and (3.12) for the conventional estimators in case of SRS, if

$$\rho \leq -\,C_x/2C_y \qquad (7.42)$$

implying $\rho \leq -.5$ when $C_x = C_y$. This yields the following statement :

The product estimator of population mean or total will be more efficient than the respective conventional estimator of mean or total in case of SRS, if $\rho \leq -\,C_x/2C_y$, which means $\rho \leq -.5$ when $C_x = C_y$.

It may be noted that the product estimator $\bar{y}_p$, defined in (7.34), can be corrected for bias. The resulting estimator

$$\bar{y}_p' = \frac{\bar{y}\,\bar{x}}{\bar{X}} - \left(\frac{N-n}{Nn}\right)\left(\frac{s_{xy}}{\bar{X}}\right)$$

is unbiased for the population mean $\bar{Y}$. Some other unbiased product type strategies have also been developed by Gupta and Adhvaryu (1982). Also, some of the procedures mentioned in remarks 7.3 and 7.4 of section 7.6 can be used to obtain unbiased product type estimators.

Example 7.8

A psychologist needs information on average duration of sleep (in hours) during night for the persons equal, or over, 50 years of age in a certain locality. Voters' list for the locality is used to prepare the list of persons aged 50 years or more. This population frame has 546 such persons. A WOR equal probability sample of 30 persons is drawn. Information regarding the age and duration of sleep gathered from the sampled persons is given in table 7.5.

Table 7.5 Age and average duration of sleep (in hours) for sample persons

Person	Age (x)	Duration of sleep (y)	Person	Age (x)	Duration of sleep (y)
1	62	7.00	16	66	7.00
2	75	5.00	17	78	5.75
3	51	8.00	18	63	6.75
4	57	7.75	19	77	5.50
5	81	5.00	20	73	4.75
6	79	5.25	21	55	7.30
7	67	7.00	22	71	6.00
8	74	6.25	23	63	6.50
9	84	4.50	24	87	4.50
10	56	7.75	25	61	6.25
11	68	7.00	26	58	6.25
12	70	7.00	27	60	6.50
13	59	7.25	28	69	6.00
14	64	6.75	29	56	6.50
15	53	8.50	30	71	5.75

The average age of persons in the target population is 70 years. Treating age as the auxiliary variable, estimate average duration of sleep in the population. Build up the confidence interval for it, and also, estimate percent relative efficiency of this estimator in relation to the SRS based usual mean estimator $\bar{y}$.

Solution

We have N= 546 and n=30. The average duration of sleep for respondents is likely to decrease with increase in age. In order to be sure that the product method of estimation could work well, we first work out

$$\bar{y} = 6.38,\ \bar{x} = 66.93,\ \bar{X} = 70,\ \hat{R} = \bar{y}/\bar{x} = .09532,$$

$$s_y^2 = 1.084,\ s_x^2 = 92.409, \text{ and } s_{xy} = -\ 8.885.$$

These figures in turn yield the sample estimate of the correlation coefficient as

$$r = \frac{s_{xy}}{\sqrt{s_y^2\ s_x^2}}$$

$$= \frac{-8.885}{\sqrt{(1.084)\,(92.409)}}$$

$$= -\ .89$$

In the light of condition (7.42), the product method of estimation could be used profitably. Thus, we get from (7.34), the estimate of average duration of sleep as

$$\bar{y}_p = \frac{\bar{y}\,\bar{x}}{\bar{X}}$$

$$= \frac{(6.38)\,(66.93)}{70}$$

$$= 6.10$$

For building up the confidence interval, we require estimate of mean square error. It is computed by using (7.37). Therefore,

$$\text{mse}\,(\bar{y}_p) = \left(\frac{1}{n} - \frac{1}{N}\right)(s_y^2 + \hat{R}^2\,s_x^2 + 2\hat{R}\,s_{xy})$$

$$= \left(\frac{1}{30} - \frac{1}{546}\right)[1.084 + (.09532)^2\,(92.409) + 2(.09532)\,(-8.885)]$$

$$= .00724$$

The confidence interval for average duration of sleep in the population is given by

$$\bar{y}_p \pm 2\sqrt{\text{mse}\,(\bar{y}_p)}$$

$$= 6.10 \pm 2\sqrt{.00724}$$

$$= 5.93,\ 6.27$$

Thus, the persons of age 50 years and more are, on the average, likely to sleep from 5.93 to 6.27 hours.

The estimated percent relative efficiency of the product estimator $\bar{y}_p$, in relation to the simple mean estimator $\bar{y}$, is given by

$$\text{RE} = \frac{v(\bar{y})}{\text{mse}\,(\bar{y}_p)}\,(100)$$

where from (3.10)

$$v(\bar{y}) = \frac{N-n}{Nn}\,s_y^2$$

$$= \frac{(546-30)}{(546)\,(30)}\,(1.084)$$

$$= .03415$$

Hence,

$$\begin{aligned} \text{RE} &= \frac{.03415}{.00724}(100) \\ &= 471.69 \end{aligned}$$

Therefore, the use of auxiliary information on age, in the form of product estimator, has resulted in increasing the relative efficiency from 100% to 471.69%. ■

7.8 DETERMINATION OF SAMPLE SIZE FOR PRODUCT ESTIMATOR

In order to arrive at the required sample size for estimating mean/total, let n_1 be the number of units selected in the preliminary sample. The value of the expression $(s_y^2 + \hat{R}^2 s_x^2 + 2\hat{R} s_{xy})$ in (7.37), is calculated from the observations on this preliminary sample. Let this value be denoted by

$$s_p^2 = (s_{y1}^2 + \hat{R}_1^2 s_{x1}^2 + 2\hat{R}_1 s_{xy1}) \qquad (7.43)$$

The mean square error estimator then becomes $\text{mse}(\bar{y}_p) = [(N\text{-}n)\ /Nn]\ s_p^2$, and $\text{mse}(\hat{Y}_p) = [N(N\text{-}n)/\ n]\ s_p^2$. Using these calculated estimates in place of $\text{mse}(\bar{y}_p)$ and $\text{mse}(\hat{Y}_p)$ in (7.37) and (7.41) respectively, and proceeding in the same way as in case of ratio estimator of mean/total, one gets the rule for choosing the required sample size for estimating the population mean/total through product method of estimation.

Sample size to estimate the population mean/total with a bound B on the error of estimation :

$$n = \frac{Ns_p^2}{ND + s_p^2} \qquad (7.44)$$

where

$$D = \frac{B^2}{4} \quad \text{(when estimating mean)}$$

$$D = \frac{B^2}{4N^2} \quad \text{(when estimating total)}$$

and s_p^2 is as defined in (7.43). When $n_1 \geq n$, no additional unit need be selected, otherwise, augment the preliminary sample by selecting $(n\text{-}n_1)$ more units.

Example 7.9

Assume that the sample of 30 persons, drawn in example 7.8, is a preliminary sample. Examine, whether this sample is sufficient for estimating mean sleeping hours with a margin of error .25 hours ?

Solution

We have assumed the sample of 30 persons drawn in example 7.8 as the preliminary sample. Obviously, all the estimated values computed in that example will be treated as estimates provided by the preliminary sample. We, therefore, get

$$N = 546,\ n_l = 30,\ \hat{R} = \hat{R}_l = .09532,\ s_y^2 = s_{yl}^2 = 1.084,$$

$$s_x^2 = s_{xl}^2 = 92.409, \text{ and } s_{xy} = s_{xyl} = -8.885.$$

Using (7.43), we first work out s_p^2. Thus, we have

$$\begin{aligned} s_p^2 &= s_{yl}^2 + \hat{R}_l^2\, s_{xl}^2 + 2\hat{R}_l\, s_{xyl} \\ &= 1.084 + (.09532)^2\,(92.409) + 2\,(.09532)\,(-8.885) \\ &= .2298 \end{aligned}$$

Now,

$$D = \frac{B^2}{4} = \frac{(.25)^2}{4} = .015625$$

The required sample size is then obtained by using (7.44). We thus have

$$\begin{aligned} n &= \frac{Ns_p^2}{ND + s_p^2} \\ &= \frac{(546)\,(.2298)}{(546)\,(.015625) + .2298} \\ &= 14.3 \\ &\approx 14 \end{aligned}$$

As the preliminary sample size $n_l = 30$ is more than the required sample size $n = 14$, the sample of size 30 already drawn is sufficient to yield an estimate of mean sleeping hours with desired accuracy. ■

The reader should note that like ratio estimator, the product estimator can also be used in stratified simple random sampling yielding separate and combined product estimators. Thus we have

$$\bar{y}_{sp} = \sum_{h=1}^{L} W_h \left(\frac{\bar{y}_h \bar{x}_h}{\bar{X}_h} \right) \tag{7.45}$$

and

$$\bar{y}_{cp} = \frac{1}{\bar{X}} \left(\sum_{h=1}^{L} W_h\, \bar{y}_h \right) \left(\sum_{h=1}^{L} W_h \bar{x}_h \right) \tag{7.46}$$

Both the above estimators are biased and the expressions for their biases are given by

$$B(\bar{y}_{sp}) = \sum_{h=1}^{L} W_h \left(\frac{N_h - n_h}{N_h n_h}\right) \frac{S_{hxy}}{\bar{X}_h} \tag{7.47}$$

and

$$B(\bar{y}_{cp}) = \frac{1}{\bar{X}} \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right) S_{hxy} \tag{7.48}$$

Expressions for mean square errors and their estimators for the above two estimators can be obtained from the corresponding expressions for separate and combined ratio estimators by replacing $(-2R_h S_{hxy})$, $(-2\hat{R}_h s_{hxy})$, $(-2RS_{hxy})$, and $(-2\hat{R}\, s_{hxy})$ in (7.24), (7.25), (7.28), and (7.29) respectively with same terms with a positive sign.

LET US DO

7.1 It is desired to estimate per acre yield of wheat crop, which is the ratio of total wheat yield to the total area under the crop, in a certain area. Discuss, how will you go about it ?

7.2 An investigator wishes to estimate sex ratio in a town. The best frame available is the list of ration depots (shops from where people get essential commodities at prices fixed by the government). Each household possesses a ration card, and is listed with a nearby depot. The details of family members, like sex, age, etc., are mentioned on the ration card. These details are also available with the ration depot. Twenty seven depots were selected from a total of 148 depots in the town, using SRS without replacement. The information in respect of family members gathered from the selected depots is given below, where M and F stand for number of males and females respectively.

Depot	M	F	Depot	M	F	Depot	M	F
1	870	630	10	911	860	19	890	763
2	500	440	11	731	601	20	525	624
3	981	670	12	508	520	21	560	574
4	893	703	13	680	570	22	674	766
5	613	688	14	713	591	23	990	720
6	380	360	15	507	569	24	863	618
7	421	473	16	984	887	25	782	614
8	671	576	17	768	703	26	828	701
9	933	956	18	635	648	27	774	540

Estimate the male : female ratio in the town, and work out confidence interval for it.

7.3 A survey was undertaken to estimate change in the per acre rental value of irrigated land in a certain district comprising of 760 villages. A WOR simple random sample of 30 villages was drawn. *Sarpanch*, the elected head of each village, was interviewed and the information on rental value of irrigated land for the current year (y) and the assessed rental value 5 years back (x) was obtained. The collected information (in '00 rupees) is given below :

Village	x	y	Village	x	y	Village	x	y
1	35	48	11	35	46	21	33	48
2	40	46	12	37	49	22	34	46
3	38	50	13	34	45	23	37	49
4	42	51	14	38	49	24	34	47
5	38	47	15	35	44	25	33	46
6	37	46	16	34	46	26	35	46
7	35	41	17	33	45	27	34	47
8	36	43	18	37	49	28	33	45
9	35	44	19	34	48	29	34	49
10	36	47	20	36	42	30	36	50

Using the sample information, estimate the change in rental value, and place confidence limits on it.

7.4 Describe ratio method for estimating population total, and state the situations where the estimator based on this method is likely to be more efficient than the usual SRS estimator. Under what condition the bias of this estimator becomes zero ?

7.5 A cattle-cake manufacturer is interested in examining the effect of a new cattle-cake on milk yield. For this purpose, a WOR simple random sample of 27 buffalos was drawn from a population of 540 buffalos. One day milk yield of these 27 buffalos was recorded individually, and the milk yield for the remaining 540-27= 513 buffalos was measured collectively. This gave the average per day milk yield for the population of 540 buffalos as 10 kg. The selected 27 buffalos were then kept on the new feed for 10 days. After the expiry of this period, the milk yield for a day of these sample buffalos was again recorded. The average milk yield (in kg) for the sample buffalos was then computed. The milk yields denoted by x and y respectively, recorded before and after the introduction of the new feed, for the selected buffalos are given as follows :

Buffalo	Milk yield		Buffalo	Milk yield		Buffalo	Milk yield	
	x	y		x	y		x	y
1	10.8	11.6	10	12.4	13.1	19	8.8	9.6
2	8.6	9.4	11	8.9	10.2	20	6.5	7.9
3	7.5	8.3	12	11.3	12.0	21	14.1	14.5
4	11.4	12.9	13	12.0	12.8	22	9.1	10.4
5	12.9	13.4	14	9.6	10.9	23	13.6	14.2
6	7.7	8.5	15	7.5	8.7	24	9.4	10.1
7	9.6	10.9	16	13.6	14.3	25	10.7	10.9
8	10.3	11.0	17	8.5	9.4	26	8.4	9.6
9	11.6	12.2	18	14.5	14.8	27	11.8	12.2

Estimate average daily milk yield per buffalo after being kept on the new feed, and place confidence limits on it. Also, work out the estimated relative efficiency of the ratio estimator of average daily milk yield with respect to the usual estimator $\bar{y}$ based on SRS without replacement.

7.6 A town consists of 138 wards. Total number of dwellings in the town are known to be 14060. An investigator desires to estimate the total number of dwellings occupied by tenants. For this purpose, a WOR simple random sample of 24 wards was selected. The information on number of dwellings (x), and the number of dwellings occupied by tenants (y), was obtained for the selected wards. This is presented in table below :

Ward	x	y	Ward	x	y	Ward	x	y
1	80	8	9	116	9	17	93	8
2	218	16	10	130	11	18	157	13
3	108	4	11	105	12	19	87	6
4	90	6	12	75	6	20	137	10
5	126	10	13	128	10	21	77	7
6	143	16	14	110	9	22	179	21
7	70	6	15	85	9	23	198	16
8	85	5	16	150	13	24	166	13

Using ratio method, estimate the total number of rented dwellings in the town, and determine confidence limits for it. Also, estimate relative efficiency of the ratio estimator of this total in relation to the usual SRS without replacement based estimator $\hat{Y}$.

7.7 Given a preliminary sample of size n_1, discuss the procedure of determining the necessary sample size for estimating population ratio R, when the margin of error that can be tolerated is 4%.

7.8 Assume that the sample of 27 depots, drawn in exercise 7.2, is a preliminary sample. Using the information presented in that exercise, verify whether this sample size is sufficient to estimate male : female ratio with a margin of error .05 ? If not, how many additional units need to be selected?

7.9 Treating the estimate of mean square error obtained in exercise 7.6 as from a preliminary sample, check whether the sample of 24 wards is sufficient to estimate the total number of dwellings occupied by tenants with a margin of error as 30 ?

7.10 The students admitted to a college this year were stratified into 3 categories - rich, middle class, and poor - depending on the income and occupation of their parents, they had stated in their admission forms. The three strata respectively consisted of $N_1 = 120$, $N_2 = 300$, and $N_3 = 390$ students. The average annual income of the parents of students belonging to rich, middle class, and poor strata was worked out from the admission forms of the students. It was found to be rupees 92, 60, and 40 thousands respectively. The investigator wishes to estimate average amount of pocket money these students had spent during the preceding 3 months. It was decided to select a WOR simple random sample of size 27 from the total population of 810 students. Using proportional allocation,

$$n_1 = \left(\frac{27}{810}\right)(120) = 4$$

$$n_2 = \left(\frac{27}{810}\right)(300) = 10$$

$$n_3 = \left(\frac{27}{810}\right)(390) = 13$$

students were selected from the first, second, and third stratum respectively. The information regarding pocket money (y) they had spent during the past three months was obtained through personal interview. Given below is the amount of pocket money spent (in rupees) and the annual income (x) of students' parents (in '000 rupees).

Rich		Middle class				Poor			
x	y	x	y	x	y	x	y	x	y
78	900	50	450	62	550	38	400	43	200
135	1250	66	700	50	450	41	350	33	250
70	750	58	650	56	500	47	450	30	200
87	800	60	500	63	650	43	400	46	400
		55	550	52	350	34	250	37	350
						35	200	39	350
						28	150		

Using combined ratio estimator, estimate the average pocket money a student had spent during the past three months, and build up the confidence interval for it.

7.11 From the data presented in exercise 7.10, estimate the average amount of pocket money spent by a student using separate ratio estimator. Also, discuss which of the two ratio estimators - separate or combined - will be more appropriate in the present situation ?

7.12 In which situation the product estimator of mean is more efficient than the simple random sample mean ? Do you agree with the statement that the product method of estimation finds application in lesser number of situations commonly encountered in practice in relation to the ratio method ? If so, why ?

7.13 Writing the y values in example 7.2 in decreasing order while retaining x values as such, compare the mean square error of product estimator $\bar{y}_p$ with the variance of the usual SRS mean estimator $\bar{y}$ based on WOR sample of size 3. Also, work out percent relative efficiency of the estimator $\bar{y}_p$ with respect to $\bar{y}$.

7.14 It is of interest to estimate weekly time spent by the undergraduate students of a university in viewing television, playing cards, gossips, or just wandering around, etc. In order to accomplish this task, a WOR simple random sample of 36 students was drawn from a population of 960 students. The overall grade point average (OGPA) was used as auxiliary variable. The mean OGPA for this population, obtained from the office of Registrar, is 2.97 (4.00 basis). The information collected in respect of OGPA (x), and the weekly time (in hours) spent on the above said nonacademic activities (y), is presented in the table below :

Student	x	y	Student	x	y	Student	x	y
1	2.35	28	13	1.91	20	25	3.75	6
2	2.00	33	14	2.84	22	26	3.35	8
3	2.75	14	15	2.36	17	27	2.95	12
4	2.80	21	16	3.62	10	28	3.09	11
5	2.95	17	17	1.83	18	29	2.81	12
6	2.70	25	18	2.69	14	30	2.71	13
7	3.43	10	19	3.29	11	31	2.92	13
8	3.82	7	20	2.41	19	32	3.36	9
9	3.56	16	21	2.07	13	33	1.96	20
10	2.55	15	22	1.86	18	34	3.66	8
11	2.85	14	23	2.79	13	35	3.23	7
12	3.44	7	24	2.71	11	36	1.81	20

Using appropriate method, estimate average weekly time (in hours) spent by a student on nonacademic activities. Also, build up the confidence interval for this average.

7.15 Using data of exercise 7.14, estimate total weekly time (in hours) spent in nonacademic activities by all the students in the population under study. Also, build up the confidence interval for this total time spent.

7.16 Using estimate of mean square error obtained in exercise 7.14, comment whether, or not, the sample of 36 students is sufficient to provide an estimate of average weekly time spent by a student on nonacademic activities with a margin of error equal to 2 hours. If not, how many more students should be included in the sample ?

CHAPTER 8

Regression Method of Estimation

8.1 INTRODUCTION

Analogous to the ratio and product estimators, the linear *regression estimator* is also designed to increase the efficiency of estimation by using information on the auxiliary variable x which is correlated with the study variable y. As stated before, the ratio method of estimation is at its best when the correlation between y and x is positive and high, and also the regression of y on x is linear through the origin. In practice, however, it is observed that even when the regression of y on x is linear, the regression line passes through a point away from the origin. The efficiency of the ratio estimator in such cases is very low, as it decreases with the increase in length of the intercept cut on y-axis by the regression line. Regression estimator is the appropriate estimator for such situations. Although this estimator requires little more calculations than the ratio estimator, it is always at least as efficient as the ratio estimator for estimating population mean or total. Similarly, the product estimator of population mean or total is never more efficient than the corresponding linear regression estimator.

Regression estimator for mean/total has been in use for quite sometime. Watson (1937) used regression of leaf area on leaf weight to estimate the average leaf area for a plant. In another interesting application presented by Yates (1960), an eye estimate of the volume of timber was made on each of the 1/10 acre plots of a population of plots. The actual timber volume was measured for a sample of plots. Using the regression estimator, total timber volume was then estimated.

In this chapter, we shall discuss two types of regression estimators - one when the value of the regression coefficient is known in advance, and the other when it is to be estimated from the sample. The estimator based on the known value of regression coefficient is termed difference estimator.

8.2 ESTIMATION OF MEAN/TOTAL USING DIFFERENCE ESTIMATOR

We assume that (y_1, x_1), (y_2, x_2), ...,(y_n, x_n) denote the observations on the study and auxiliary variables on an SRS without replacement sample of n units selected from the finite population of N units. Also, let $\bar{y}$ and $\bar{x}$ denote the corresponding sample means, whereas $\bar{X}$ is the population mean for the auxiliary variable. Let α be a predetermined constant. Then, the *difference estimator* of population mean $\bar{Y}$ is defined as

$$\bar{y}_d = \bar{y} + \alpha(\bar{X} - \bar{x})$$

The estimator $\bar{y}_d$ is unbiased for $\bar{Y}$, whatever be the value of α. It is easy to see that the variance $V(\bar{y}_d)$ is minimum when

$$\alpha = \frac{S_{xy}}{S_x^2} = \beta \text{ (say)}$$

where β is the population regression coefficient of y on x. In practice, the actual value of β will not be available. Efficiency considerations will, however, require that a value of α very close to β be chosen for building up the difference estimator.

In repeated surveys, analysis of survey data over time may indicate that the population regression coefficient remains fairly constant. It may help us in choosing an appropriate value for the regression coefficient of y on x, in advance. The difference estimators, based on predetermined value β_o of the regression coefficient β, are simple and informative. We first consider these estimators. The subscript (d) used with the estimator stands for the difference estimator.

Unbiased difference estimator of population mean $\bar{Y}$:

$$\bar{y}_d = \bar{y} + \beta_o(\bar{X} - \bar{x}) \tag{8.1}$$

Variance of estimator $\bar{y}_d$:

$$\left.\begin{aligned} V(\bar{y}_d) &= \frac{N-n}{Nn}(S_y^2 + \beta_o^2 S_x^2 - 2\beta_o S_{xy}) \\ &= \frac{N-n}{Nn} S_y^2(1-\rho^2) \quad \text{when } \beta_o = \beta \end{aligned}\right] \tag{8.2}$$

Estimator of variance $V(\bar{y}_d)$:

$$v(\bar{y}_d) = \frac{N-n}{Nn}(s_y^2 + \beta_o^2 s_x^2 - 2\beta_o s_{xy}) \tag{8.3}$$

The parameters and their respective estimators involved in (8.1) to (8.3), have already been defined in the preceding chapter.

From $V(\bar{y}_d)$ in (8.2) and $V(\bar{y})$ in (3.9), it is obvious that regression estimator using known β_o is always more efficient than the usual simple random sample mean when $\beta_o = \beta$. The efficiency of the estimator $\bar{y}_d$ in (8.1) will decrease, as the difference between the guess value β_o and the actual value β of the regression coefficient increases.

As stated earlier, the corresponding expressions for estimating population total can be easily obtained.

Example 8.1

A medical student was given an assignment to estimate the average systolic blood pressure (BP) for teachers between 30 and 60 years of age in a certain university. The objective was to compare it with the systolic BP of a part of the population engaged in manual work. A WOR simple random sample of 24 teachers was drawn from the frame consisting of 961 teachers. Age of teachers was taken as the auxiliary variable. Average

age for the population of teachers was calculated from university records as 42.7 years. The regression coefficient of systolic BP on age, available from an earlier study conducted 5 years ago, is .8952. Using difference estimator, estimate the average systolic blood pressure, and build up confidence interval for it.

Table 8.1 Age (in completed years) and systolic blood pressure for sample teachers

Teacher	BP (y)	Age (x)	Teacher	BP (y)	Age (x)	Teacher	BP (y)	Age (x)
1	130	38	9	130	58	17	146	55
2	128	41	10	124	36	18	144	49
3	147	55	11	150	57	19	126	35
4	130	37	12	125	31	20	124	32
5	128	39	13	121	36	21	148	58
6	120	30	14	115	47	22	140	34
7	151	48	15	163	59	23	133	43
8	140	44	16	141	47	24	143	50

Solution

Here, we have N = 961, n = 24, $\overline{X} = 42.7$, and $\beta_o = .8952$. We first work out intermediate sample values to be used later. These are

$$\begin{aligned}
\overline{y} &= \frac{1}{24}(130 + 128 + ... + 143) \\
&= 135.29 \\
\overline{x} &= \frac{1}{24}(38 + 41 + ... + 50) \\
&= 44.13 \\
s_y^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right) \\
&= \frac{1}{23}[(130)^2 + (128)^2 + ... + (143)^2 - 24(135.29)^2] \\
&= 146.08 \\
s_x^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) \\
&= \frac{1}{23}[(38)^2 + (41)^2 + ... + (50)^2 - 24(44.13)^2] \\
&= 89.13
\end{aligned}$$

$$
\begin{aligned}
s_{xy} &= \frac{1}{n-1}\,(\sum_{i=1}^{n} x_i\, y_i - n\,\bar{x}\,\bar{y}) \\
&= \frac{1}{23}\,[(38)\,(130) + (41)\,(128) + \ldots + (50)\,(143) - 24\,(44.13)\,(135.29)] \\
&= \frac{1}{23}\,(145194 - 143288.34) \\
&= 82.85
\end{aligned}
$$

For working out estimate of average systolic BP, we use (8.1). That means,

$$
\begin{aligned}
\bar{y}_d &= \bar{y} + \beta_o\,(\bar{X} - \bar{x}) \\
&= 135.29 + .8952\,(42.7 - 44.13) \\
&= 134.01
\end{aligned}
$$

Next step is to estimate the variance of the estimator $\bar{y}_d$. For this, we use (8.3). Thus,

$$
\begin{aligned}
v(\bar{y}_d) &= \frac{N-n}{Nn}\,(s_y^2 + \beta_0^2\, s_x^2 - 2\beta_o\, s_{xy}) \\
&= \frac{961-24}{(961)\,(24)}\,[146.08 + (.8952)^2\,(89.13) - 2\,(.8952)\,(82.85)] \\
&= 2.8102
\end{aligned}
$$

The confidence limits, in which the average systolic BP is likely to fall, are

$$
\begin{aligned}
& \bar{y}_d \pm 2\sqrt{v(\bar{y}_d)} \\
&= 134.01 \pm 2\,\sqrt{2.8102} \\
&= 134.01 \pm 3.35 \\
&= 130.66,\ 137.36
\end{aligned}
$$

Hence, the average systolic blood pressure for the population of 961 teachers is most likely to take a value in the closed interval [130.66, 137.36].

If the information on the auxiliary variable x (age) was not used, the average systolic blood pressure could be estimated through the usual simple random sample mean $\bar{y}$. The variance of this estimator for WOR sampling would have been estimated from (3.10) as

$$
\begin{aligned}
v(\bar{y}) &= \frac{N-n}{Nn}\, s_y^2 \\
&= \frac{961-24}{(961)\,(24)}\,(146.08) \\
&= 5.9347
\end{aligned}
$$

which is larger than $v(\bar{y}_d)$. The percent relative efficiency of $\bar{y}_d$ with respect to $\bar{y}$, is estimated by

$$\begin{aligned} RE &= \frac{v(\bar{y})}{v(\bar{y}_d)}(100) \\ &= \frac{5.9347}{2.8102}(100) \\ &= 211.18 \end{aligned}$$

Thus, the use of auxiliary information on age, in the form of difference estimator, has increased the efficiency more than two times for the estimation of average systolic blood pressure. ■

8.3 ESTIMATION OF MEAN/TOTAL USING ESTIMATED REGRESSION COEFFICIENT

The estimator $\bar{y}_d$ considered in the preceding section is unbiased for $\bar{Y}$and the expression for its variance is exact. The estimator of variance is also unbiased. The estimator itself doesn't require heavy computations and is easy to apply in practice. Its relative precision depends on the accuracy with which the value of the regression coefficient β has been guessed. However, situations are frequently encountered where it is not possible to make a reliable guess for β. Then, the only alternative left is to estimate it from the sample itself and use the estimated value in place of β_o, in (8.1). The estimator of population mean $\bar{Y}$, so obtained, is called *linear regression estimator.*

From least square principle, the sample estimate for β is given by

$$\hat{\beta} = \frac{s_{xy}}{s_x^2}$$

Since this estimator $\hat{\beta}$ is a random variable, exact expressions for the expected value and the mean square error of the regression estimator are bit difficult to obtain. We restrict again to the WOR simple random sampling and assume that the sample is large enough, so that, the approximate expressions for $MSE(\bar{y}_{lr})$ and its estimator in (8.6) and (8.7) are accurate enough.

Regression estimator of population mean $\bar{Y}$:

$$\bar{y}_{lr} = \bar{y} + \hat{\beta}(\bar{X} - \bar{x}) \qquad (8.4)$$

Bias of estimator $\bar{y}_{lr}$:

$$B(\bar{y}_{lr}) = -\operatorname{Cov}(\bar{x}, \hat{\beta}) \qquad (8.5)$$

Approximate mean square error of estimator $\bar{y}_{lr}$:

$$\left.\begin{aligned} MSE(\bar{y}_{lr}) &= \frac{N-n}{Nn}(S_y^2 + \beta^2 S_x^2 - 2\beta S_{xy}) \\ &= \frac{N-n}{Nn} S_y^2(1-\rho^2) \end{aligned}\right] \qquad (8.6)$$

Estimator of mean square error MSE ($\bar{y}_{lr}$) :

$$\left.\begin{aligned} \text{mse}(\bar{y}_{lr}) &= \frac{N-n}{Nn}\,(s_y^2 + \hat{\beta}^2 s_x^2 - 2\hat{\beta}\, s_{xy}) \\ &= \frac{N-n}{Nn}\, s_y^2\,(1-r^2) \end{aligned}\right] \qquad (8.7)$$

where r is the sample correlation coefficient.

The corresponding expressions for the estimator of population total can be obtained from the above expressions in the usual manner.

Example 8.2

A physiologist has undertaken a project to estimate the average leaf area of a newly developed strain of wheat of which 120 plants were raised. In all, there were 2106 leaves and their total weight was 242.078 gm. Since measuring of area of all the 2106 leaves is difficult, a WOR simple random sample of 33 leaves was drawn. The area and weight of each sampled leaf were recorded. These are given in table 8.2 below. Estimate the average leaf area for the population under consideration. Also, work out the confidence interval for the actual value of this population parameter.

Table 8.2 Area (in sq cm) and weight (in mg) of leaves in the sample

Leaf	Area (y)	Weight (x)	Leaf	Area (y)	Weight (x)	Leaf	Area (y)	Weight (x)
1	27.37	105	12	24.18	106	23	14.31	78
2	30.21	109	13	35.72	125	24	39.28	128
3	22.18	100	14	19.76	97	25	24.16	102
4	36.76	125	15	33.46	125	26	26.51	114
5	28.51	116	16	43.62	131	27	29.69	119
6	30.34	118	17	16.11	85	28	20.03	101
7	21.81	104	18	21.07	112	29	18.41	93
8	29.11	118	19	26.71	117	30	35.72	124
9	38.90	129	20	18.51	96	31	29.33	117
10	17.21	90	21	23.43	103	32	21.88	107
11	42.44	130	22	31.66	121	33	21.29	103

Solution

We have N = 2106, n = 33, and X = 242.078 gm. This implies that

$$\begin{aligned} \bar{X} &= \frac{242.078}{2106} \\ &= .1149 \text{ gm} \\ &= 114.9 \text{ mg} \end{aligned}$$

As in examples 7.2, 7.3, and 8.1, we first work out the intermediate sample estimates. These are

$\bar{y}$ = 27.263, $\bar{x}$ = 110.545, s_y^2 = 61.003, $s^2{}_x$= 187.631, and s_{xy}=100.312

Thus,

$$r = \frac{s_{xy}}{s_y\, s_x}$$

$$= \frac{100.312}{\sqrt{(61.003)\,(187.631)}}$$

$$= .9376$$

$$\hat{\beta} = \frac{s_{xy}}{s_x^2}$$

$$= \frac{100.312}{187.631} = .5346$$

Now, we compute the estimate of average leaf area by using (8.4). Hence,

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

$$= 27.263 + .5346\,(114.9 - 110.545)$$

$$= 29.591$$

Estimate of mean square error is calculated from (8.7), as

$$\text{mse}(\bar{y}_{lr}) = \frac{N-n}{Nn}\,(s_y^2 + \hat{\beta}^2\, s_x^2 - 2\hat{\beta}\, s_{xy})$$

$$= \frac{2106-33}{(2106)\,(33)}\,[61.003 + (.5346)^2\,(187.631) - 2(.5346)\,(100.312)]$$

$$= .2200$$

Alternatively, it can also be computed by using second version of (8.7). This yields

$$\text{mse}\,(\bar{y}_{lr}) = \frac{N-n}{Nn}\,s_y^2\,(1-r^2)$$

$$= \frac{2106-33}{(2106)\,(33)}\,(61.003)\,[1-(.9376)^2]$$

$$= .2200$$

The range, in which the average leaf area for the population would probably lie, is determined by using confidence limits. These are obtained as

$$\bar{y}_{lr} \pm 2\sqrt{\text{mse}\,(\bar{y}_{lr})}$$

$$= 29.591 \pm 2\sqrt{.2200}$$

$$= 28.65,\ 30.53$$

Thus, the average leaf area is expected to be in the range 28.65 to 30.53 sq cm, with probability approximately equal to .95.

It should be noted here that if the information on the auxiliary variable x was not used in the form of the regression estimator and the average leaf area was estimated using the simple sample mean $\bar{y}$, the estimate of the variance $V(\bar{y})$ from (3.10) would have been

$$
\begin{aligned}
v(\bar{y}) &= \frac{N-n}{Nn}\, s_y^2 \\
&= \frac{2106-33}{(2106)\,(33)}\,(61.003) \\
&= 1.8196
\end{aligned}
$$

which is much larger than $mse(\bar{y}_{lr})$. The estimated percent relative efficiency of the linear regression estimator, with respect to the simple mean $\bar{y}$, is given by

$$
\begin{aligned}
RE &= \frac{v(\bar{y})}{mse\,(\bar{y}_{lr})}\,(100) \\
&= \frac{1.8196}{.2200}\,(100) \\
&= 827.09
\end{aligned}
$$

Hence, we find that the linear regression estimator is over 8 times more efficient than the simple mean $\bar{y}$. One could also say, that the use of auxiliary information, in the form of regression estimator, has reduced the error in the estimate of average leaf area by over eight times. ■

8.4 SAMPLE SIZE DETERMINATION FOR ESTIMATING MEAN / TOTAL

For estimating the size of the sample required to estimate population mean/total with a specified bound B on the error of estimation, we proceed as in chapter 7. Let a preliminary sample of n_l units be selected using SRS without replacement. The values of $(s_y^2 + \beta_o^2 s_x^2 - 2\beta_o s_{xy})$ and $(s_y^2 + \hat{\beta}^2 s_x^2 - 2\hat{\beta}\, s_{xy})$ are obtained using data from this preliminary sample, and are redenoted as

$$
\begin{aligned}
s_{l1}^2 &= s_{yl}^2 + \beta_o^2\, s_{xl}^2 - 2\beta_o\, s_{xyl} \\
s_{l2}^2 &= s_{yl}^2 + \hat{\beta}_l^2\, s_{xl}^2 - 2\hat{\beta}_l\, s_{xyl}
\end{aligned}
$$

respectively. These values are then used in $v(\bar{y}_d)$ and $mse(\bar{y}_{lr})$ given in (8.3) and (8.7) respectively, in place of the corresponding values that would have been obtained from the sample of n units. The equations

$$2\sqrt{v(\bar{y}_d)} = B \qquad \text{(for difference estimator)} \tag{8.8}$$

$$2\sqrt{mse\,(\bar{y}_{lr})} = B \qquad \text{(for linear regression estimator)} \tag{8.9}$$

are then solved for n.

For estimating total, $v(\bar{y}_d)$ and $mse(\bar{y}_{lr})$ in the above two equalities are multiplied by N^2. Depending on the estimator used, the solution of (8.8) or (8.9) for n, gives the formula for determining the required sample size when estimating population mean/total. Thus we have :

Sample size required to estimate the population mean/total with a bound B on the error of estimation :

$$n = \frac{Ns_{li}^2}{ND + s_{li}^2},\ i = 1, 2 \qquad (8.10)$$

where

$$D = \frac{B^2}{4} \qquad \text{(when estimating mean)}$$

$$D = \frac{B^2}{4N^2} \qquad \text{(when estimating total)}$$

$$s_{l1}^2 = s_{yl}^2 + \beta_o^2\, s_{xl}^2 - 2\beta_o\, s_{xyl} \qquad \text{(in case of difference estimator)}$$

$$s_{l2}^2 = s_{yl}^2 + \hat{\beta}_l^2\, s_{xl}^2 - 2\hat{\beta}_l\, s_{xyl} \qquad \text{(for linear regression estimator)}$$

If $n_l \geq n$, the sample size n_l is sufficient. Otherwise, $(n-n_l)$ more units will have to be selected in the sample.

Example 8.3

Assume that the sample drawn in example 8.2 is a preliminary sample. Based on information from this sample, determine the sample size required to estimate the average leaf area with bound on the error of estimation as one sq cm.

Solution

Here, we have N=2106 and B=1. The sample of 33 plants selected in example 8.2 is now taken as the preliminary sample. All the sample estimates computed in example 8.2 will, therefore, be treated as preliminary sample estimates. Thus,

$$n_l = 33,\ \hat{\beta} = \hat{\beta}_l = .5346,\ s_y^2 = s_{yl}^2 = 61.003,$$

$$s_x^2 = s_{xl}^2 = 187.631, \text{ and } s_{xy} = s_{xyl} = 100.312.$$

Then we compute

$$\begin{aligned} s_{l2}^2 &= s_{yl}^2 + \hat{\beta}_l^2\, s_{xl}^2 - 2\hat{\beta}_l\, s_{xyl} \\ &= 61.003 + (.5346)^2\,(187.631) - 2(.5346)\,(100.312) \\ &= 7.374 \end{aligned}$$

Further,

$$D = \frac{B^2}{4} = \frac{1}{4} = .25$$

The required sample size can then be obtained from (8.10). This means,

$$n = \frac{Ns_{12}^2}{ND + s_{12}^2}$$

$$= \frac{(2106)(7.374)}{(2106)(.25) + 7.374}$$

$$= 29.1$$

$$\approx 29$$

As $n_l > n$, the already selected sample of 33 leaves is sufficient to achieve the required margin of error. ■

8.5 SEPARATE AND COMBINED REGRESSION ESTIMATORS

As with the ratio estimator, two types of regression estimators can be developed for stratified random sampling. In the first estimator $\bar{y}_{lrs}$, a *separate regression estimate* is computed for each stratum mean and then their weighted average is taken. This estimator is appropriate, when one suspects that the true regression coefficients β_h, h=1,2,..., L, vary from stratum to stratum. As in chapters 5 and 7, let n_h be the number of units selected in the sample from the h-th stratum, containing N_h units, using SRS without replacement. Let

$$\rho_h = \frac{S_{hxy}}{S_{hy} S_{hx}}$$

$$r_h = \frac{s_{hxy}}{s_{hy} s_{hx}}$$

denote the correlation coefficients for the h-th stratum computed from N_h and n_h units respectively. Further, let

$$\beta_h = \frac{S_{hxy}}{S_{hx}^2}$$

be the unknown regression coefficient of y on x in the h-th stratum, and

$$\hat{\beta}_h = \frac{s_{hxy}}{s_{hx}^2} \tag{8.11}$$

its least square estimator from the sample of size n_h, h = 1, 2,..., L. We shall also assume that the sample sizes $\{n_h\}$ are large enough, so that, the approximations used in $B(\bar{y}_{lrs})$, $MSE(\bar{y}_{lrs})$, and estimator $mse(\bar{y}_{lrs})$ are sufficiently accurate.

Separate regression estimator of population mean $\bar{Y}$:

$$\begin{aligned}\bar{y}_{lrs} &= \sum_{h=1}^{L} W_h \, \bar{y}_{hlr} \\ &= \sum_{h=1}^{L} W_h \, [\, \bar{y}_h + \hat{\beta}_h \, (\bar{X}_h - \bar{x}_h)]\end{aligned} \tag{8.12}$$

Bias of estimator $\bar{y}_{lrs}$:

$$B(\bar{y}_{lrs}) = - \sum_{h=1}^{L} W_h \, \mathrm{Cov}(\bar{x}_h, \hat{\beta}_h) \tag{8.13}$$

Approximate mean square error of estimator $\bar{y}_{lrs}$:

$$\begin{aligned}MSE(\bar{y}_{lrs}) &= \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h \, n_h}\right) (S_{hy}^2 + \beta_h^2 \, S_{hx}^2 - 2\beta_h S_{hxy}) \\ &= \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h \, n_h}\right) S_{hy}^2 \, (1 - \rho_h^2)\end{aligned} \tag{8.14}$$

Estimator of mean square error MSE $(\bar{y}_{lrs})$:

$$\begin{aligned}mse(\bar{y}_{lrs}) &= \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h \, n_h}\right) (s_{hy}^2 + \hat{\beta}_h^2 \, s_{hx}^2 - 2\hat{\beta}_h \, s_{hxy}) \\ &= \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h \, n_h}\right) s_{hy}^2 \, (1 - r_h^2)\end{aligned} \tag{8.15}$$

However, if a predetermined value β_{ho} is used for β_h, the bias B $(\bar{y}_{lrs})$ will be zero and $\hat{\beta}_h$ in mse $(\bar{y}_{lrs})$ will be replaced by β_{ho}. In this situation, the first expressions in (8.14) and (8.15) will not involve any approximation and will be exact. Also, these will not be equal to the corresponding second expressions, which will cease to hold in this case.

Example 8.4

A refrigerator manufacturing company contemplates to review its existing marketing policy. It has, therefore, decided to estimate the total number of refrigerators expected to be sold during the current summer season, half of which is almost over. Keeping various relevant factors in view, the whole country is divided into 4 zones. The number of registered dealers in these four zones are 400, 216, 364, and 274, whereas the total number of refrigerators sold by them during last summer are 29100, 12060, 26567, and 18111 respectively. Treating zones as strata, it was decided to select an overall sample of 42 dealers. Neyman allocation method was used to allocate it to different strata. The population mean squares for the number of refrigerators sold by a dealer during last summer season, for the four zones respectively, are 207.36, 282.24, 184.96, and 127.69. The samples of sizes 14, 9, 12, and 7 dealers were then selected from zones I, II, III, and IV respectively (the procedure of allocation is explained in solution). The data in respect of number of refrigerators sold during last summer season and expected to be sold during current season are given in table 8.3 below.

Table 8.3 The number of refrigerators sold during last summer (LS) and expected sale for the current summer (CS)

Zone I		Zone II		Zone III		Zone IV	
LS (x)	CS (y)	LS (x)	CS (y)	LS (x)	CS (y)	LS (x)	CS (y)
53	69	44	52	60	67	58	52
84	80	67	73	76	86	65	71
93	87	84	78	78	75	56	62
66	72	52	60	68	77	48	44
77	81	48	42	55	64	73	77
82	94	62	56	48	45	85	80
68	64	56	50	86	98	61	66
84	88	70	76	91	95		
79	72	40	48	69	76		
98	110			70	79		
50	62			79	92		
78	70			49	66		
92	85						
63	77						

Solution

We are given that

$$n = 42,\ N_1 = 400,\ N_2 = 216,\ N_3 = 364,\ N_4 = 274,\ N = \sum_{h=1}^{4} N_h = 1254,$$

$$S_1^2 = 207.36,\ S_2^2 = 282.24,\ S_3^2 = 184.96,\ S_4^2 = 127.69, \text{ and } X = \sum_{h=1}^{4} X_h = 85838.$$

The first step is to work out the sizes of samples to be selected from different strata. For this we need

$$\sum_{h=1}^{4} N_h S_h = (400)(14.4) + (216)(16.8) + (364)(13.6) + (274)(11.3)$$

$$= 17435.4$$

Then according to Neyman allocation method given in (5.14), the sample size for the h-th stratum is given by

$$n_h = n \frac{N_h S_h}{\Sigma N_h S_h}, \; h = 1, 2, 3, 4$$

This gives

$$n_1 = (42) \frac{(400)\cdot(14.4)}{17435.4} = 13.9 \approx 14$$

$$n_2 = (42) \frac{(216)\,(16.8)}{17435.4} = 8.7 \approx 9$$

$$n_3 = (42) \frac{(364)\,(13.6)}{17435.4} = 11.9 \approx 12$$

$$n_4 = (42) \frac{(274)\,(11.3)}{17435.4} = 7.4 \approx 7$$

As in chapters 5 and 7, the sample means, sample mean squares, and products for each of the four strata have been computed, and are given in table 8.4 along with certain other sample values.

Table 8.4 Different population and sample values for the four strata

Zone I	Zone II	Zone III	Zone IV
n_1 = 14	n_2 = 9	n_3 = 12	n_4 = 7
N_1 = 400	N_2 = 216	N_3 = 364	N_4 = 274
W_1 = .3190	W_2 = .1722	W_3 = .2903	W_4 = .2185
X_1 = 29100	X_2 = 12060	X_3 = 26567	X_4 = 18111
$\bar{X}_1$ = 72.8	$\bar{X}_2$ = 55.8	$\bar{X}_3$ = 73.0	$\bar{X}_4$ = 66.1
$\bar{y}_1$ = 79.4	$\bar{y}_2$ = 59.4	$\bar{y}_3$ = 76.7	$\bar{y}_4$ = 64.6
$\bar{x}_1$ = 76.2	$\bar{x}_2$ = 58.1	$\bar{x}_3$ = 69.1	$\bar{x}_4$ = 63.7
s^2_{1y} = 166.7	s^2_{2y} = 174.3	s^2_{3y} = 226.6	s^2_{4y} = 170.6
s^2_{1x} = 211.1	s^2_{2x} = 197.1	s^2_{3x} = 193.0	s^2_{4x} = 147.9
s_{1xy} = 146.5	s_{2xy} = 164.8	s_{3xy} = 188.3	s_{4xy} = 142.8
$\hat{\beta}_1$ = .6940	$\hat{\beta}_2$ = .8361	$\hat{\beta}_3$ = .9756	$\hat{\beta}_4$ = .9655
r_1 = .7810	r_2 = .8891	r_3 = .9004	r_4 = .8990

Since the β_h, h = 1, 2, 3, 4, values are likely to differ from stratum to stratum, we go for separate regression estimate for estimating the total expected sale of refrigerators. We, therefore, work out estimator $\hat{Y}_{lrs} = N\bar{y}_{lrs}$ of population total, where the estimator $\bar{y}_{lrs}$ is defined in (8.12). This yields

$$\begin{aligned}
\hat{Y}_{lrs} &= N\bar{y}_{lrs} = N \sum_{h=1}^{L} W_h [\,\bar{y}_h + \hat{\beta}_h (\bar{X}_h - \bar{x}_h)] \\
&= 1254\,[.3190\,\{79.4 + (.6940)\,(72.8 - 76.2)\} \\
&\quad + .1722\,\{59.4 + (.8361)\,(55.8 - 58.1)\} \\
&\quad + .2903\,\{76.7 + (.9756)\,(73.0 - 69.1)\} \\
&\quad + .2185\,\{64.6 + (.9655)\,(66.1 - 63.7)\}] \\
&= 1254\,(24.58 + 9.90 + 23.37 + 14.62) \\
&= 90877.38 \\
&\approx 90877
\end{aligned}$$

For estimating mean square error of $\hat{Y}_{lrs}$, we use second version of (8.15). Thus,

$$\begin{aligned}
\text{mse}\,(\hat{Y}_{lrs}) &= N^2\text{mse}\,(\bar{y}_{lrs}) = N^2 \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h\, n_h}\right) s_{hy}^2\,(1 - r_h^2) \\
&= (1254)^2 \left[(.3190)^2 \left(\frac{400-14}{(400)\,(14)}\right)(166.7)\,\{1-(.7810)^2\}\right. \\
&\quad + (.1722)^2 \left(\frac{216-9}{(216)\,(9)}\right)(174.3)\,\{1-(.8891)^2\} \\
&\quad + (.2903)^2 \left(\frac{364-12}{(364)\,(12)}\right)(226.6)\,\{1-(.9004)^2\} \\
&\quad \left. + (.2185)^2 \left(\frac{274-7}{(274)\,(7)}\right)(170.6)\,\{1-(.8990)^2\}\right] \\
&= (1254)^2\,(.4561 + .1153 + .2913 + .2175) \\
&= 1698631.7
\end{aligned}$$

Below we work out the confidence limits in which the total number of refrigerators, expected to be sold during the current summer season, is likely to fall. These limits are given by

$$\begin{aligned}
&\hat{Y}_{lrs} \pm 2\sqrt{\text{mse}\,(\hat{Y}_{lrs})} \\
&= 90877.38 \pm 2\sqrt{1698631.7} \\
&= 90877.38 \pm 2606.63 \\
&= 88270.75,\ 93484.01 \\
&\approx 88271,\ 93484
\end{aligned}$$

In case, the information on refrigerators sold during the last summer season was not available, or not used, the expected total sale of refrigerators for this summer would have been estimated by using stratified simple random sampling WOR estimator given in (5.1), as

$$\hat{Y}_{st} = N \sum_{h=1}^{L} W_h \bar{y}_h = \sum_{h=1}^{L} N_h \bar{y}_h$$

The estimated variance of the estimator $\hat{Y}_{st}$, following (5.3), would have been

$$\begin{aligned} v(\hat{Y}_{st}) &= N^2 \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h \, n_h}\right) s_{hy}^2 \\ &= \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{n_h} s_{hy}^2 \\ &= \frac{400\,(400-14)}{14}(166.7) + \frac{216\,(216-9)}{9}(174.3) \\ &\quad + \frac{364\,(364-12)}{12}(226.6) + \frac{274\,(274-7)}{7}(170.6) \\ &= 6906833.9 \end{aligned}$$

Hence, the estimated percent relative efficiency of the separate linear regression estimator $\hat{Y}_{lrs}$ with respect to the stratified SRS without replacement estimator $\hat{Y}_{st}$, is given by

$$\begin{aligned} RE &= \frac{v(\hat{Y}_{st})}{mse\,(\hat{Y}_{lrs})}(100) \\ &= \frac{6906833.9}{1698631.7}(100) \\ &= 406.6 \end{aligned}$$

The increased efficiency of estimation can, therefore, be attributed to the use of auxiliary information through separate linear regression estimator. ■

When there is evidence that $\{\beta_h\}$ values do not differ much from stratum to stratum, we can use an alternative estimator $\bar{y}_{lrc}$, which is termed as *combined regression estimator.* Once again we assume that the sample sizes $\{n_h\}$ are sufficiently large, so that, the approximate expressions given here are close to the actual values. Now define

$$\beta_c = \frac{\sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h \, n_h}\right) S_{hxy}}{\sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h \, n_h}\right) S_{hx}^2}$$

$$\hat{\beta}_c = \frac{\sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h\, n_h}\right) s_{hxy}}{\sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h\, n_h}\right) s_{hx}^2} \tag{8.16}$$

Then we have the results (8.17) through (8.20).

Combined regression estimator of population mean $\overline{Y}$:

$$\left.\begin{aligned} \overline{y}_{lrc} &= \overline{y}_{st} + \hat{\beta}_c\,(\overline{X} - \overline{x}_{st}) \\ &= \sum_{h=1}^{L} W_h \overline{y}_h + \hat{\beta}_c\,(\overline{X} - \sum_{h=1}^{L} W_h \overline{x}_h) \end{aligned}\right] \tag{8.17}$$

Bias of estimator $\overline{y}_{lrc}$:

$$B\,(\overline{y}_{lrc}) = -\,\mathrm{Cov}\,(\overline{x}_{st},\, \hat{\beta}_c) \tag{8.18}$$

Approximate mean square error of estimator $\overline{y}_{lrc}$:

$$\mathrm{MSE}(\overline{y}_{lrc}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h\, n_h}\right) (S_{hy}^2 + \beta_c^2\, S_{hx}^2 - 2\beta_c\, S_{hxy}) \tag{8.19}$$

Estimator of $\mathrm{MSE}(\overline{y}_{lrc})$:

$$\mathrm{mse}(\overline{y}_{lrc}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h\, n_h}\right) (s_{hy}^2 + \hat{\beta}_c^2\, s_{hx}^2 - 2\hat{\beta}_c\, s_{hxy}) \tag{8.20}$$

The bias in the separate regression estimator, as compared to the combined estimator, is large if the sample sizes $\{n_h\}$ are rather small. As stated earlier, if the regression coefficients $\{\beta_h\}$ do not seem to vary appreciably from stratum to stratum, a combined regression estimator should be preferred. However, if $\{\beta_h\}$ do vary from stratum to stratum, one should go for the separate estimator.

Example 8.5

In order to arrive at a more logical estimate of average leaf area for the newly developed strain of wheat (example 8.2), the physiologist raised 40 plants at each of the 3 different locations. This gave rise to 640, 710, and 769 leaves respectively. The total weight of these leaves, recorded at three locations, was found to be 69.000, 81.137, and 78.009 gm respectively. A WOR stratified random sample of 39 leaves was selected using proportional allocation (details given in solution). The observations recorded on sample leaves, in respect of area (in sq cm) and weight (in mg), are given in table 8.5. Estimate the average leaf area, and also build up the confidence interval for it.

Table 8.5 Area and weight for sample leaves

	Locations					
	I		II		III	
Leaf	Area (y)	Weight (x)	Area (y)	Weight (x)	Area (y)	Weight (x)
1	21.08	97	41.07	130	26.01	103
2	25.70	103	26.13	107	18.00	89
3	34.23	119	28.05	109	17.92	91
4	26.16	107	33.71	117	26.73	105
5	19.37	99	28.56	112	24.81	101
6	28.00	103	29.43	110	28.30	107
7	24.03	91	22.41	105	16.07	81
8	36.61	123	32.06	113	29.41	111
9	34.09	117	27.64	108	21.09	104
10	19.84	96	21.00	102	35.47	121
11	22.18	102	34.78	122	31.57	113
12	17.76	84	23.17	106	39.06	129
13			28.21	101	20.66	99
14					26.70	106

Solution

The number of units at three locations (strata) are N_1 = 640, N_2=710, and N_3=769, so that, $N=N_1+N_2+N_3$ = 640+710+769 = 2119. The total sample size is n=39 leaves. The number of units to be selected from each stratum are calculated by using (5.9). This gives

$$n_1 = \left(\frac{n}{N}\right)N_1 = \left(\frac{39}{2119}\right)640 = 11.8 \approx 12$$

$$n_2 = \left(\frac{n}{N}\right)N_2 = \left(\frac{39}{2119}\right)710 = 13.1 \approx 13$$

$$n_3 = \left(\frac{n}{N}\right)N_3 = \left(\frac{39}{2119}\right)769 = 14.2 \approx 14$$

The observations on area and weight from these selected leaves are given in table 8.5. The sample estimates $\bar{y}_h$, $\bar{x}_h$, s^2_{hy}, s^2_{hx}, and s_{hxy}, h=1,2,3, can be computed as in chapter 7. The estimates of regression coefficients are obtained from (8.11). The necessary computations are presented in table 8.6.

Table 8.6 Computations for samples selected from different strata

Location I	Location II	Location III
$n_1 = 12$	$n_2 = 13$	$n_3 = 14$
$N_1 = 640$	$N_2 = 710$	$N_3 = 769$
$W_1 = .3020$	$W_2 = .3351$	$W_3 = .3629$
$X_1 = 69000$	$X_2 = 81137$	$X_3 = 78009$
$\bar{y}_1 = 25.75$	$\bar{y}_2 = 28.94$	$\bar{y}_3 = 25.84$
$\bar{x}_1 = 103.4$	$\bar{x}_2 = 110.9$	$\bar{x}_3 = 104.3$
$\bar{X}_1 = 107.8$	$\bar{X}_2 = 114.3$	$\bar{X}_3 = 101.4$
$s^2_{1y} = 40.148$	$s^2_{2y} = 30.334$	$s^2_{3y} = 45.446$
$s^2_{1x} = 133.90$	$s^2_{2x} = 66.24$	$s^2_{3x} = 154.99$
$s_{1xy} = 68.410$	$s_{2xy} = 41.758$	$s_{3xy} = 81.065$
$\hat{\beta}_1 = .5109$	$\hat{\beta}_2 = .6304$	$\hat{\beta}_3 = .5230$
$r_1 = .933$	$r_2 = .932$	$r_3 = .966$

Calculations in table 8.6 indicate that the regression coefficients $\{\beta_h\}$ do not differ much from stratum to stratum. In such a situation, the combined regression estimator is more appropriate. We first work out stratified estimates of means for variables y and x. Thus from (5.1)

$$\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h\bar{y}_h = \sum_{h=1}^{L} W_h\bar{y}_h$$

$$= (.3020)(25.75) + (.3351)(28.94) + (.3629)(25.84)$$

$$= 26.85$$

Similarly,

$$\bar{x}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h\bar{x}_h = \sum_{h=1}^{L} W_h\bar{x}_h$$

$$= (.3020)(103.4) + (.3351)(110.9) + (.3629)(104.3)$$

$$= 106.24$$

Also,

$$\sum_{h=1}^{L} W_h^2\left(\frac{N_h - n_h}{N_h\ n_h}\right)s_{hxy} = (.3020)^2\left(\frac{640-12}{(640)(12)}\right)(68.410)$$

$$+ (.3351)^2\left(\frac{710-13}{(710)(13)}\right)(41.758)$$

$$+ (.3629)^2\left(\frac{769-14}{(769)(14)}\right)(81.065)$$

$$= .5102 + .3541 + .7487$$

$$= 1.6130$$

$$\sum_{h=1}^{L} W_h^2\left(\frac{N_h - n_h}{N_h\ n_h}\right)s^2_{hx} = (.3020)^2\left(\frac{640-12}{(640)(12)}\right)(133.90)$$

$$+ (.3351)^2 \left(\frac{710-13}{(710)\ (13)}\right) (66.24)$$

$$+ (.3629)^2 \left(\frac{769-14}{(769)\ (14)}\right) (154.99)$$

$$= .9986 + .5617 + 1.4314$$

$$= 2.9917$$

Then on using the above computed values in (8.16), we obtain the estimate of combined regression coefficient β_c as

$$\hat{\beta}_c = \frac{1.6130}{2.9917} = .5392$$

Now,

$$X = 69000 + 81137 + 78009 = 228146$$

$$\overline{X} = \frac{228146}{2119} = 107.67$$

Using combined regression estimator defined in (8.17), the estimate of average leaf area is

$$\overline{y}_{lrc} = \overline{y}_{st} + \hat{\beta}_c (\overline{X} - \overline{x}_{st})$$

$$= 26.85 + (.5392)(107.67 - 106.24)$$

$$= 27.62$$

From (8.20), the estimate of mean square error of $\overline{y}_{lrc}$ is obtained as

$$\text{mse}(\overline{y}_{lrc}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h\ n_h}\right) (s_{hy}^2 + \hat{\beta}_c^2 s_{hx}^2 - 2\hat{\beta}_c s_{hxy})$$

$$= (.3020)^2 \left(\frac{640-12}{(640)\ (12)}\right) [40.148 + (.5392)^2 (133.90) - 2(.5392)\ (68.410)]$$

$$+ (.3351)^2 \left(\frac{710-13}{(710)\ (13)}\right) [30.334 + (.5392)^2 (66.24) - 2(.5392)(41.758)]$$

$$+ (.3629)^2 \left(\frac{769-14}{(769)\ (14)}\right) [45.446 + (.5392)^2 (154.99) - 2(.5392)(81.065)]$$

$$= .0396 + .0387 + .0285$$

$$= .1068$$

The confidence interval for the actual average leaf area is now obtained from

$$\overline{y}_{lrc} \pm 2 \sqrt{\text{mse}(\overline{y}_{lrc})}$$

$$= 27.62 \pm 2\sqrt{.1068}$$

$$= 27.62 \pm .65$$

$$= 26.97,\ 28.27$$

If the information on the leaf weight was not available, or it was not used, the average leaf area could be estimated by using the stratified SRS without replacement estimator $\bar{y}_{st}$ defined in (5.1). The estimator of variance for $\bar{y}_{st}$ is given by (5.3), which in this case becomes

$$\begin{aligned} v(\bar{y}_{st}) &= \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h\, n_h}\right) s_{hy}^2 \\ &= (.3020)^2 \left(\frac{640-12}{(640)\,(12)}\right)(40.148) + (.3351)^2 \left(\frac{710-13}{(710)\,(13)}\right)(30.334) \\ &\quad + (.3629)^2 \left(\frac{769-14}{(769)\,(14)}\right)(45.446) \\ &= .2994 + .2572 + .4197 \\ &= .9763 \end{aligned}$$

Therefore, the estimated percent relative efficiency of the combined regression estimator, in relation to the stratified without replacement SRS estimator, is given by

$$\begin{aligned} RE &= \frac{v(\bar{y}_{st})}{\text{mse}\,(\bar{y}_{lrc})}\,(100) \\ &= \frac{.9763}{.1068}\,(100) \\ &= 914.14 \end{aligned}$$

Hence, the use of information on leaf weight through the combined regression estimator has reduced the error of estimation for average leaf area to about one ninth of the stratified estimator which does not use any auxiliary information. ■

8.6 SOME FURTHER REMARKS

8.1 *Unbiased regression estimators* have been developed by Mickey (1959) and Williams (1963). Singh and Srivastava (1980) have proposed a sampling scheme for which the usual regression estimator becomes unbiased. Rao (1969) has found Mickey's estimator usually less efficient as compared to the usual regression estimator, in natural populations.

8.2 Often, the data are available on several auxiliary characteristics. In such cases, it will be beneficial to build up regression estimator which uses all the available information. Ghosh (1947) has proposed an estimator of the type

$$\bar{y}_{lrg} = \bar{y} + \sum_{i=1}^{k} \hat{\beta}_i\,(\bar{X}_i - \bar{x}_i)$$

where the symbols stand for their usual meaning, and k is the number of auxiliary variables on which the information is available.

8.3 Des Raj (1965) has proposed a *multivariate difference estimator* as

$$\bar{y}_{lrd} = \sum_{i=1}^{k} W_i [\,\bar{y} + \tau_i (\bar{X}_i - \bar{x}_i)]$$

where the weights W_i add up to unity and τ_i's are known constants.

LET US DO

8.1 Describe difference and regression methods of estimation for estimating population mean/total.

8.2 Using the data for a hypothetical population given in example 7.2, work out the relative efficiency of regression estimator $\bar{y}_{lr}$ with respect to the usual mean estimator $\bar{y}$, for a WOR simple random sample of size 3.

8.3 The total number of households in a development block consisting of 150 villages is 4066. A social scientist, interested in estimating total number of TV sets in the block, selected 30 villages using SRS without replacement procedure. Information on number of households and the number of TV sets for the 30 sample villages is given below :

Village	House-holds	TV sets	Village	House-holds	TV sets	Village	House-holds	TV sets
1	500	158	11	170	58	21	560	201
2	206	60	12	110	37	22	410	137
3	373	107	13	280	76	23	380	121
4	120	35	14	440	138	24	109	33
5	470	135	15	95	42	25	406	160
6	310	108	16	396	128	26	220	66
7	425	138	17	333	147	27	380	117
8	610	198	18	178	52	28	310	122
9	204	60	19	270	87	29	580	213
10	370	116	20	343	108	30	76	22

Estimate total number of TV sets in the block by using difference estimator, and also build up confidence interval for it. From an earlier survey, the value of the regression coefficient is known to be .35.

8.4 A campaign was launched to make the people aware of the fact that overweight persons run a heavy risk to their lives. As a result, 1000 overweight women got registered with a *Yoga Ashram* (an institution that conducts yoga lessons) for the purpose of reducing their weight and become more fit. At the time of registration, weight (in kg) of each woman was recorded and the average weight of all the 1000 women came to be 65 kg. All the registered women were given lessons on the physical exercises to be undertaken daily. After 3 months, to determine the impact of the weight reducing program, the organizers selected a sample of 30 women and weighed them individually again. Present weight (y) and the initial weight (x) for the sample women are presented in the following table:

Woman	x	y	Woman	x	y	Woman	x	y
1	67.5	66.0	11	68.7	66.9	21	73.2	69.8
2	70.6	67.3	12	66.4	64.1	22	67.6	66.3
3	60.4	58.7	13	70.1	68.7	23	68.0	65.8
4	68.9	66.3	14	65.5	63.2	24	64.3	60.6
5	72.0	69.1	15	72.8	69.6	25	69.0	67.4
6	66.3	64.8	16	63.5	61.0	26	70.8	67.3
7	69.7	67.4	17	66.7	64.7	27	63.5	60.5
8	64.8	63.5	18	69.1	67.8	28	68.1	65.0
9	71.6	68.2	19	64.3	63.5	29	62.8	60.0
10	65.9	63.6	20	70.6	69.1	30	64.6	63.2

Using regression method of estimation, estimate the average present weight of a woman, and also build up the confidence interval for it. Also, compute its estimated relative efficiency in relation to the usual SRS estimator $\bar{y}$.

8.5 Employing regression method, estimate the total number of dwellings occupied by renters from the survey data given in exercise 7.6. Also, obtain confidence limits for it.

8.6 Using data of exercise 8.3, estimate the total number of TV sets in the block, and place confidence limits on it. Assume that the value of regression coefficient is not available in advance.

8.7 Suppose you are to estimate population total using regression method of estimation. Discuss how would you determine the optimal sample size from the information provided by a preliminary sample of size n_1?

8.8 Suppose that the sample of 30 villages, drawn from a total of 150 villages in exercise 8.3, is a preliminary sample. Using data of this exercise, examine whether this sample size is sufficient, or it has to be supplemented by selecting additional units, if one is interested in estimating total number of TV sets in the block with a margin of error equal to 50 TV sets ?

8.9 Assume that the sample of 30 women, drawn in exercise 8.4, is the preliminary sample. Using the data for this sample, comment whether, or not, this sample size is sufficient if the present average weight is to be estimated with a margin of error of 2 kg ? If not, what will you suggest ?

8.10 A plant breeder had limited amount of seed for 3 newly developed strains of sugarcane. He raised 36, 72, and 42 plants of strains 1, 2, and 3 respectively. Being in possession of only a limited quantity of seed for the valuable strains, he could not afford to crush all the 150 canes to estimate its juice content. Using proportional allocation, he selected 6, 12, and 7 plants of strains 1, 2, and 3 respectively through simple random sampling WOR, so that, the overall sample was of 25 plants. Assume that the total weight of all the 150 canes is 70 kg. The quantity of juice and the weight of respective selected canes, both in grams, are given below :

Strain 1		Strain 2				Strain 3	
Cane	Juice	Cane	Juice	Cane	Juice	Cane	Juice
300	125	270	90	320	100	250	80
450	150	320	105	340	110	320	90
360	130	410	135	310	100	300	70
340	135	360	110	260	80	310	75
400	140	290	90	280	80	420	100
350	130	270	95	300	95	340	80
						280	70

Estimate juice quantity per cane by using separate regression estimator, and also determine the lower and upper confidence limits for it.

8.11 The Department of Animal Husbandry of a state government has undertaken a project on feeding and management practices of milch cows in a district comprising of 3 development blocks. These blocks, consisting of 70, 120, and 50 villages respectively, were treated as strata. A WOR random sample of 24 villages was drawn using proportional allocation. That means,

$$n_1 = \left(\frac{24}{240}\right) 70 = 7$$

$$n_2 = \left(\frac{24}{240}\right) 120 = 12$$

$$n_3 = \left(\frac{24}{240}\right) 50 = 5$$

villages were selected from strata I, II, and III respectively. The following table presents the total number of milch cows in 7, 12, and 5 randomly selected villages of strata I, II, and III respectively, during March 1993 and as per 1990 livestock census.

Stratum I			Stratum II			Stratum III	
1990	1993	1990	1993	1990	1993	1990	1993
17	21	21	18	18	24	16	21
19	22	14	21	8	15	21	18
9	11	16	20	16	11	13	16
13	14	18	24	26	21	19	25
22	18	26	20	11	16	20	23
16	21	13	20				
11	15	20	25				

From the 1990 census records, total number of milch cows in strata I, II, and III were 1260, 2400, and 1150 respectively. Using combined regression estimator, estimate the total number of milch cows in the district in March 1993. Also, obtain lower and upper confidence limits for it.