CHAPTER 9

Two-Phase Sampling

9.1 NEED FOR TWO-PHASE SAMPLING

The discussion in some of the previous chapters has revealed that the prior information on an auxiliary variable could be used to enhance the precision of an estimator. Ratio, product, and regression estimators require the knowledge of population mean $\bar{X}$ (or equivalently of total X) for the auxiliary variable x. For stratifying the population on the basis of the auxiliary variable, knowledge of its frequency distribution is required. When such information is lacking, it is some times less expensive to select a large sample (called *first-phase sample* or *initial sample*) on which auxiliary variable alone is observed. The purpose of this is to furnish a good estimate of $\bar{X}$, or equivalently of X. Frequency distribution for the auxiliary variable can also be estimated from observations made on the first-phase sample. A subsample (also called *final sample* or *second-phase sample*) from the initial sample is selected for observing the variable of interest. Information collected on the two samples is then used to construct estimators for the parameter under consideration.

As an illustration, let us consider the problem of estimating total production of cow milk in a certain region. For this purpose, we take village as the sampling unit and the number of milch cows in a village as the auxiliary variable. Since the total number of milch cows in all the villages of the region may not be available, the investigator could decide to take a large initial sample of villages, and collect information on number of milch cows in the sample villages. This information is then used to build up an estimate of X, the total number of milch cows in the region. The estimate of X, so obtained, could then be used in place of X in the ratio or regression estimate of total production of cow milk in the region. A subsample of villages is selected from the first-phase sample to observe the study variable, viz., cow milk yield in the village.

As yet another illustration, let us consider the situation of example 8.2, where a physiologist wanted to estimate average leaf area for a new strain of wheat. It may, sometimes, not be desirable to pluck all the leaves in the population of 120 plants and obtain total weight X for building a regression estimator of the average leaf area. It will, therefore, be more appropriate to select a large first-phase sample of leaves and measure weight for the sample leaves. A subsample from this initial sample of leaves could then be selected to determine leaf area. An estimate for the average weight $\bar{X}$ of leaves from

all the 120 plants could then be obtained from the observations made on the initial sample. This estimate can then be used in place of $\bar{X}$ in the regression estimate of the average leaf area.

The main point of deviation from the previously discussed procedures is that the sample is now drawn in two phases - first a large initial sample and then a subsample from this initial sample. Hence the name of the procedure.

Definition 9.1 *Two-phase sampling* (or *double sampling*) is a procedure where the lacking information on the auxiliary variable is collected from a large first-phase sample, and the study variable is observed on a smaller subsample selected from the first-phase sample.

When the sampling procedure is completed in three or more phases, the sampling procedure is termed as *multiphase sampling*. This procedure differs from the multistage sampling in the sense that the former requires a complete sampling frame of the ultimate sampling units, whereas in the latter a frame of the next stage units is necessary only for the sample units selected at that stage. Also, the sampling unit in case of multiphase sampling remains same at each phase of sampling, whereas it changes in case of multistage sampling. For example, if the sampling unit is a household, then in two-phase sampling, both first and second phase samples will be samples of households, whereas in case of multistage sampling if the first-stage sample is a sample of villages, the second-stage sample may be the sample of households.

It should be noted that in case of two-phase sampling, the size of the second sample on which we observe the study variable, will be reduced for the fixed total budget. This is because we had to spend part of the budget to observe auxiliary variable on the initial sample which could have otherwise been used to observe study variable on a comparatively larger sample. The technique is, therefore, beneficial only if the gain in precision is more than the loss in precision due to the reduction in the size of the final sample.

Two-phase sampling has been used in different ways by research workers. We shall, however, restrict our discussion on double sampling to ratio, product, regression, and PPS estimation procedures only.

9.2 TWO-PHASE SAMPLING IN RATIO, PRODUCT, AND REGRESSION METHODS OF ESTIMATION

While estimating mean/total through ratio, product, and regression methods of estimation, it was assumed that the population mean $\bar{X}$, or total X, for the auxiliary variable is known. If this information is lacking, the technique of double sampling provides an alternative. A large first-phase sample of n' units is drawn to estimate $\bar{X}$. Then, a final sample of n units is taken from n' units of the first-phase sample to observe the estimation variable y. If both the first-phase and the final samples are drawn using WOR equal probabilities

sampling, the different estimators and the other related results are given in sections 9.2.1, 9.2.2, and 9.2.3.

9.2.1 *Double Sampling for Ratio Method of Estimation*

Estimator of population mean $\bar{Y}$:

$$\bar{y}_{rd} = \frac{\bar{y}}{\bar{x}}\,\bar{x}' \tag{9.1}$$

where $\bar{x}' = (\Sigma x_i)/n'$, $i = 1, 2, \ldots n'$, is the mean for auxiliary variable x based on the initial sample of size n'.

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

$$B(\bar{y}_{rd}) = \left(\frac{1}{n} - \frac{1}{n'}\right)\left(\frac{1}{\bar{X}}\right)(RS_x^2 - S_{xy}) \tag{9.2}$$

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

$$MSE(\bar{y}_{rd}) = \left(\frac{1}{n} - \frac{1}{n'}\right)(S_y^2 + R^2 S_x^2 - 2RS_{xy}) + \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 \tag{9.3}$$

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

$$mse(\bar{y}_{rd}) = \left(\frac{1}{n} - \frac{1}{n'}\right)(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}\, s_{xy}) + \left(\frac{1}{n'} - \frac{1}{N}\right) s_y^2 \tag{9.4}$$

The symbols used have the same meaning as in chapters 7 and 8.

Example 9.1

The Sugar Mills Association of the state of Uttar Pradesh (India) wanted to estimate total man-hours lost due to strikes, power failure, breakdowns, etc., during February, 1992. The number of employees in a mill was taken as the auxiliary variable. The population total for this variable was, however, not available at the central office. A first-phase simple random WOR sample of 51 mills was, therefore, drawn from the population of 671 mills. Number of employees for each of these selected mills were recorded. A subsample of 24 mills was drawn from this initial sample using the same sampling scheme, and the man-hours lost were observed for the units included in the subsample. The information gathered on these two characters, for the units in the two samples, is given in the following table.

Table 9.1 Number of employees (x) and total man-hours lost (y) in '000 hour units

Mill	y	x	Mill	y	x	Mill	y	x
1		279	18	30.780	552	35	55.771	982
2	16.711	366	19		270	36		716
3		791	20	29.900	661	37		658
4	9.419	180	21		350	38	34.104	790
5		1001	22		690	39		381
6	14.370	291	23		570	40		280
7	20.609	420	24		240	41	21.680	411
8		371	25	41.460	691	42	9.413	150
9		687	26		524	43		398
10	9.358	196	27	41.220	832	44	11.639	241
11		792	28	10.004	266	45		336
12	52.024	1146	29		441	46		619
13		351	30		395	47	20.580	403
14		460	31	66.786	1246	48		186
15	20.113	370	32		179	49	36.009	864
16	13.200	220	33	18.632	413	50	12.600	316
17		485	34	10.852	286	51		298

Estimate the parameter in question, and place confidence limits on its population value.

Solution

From the statement of the problem, we have N=671, n'=51, and n= 24. Using table 9.1, we first work out the total number of employees in the sugar mills included in the initial sample. Thus,

$$\begin{aligned} x' &= x_1 + x_2 + \ldots + x_{51} \\ &= 279 + 366 + \ldots + 298 \\ &= 25041 \end{aligned}$$

It gives

$$\bar{x}' = \frac{25041}{51} = 491$$

Further, from the final subsample of n=24 mills, we compute the following sample estimates. These intermediate estimates will be used later. The calculations are analogous to those in chapters 7 and 8. Thus,

$$\bar{y} = \frac{1}{24}(16.711 + 9.419 + \ldots + 12.600)$$
$$= 25.3014$$

$$\bar{x} = \frac{1}{24}(366 + 180 + \ldots + 316)$$
$$= 512.2$$

$$\hat{R} = \frac{25.3014}{512.2} = .0494$$

Also,

$$s_y^2 = \frac{1}{24-1}[(16.711)^2 + (9.419)^2 + \ldots + (12.600)^2 - 24(25.3014)^2]$$
$$= 266.9345$$

$$s_x^2 = \frac{1}{24-1}[(366)^2 + (180)^2 + \ldots + (316)^2 - 24(512.2)^2]$$
$$= 100067.21$$

$$s_{xy} = \frac{1}{24-1}[(366)(16.711) + (180)(9.419) + \ldots + (316)(12.600)$$
$$- 24(512.2)(25.3014)]$$
$$= 5044.3834$$

Then, the estimate of correlation coefficient will be

$$r = \frac{s_{xy}}{s_x s_y} = \frac{5044.3834}{(\sqrt{100067.21})(\sqrt{266.9345})} = .976$$

As the man-hours lost and the number of employees in a mill are highly positively correlated, we decide to use ratio method of estimation for estimating total man-hours lost in February, 1992. Thus from (9.1), we have

$$\hat{Y}_{rd} = N\bar{y}_{rd} = \frac{N\bar{y}\,\bar{x}'}{\bar{x}}$$
$$= \frac{(671)(25.3014)(491)}{512.2}$$
$$= 16274.550$$

The total man-hours lost are, therefore, estimated as 16274550.

The estimate of mean square error $MSE(\hat{Y}_{rd})$ is worked out by using (9.4). It yields

$$\begin{aligned}
mse(\hat{Y}_{rd}) &= N^2\, mse(\bar{y}_{rd}) \\
&= N^2\left[\left(\frac{1}{n} - \frac{1}{n'}\right)(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}) + \left(\frac{1}{n'} - \frac{1}{N}\right)s_y^2\right] \\
&= (671)^2\left[\left(\frac{1}{24} - \frac{1}{51}\right)\{266.9345 + (.0494)^2\,(100067.21) \right. \\
&\qquad \left. - 2(.0494)\,(5044.3834)\} + \left(\frac{1}{51} - \frac{1}{671}\right)(266.9345)\right] \\
&= 126624.68 + 2177452.7 \\
&= 2304077.4
\end{aligned}$$

Making use of the $mse(\hat{Y}_{rd})$ computed above, the required confidence limits for the population total are worked out as

$$\begin{aligned}
&\hat{Y}_{rd} \pm 2\sqrt{mse(\hat{Y}_{rd})} \\
&= 16274.550 \pm 2\sqrt{2304077.4} \\
&= 16274.550 \pm 3035.838 \\
&= 13238.712,\ 19310.388
\end{aligned}$$

Thus the Association could infer that the total man-hours lost during February, 1992, for the entire population of 671 sugar mills, are likely to fall in the closed interval [13238.712, 19310.388] thousand hours. ■

9.2.2 *Double Sampling for Product Method of Estimation*

Estimator of population mean $\bar{Y}$:

$$\bar{y}_{pd} = \frac{\bar{y}\,\bar{x}}{\bar{x}'} \tag{9.5}$$

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

$$B(\bar{y}_{pd}) = \left(\frac{1}{n} - \frac{1}{n'}\right)\frac{S_{xy}}{\bar{X}} \tag{9.6}$$

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

$$MSE(\bar{y}_{pd}) = \left(\frac{1}{n} - \frac{1}{n'}\right)(S_y^2 + R^2 S_x^2 + 2RS_{xy}) + \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 \tag{9.7}$$

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

$$\text{mse}(\bar{y}_{pd}) = \left(\frac{1}{n} - \frac{1}{n'}\right)(s_y^2 + \hat{R}^2 s_x^2 + 2\hat{R} s_{xy}) + \left(\frac{1}{n'} - \frac{1}{N}\right) s_y^2 \quad (9.8)$$

The symbols used have the same meaning as before.

Example 9.2

A graduate student of statistics was asked to estimate average time per week for which the undergraduate students of a certain university view television (TV). The overall grade point average (OGPA) of the students was taken as the auxiliary variable. As the investigator found it difficult to record OGPA of all the 1964 undergraduate students, a first-phase sample of 150 students was selected. The OGPA of the students included in this initial sample were recorded from their personal files in the Registrar's office. The average OGPA for the first-phase sample was computed as 2.870 (on 4.00 basis). A subsample of 36 students was then selected from the first-phase sample. The students selected in the subsample were contacted personally to find the total time for which they view television in a week. The data in respect of both the characters, for the units selected in the subsample, are given in table 9.2.

Table 9.2 The OGPA (x) and number of hours per week (y) devoted to TV viewing

Student	y	x	Student	y	x	Student	y	x
1	8	2.51	13	14	2.86	25	5	3.25
2	3	3.41	14	11	2.24	26	3	3.49
3	1	3.25	15	3	3.43	27	8	2.63
4	5	3.04	16	6	3.25	28	4	3.61
5	12	2.73	17	7	2.73	29	13	2.17
6	6	3.10	18	5	2.91	30	14	3.10
7	9	2.58	19	4	3.07	31	6	3.01
8	2	3.46	20	10	2.61	32	8	2.58
9	0	3.69	21	8	2.48	33	9	2.41
10	8	2.83	22	12	3.39	34	4	2.96
11	9	2.91	23	6	2.95	35	5	2.85
12	6	3.06	24	1	3.77	36	10	3.74

Solution

In this problem, we are given that N = 1964, n' =150, n = 36, and $\bar{x}'$ = 2.870. From table 9.2, we have

$$\bar{y} = \frac{1}{36}(8+3+\ldots+10)$$

$$= 6.806$$

$$\bar{x} = \frac{1}{36}(2.51+3.41+\ldots+3.74)$$

$$= 3.002$$

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{6.806}{3.002} = 2.267$$

$$s_y^2 = \frac{1}{36-1}[8^2+3^2+\ldots+10^2-36(6.806)^2]$$

$$= 13.5897$$

$$s_x^2 = \frac{1}{36-1}[(2.51)^2+(3.41)^2+\ldots+(3.74)^2-36(3.002)^2]$$

$$= .1762$$

$$s_{xy} = \frac{1}{36-1}[(2.51)(8)+(3.41)(3)+\ldots+(3.74)(10)-36(3.002)(6.806)]$$

$$= -.9114$$

Then, the estimated correlation coefficient will be

$$r = \frac{-.9114}{(\sqrt{.1762})(\sqrt{13.5897})} = -.589$$

Since OGPA of a student and the time devoted by him to viewing TV, are negatively correlated, we shall use double sampling product estimator. The estimate of mean is then worked out by using (9.5). Thus,

$$\bar{y}_{pd} = \frac{\bar{y}\,\bar{x}}{\bar{x}'}$$

$$= \frac{(6.806)(3.002)}{2.870}$$

$$= 7.119$$

From (9.8), the estimate of mean square error of $\bar{y}_{pd}$ is obtained as

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

$$= \left(\frac{1}{36}-\frac{1}{150}\right)[13.5897+(2.267)^2(.1762)+2(2.267)(-.9114)]$$

$$+\left(\frac{1}{150}-\frac{1}{1964}\right)(13.5897)$$

$$= .2188+.0837$$

$$= .3025$$

We now work out the confidence interval for population mean. This is defined by the limits

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

$$= 7.119 \pm 2\sqrt{.3025}$$

$$= 7.119 \pm 1.100$$

$$= 6.019, 8.219$$

To conclude, an undergraduate student devotes, on the average, 7.119 hours per week to TV viewing. Also, the student conducting the survey is reasonably sure that had the information been collected for all the 1964 students, the per week average TV viewing time would have taken a value in the range 6.019 to 8.219 hours. ■

9.2.3 *Double Sampling for Regression Method of Estimation*

Estimator of population mean $\bar{Y}$:

$$\bar{y}_{lrd} = \bar{y} + \hat{\beta}(\bar{x}' - \bar{x}) \tag{9.9}$$

where $\hat{\beta} = s_{xy}/s_x^2$ as defined in section 8.3.

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

$$B(\bar{y}_{lrd}) = -\beta\left(\frac{1}{n} - \frac{1}{n'}\right)\left(\frac{\mu_{21}}{S_{xy}} - \frac{\mu_{30}}{S_x^2}\right) \tag{9.10}$$

where $\mu_{21} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2(Y_i - \bar{Y})$ and $\mu_{30} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^3$.

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

$$\left.\begin{aligned} MSE(\bar{y}_{lrd}) &= \left(\frac{1}{n} - \frac{1}{n'}\right)(S_y^2 + \beta^2 S_x^2 - 2\beta S_{xy}) + \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 \\ &= \left(\frac{1}{n} - \frac{1}{n'}\right)S_y^2(1-\rho^2) + \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 \end{aligned}\right] \tag{9.11}$$

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

$$mse(\bar{y}_{lrd}) = \left(\frac{1}{n} - \frac{1}{n'}\right)s_y^2(1-r^2) + \left(\frac{1}{n'} - \frac{1}{N}\right)s_y^2 \tag{9.12}$$

Symbols above have their usual meaning.

Example 9.3

Assume that in example 8.2, it was not possible to pluck all the 2106 leaves for weighing. Thus a first-phase WOR random sample of 120 leaves was selected and the sampled leaves were then plucked. The total weight of these 120 leaves was recorded as 13,400 mg. From this initial sample, a subsample of 33 leaves was drawn using same sampling scheme. Let us suppose that this subsample is same as the one considered in example 8.2. Therefore, using the sample data of table 8.2, estimate the average leaf area for the target population, and also build up confidence interval for it.

Solution

We have in this case, n' = 120 and $\bar{x}'$ = 13,400/120 = 111.667. Certain other sample estimates have also been calculated in example 8.2. These are reproduced below :

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

$r = .9376$, and $\hat{\beta} = .5346$.

Since only the first-phase sample mean is available in place of the population mean, we go for regression estimator using double sampling. The estimator of mean in this case is given by (9.9). Thus, the estimate of average leaf area is

$$\begin{aligned}\bar{y}_{lrd} &= \bar{y} + \hat{\beta}(\bar{x}' - \bar{x}) \\ &= 27.263 + .5346(111.667 - 110.545) \\ &= 27.863\end{aligned}$$

Estimator of mean square error of $\bar{y}_{lrd}$ is obtained by using (9.12). Therefore,

$$\begin{aligned}mse(\bar{y}_{lrd}) &= \left(\frac{1}{n} - \frac{1}{n'}\right) s_y^2 (1 - r^2) + \left(\frac{1}{n'} - \frac{1}{N}\right) s_y^2 \\ &= \left(\frac{1}{33} - \frac{1}{120}\right)(61.003)\{1 - (.9376)^2\} + \left(\frac{1}{120} - \frac{1}{2106}\right)(61.003) \\ &= .6414\end{aligned}$$

The confidence interval in which average leaf area for the population under consideration is likely to fall, with probability approximately .95, is determined by the limits

$$\begin{aligned}&\bar{y}_{lrd} \pm 2\sqrt{mse(\bar{y}_{lrd})} \\ &= 27.863 \pm 2\sqrt{.6414} \\ &= 27.863 \pm 1.6017 \\ &= 26.261, 29.465 \blacksquare\end{aligned}$$

9.3 SAMPLE SIZE DETERMINATION FOR RATIO, PRODUCT, AND REGRESSION ESTIMATORS

Let the cost function for two-phase sampling be

$$C = c_o + cn + c' n' \tag{9.13}$$

where c_o is the overhead cost, and c and c' are the per unit costs for observing the study and auxiliary variables respectively. For the sake of simplicity, we assume the population to be large so that 1/N is negligibly small. On minimizing the cost for the fixed variance, one gets the optimum values of n' and n. The values of n' and n thus obtained, are given in (9.14) and (9.15) for ratio, product, and regression estimators. The relations (9.16) and (9.17) give alternative formulas for the regression estimator.

First-phase and second-phase sample sizes for estimating mean using ratio, product, and regression estimators:

$$n' = \frac{S_y^2 - A}{V_o}\left[1 + \left\{\left(\frac{c}{c'}\right)\left(\frac{A}{S_y^2 - A}\right)\right\}^{\frac{1}{2}}\right] \tag{9.14}$$

$$n = n'\left[\left(\frac{c'}{c}\right)\left(\frac{A}{S_y^2 - A}\right)\right]^{\frac{1}{2}} \tag{9.15}$$

where

$A = S_y^2 + R^2 S_x^2 - 2RS_{xy}$ (for ratio estimator)

$A = S_y^2 + R^2 S_x^2 + 2RS_{xy}$ (for product estimator)

$A = S_y^2 + \beta^2 S_x^2 - 2\beta S_{xy}$ (for regression estimator)

Alternative formulas for regression estimator :

$$n' = \frac{\rho^2 S_y^2}{V_o}\left[1 + \left\{\frac{c(1-\rho^2)}{c'\,\rho^2}\right\}^{\frac{1}{2}}\right] \tag{9.16}$$

$$n = n'\left[\frac{c'\,(1-\rho^2)}{c\rho^2}\right]^{\frac{1}{2}} \tag{9.17}$$

where ρ is the population correlation coefficient between y and x.

In surveys, depending on the precision required, the value of the variance V_o is fixed in advance. The costs c and c' are also known. Using the guess values of parameters S_y^2, S_x^2, S_{xy}, R, and ρ, one can arrive at optimum n' and n. In case the guess values of above said parameters are not available, then as in the previous chapters, a preliminary sample of n_1 units is selected and estimates of these parameters are obtained from these n_1 observations. These estimates are then used in place of the parameters involved in (9.14) to (9.17).

However, if the margin of error is to be fixed in terms of permissible error B, instead of the variance, then V_o in the above expressions will be replaced by $B^2/4$.

The results (9.14) to (9.17) can also be used for the estimation of population total by taking $V_o = (1/N^2)$ times the value of the variance fixed for the estimator of population total, or $V_o = B^2/4N^2$ when the margin of error is specified in terms of the permissible error B.

Example 9.4

Assume that the subsample drawn in example 9.2, is the preliminary sample of size $n_1 = 36$ students drawn from the population of 1964 students. Taking the costs of collecting information on OGPA and the time for which the students view TV as \$ .20 and \$ 1.25 per student respectively, determine the required subsample and the first-phase sample sizes if the variance is fixed at .20.

Solution

Here we have $n_1 = 36$, c' = \$ .20, c = \$ 1.25, and $V_o = .20$. Since we are using the observations on a preliminary sample for the determination of sample size, let us recall the sample values computed in example 9.2. Using the symbols of preceding chapters, we, therefore, write

$$s_{y1}^2 = 13.5897,\ s_{x1}^2 = .1762,\ s_{xy1} = -.9114,\ \text{and}\ \hat{R}_1 = 2.267.$$

Then,

$$\begin{aligned} A_1 &= s_{y1}^2 + \hat{R}_i^2 s_{x1}^2 + 2\hat{R}_1 s_{xy1} \\ &= 13.5897 + (2.267)^2 (.1762) + 2(2.267)(-.9114) \\ &= 10.3630 \end{aligned}$$

The optimum value of n' is given by (9.14). Using the analogous preliminary sample based values, it can be written as

$$n' = \frac{s_{y1}^2 - A_1}{V_o}\left[1 + \left\{\left(\frac{c}{c'}\right)\left(\frac{A_1}{s_{y1}^2 - A_1}\right)\right\}^{\frac{1}{2}}\right]$$

Substituting numerical values for different terms, we get

$$n' = \frac{13.5897 - 10.3630}{.20}\left[1 + \left\{\left(\frac{1.25}{.20}\right)\left(\frac{10.3630}{13.5897 - 10.3630}\right)\right\}^{\frac{1}{2}}\right]$$

$$= 88.4$$

$$\approx 88$$

Optimum value of n is then worked out through sample analog of (9.15), based on n_1 units. Thus, on redenoting the terms, it can be put as

$$n = n'\left[\left(\frac{c'}{c}\right)\left(\frac{A_1}{s_{y1}^2 - A_1}\right)\right]^{\frac{1}{2}}$$

On making substitutions, one gets

$$n = (88.4)\left[\left(\frac{.20}{1.25}\right)\left(\frac{10.3630}{13.5897 - 10.3630}\right)\right]^{\frac{1}{2}}$$

$$= 63.4$$

$$\approx 63$$

Thus, for getting an initial sample of 88 students, the preliminary sample of 36 students should be augmented by another SRS without replacement sample of 88-36=52 students selected from the population of 1964-36=1928 students left after the selection of the preliminary sample. These newly selected 52 students will be observed for their OGPA. To get a second-phase sample of 63 students, a subsample of 63-36= 27 students will be selected from the 52 newly selected students. These 27 students will also be observed for their TV viewing time. ■

9.4 TWO-PHASE PPS SAMPLING

In PPS sampling, the sample units are drawn with probability proportional to the size measure x. If the information on x is lacking for the population units, one can opt for two-phase sampling procedure. A first-phase sample of n' units is drawn from the given population of N units using SRS without replacement. The auxiliary variable is observed on these n' units. From this first-phase sample, a subsample of n units is selected by PPS with replacement method. The study variable is then measured on the subsample units. For i = 1, 2,..., n', let

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

$$p_i' = \frac{x_i}{x'} \tag{9.18}$$

We then have the results (9.19) through (9.21).

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

$$\bar{y}_{dp} = \frac{1}{n'n} \sum_{i=1}^{n} \frac{y_i}{p_i'} \tag{9.19}$$

Variance of estimator $\bar{y}_{dp}$:

$$V(\bar{y}_{dp}) = \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \frac{n'-1}{nn'N(N-1)} \sigma_z^2 \tag{9.20}$$

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

$$v(\bar{y}_{dp}) = \frac{1}{N(n-1)(n'-1)} \left[\frac{N-1}{nn'} \sum_{i=1}^{n} \frac{y_i^2}{p_i'^2} + \frac{(n-1)(N-n')}{nn'} \left(\sum_{i=1}^{n} \frac{y_i^2}{p_i'} \right) - \{N(n'+n-1) - nn'\} \bar{y}_{dp}^2 \right] \tag{9.21}$$

where S_y^2 has been defined in (7.2), and $\sigma_z^2 = \sum_{i=1}^{N} P_i \left(\frac{Y_i}{P_i} - Y \right)^2$ with $P_i = X_i/X$.

Example 9.5

An investigator is interested in estimating the average total money spent in a year on all the important festivals by a family in a certain locality consisting of 671 households. It is felt that the amount spent by a household on a festival depends on the income of the family, which in turn is related to the market price of the house where the family lives. Thus, the eye estimated price of the house was taken as the auxiliary variable. Since the determination of total eye estimated price of all the houses of the locality was time consuming, two-phase sampling was used to select a sample of households. A first-phase WOR simple random sample of 40 households was drawn. The eye estimated price of houses in the first-phase sample was determined and it totalled to 4980 thousand rupees. A subsample of 18 households was drawn, using PPS with replacement sampling, from 40 households of the first-phase sample. The data collected for these 18 households in respect of the two characters is given in table 9.3 as follows :

Table 9.3 Annual expenditure (y) on festivals by a household and the eye estimated price (x) of the house

Household	y	x	p'_i	Household	y	x	p'_i
1	2.50	70	.0141	21	9.00	295	.0592
2		25		22	4.80	170	.0341
3	3.70	110	.0221	23		84	
4		105		24	4.00	135	.0271
5		66		25		62	
6		73		26		30	
7	1.00	34	.0068	27	6.60	192	.0386
8	5.00	140	.0281	28	2.30	76	.0153
9		96		29		59	
10	6.50	260	.0522	30		173	
11		117		31	3.50	90	.0181
12		47		32	7.00	286	.0574
13	4.10	136	.0273	33		241	
14		126		34	2.70	54	.0108
15	2.65	80	.0161	35		134	
16		64		36		28	
17		182		37		85	
18		96		38	6.80	261	.0524
19		119		39	3.40	103	.0207
20	11.00	410	.0823	40		66	
Total						4980	

The units for the variables y and x in the table are in '00 rupees and '000 rupees respectively.

Estimate the annual expenditure on festivals per family, and build up confidence interval for it.

Solution

The statement of the problem gives N = 671, n' = 40, n = 18, and $x' = \Sigma x_i = 4980$.

The selection probabilities p'_i for the units included in the subsample are computed, by using (9.18), as

$$p'_1 = \frac{70}{4980} = .0141$$

$$p'_2 = \frac{110}{4980} = .0221$$

$$\vdots$$

$$p'_{18} = \frac{103}{4980} = .0207$$

These probabilities are given in table 9.3. Then,

$$\sum_{i=1}^{n} \frac{y_i}{p'_i} = \frac{2.50}{.0141} + \frac{3.70}{.0221} + \ldots + \frac{3.40}{.0207} = 2863.72$$

$$\sum_{i=1}^{n} \left(\frac{y_i}{p'_i}\right)^2 = \left(\frac{2.50}{.0141}\right)^2 + \left(\frac{3.70}{.0221}\right)^2 + \ldots + \left(\frac{3.40}{.0207}\right)^2 = 470900.37$$

$$\sum_{i=1}^{n} \frac{y_i^2}{p'_i} = \frac{(2.50)^2}{.0141} + \frac{(3.70)^2}{.0221} + \ldots + \frac{(3.40)^2}{.0207} = 13185.92$$

These sample values will be used for working out the estimate of mean and also the variance estimate $v(\bar{y}_{dp})$. From (9.19), the estimate of the average amount spent annualy on festivals by a family is given as

$$\bar{y}_{dp} = \frac{1}{n' n} \sum_{i=1}^{n} \frac{y_i}{p'_i}$$

$$= \frac{2863.72}{(40)(18)}$$

$$= 3.9774$$

Now for computing the estimate of variance $V(\bar{y}_{dp})$, we use (9.21). Thus,

$$v(\bar{y}_{dp}) = \frac{1}{N(n-1)(n'-1)} \left[\frac{N-1}{nn'} \sum_{i=1}^{n} \frac{y_i^2}{p'^2_i} + \frac{(n-1)(N-n')}{nn'} \left(\sum_{i=1}^{n} \frac{y_i^2}{p'_i} \right) - \{N(n'+n-1) - nn'\} \bar{y}^2_{dp} \right]$$

$$= \frac{1}{(671)(18-1)(40-1)} \left[\frac{671-1}{(18)(40)} (470900.37) + \frac{(18-1)(671-40)}{(18)(40)} (13185.92) - \{671(40+18-1) - (18)(40)\} (3.9774)^2 \right]$$

$$= \frac{1}{(671)(18-1)(40-1)} (438198.95 + 196451.89 - 593666.28)$$

$$= .09213$$

The required confidence limits will then be given by

$$\bar{y}_{dp} \pm 2\sqrt{v(\bar{y}_{dp})}$$

$$= 3.9774 \pm 2\sqrt{.09213}$$

$$= 3.3703, 4.5845$$

The investigator is thus reasonably sure that had all the 671 households been examined, the per family annual expenditure on festivals would have taken a value in the range from 337.03 to 458.45 rupees. ■

For determining optimum sample sizes n and n' in two-phase PPS with replacement sampling, one can proceed in the usual manner. However, the expression for the estimator of variance involved is somewhat complicated, and it makes determination of optimum n and n' quite difficult. Keeping the level of the book in mind, this topic is, therefore, not considered.

9.5 SAMPLING ON TWO OCCASIONS

In case of populations that are changing fast, census at long and infrequent intervals are not of much use. In such cases, it is desirable that the population is sampled at annual, or even at shorter, intervals regularly. Similarly, one may have to resort to repeated sampling of populations for which several kinds of data are to be collected and published at regular intervals. The method of sampling of units from the same population on successive occasions is called *multiple sampling* or *successive sampling*. This kind of sampling involves selection of samples on different occasions such that they have none, some, or all, units common with samples selected on previous occasions. In such surveys, it is possible to use information collected on previous occasions to improve the efficiency of estimators for subsequent occasions.

While resorting to repeated surveys, the objective of the investigator might be the estimation of one, or more, of the following parameters:

1. Average of population means over occasions.
2. Change in population mean value from one occasion to the next.
3. The population mean for the current occasion.

For estimating the average of population means over different occasions, the most appropriate strategy is to draw a fresh sample on each occasion. If it is desired to estimate the change in population mean from one occasion to the next, the same sample should be retained through all occasions. In case, the interest is to estimate the population average on the most recent occasion, the retaining of the part of the sample over occasions provides efficient estimates as compared to the other alternatives. As the theory for more than two occasions becomes complicated, we shall restrict our discussion to two occasions only. The problem of estimation in cases (1) and (2) above, is theoretically straightforward. In this section, we shall, therefore, only consider the problem of estimating population mean on the current occasion. For this, we shall assume that sample size on the two occasions is same.

Consider a population of N units and assume that the size of the population remains same over occasions and only the value of the study variable for different population units may get changed from occasion to occasion. Suppose a WOR simple random sample of n units is drawn on the first occasion. Out of this sample, m randomly selected units are retained on the second occasion. This sample is then augmented by selecting u additional units from the remaining population of (N-n) units using equal probability WOR sampling, so that, n = m+u. For h = 1, 2, let

$\bar{Y}_2$ = the population mean on the second occasion
$\bar{y}_h$ = the sample mean based on n units observed on the h-th occasion
$\bar{y}_{mh}$= the sample mean based on m matched units observed on the h-th occasion

$\bar{y}_{uh}$ = the sample mean based on u units drawn afresh on the h-th occasion
S_2^2 = the population mean square for second occasion
s_{m2}^2 = the sample mean square based on m matched units drawn on the second occasion
s_{u2}^2 = the sample mean square based on u units drawn afresh on the second occasion
s_2^2 = the pooled mean square based on matched and unmatched sample mean squares for the second occasion
ρ = the correlation coefficient between the variate of the second occasion and of the first occasion in the population
r = estimate of ρ based on matched part of the sample
$\hat{\beta}$ = estimate of regression coefficient β for the regression of variate of the second occasion on the variate of the first occasion

For the sake of simplicity, we take

$$p = \frac{m}{n} \text{ and } q = \frac{u}{n}$$

Let us further define

$$\frac{1}{W_u} = \frac{s_{u2}^2}{u} \tag{9.22}$$

$$\frac{1}{W_m} = \frac{(1-r^2)\, s_{m2}^2}{m} + \frac{r^2 s_{m2}^2}{n} \tag{9.23}$$

Then, we present the estimator of population mean on the second occasion along with its variance and estimator of varinace. Thus for large population, we have the following :

Estimator of population mean $\bar{Y}_2$:

$$\bar{y}_2 = \left(\frac{1}{W_m + W_u}\right)[W_u\, \bar{y}_{u2} + W_m \{\bar{y}_{m2} + \hat{\beta}\,(\bar{y}_1 - \bar{y}_{m1})\}] \tag{9.24}$$

Approximate variance of estimator $\bar{y}_2$:

$$V(\bar{y}_2) = \left(\frac{1-\rho^2 q}{1-\rho^2 q^2}\right)\frac{S_2^2}{n} \tag{9.25}$$

An estimator of variance V $(\bar{y}_2)$:

$$v(\bar{y}_2) = \left(\frac{1-r^2 q}{1-r^2 q^2}\right)\frac{s_2^2}{n} \tag{9.26}$$

where $s_2^2 = [(m-1)\, s_{m2}^2 + (u-1)\, s_{u2}^2]/(n-2)$.

The optimum value of q that minimizes the variance $V(\bar{y}_2)$, can be derived as

$$q_o = \frac{1}{1+\sqrt{1-\rho^2}} \tag{9.27}$$

This gives the fraction of the sample on the first occasion to be replaced on the second occasion, so that, one may achieve maximum precision. On substituting optimal q from (9.27) in (9.25), the minimum variance works out to be

$$V_0(\bar{y}_2) = [1+\sqrt{1-\rho^2}]\,\frac{S_2^2}{2n} \tag{9.28}$$

The estimate of $V_0(\bar{y}_2)$ is obtained by replacing ρ and S_2^2 in (9.28) by r and s_2^2 respectively. For further details, the reader may refer to Cochran (1977).

Example 9.6

A WOR simple random sample of 25 professors was drawn from 528 professors of a university, during the financial year 1991-92, to estimate average amount of money spent in buying National Saving Certificates (NSC). The government decided that from the financial year 1992-93, the interest on the amount of NSC purchased will be deducted from the total income for income-tax calculations, whereas, before 1992-93 this interest was counted towards personal income. Of the 25 professors selected in 1991-92, 12 were retained which constituted the matched part of the sample for the survey to be undertaken during 1992-93. A fresh sample of 13 professors was drawn using SRS without replacement from the remaining 528-25 = 503 professors. The information collected on both the occasions is presented in table 9.4.

Table 9.4 Amount (in '000 rupees) of NSC purchased

Professor	Amount of NSC		Professor	Amount of NSC	
	1991-92	1992-93		1991-92	1992-93
1	7.05	5.40	20	.50	
2	.40	.50	21	5.80	
3	1.00	2.02	22	2.70	
4	1.70	2.20	23	1.14	
5	1.32	.80	24	8.60	
6	0	.30	25	1.70	
7	1.50	1.00	26		1.35
8	.92	.50	27		1.80
9	1.20	.80	28		.90
10	.34	.20	29		4.00
11	3.10	5.20	30		5.20
12	1.40	2.00	31		.78
13	.98		32		2.50
14	.67		33		1.40
15	1.12		34		.75
16	.45		35		7.20
17	2.10		36		2.60
18	.90		37		3.00
19	1.10		38		2.40

Estimate the average amount spent by a professor for the purchase of NSC during 1992-93, and place confidence limits on it.

Solution

We are given that N=528, n=25, m=12, and u=13. This gives

$$p = \frac{12}{25} = .48 \text{ and } q = \frac{13}{25} = .52$$

Next, we compute the following averages :

$$\bar{y}_{m1} = \frac{1}{12}(7.05 + .40 + \ldots + 1.40)$$
$$= 1.661$$

$$\bar{y}_{m2} = \frac{1}{12}(5.40 + .50 + \ldots + 2.00)$$
$$= 1.743$$

$$\bar{y}_{u2} = \frac{1}{13}(1.35 + 1.80 + \ldots + 2.40)$$
$$= 2.606$$

$$\bar{y}_{1} = \frac{1}{25}(7.05 + .40 + \ldots + 1.70)$$
$$= 1.908$$

For working out r, we need sample mean squares and sample mean product based on matched part of the sample on first and second occasions. We, therefore, have

$$s^2_{m1} = \frac{1}{12-1}[(7.05)^2 + (.40)^2 + \ldots + (1.40)^2 - (12)(1.661)^2]$$
$$= \frac{1}{12-1}[71.7169 - 12(1.661)^2]$$
$$= 3.5100$$

$$s^2_{m2} = \frac{1}{12-1}[(5.40)^2 + (.50)^2 + \ldots + (2.00)^2 - (12)(1.743)^2]$$
$$= \frac{1}{12-1}[72.0304 - 12(1.743)^2]$$
$$= 3.2340$$

$$s_{m12} = \frac{1}{12-1}[(7.05)(5.40) + (.40)(.50) + \ldots + (1.40)(2.00)$$
$$- (12)(1.661)(1.743)]$$

$$= \frac{1}{12-1}\,[66.9940 - 12(1.661)\,(1.743)]$$

$$= 2.9320$$

This yields

$$\hat{\beta} = \frac{s_{m12}}{s_{m1}^2} = \frac{2.9320}{3.5100} = .8353$$

$$r = \frac{s_{m12}}{\sqrt{s_{m1}^2\, s_{m2}^2}} = \frac{2.9320}{\sqrt{(3.5100)\,(3.2340)}} = .8702$$

For obtaining the weights W_u and W_m, we need the values of s_{u2}^2 and s_{m2}^2 respectively. The value of s_{m2}^2 has already been computed. The value of s_{u2}^2 is obtained as

$$s_{u2}^2 = \frac{1}{13-1}\,[(1.35)^2 + (1.80)^2 + \ldots + (2.40)^2 - 13(2.606)^2]$$

$$= \frac{1}{13-1}\,[131.6534 - 13\,(2.606)^2]$$

$$= 3.6139$$

Thus, from (9.22) and (9.23), it follows that

$$\frac{1}{W_u} = \frac{s_{u2}^2}{u} = \frac{3.6139}{13} = .2780$$

$$\frac{1}{W_m} = \left[\frac{1-r^2}{m} + \frac{r^2}{n}\right] s_{m2}^2$$

$$= \left[\frac{1-(.8702)^2}{12} + \frac{(.8702)^2}{25}\right](3.2340)$$

$$= .1634$$

Hence, (9.24) yields

$$\bar{y}_2 = \left(\frac{1}{6.1200+3.5971}\right)[(3.5971)\,(2.606) + (6.1200)\,\{1.743 + (.8353)$$

$$(1.908-1.661)\}]$$

$$= \frac{21.3110}{9.7208}$$

$$= 2.192$$

From the sample information, it is thus estimated that, on the average, each professor spent Rs 2192 for purchasing NSC in 1992-93.

We now obtain the estimate of variance $V(\bar{y}_2)$ from (9.26). For this, we first calculate s_2^2.

$$s_2^2 = \frac{1}{25-2}[(12-1)(3.2340) + (13-1)(3.6139)]$$
$$= 3.4322$$

Thus,

$$v(\bar{y}_2) = \left[\frac{1-(.8702)^2(.52)}{1-(.8702)^2(.52)^2}\right]\frac{3.4322}{25}$$
$$= .1047$$

The confidence limits for the mean on second occasion are given by

$$\bar{y}_2 \pm 2\sqrt{v(\bar{y}_2)}$$
$$= 2.192 \pm 2\sqrt{.1047}$$
$$= 2.192 \pm .647$$
$$= 1.545, 2.839$$

Hence, one can be reasonably sure that, on the average, each professor spent rupees 1545 to 2839 on purchasing NSC during 1992-93. ■

It may be pointed out that the unmatched part of the sample on second occasion can also be selected from the entire population of N units, or from entire population minus the matched part of the sample. Both these strategies, however, yield less efficient estimators of population mean on second occasion, as compared to the one discussed in this section. For details, reader may refer to Ghangurde and Rao (1969), Singh (1972), and Cochran (1977).

9.6 SOME FURTHER REMARKS

9.1 Double sampling with regression estimator has been extended by Khan and Tripathi (1967) to the case where k auxiliary variables are observed on both the samples, and the population mean is estimated using multiple linear regression of the estimation variable on these k auxiliary variables.

9.2 Neyman (1938) was the first to discuss two-phase sampling for stratification. In stratified sampling, exact knowledge of strata sizes may sometimes be lacking. For instance, if the strata are based on old census data which do not indicate true situation for the current survey, double sampling provides an alternative in such situations. A first-phase sample is selected to collect information on the auxiliary variable. On the basis of this information, one can construct the strata and also estimate the sizes of different strata. Srinath (1971) and Rao (1973) have discussed the utility of double sampling for dealing with the problem of optimum allocation. For details, the reader may refer to Cochran (1977), Singh and Chaudhary (1989), and Des Raj (1968).

9.3 The second-phase sample could also be taken, independently of the first-phase sample, from the whole population. This practice is followed when, for instance, information on auxiliary variable is available with one agency, and information on both the study and auxiliary variables has been collected on a small independent sample by another agency. The theory for this case has been discussed by Des Raj (1968), Singh and Chaudhary (1989), and Cochran (1977).

LET US DO

9.1 In what kind of situations does the use of double sampling become necessary ?

9.2 What are the negative features of double sampling? Discuss.

9.3 "The two-phase sampling is beneficial if the estimate based on the information provided by initial and final sample is more precise per unit cost than the one based on a sample for estimation variable alone." Comment.

9.4 Give expressions for estimator $\bar{y}_{rd}$ of population mean $\bar{Y}$ and the estimator for variance $V(\bar{y}_{rd})$ in case of double sampling for ratio method of estimation. Assume that both the samples are drawn using SRS without replacement.

9.5 A total of 300 pieces of barren land in a district were marked for making them cultivable by applying gypsum to them. After 5 years of continuous application of gypsum, it was decided to estimate the total area from these pieces that has been brought under cultivation. A WOR random sample of 48 pieces of barren land was drawn, and eye estimates of the area of these pieces (x) were made. A subsample of 16 of these pieces was selected, and the area brought under cultivation (y) was measured for all the pieces in the subsample. The areas recorded in hectares are given in the following table.

Land piece	x	y	Land piece	x	y	Land piece	x	y
1	2.5		17	3.5		33	2.5	
2	1.0	0.6	18	2.5	2.2	34	3.0	
3	3.0		19	4.0		35	2.5	
4	4.0		20	5.0		36	3.0	2.4
5	4.0		21	5.5	4.9	37	1.0	
6	8.0	8.3	22	4.0		38	6.0	6.5
7	4.5		23	4.5		39	1.5	1.3
8	4.0	3.7	24	1.0		40	3.5	
9	5.0		25	5.2	4.7	41	4.0	
10	3.5		26	.5		42	2.0	
11	4.0		27	2.5		43	5.0	
12	6.5	5.8	28	9.0	8.4	44	6.5	
13	5.5		29	1.8		45	11.0	10.0
14	3.0		30	7.0	6.5	46	7.0	6.8
15	4.5		31	5.5		47	1.0	
16	4.5	4.1	32	2.5		48	2.5	1.7

Using double sampling ratio estimator, estimate the total area of barren land that has come under cultivation. Work out standard error of your estimate, and also build up the confidence interval for the population value.

9.6 In case of double sampling with simple random sampling WOR used for drawing both the samples, explain, how will you obtain expressions for the estimator $\hat{Y}_{pd}$ of population total Y and the estimator for variance $V(\hat{Y}_{pd})$ from the corresponding expressions for the mean estimator $\bar{y}_{pd}$?

9.7 A small survey study was conducted to estimate the total amount of money spent, during a year, on medical treatment by the faculty and staff of a certain Indian university. It is known that better paid employees and their dependents visit doctor/hospital for treatment less frequently. The reason is possibly their smaller family size and better living conditions. As the computation of total monthly salary paid to all the 4000 employees of the university is cumbersome, a preliminary WOR simple random sample of n'=100 employees was drawn. The total of monthly salaries (x) for these selected employees worked out to be Rs. 3,00,000 . A subsample of n=30 employees was then drawn using SRS without replacement. The amount of money spent on treatment during the preceding year (y) was obtained for each of the thirty selected employees. The information thus collected on x and y is given in the following table in hundred rupee units.

Employee	x	y	Employee	x	y	Employee	x	y
1	22.00	4.16	11	16.90	13.81	21	56.70	2.20
2	70.00	3.84	12	44.70	4.62	22	72.20	0
3	16.50	9.84	13	31.85	3.14	23	26.70	6.09
4	33.14	4.07	14	25.13	4.15	24	62.60	1.70
5	45.50	3.26	15	66.07	3.15	25	48.08	3.44
6	30.10	2.80	16	28.10	4.10	26	17.05	10.80
7	14.18	12.00	17	55.10	2.08	27	25.35	4.50
8	62.00	.60	18	15.80	9.70	28	80.10	0
9	48.60	1.76	19	18.20	11.10	29	70.96	2.50
10	90.30	2.00	20	82.30	.90	30	24.30	8.18

Using an appropriate estimator, determine the total amount of money spent on medical treatment during the preceding year by all the university employees. Also, place confidence limits on the population value.

9.8 For the data considered in exercise 9.5, estimate the total barren land area that has come under cultivation. Use double sampling based regression estimator for the purpose. Also, construct the confidence interval for the population value.

9.9 In the problem considered in exercise 8.4, assume that all the 1000 overweight women had registered for *yoga* excercises by mail. Since it was difficult to weigh all the women who got registered, a preliminary sample of 120 women was drawn using WOR simple random sampling, and initial weight of each of these women was recorded. The average initial weight of women in the preliminary sample came out to be 63.7 kg. All the 1000 women were provided video tapes and other literature containing lessons on the physical exercises to be undertaken. In order to determine the impact of the weight reducing program, the organizers selected a subsample of 30 women from the preliminary sample after 3 months of yoga participation. Each of these women was again weighed individually. Present weight (y) and the initial weight (x) for the women included in the subsample are as given in the table of exercise 8.4. Using double sampling based regression estimator, estimate the present average weight of a woman participant. Also, place the confidence limits on the present average weight of 1000 women registered for *yoga*.

9.10 Discuss, how will you determine the required initial and final sample sizes for ratio estimator in case of double sampling for fixed variance V_0 ? Assume that c_0 is the overhead cost, and c and c' are per unit costs for observing the study and auxiliary variables respectively. The cost function is as given in (9.13).

9.11 Assume that the subsample of n_1=24 units drawn in example 9.1, is a preliminary SRS without replacement sample from the population of 671 mills. Different estimates obtained from this subsample in example 9.1 are, therefore, to be treated as preliminary sample estimates. Thus, $s_y^2 = s_{y1}^2 = 266.9345$, $s_x^2 = s_{x1}^2 = 100067.21$,

$s_{xy} = s_{xyl} = 5044.3834$, and $\hat{R} = \hat{R}_l = .0494$. Let the costs of collecting information on number of employees per mill and the man-hours lost be Rs 10 and Rs 50 respectively. Determine the required initial and final sample sizes if the variance is to be fixed at 24,00,000 square man-hours.

9.12 Give expressions for the estimator of population mean and the estimator for its variance in case of two-phase PPS sampling. From these expressions, how will you write corresponding expressions for the estimator of population total and its variance estimator ?

9.13 A survey was conducted during 1987 to estimate the cattle population of 800 villages comprising a district. For this purpose, a WOR simple random sample of 54 villages was drawn. Again in 1992, taking number of cattle as the size variable, it was decided to select a PPS wih replacement subsample of 22 villages from the 54 villages selected in 1987. The number of cattle recorded in both the surveys, for the selected villages, is given below :

Village	1987	1992	Village	1987	1992	Village	1987	1992
1	980		19	1477	1610	37	354	
2	219		20	644	790	38	708	
3	640	670	21	340		39	250	285
4	710	775	22	887		40	790	842
5	572		23	317		41	1015	1160
6	693		24	166		42	317	
7	344		25	590		43	209	
8	599	620	26	270	310	44	144	195
9	412		27	621		45	990	1060
10	226	340	28	471		46	716	
11	170	210	29	156		47	681	763
12	488		30	1286	1340	48	550	
13	376		31	890		49	880	935
14	110		32	1364	1492	50	660	
15	1280	1385	33	570		51	505	
16	200	281	34	403	490	52	302	407
17	1170		35	780		53	290	376
18	918		36	291		54	170	

Estimate the total cattle population for the year 1992, and also place confidence limits on it.

9.14 A WOR simple random sample of 22 villages was drawn in 1984 from a development block of 160 villages in the Bathinda district of southern Punjab, to estimate camel population in the block. It is known that introduction of agricultural mechanization has adversely affected the camel population. In a subsequent survey conducted in 1992, out of 22 villages selected in the previous survey 10 were

retained to constitute the matched part of the sample for the survey of 1992. A fresh sample of 12 villages was drawn, using SRS without replacement, from the 138 villages not selected in the 1984 survey. The information on camel population, collected on both the occasions, is given in the table below:

Village	Camel population		Village	Camel population	
	1984	1992		1984	1992
1	40	13	18	39	
2	65	27	19	31	
3	31	10	20	46	
4	21	8	21	59	
5	67	29	22	63	
6	56	19	23		17
7	27	8	24		21
8	44	13	25		9
9	58	18	26		26
10	49	15	27		28
11	36		28		7
12	55		29		13
13	29		30		23
14	42		31		16
15	31		32		6
16	19		33		27
17	57		34		11

Estimate the camel population of the block in 1992, and also place confidence limits on it.