CHAPTER 5

Stratified Sampling

5.1 INTRODUCTION

The precision of an estimate of the population mean or total, besides sample size, also depends on the variability among the units of the population. Therefore, apart from increasing the sample size, another possible way to increase the precision of the estimate could be to divide the population units into certain number of groups, such that the variability within the groups is minimum whereas it is maximum between the groups. Smaller samples could then be selected from each of the groups so formed, such that the total number of sampled units over all the groups equal the required overall sample size. The groups thus formed are called *strata*, and the process of forming strata is known as *stratification*. In this connection, we have the following definitions :

Definition 5.1 The procedure of partitioning the population into groups, called strata, and then drawing a sample independently from each stratum, is known as *stratified sampling*.

Definition 5.2 If the sample drawn from each stratum is random one, the procedure is then termed as *stratified random sampling*.

In case of stratified simple random sampling, since the samples from different strata are selected independently, each stratum can, therefore, be treated as a separate population. All the results given in chapter 3 can thus be applied to each stratum.

The stratified mean estimator will be more efficient than the usual simple random sample mean if variation between the strata means is sufficiently large in relation to within stratum variation. The extent of gain in precision, however, also depends on the method used for selecting the units from each stratum. Once the procedure of selecting units from the strata is finalized, the other points that need careful consideration are :

1. determining the number of strata to be constructed,
2. allocation of total sample size to different strata, and
3. the choice of strata.

The answers to the above points are to be such that they minimize sampling variance for a given cost, or the cost is minimized for a specified precision. The exact solutions to these problems depend on the values of study variable (also called *estimation variable)* for all population units, which are never available. Hence, the solutions are to be based on the similar data available for some suitable supplementary variable (called *stratification* *variable* when strata are constructed on it), and on the knowledge regarding the relationship between this variable and the estimation variable.

Given below are some broad *principles* that should be kept in mind, while going for stratified sampling.

1. The strata should be nonoverlapping, and should together comprise the whole population.
2. The units forming any stratum should be similar with respect to the study variable, so that, the variability within each stratum is reduced.
3. When it is difficult to stratify the population with respect to study variable, or a highly correlated auxiliary variable, the administrative convenience may be considered as the basis for stratification . However, the gain in precision can not be guaranteed in this case, since the stratification chosen purely for administrative convenience will not necessarily yield the relative homogeneity within the strata.

We now point out some of the *advantages*, that the stratified random sampling enjoys over unstratified sampling. These are briefly discussed below :

1. Since the population is first divided into various strata, and then samples are drawn from each stratum, there is little possibility of any essential group of population being completely excluded. Hence, stratification ensures that a better cross section of the population is represented in the sample as compared to that under unstratified sampling.
2. The stratification makes it possible to use different sampling designs in different strata thereby enabling effective utilization of the available auxiliary information. It is particularly true in cases, where the extent and nature of the available information vary from stratum to stratum. Separate estimates obtained for different strata can be combined into a precise estimate for the whole population.
3. When a survey organization has field offices in several zones, it might be desirable to treat the zones as strata from the point of view of administrative convenience, as it will facilitate the supervision and organization of field work.
4. When there are extreme values in population, these can be grouped into a separate stratum thereby reducing the variability within other strata.
5. The geographical and topographical considerations may also be the reason for resorting to stratification. There may be different types of sampling problems in plains, deserts, and hilly areas. These may need different approaches for their resolution. Hence, it would be advantageous to form separate stratum for each of such areas.
6. Since the variability within strata is considerably reduced, the stratification normally provides more efficient estimates than the usual unstratified sampling.
7. The cost of conducting the survey is expected to be less for stratified sampling, when strata are formed keeping administrative convenience in mind.

From the foregoing discussion, we thus conclude that the stratification might be an aid to efficient estimation. It is, therefore, worth considering the procedure in greater detail.

5.2 NOTATIONS

Unless specified otherwise, throughout this chapter, we shall assume the sampling within each stratum to be simple random sampling WOR. The suffix h stands for h-th stratum, h=1,2,...,L, where L denotes the total number of strata into which the population has been divided. Similarly, the suffix i will indicate the i-th unit within the stratum. All the following symbols refer to the h-th stratum :

N_h = total number of units in the stratum

n_h = number of units selected in the sample from the stratum

$W_h = N_h/N$ = proportion of the population units falling in the stratum

$f_h = n_h/N_h$ = sampling fraction for the stratum

Y_{hi} = the value of study variable for the i-th unit in the stratum, $i=1,2,\dots,N_h$

$Y_h = \sum_{i=1}^{N_h} Y_{hi}$ = stratum total for the estimation variable based on N_h units

$\overline{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}$ = mean for the estimation variable in the stratum

$\overline{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$ = stratum sample mean for the estimation variable

$\sigma_h^2 = \frac{1}{N_h} (\sum_{i=1}^{N_h} Y_{hi}^2 - N_h\overline{Y}_h^2)$ = stratum variance based on N_h units

$S_h^2 = \frac{N_h}{N_h - 1} \sigma_h^2$ = stratum mean square based on N_h units

$s_h^2 = \frac{1}{n_h - 1} (\sum_{i=1}^{n_h} y_{hi}^2 - n_h\overline{y}_h^2)$ =sample mean square based on n_h sample units drawn from the stratum

We now consider the problem of estimating population mean or total from a stratified simple random sample.

5.3 ESTIMATION OF MEAN AND TOTAL USING SIMPLE RANDOM SAMPLING

In stratified sampling, it is important to keep in mind that the samples are drawn independently from each stratum, so that, the strata estimates are not correlated. From chapter 3, we know that the sample mean $\overline{y}_h$ is an unbiased estimator of the stratum mean $\overline{Y}_h$, which implies that $N_h\overline{y}_h$ is an unbiased estimator of stratum total $N_h\overline{Y}_h$. It is, therefore, reasonable to arrive at the following estimator of population mean $\bar{Y}$. We denote this estimator by $\bar{y}_{st}$, where the subscript (st) stands for stratified.

5.3.1 *Stratified SRS Without Replacement*

The unbiased estimator of population mean and the other related expressions for this case are listed below :

Unbiased estimator of population mean :

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \tag{5.1}$$

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

$$\left.\begin{aligned} V(\bar{y}_{st}) &= \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right) S_h^2 \\ &= \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} \\ &= \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h - 1}{N_h - 1}\right) \frac{\sigma_h^2}{n_h} \end{aligned}\right] \tag{5.2}$$

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

$$v(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right) s_h^2 \tag{5.3}$$

5.3.2 *Stratified SRS With Replacement*

If the sample from each stratum is selected by SRS with replacement, the expressions for variance and estimator of variance for the stratified estimator of population mean follow from relations (5.2) and (5.3), as explained in section 3.3.2, by taking fpc = $[1- (n_h-1)/(N_h-1)]$ equal to one and the sampling fraction $f_h = n_h/N_h$ as zero, for $h = 1, 2,...,L$. Thus we have :

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

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 \sigma_h^2}{n_h} \tag{5.4}$$

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

$$v(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 s_h^2}{n_h} \tag{5.5}$$

As mentioned in the earlier chapters, the estimator $\hat{Y}_{st}$ of population total Y can be obtained by multiplying the estimator of mean $\bar{y}_{st}$ by N. Also, the variance and estimator of variance expressions for $\hat{Y}_{st}$ are obtained by multiplying $V(\bar{y}_{st})$ and $v(\bar{y}_{st})$ respectively by N^2, using the expressions of $V(\bar{y}_{st})$ and $v(\bar{y}_{st})$ for WOR or WR sampling as the case may be. Accordingly, the lower and upper confidence limits for the population mean are to be multiplied by N to yield the corresponding expressions for the total Y.

Example 5.1

An assignment was given to four students attending a sample survey course. The problem was to estimate the average time per week devoted to study in Punjab Agricultural University (PAU) library by the students of this university. The university is running undergraduate, master's degree and doctoral programs. Number of students registered for the three programs is 1300, 450, and 250 respectively. Since the value of the study variable is likely to differ considerably with the program, the investigator divided the population of students into 3 strata: undergraduate program (stratum I), master's program (stratum II), and doctoral program (stratum III). First of the four students selected WOR simple random samples of sizes 20, 10, and 12 students from strata I, II, and III respectively, so that, the total sample is of size 42. The information about weekly time devoted in library is given in table 5.1.

Table 5.1 Time (in hours) devoted to study in the university library during a week

Stratum I			Stratum II		Stratum III	
0	1	9	12	6	10	24
4	4	4	9	10	14	15
3	3	6	11	9	20	14
5	6	1	13	11	11	18
2	8	2	8	7	16	19
0	10	3			13	20
3	2					

Estimate the average time per week devoted to study by a student in PAU library. Also, build up the confidence interval for this average.

Solution

Proceeding with the solution, we first prepare table 5.2 presenting calculated values of strata sample means and sample mean squares.

Table 5.2 Calculated values of strata weights, sample means, and sample mean squares

Stratum I	Stratum II	Stratum III
$n_1 = 20$	$n_2 = 10$	$n_3 = 12$
$N_1 = 1300$	$N_2 = 450$	$N_3 = 250$
$W_1 = .650$	$W_2 = .225$	$W_3 = .125$
$\bar{y}_1 = 3.800$	$\bar{y}_2 = 9.600$	$\bar{y}_3 = 16.167$
$s_1^2 = 7.958$	$s_2^2 = 4.933$	$s_3^2 = 17.049$

The stratum weight W_h,. sample mean $\bar{y}_h$, and sample mean square s_h^2 have been defined earlier in section 5.2. For calculation of $\bar{y}_h$ and s_h^2, one is to proceed in the same way as for $\bar{y}$ and s^2 in chapter 3. Now, the estimate of average time (in hours) per week devoted to study by a student in the university library, is

$$\begin{aligned}\bar{y}_{st} &= \frac{1}{N}(N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3)\\ &= \frac{1}{2000}[1300\,(3.800) + 450\,(9.600) + 250\,(16.167)]\\ &= 6.651\end{aligned}$$

Also, the estimate of variance is computed from (5.3) as

$$\begin{aligned}v(\bar{y}_{st}) &= \frac{W_1^2(N_1-n_1)\,s_1^2}{N_1n_1} + \frac{W_2^2\,(N_2-n_2)\,s_2^2}{N_2n_2} + \frac{W_3^2(N_3-n_3)\,s_3^2}{N_3n_3}\\ &= \frac{(.650)^2\,(1300-20)\,(7.958)}{(1300)\,(20)} + \frac{(.225)^2\,(450-10)\,(4.933)}{(450)\,(10)}\\ &\quad + \frac{(.125)^2\,(250-12)\,(17.049)}{(250)\,(12)}\\ &= .16553 + .02442 + .02113\\ &= .21108\end{aligned}$$

Using (2.8), we obtain the limits of confidence interval as

$$\begin{aligned}&\bar{y}_{st} \pm 2\sqrt{v(\bar{y}_{st})}\\ &= 6.651 \pm 2\sqrt{.21108}\\ &= 5.732,\ 7.570\end{aligned}$$

Thus, the average time per week devoted to study by a student in PAU library, falls in the closed interval [5.732, 7.570] hours, with probability approximately equal to .95. ■

5.4 ALLOCATION OF SAMPLE SIZE

Although the total sample size n is generally limited by the budget available for a survey, the allocation of the total sample to the strata remains at the discretion of the investigator. The precision of the estimator of population mean based on stratified sample also depends on the allocation of sample to different strata. Arbitrary allocation of the overall sample to different strata, as considered in example 5.1, is not based on any criterion, and hence does not seem reasonable. Intuitively, one may feel that the principal factors that should be kept in mind in this case are the stratum size, variability within stratum, and the cost of taking observations per sampling unit in the stratum. So far as the cost aspect is concerned, we consider one of the simplest cost functions as

$$C = c_o + \sum_{h=1}^{L} c_h n_h \tag{5.6}$$

where c_o is the overhead cost, which includes the cost of designing the questionnaire, selection of the sample, and analysis of survey data, etc. Also, c_h is the cost of observing study variable y for each unit selected in the sample from h-th stratum, h=1, 2,...,L.

A good allocation is one, where maximum precision is obtained with minimum resources. Various workers have proposed different methods to achieve this aim. However, we shall discuss only the commonly used methods of sample allocation.

Methods of sample allocation to different strata :

1. Equal allocation
2. Proportional allocation
3. Optimum allocation

We now briefly discuss these methods of sample allocation.

5.4.1 *Equal Allocation*

In case of *equal allocation*, number of sampling units selected from each stratum is equal. Thus for h = 1, 2,...,L,

$$n_h = \frac{n}{L}$$

units will be selected from each stratum. On substituting the above value of n_h in the cost constraint (5.6), one gets the required total sample size. This would be

$$n = \frac{L\,(C - c_0)}{\sum\limits_{h=1}^{L} c_h}$$

Sample size for h-th stratum in case of equal allocation :

$$n_h = \frac{n}{L} \tag{5.7}$$

Total sample size for fixed total cost :

$$n = \frac{L\,(C - c_0)}{\sum\limits_{h=1}^{L} c_h} \tag{5.8}$$

On substituting the value of $n_h = n/L$ in expressions from (5.2) to (5.5) for variance and estimator of variance, one may get the expressions appropriate for equal allocation.

It may be pointed out here, that this method of sample allocation is used when strata sizes do not differ much from each other, and the information about the variation within the strata is lacking.

Example 5.2

Second student in the group of four, was asked to independently take up the estimation problem given in example 5.1, using equal allocation. He was provided with \$150, including overhead cost of \$ 24. The cost of contacting the students, and collecting information is \$ 3 per student. How many students would he select in the sample, for collecting the desired information ?

Solution

The given details are: N_1=1300, N_2 = 450, N_3 = 250, L = 3, C = \$150, c_0 = \$24, and $c_1 = c_2 = c_3$ = \$3. The total number of students that could be included in sample is given by (5.8). Thus,

$$\begin{aligned} n &= \frac{L\,(C - c_0)}{\sum\limits_{h=1}^{L} c_h} \\ &= \frac{3\,(150 - 24)}{3 + 3 + 3} \\ &= 42 \blacksquare \end{aligned}$$

Example 5.3

In example 5.2, the second student from the group of four determined that 42 students could be selected and examined, with the funds available, to estimate the parameters of the problem given in example 5.1. Using equal allocation method, he selected n_h = n/L = 42/3 = 14 students from each stratum by using WOR simple random sampling. The information so obtained from the selected students is given in the following table :

Table 5.3 Time (in hours) devoted to study in university library during a week

Stratum I		Stratum II		Stratum III	
0	10	7	14	15	24
2	0	8	6	17	14
1	7	11	4	9	8
3	8	5	6	18	20
5	3	9	12	24	11
6	8	10	6	22	21
8	4	12	13	23	16

Estimate the parameters of example 5.1 from the above data.

Solution

Using the data given in table 5.3 above, we prepare the following table :

Table 5.4 Values of various statistics calculated from data given in table 5.3

Stratum I	Stratum II	Stratum III
$n_1 = 14$	$n_2 = 14$	$n_3 = 14$
$N_1 = 1300$	$N_2 = 450$	$N_3 = 250$
$W_1 = .650$	$W_2 = .225$	$W_3 = .125$
$\bar{y}_1 = 4.643$	$\bar{y}_2 = 8.786$	$\bar{y}_3 = 17.286$
$s_1^2 = 10.707$	$s_2^2 = 10.484$	$s_3^2 = 29.132$

From expression (5.1) and table 5.4

$$\begin{aligned}\bar{y}_{st} &= \frac{1}{N}(N_1\,\bar{y}_1 + N_2\,\bar{y}_2 + N_3\,\bar{y}_3)\\ &= \frac{1}{2000}\,[1300\,(4.643) + 450\,(8.786) + 250\,(17.286)]\\ &= 7.156\end{aligned}$$

is the estimate of the weekly average time, in hours, devoted to study by a student in PAU library. Also from (5.3), the estimate of variance is

$$\begin{aligned}v(\bar{y}_{st}) &= \frac{W_1^2(N_1 - n_1)\,s_1^2}{N_1 n_1} + \frac{W_2^2\,(N_2 - n_2)\,s_2^2}{N_2 n_2} + \frac{W_3^2(N_3 - n_3)\,s_3^2}{N_3 n_3}\\ &= \frac{(.650)^2\,(1300 - 14)\,(10.707)}{(1300)\,(14)} + \frac{(.225)^2\,(450 - 14)\,(10.484)}{(450)\,(14)}\end{aligned}$$

$$+ \frac{(.125)^2\,(250-14)\,(29.132)}{(250)\,(14)}$$

$$= .3196 + .0367 + .0307$$

$$= .3870$$

Using (2.8), we obtain the lower and upper limits of the confidence interval as

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

$$= 7.156 \pm 2\sqrt{.3870}$$

$$= 5.912,\ 8.400$$

To summarize, the estimate of the average time per week devoted to study by a student in PAU library is 7.156 hours. We are confident, with probability approximately equal to .95, that the actual average library study time per week for the PAU students will lie between 5.912 and 8.400 hours. ■

It may be noted that in case of equal allocation, no population characteristic is taken into consideration for determining the sample sizes for different strata. One needs to know only the number of strata to be constructed for finding the values of n_h, $h = 1,2,\ldots, L$. However, it seems only reasonable that the importance of characteristics like strata sizes, variability, and per unit cost of observing study variable, which may change from stratum to stratum, be recognized and given due weight while determining the sample sizes $\{n_h\}$. The methods of sample allocation that we shall discuss now, are based on these considerations.

5.4.2 *Proportional Allocation*

This allocation was first proposed by Bowley (1926). When no other information except N_h, $h = 1,2,\ldots,L$, is available, the size of strata is taken into account, and the number of units are drawn in proportion to the size of strata. This means $n_h \propto N_h$, implying that $n_h = (n/N)N_h$, $h = 1,2,\ldots,L$. On substituting in (5.6) the value of n_h thus obtained, one gets the total sample size that can be selected and observed with the available money.

Sample size for h-th stratum in case of proportional allocation :

$$n_h = \frac{n}{N} N_h \qquad (5.9)$$

Total sample size for fixed total cost :

$$n = \frac{C - c_0}{\sum\limits_{h=1}^{L} W_h c_h} \qquad (5.10)$$

Because of its simplicity, this procedure of allocation is often resorted to in practice. The allocation is likely to be nearly optimum for a fixed sample size, when the strata variances are almost same. On using the allocation $n_h = (n/N)N_h$, $h = 1,2,...,L$, in (5.2) to (5.5), one gets corresponding expressions for variance and estimator of variance for the estimator of population mean under proportional allocation.

Example 5.4

The third student of the group of four, was independently assigned the estimation problem of example 5.1, and was asked to use proportional allocation method . Using the budget and cost information of example 5.2, determine the total number of students that he could afford to select. Also, allocate the sample units to different strata.

Solution

In this case, we have $N_1 = 1300$, $N_2 = 450$, $N_3 = 250$, $N = 2000$, $L = 3$, $C = \$150$, $c_0 = \$24$, and $c_1 = c_2 = c_3 = \$3$. The total sample size that could be possible with the given information, is obtained by using (5.10). As $W_h = N_h/N$, the expression (5.10) can be written as

$$n = \frac{N(C - c_0)}{\sum_{h=1}^{L} N_h c_h}$$

$$= \frac{(2000)\,(150-24)}{(1300)\,(3) + (450)\,(3) + (250)\,(3)}$$

$$= 42$$

The total number of 42 students to be included in the sample are allocated to each of the 3 strata through (5.9). Thus,

$$n_1 = \left(\frac{n}{N}\right)N_1 = \left(\frac{42}{2000}\right)(1300) = 27.3 \approx 27$$

$$n_2 = \left(\frac{n}{N}\right)N_2 = \left(\frac{42}{2000}\right)(450) = 9.5 \approx 10$$

$$n_3 = \left(\frac{n}{N}\right)N_3 = \left(\frac{42}{2000}\right)(250) = 5.3 \approx 5$$

Therefore, 27, 10, and 5 students would be selected from strata I, II, and III respectively. ■

Example 5.5

In order to estimate the parameters of example 5.1, the total sample size that could be possible with the given budget has been obtained, in example 5.4, as 42 along with its proportional allocation to different strata. Accordingly, the student investigator selected 27 students from stratum I, 10 from stratum II, and 5 from stratum III. The information collected from the students in the sample is given in table 5.5.

Table 5.5 Time (in hours) devoted to study in library during a week

Stratum I					Stratum II		Stratum III
4	5	11	3	2	5	12	18
3	6	0	8	1	9	8	20
10	4	1	2	7	7	10	17
6	9	10	4		16	17	23
8	3	5	6		11	7	10
1	12	4	5				

Estimate the parameters of example 5.1.

Solution

For computations, we prepare table 5.6.

Table 5.6 Calculated values of various statistics for data given in table 5.5

Stratum I	Stratum II	Stratum III
$n_1 = 27$	$n_2 = 10$	$n_3 = 5$
$N_1 = 1300$	$N_2 = 450$	$N_3 = 250$
$W_1 =$ ..650	$W_2 = .225$	$W_3 = .125$
$\bar{y}_1 = 5.185$	$\bar{y}_2 = 10.200$	$\bar{y}_3 = 17.600$
$s_1^2 = 10.851$	$s_2^2 = 15.289$	$s_3^2 = 23.300$

Using (5.1), and various values from table 5.6, we find

$$\bar{y}_{st} = \frac{1}{N}(N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3)$$

$$= \frac{1}{2000}[1300(5.185) + 450(10.200) + 250(17.600)]$$

$$= 7.865$$

The estimate of variance $V(\bar{y}_{st})$ is computed by using (5.3) as

$$v(\bar{y}_{st}) = \frac{W_1^2(N_1 - n_1)s_1^2}{N_1 n_1} + \frac{W_2^2(N_2 - n_2)s_2^2}{N_2 n_2} + \frac{W_3^2(N_3 - n_3)s_3^2}{N_3 n_3}$$

$$= \frac{(.650)^2(1300-27)(10.851)}{(1300)(27)} + \frac{(.225)^2(450-10)(15.289)}{(450)(10)}$$

$$+ \frac{(.125)^2(250-5)(23.300)}{(250)(5)}$$

$$= .1663 + .0757 + .0714$$

$$= .3134$$

Utilizing (2.8), we work out the confidence interval from

$$\begin{aligned}&\bar{y}_{st} \pm 2\sqrt{v(\bar{y}_{st})}\\&= 7.865 \pm 2\sqrt{.3134}\\&= 6.745,\ 8.985\end{aligned}$$

Thus, on the average, a PAU student devotes 7.865 hours per week to study in the library. So far as the actual population average is concerned, we are reasonably confident that it will fall in the closed interval [6.745, 8.985] hours. ■

5.4.3 *Optimum/Neyman Allocation*

Very often, a survey statistician has to work within a fixed budget. In such a situation, he is expected to minimize the variance of the estimator subject to the cost constraint. In certain other cases, from the point of view of the results of the survey, it might be possible to state an acceptable value of the variance. The problem then is to minimize the cost, subject to the constraint that the variance of the estimator does not exceed the stated value. First we consider the former situation.

Case I. In this case, the total cost of the survey is fixed. The aim is to find the sample allocation $\{n_h\}$ such that the variance of the estimator is minimum. The allocation $\{n_h\}$, which minimizes the variance in (5.2) for a given cost C in (5.6), is called *optimum allocation*. The sample allocation, so obtained, is given in (5.11). The overall sample size for the optimum allocation can, however, be found by adding the n_h values obtained under this allocation.

Fixed total cost - minimum variance allocation :

$$n_h = \frac{(C - c_0)\, W_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} W_h S_h \sqrt{c_h}} \tag{5.11}$$

Total sample size :

$$n = \sum_{h=1}^{L} n_h \tag{5.12}$$

where n_h has been given in (5.11).

We thus see that the stratum sample size will be proportional to the stratum size and stratum standard deviation, but inversely proportional to the square root of the cost per sampling unit in each stratum. It means, large strata with greater variability and low per unit observation cost will lead to larger samples in relation to those from other strata. In case the sampling is WR, the sample allocation $\{n_h\}$ that minimizes the variance in (5.4) is obtained by replacing S_h by σ_h in (5.11).

Example 5.6

A car manufacturing company has sold 2000 cars to the public through licensed dealers. The company is now interested in finding out the average distance travelled per week by a car manufactured by the company. This information is likely to be helpful in fixing the warranty period for certain parts of the car. The addresses and telephone numbers, if installed, of all the buyers along with their occupations are available at the head office of the company. Since the distance travelled by a car is likely to vary with the profession of the buyer, the investigator divides the population into 3 groups - the businessmen (stratum I), employees (stratum II), and others (stratum III) which includes farmers, etc. Out of 2000 buyers, 825 are businessmen, 700 employees, and 475 others. The average per unit cost for collecting information is expected to be \$ 4 for businessmen, \$ 5.5 for employees, and \$ 6.5 for persons from other category. The total budget at hand is \$ 1550 which includes the overhead cost of \$1000. On using optimum allocation formula given in (5.11), the investigator arrived at allocation of sample size n_1 = 53 buyers to stratum I, n_2 = 34 buyers to stratum II, and n_3 = 23 buyers to stratum III (procedure of determining these allocations is explained in the solution). The observations on the study variable obtained from these three WOR simple random samples are given in table 5.7.

Table 5.7 Average distance (in km) per week covered by cars included in the sample

Stratum I				Stratum II			Stratum III	
656	301	575	666	470	281	685	712	236
400	870	525	715	351	410	492	679	824
526	813	310	691	625	240	206	665	385
774	861	650	480	388	636	579	319	650
780	722	470	680	566	422	358	840	585
812	705	460	841	421	517	385	421	496
805	831	483	825	398	451		666	704
525	748	310	488	881	380		848	569
401	446	489	330	434	326		410	614
806	856	576	580	405	595		549	
828	387	615	811	693	401		602	
746	399	704		615	612		253	
560	635	774		375	564		777	
475	560	533		469	343		411	

The information on strata mean squares, from a similar survey carried out in the past for another car model, is given for strata I, II, and III respectively as S^2_1 = 30505, S^2_2 = 24008, and S^2_3 =29215.

Solution

Here we have

$$C = \$\ 1550,\ c_o = \$\ 1000,\ c_1 = \$\ 4,\ c_2 = \$\ 5.5,\ c_3 = \$\ 6.5,$$
$$N_1 = 825,\ N_2 = 700,\ N_3 = 475,\ N = 2000,\ W_1 = .4125,\ W_2 = .3500,$$
$$W_3 = .2375,\ S_1^2 = 30505,\ S_2^2 = 24008,\text{ and } S_3^2 = 29215.$$

The sample size allocation to different strata from (5.11) will be

$$n_h = \frac{(C - c_0)\, W_hS_h/\sqrt{c_h}}{\sum_{h=1}^{L} W_hS_h\sqrt{c_h}},\quad h = 1,2, ..., L$$

Now, we first compute

$$\sum_{h=1}^{L} W_hS_h\sqrt{c_h} = (.4125)(\sqrt{30505})(\sqrt{4}) + (.3500)(\sqrt{24008})(\sqrt{5.5})$$
$$+ (.2375)(\sqrt{29215})(\sqrt{6.5})$$
$$= 374.77$$

Then, the sample size allocation for the three strata will be

$$n_1 = \frac{(1550 - 1000)(.4125)\sqrt{30505}}{(374.77)\sqrt{4}} = 52.87 \approx 53$$

$$n_2 = \frac{(1550 - 1000)(.3500)\sqrt{24008}}{(374.77)\sqrt{5.5}} = 33.94 \approx 34$$

$$n_3 = \frac{(1550 - 1000)(.2375)\sqrt{29215}}{(374.77)\sqrt{6.5}} = 23.37 \approx 23$$

The information collected from WOR simple random samples of 53, 34, and 23 respondents selected from strata I, II, and III is given in table 5.7 above. The sample means and sample mean squares for the 3 strata are then calculated. These are given below :

$$\bar{y}_1 = 619.038 \qquad \bar{y}_2 = 469.824 \qquad \bar{y}_3 = 574.565$$
$$s_1^2 = 28531.190 \qquad s_2^2 = 20696.634 \qquad s_3^2 = 32871.256$$

The estimate of mean is now computed from (5.1). Thus,

$$\bar{y}_{st} = (W_1\bar{y}_1 + W_2\bar{y}_2 + W_3\bar{y}_3)$$
$$= \frac{1}{N}(N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3)$$
$$= \frac{1}{2000}[(825)(619.038) + (700)(469.824) + (475)(574.565)]$$
$$= 556.251$$

Hence, the estimate of the average distance per weak covered by a car, manufactured by the company, is 556.251 km.

For calculating the estimate of variance we use (5.3), and get

$$v(\bar{y}_{st}) = \frac{W_1^2(N_1-n_1)\,s_1^2}{N_1n_1} + \frac{W_2^2\,(N_2-n_2)\,s_2^2}{N_2n_2} + \frac{W_3^2(N_3-n_3)\,s_3^2}{N_3n_3}$$

$$= \frac{(.4125)^2\ (825-53)\ (28531.190)}{(825)\ (53)} + \frac{(.3500)^2\ (700-34)\ (20696.634)}{(700)\ (34)}$$

$$+ \frac{(.2375)^2\ (475-23)\ (32871.256)}{(475)\ (23)}$$

$$= 85.7147 + 70.9468 + 76.7115$$

$$= 233.373$$

Further, we have

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

$$= 556.251 \pm 2\sqrt{233.373}$$

$$= 525.698,\ 586.804$$

Thus, the confidence interval for the population value of the average distance covered per week by a car is obtained as [525.698, 586.804] km. ■

If the cost per unit is same for all the strata, that is, $c_h = c'$ for each h, optimum allocation is known as *Neyman allocation*, after Neyman (1934). For this case, the cost function (5.6) takes more simpler form

$$C = c_0 + c'n \qquad (5.13)$$

Then, the strata sample sizes for Neyman allocation are given as :

Minimum variance – Neyman allocation :

$$\left.\begin{aligned} n_h &= n\,\frac{W_h\,S_h}{\sum\limits_{h=1}^{L} W_hS_h} \\ &= n\,\frac{N_hS_h}{\sum\limits_{h=1}^{L} N_hS_h} \end{aligned}\right] \qquad (5.14)$$

Total sample size :

$$n = \frac{C-c_o}{c'} \qquad (5.15)$$

Example 5.7

The fourth student, in the group of four, was asked to undertake the estimation of parameters considered in examples 5.1 to 5.5 by using Neyman allocation. Again, the cost for contacting the students and gathering the required information is \$ 3 per student. The total budget at disposal was \$150 including the overhead cost of \$24. Using Neyman allocation, the samples of sizes 25, 10, and 7 students were selected from strata I, II, and III respectively (procedure of determining these sample sizes is explained in the solution). The data collected from these three WOR simple random samples, selected from three strata, are presented in table 5.8. Using this information, estimate the average time per week devoted to study in library by a student. Also, set up confidence interval for the population mean. The information on strata mean squares is to be used from example 5.5.

Solution

We are given that

$C = \$150$, $c_o = \$24$, $c' = \$\ 3$, $N_1 = 1300$, $N_2 = 450$, $N_3 = 250$, $S_1^2 = 10.851$, $S_2^2 = 15.289$, and $S_3^2 = 23.300$.

Taking cost into account, the total sample size from (5.15) will be

$$n = \frac{C - c_o}{c'} = \frac{150 - 24}{3} = 42$$

Now,

$$\sum_{h=1}^{3} N_h S_h = N_1 S_1 + N_2 S_2 + N_3 S_3$$

$$= (1300)(\sqrt{10.851}) + (450)(\sqrt{15.289}) + (250)(\sqrt{23.300})$$

$$= 7248.615$$

The sample sizes for different strata are then determined using (5.14), where

$$n_h = n \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h}, \quad h = 1, 2, \ldots, L$$

Thus,

$$n_1 = (42)\frac{(1300)\left(\sqrt{10.851}\right)}{7248.615} = 24.81 \approx 25$$

$$n_2 = (42)\frac{(450)\left(\sqrt{15.289}\right)}{7248.615} = 10.20 \approx 10$$

$$n_3 = (42)\frac{(250)\left(\sqrt{23.300}\right)}{7248.615} = 6.99 \approx 7$$

The observations recorded from these selected students are given below in table 5.8.

Table 5.8 Time (in hours) devoted to study in library by selected students during a week

Stratum I					Stratum II		Stratum III	
9	6	8	6	10	9	16	24	25
1	7	3	9	5	14	6	18	22
3	2	5	2	3	13	8	11	
5	4	4	5	4	8	12	19	
4	6	4	1	7	12	11	16	

For convenience, we compute sample estimates for mean and mean square error for each stratum. These are given along with other required information in table 5.9.

Table 5.9 Necessary computations for strata I, II, and III

Stratum I	Stratum II	Stratum III
$n_1 = 25$	$n_2 = 10$	$n_3 = 7$
$N_1 = 1300$	$N_2 = 450$	$N_3 = 250$
$W_1 = .650$	$W_2 = .225$	$W_3 = .125$
$\bar{y}_1 = 4.920$	$\bar{y}_2 = 10.900$	$\bar{y}_3 = 19.286$
$s_1^2 = 5.993$	$s_2^2 = 9.656$	$s_3^2 = 23.892$

By using figures from table 5.9 in (5.1), we work out the estimate of weekly average study time in the library as

$$\begin{aligned} \bar{y}_{st} &= \frac{1}{N}(N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3) \\ &= \frac{1}{2000}[1300(4.920) + 450(10.900) + 250(19.286)] \\ &= 8.061 \end{aligned}$$

Next we compute estimate of variance $V(\bar{y}_{st})$ from (5.3) as

$$\begin{aligned} v(\bar{y}_{st}) &= \frac{W_1^2(N_1 - n_1)\,s_1^2}{N_1n_1} + \frac{W_2^2\,(N_2 - n_2)\,s_2^2}{N_2n_2} + \frac{W_3^2(N_3 - n_3)\,s_3^2}{N_3n_3} \\ &= \frac{(.650)^2\,(1300 - 25)\,(5.993)}{(1300)\,(25)} + \frac{(.225)^2\,(450 - 10)\,(9.656)}{(450)\,(10)} \\ &\quad + \frac{(.125)^2\,(250 - 7)\,(23.892)}{(250)\,(7)} \\ &= .0993 + .0478 + .0518 \\ &= .1989 \end{aligned}$$

The required confidence interval for population mean is then obtained from

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

$$= 8.061 \pm 2\sqrt{.1989}$$

$$= 7.169,\ 8.953$$

One can claim that there is approximately 95% chance that the population value of the average weekly study time in the university library will be covered by the closed interval [7.169, 8.953] hours. ■

Case II. Here, we fix the precision of the estimator at a specified level and minimize the total cost of survey. The desired level of precision can be specified in two ways. It could be done either by fixing the value of variance $V(\bar{y}_{st})$ at V_o, or by specifying the bound B on the error for the estimator $\bar{y}_{st}$. These two modes of precision specification are, however, related to each other through the equation

$$2\sqrt{V(\bar{y}_{st})} = B$$

implying

$$V(\bar{y}_{st}) = V_o = \frac{B^2}{4}$$

Above mentioned relation can, therefore, be used to convert one mode of specification to the other. Thus for any given value of B, we can find the corresponding value of V_o. So, if it is required to ensure a specified value V_o of the variance of the estimator, then on using the cost function (5.6), the sample size for the h-th stratum, h=1,2,..,L, is found as the one given in (5.16). The total sample size is again the sum of n_h values obtained under this allocation.

Fixed variance - minimum cost allocation :

$$n_h = \frac{\left(W_h S_h/\sqrt{c_h}\right) \sum_{h=1}^{L} W_h S_h \sqrt{c_h}}{V_o + \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2} \tag{5.16}$$

where V_o is the value at which the variance of the estimator $\bar{y}_{st}$ is fixed.

Total sample size :

$$n = \sum_{h=1}^{L} n_h \tag{5.17}$$

where n_h has been obtained in (5.16).

For the cost function in (5.13), the above allocation reduces to minimum cost-Neyman allocation.

Minimum cost - Neyman allocation :

$$n_h = \frac{(W_hS_h) \sum_{h=1}^{L} W_hS_h}{V_o + \frac{1}{N} \sum_{h=1}^{L} W_hS_h^2} \qquad (5.18)$$

Total sample size :

$$n = \sum_{h=1}^{L} n_h \qquad (5.19)$$

where n_h has been obtained in (5.18).

In order to obtain sample allocation and total sample size for estimation of mean/total for a given variance V_o in case of WR sampling, take the factor $(\Sigma W_h S^2_h)/N$, h=1,2,...,L, as zero and replace S_h^2 by σ_h^2 in (5.16) and (5.18).

In both the cases (fixed cost or fixed variance) of optimum/ Neyman allocation, it is necessary to have approximate values for S_h^2 (or σ_h^2) for each stratum to be able to use the allocation formulas. These values will not be normally available in practice. Approximate values of these strata mean squares can be had from some similar studies conducted in the past, or from the ranges of the observations within each stratum, if available. The range of a set of data is known to be approximately four times its standard deviation. Alternatively, a pilot survey, based on a small sample, could also be conducted to obtain sample estimates s_h^2, of S_h^2 in case of WOR sampling and of σ_h^2 if the sampling is to be with replacement. These estimated values of S_h^2 (or σ_h^2) can then be used to determine sample sizes for different strata.

Example 5.8

If the car manufacturer wishes to estimate the parameters of example 5.6 with a predetermined variance V_o = 170, what will be the total sample size and allocation of sample to different strata, so that, the total cost is minimum for the same per unit costs as in example 5.6 ? The estimates of strata mean squares (s_1^2 = 28531.190, s_2^2 = 20696.634, and s_3^2 = 32871.256) obtained in example 5.6 are to be used as known strata mean squares for the purpose of present allocation.

Solution

From the statement of the example, we have V_o = 170. The other information from example 5.6 is

c_1 = \$4, c_2 = \$5.5, c_3 = \$6.5, N_1 = 825, N_2 = 700,

N_3 = 475, W_1 = .4125, W_2 = .3500, and W_3 = .2375.

The estimated values of S_h^2 in example 5.6 are now taken as good approximations of S_h^2, h = 1, 2, 3. Therefore,

$S_1^2 = 28531.190$, $S_2^2 = 20696.634$, and $S_3^2 = 32871.256$.

For a fixed variance and minimum cost, the sample allocation is given by (5.16) as

$$n_h = \frac{\left(W_h S_h / \sqrt{c_h}\right) \sum_{h=1}^{L} W_h S_h \sqrt{c_h}}{V_o + \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2}$$

First, we calculate two terms that are to be used later on. These terms are

$$\begin{aligned} \sum_{h=1}^{L} W_h S_h \sqrt{c_h} &= (.4125)(\sqrt{28531.190})(\sqrt{4}) + (.3500)(\sqrt{20696.634})(\sqrt{5.5}) \\ &\quad + (.2375)(\sqrt{32871.256})(\sqrt{6.5}) \\ &= 367.220 \end{aligned}$$

$$\begin{aligned} \sum_{h=1}^{L} W_h\, S_h^2 &= (.4125)(28531.190) + (.3500)(20696.634) \\ &\quad + (.2375)(32871.256) \\ &= 26819.861 \end{aligned}$$

It follows that

$$\begin{aligned} V_o + \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2 &= 170 + \frac{26819.861}{2000} \\ &= 183.410 \end{aligned}$$

Then, the sample allocation is seen to be

$$n_1 = \frac{(.4125)(\sqrt{28531.190})(367.220)}{(\sqrt{4})(183.410)} = 69.8 \approx 70$$

$$n_2 = \frac{(.3500)(\sqrt{20696.634})(367.220)}{(\sqrt{5.5})(183.410)} = 43$$

$$n_3 = \frac{(.2375)(\sqrt{32871.256})(367.220)}{(\sqrt{6.5})(183.410)} = 33.8 \approx 34$$

The total sample size required will, therefore, be

$$\begin{aligned} n &= n_1 + n_2 + n_3 \\ &= 70 + 43 + 34 \\ &= 147 \; ■ \end{aligned}$$

After discussing the various popular methods of sample allocation to different strata, we now attempt to answer the question whether a particular stratification and sample allocation combination will at all be advantageous in relation to the unstratified simple random sampling ?

5.5 RELATIVE EFFICIENCY OF STRATIFIED ESTIMATOR

For examining the usefulness of stratification, we need the sampling variances of the estimators of population mean/total for stratified and unstratified population. The percent relative efficiency of the estimator $\bar{y}_{st}$, with respect to the usual unstratified estimator $\bar{y}$, is then given by

$$RE = \frac{V(\bar{y})}{V(\bar{y}_{st})} (100)$$

where $V(\bar{y})$ and $V(\bar{y}_{st})$, for WOR sampling, are defined in (3.9) and (5.2) respectively. We illustrate below, the various steps involved in the calculation of the above said actual percent relative efficiency for three commonly used sample allocation methods. A relative efficiency figure of well over 100 indicates that the stratification of the population would be effective in reducing the estimation error. For this purpose, we consider a hypothetical situation where the study variable values are known for all population units.

Example 5.9

All the 80 farms in a population are stratified by farm size. The expenditure on the insecticides used during the last year by each farmer is presented in table 5.10 below :

Table 5.10 Expenditure (in '00 rupees) on insecticides used

Large farmers		Medium farmers				Small farmers		
75	76	55	40	51	28	35	31	26
65	79	45	38	55	47	28	38	32
86	62	35	33	41	61	36	42	18
57	92	30	43	48	35	40	33	16
45	50	42	53	54	31	25	29	
69	48	38	37	36	23	18	25	
48	77	40	52	44		28	35	
60	60	36	39	47		32	26	
55	64	48	46	39		13	30	
66	58	46	42	41		19	37	

Select a stratified sample of 24 farmers by using equal allocation, proportional allocation, and Neyman allocation. Compute the overall population mean $\bar{Y}$ and the population mean square S^2. Work out the relative efficiency of stratified sample mean $\bar{y}_{st}$, based on each of the above mentioned allocations, with respect to the simple random sample mean $\bar{y}$ for the same total sample size. Assume that the sampling is WOR.

Solution

It is given that n = 24, $N_1 = 20$, $N_2 = 36$, and $N_3 = 24$. Hence $W_1 = .25$, $W_2 = .45$, and $W_3 = .30$. First, we calculate overall population mean $\overline{Y}$ and population mean square S^2 based on all the 80 farms. Thus,

$$\begin{aligned}\overline{Y} &= \frac{1}{80}(75+65+\ldots+16)\\ &= 43.7875\end{aligned}$$

$$\begin{aligned}S^2 &= \frac{1}{80-1}[(75)^2 + (65)^2 + \ldots + (16)^2 - (80)(43.7875)^2]\\ &= 268.6758\end{aligned}$$

From (3.9), the sampling variance of mean in case of usual SRS without replacement is given by

$$\begin{aligned}V(\overline{y}) &= \frac{N-n}{Nn}S^2\\ &= \frac{80-24}{(80)(24)}(268.6758)\\ &= 7.8364\end{aligned}$$

Analogous to S^2, the stratum mean square S_h^2 is computed separately for each stratum. Hence, one gets

$$S_1^2 = 169.5158,\ S_2^2 = 70.5611, \text{ and } S_3^2 = 61.4493.$$

We now obtain the value of $V(\overline{y}_{st})$ under three sample allocation methods.

Equal allocation. In this case, the number of units to be selected from each stratum will be $n_h = 24/3 = 8$. The sampling variance for stratified mean $\overline{y}_{st}$ from (5.2), will be

$$\begin{aligned}V(\overline{y}_{st}) &= \sum_{h=1}^{L} W_h^2\left(\frac{N_h - n_h}{N_h n_h}\right)S_h^2\\ &= (.25)^2\left(\frac{20-8}{(20)(8)}\right)(169.5158) + (.45)^2\left(\frac{36-8}{(36)(8)}\right)(70.5611)\\ &\quad + (.30)^2\left(\frac{24-8}{(24)(8)}\right)(61.4493)\\ &= .7946 + 1.3892 + .4609\\ &= 2.6447\end{aligned}$$

The percent relative efficiency of equal allocation based mean $\overline{y}_{st}$, with respect to usual mean $\overline{y}$, is obtained as

$$\begin{aligned} \text{RE} &= \frac{V(\bar{y})}{V(\bar{y}_{st})}(100) \\ &= \frac{7.8364}{2.6447}(100) \\ &= 296.3 \end{aligned}$$

Proportional allocation. The number of units to be selected from each stratum, using proportional allocation defined in (5.9), will be

$$n_1 = \left(\frac{n}{N}\right)N_1 = \left(\frac{24}{80}\right)(20) = 6$$

$$n_2 = \left(\frac{n}{N}\right)N_2 = \left(\frac{24}{80}\right)(36) = 10.8 \approx 11$$

$$n_3 = \left(\frac{n}{N}\right)N_3 = \left(\frac{24}{80}\right)(24) = 7.2 \approx 7$$

In place of sample size 8, used for equal allocation for calculating variance $V(\bar{y}_{st})$, here we use 6, 11, and 7 for strata I, II, and III respectively. Therefore,

$$\begin{aligned} V(\bar{y}_{st}) &= (.25)^2\left(\frac{20-6}{(20)(6)}\right)(169.5158) + (.45)^2\left(\frac{36-11}{(36)(11)}\right)(70.5611) \\ &\quad + (.30)^2\left(\frac{24-7}{(24)(7)}\right)(61.4493) \\ &= 1.2361 + .9021 + .5596 \\ &= 2.6978 \end{aligned}$$

The percent relative efficiency of proportional allocation based stratified mean estimator $\bar{y}_{st}$, with respect to usual mean estimator $\bar{y}$, is given by

$$\begin{aligned} \text{RE} &= \frac{7.8364}{2.6978}(100) \\ &= 290.5 \end{aligned}$$

Neyman allocation. To arrive at the number of units to be selected from each stratum under Neyman allocation, we first work out $\Sigma N_h S_h$, $h = 1, 2, 3$. Thus,

$$\begin{aligned} \Sigma N_h S_h &= (20)(\sqrt{169.5158}) + (36)(\sqrt{70.5611}) + (24)(\sqrt{61.4493}) \\ &= 260.4 + 302.4 + 188.1 \\ &= 750.9 \end{aligned}$$

Then from (5.14), for h = 1, 2 ,..., L, we have

$$n_1 = n\frac{N_1S_1}{\Sigma N_hS_h} = (24)\frac{260.4}{750.9} = 8.3 \approx 8$$

$$n_2 = n\frac{N_2S_2}{\Sigma N_hS_h} = (24)\frac{302.4}{750.9} = 9.7 \approx 10$$

$$n_3 = n\frac{N_3S_3}{\Sigma N_hS_h} \doteq (24)\frac{188.1}{750.9} = 6$$

On using the above allocated sample sizes, the variance of mean estimator becomes

$$V(\bar{y}_{st}) = (.25)^2\left(\frac{20-8}{(20)(8)}\right)(169.5158) + (.45)^2\left(\frac{36-10}{(36)(10)}\right)(70.5611)$$

$$+ (.30)^2\left(\frac{24-6}{(24)(6)}\right)(61.4493)$$

$$= .7946 + 1.0320 + .6913$$

$$= 2.5179$$

The percent relative efficiency of Neyman allocation based mean estimator, with respect to usual simple random sampling estimator of mean, is obtained as

$$RE = \frac{7.8364}{2.5179}(100)$$

$$= 311.2 \blacksquare$$

It will be observed in the above example, that the calculated $\{n_h\}$ values often involve approximations due to the rounding off to the nearest integer. These approximations ultimately affect the sampling variance. The more exact values of the variance $V(\bar{y}_{st})$ can be obtained by using alternative forms of variance expressions. These forms of variances do not directly involve $\{n_h\}$ values. Hence no rounding off approximations are involved. These expressions can be obtained by putting respective $\{n_h\}$ values (in expression form) in (5.2). We shall here consider such variance expressions only for proportional and Neyman allocation methods. These expressions, given in (5.20) and (5.21), can respectively be obtained by substituting $\{n_h\}$ values from (5.9) and (5.14) in (5.2). Similar variance expression for the equal allocation method can be easily obtained by putting $n_h = n/L$, h = 1, 2,..., L, in (5.2). The reader must note that such type of alternative expressions can not be obtained for estimators of variance $V(\bar{y}_{st})$, since each s_h^2 has to be calculated from samples of sizes that are whole numbers.

Example 5.10

For the data of example 5.9, work out variances of stratified sample mean $\bar{y}_{st}$, under proportional and Neyman allocations, by using the expressions

$$V_p(\bar{y}_{st}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{L} W_h S_h^2 \tag{5.20}$$

$$V_n(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 \tag{5.21}$$

respectively.

Solution

Using different values already computed in example 5.9, we calculate the following two terms :

$$\begin{aligned}\sum_{h=1}^{L} W_h S_h^2 &= (.25)(169.5158) + (.45)(70.5611) + (.30)(61.4493)\\ &= 92.5662\end{aligned}$$

$$\begin{aligned}\sum_{h=1}^{L} W_h S_h &= (.25)(\sqrt{169.5158}) + (.45)(\sqrt{70.5611}) + (.30)(\sqrt{61.4493})\\ &= 9.3867\end{aligned}$$

On substituting the different values in expressions of $V_p(\bar{y}_{st})$ and $V_n(\bar{y}_{st})$, one gets

$$\begin{aligned}V_p(\bar{y}_{st}) &= \left(\frac{1}{24} - \frac{1}{80}\right)(92.5662)\\ &= 2.6998\end{aligned}$$

$$\begin{aligned}V_n(\bar{y}_{st}) &= \frac{1}{24}(9.3867)^2 - \frac{1}{80}(92.5662)\\ &= 2.5142 \blacksquare\end{aligned}$$

For examples 5.9 and 5.10, it was assumed that the values of the study variable are available for all population units. In practice, however, the situation is different. The investigator has observations on the study variable for the stratified sample only. It is from this sample data that one has to estimate the variances $V(\bar{y}_{st})$ and $V(\bar{y})$. These estimated variances are then used to estimate the relative efficiency of the estimator $\bar{y}_{st}$ with respect to the estimator $\bar{y}$.

The estimator of $V(\bar{y}_{st})$ from a stratified simple random WOR sample is available in (5.3), while the estimator $v_{st}(\bar{y})$ of $V(\bar{y})$ from the stratified sample is obtained in (5.22) as

$$v_{st}(\bar{y}) = \frac{N-n}{Nn(N-1)}\left[\sum_{h=1}^{L}\frac{N_h}{n_h}\left(\sum_{i=1}^{n_h} y_{hi}^2\right) - N\{\bar{y}_{st}^2 - v(\bar{y}_{st})\}\right] \tag{5.22}$$

where $\bar{y}_{st}$ and $v(\bar{y}_{st})$ are defined in (5.1) and (5.3) respectively.

The estimated percent relative efficiency is then given by

$$RE = \frac{v_{st}(\bar{y})}{v(\bar{y}_{st})}(100)$$

Various steps involved in calculating the estimate of RE are explained in example 5.11.

Example 5.11

The sample data obtained by using proportional allocation are given in example 5.5. Estimate the relative efficiency of proportional allocation based stratified estimator $\bar{y}_{st}$, in relation to usual unstratified simple mean estimator $\bar{y}$, from the above referred stratified sample observations.

Solution

From example 5.5, we have $N_1 = 1300$, $N_2 = 450$, $N_3 = 250$, $N = 2000$, $n_1 = 27$, $n_2 = 10$, $n_3 = 5$, $\bar{y}_{st} = 7.865$, and $v(\bar{y}_{st}) = .3133$. Now from table 5.5, we compute the term $\sum_{i=1}^{n_h} y_{hi}^2$ for h = 1,2,3. Thus,

$$\sum_{i=1}^{27} y_{1i}^2 = 4^2 + 3^2 + \ldots + 7^2 = 1008$$

$$\sum_{i=1}^{10} y_{2i}^2 = 5^2 + 9^2 + \ldots + 7^2 = 1178$$

$$\sum_{i=1}^{5} y_{3i}^2 = 18^2 + 20^2 + \ldots + 10^2 = 1642$$

Using above computed values, we calculate

$$\sum_{h=1}^{L} \frac{N_h}{n_h}\left(\sum_{i=1}^{n_h} y_{hi}^2\right) = \frac{1300}{27}(1008) + \frac{450}{10}(1178) + \frac{250}{5}(1642)$$

$$= 183643.33$$

Now on making substitutions in (5.22), we obtain the estimated variance of the usual SRS estimator $\bar{y}$ from the stratified sample. Thus,

$$v_{st}(\bar{y}) = \frac{2000-42}{(2000)(42)(2000-1)}[183643.33 - (2000)\{(7.865)^2 - .3133\}]$$

$$= .7061$$

The required percent relative efficiency is then estimated as

$$RE = \frac{v_{st}(\bar{y})}{v(\bar{y}_{st})}(100)$$

$$= \frac{.7061}{.3133}(100)$$

$$= 225.38 \blacksquare$$

5.6 ESTIMATION OF POPULATION PROPORTION

So far, we have dealt with the estimation of population mean $\bar{Y}$ and population total Y on the basis of with and without replacement stratified simple random samples. The results for the estimation of $\bar{Y}$ can easily be extended for the estimation of population proportion P. For this, we take the value of y_{hi} as 1 or 0 according as the unit belongs to the class of interest or not. In this case, $\bar{y}_h$, $\bar{Y}_h$, and $\bar{Y}$ reduce to h-th stratum sample proportion p_h, the h-th stratum proportion P_h, and the overall population proportion P respectively. One can also see that for this case $\sigma_h^2 = P_h\ Q_h$, where $Q_h = 1 - P_h$, h = 1, 2, ..., L. Thus we have (5.23).

Unbiased estimator of population proportion P :

$$p_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h p_h = \sum_{h=1}^{L} W_h p_h \qquad (5.23)$$

The expressions for variance $V(p_{st})$ and its estimator $v(p_{st})$, for stratified simple random sampling WOR, are given below :

Variance of estimator p_{st} :

$$V(p_{st}) = \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h - 1}{N_h - 1}\right) \frac{P_h Q_h}{n_h} \qquad (5.24)$$

Estimator of variance $V(p_{st})$:

$$v(p_{st}) = \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{p_h q_h}{n_h - 1} \qquad (5.25)$$

In case of WR sampling, the fpc = [1- $(n_h$-1) / $(N_h$-1)] and the sampling fraction $f_h = n_n/N_h$ are respectively taken as 1 and 0. The expressions (5.24) and (5.25) then reduce to their counterparts in WR method.

So far as the allocation and determination of sample size for estimation of population proportion P is concerned, the discussion of section 5.4 still holds, except for the difference that σ_h^2 and S_h^2 will be replaced by P_hQ_h and $N_hP_hQ_h/(N_h-1)$ respectively, whereas s_h^2 will be replaced by $n_hp_hq_h/(n_h-1)$.

Example 5.12

The management of a local newspaper is to decide whether it should continue with the publication of 'Children Column', which had been introduced on experimental basis. For this purpose, it is imperative to estimate the proportion of readers who would favor its continuance. The frame consists of readers who had stayed with the paper for the last six months. The addresses of these readers are available in the office of the newspaper. Since different attitudes are expected from the urban and rural readers, it is reasonable to stratify the population into urban readers and rural readers. In the population, there are 73000 urban readers and 30280 rural readers. The total budget at hand is \$3000 only. The overhead cost is \$820, and the per unit cost for urban and rural readers is expected to be \$2 and \$2.5 respectively. Using proportional allocation method, explained in solution, the investigator selected WOR simple random samples of 718 respondents from stratum I (urban readers) and 298 readers from stratum II (rural readers). The number of individuals who favor continuation of the column was 570 from stratum I and 143 from stratum II. Estimate the proportion of readers interested in the continuation of the said column. Also, build up confidence interval for the population proportion.

Solution

Here we have,

$N_1 = 73000$, $N_2 = 30280$, $N = 103280$, $C = \$3000$, $c_o = \$820$,
$c_1 = \$2$, $c_2 = \$2.5$, $W_1 = N_1/N = .7068$, and $W_2 = N_2/N = .2932$.

Total number of respondents who can be surveyed is given by (5.10) as

$$n = \frac{C - c_o}{\sum\limits_{h=1}^{L} W_h c_h} = \frac{(3000 - 820)}{(.7068)(2) + (.2932)(2.5)} = 1015.6 \approx 1016$$

Using (5.9), the number of respondents to be selected from stratum I and stratum II are

$$n_1 = \left(\frac{n}{N}\right) N_1 = \left(\frac{1016}{103280}\right)(73000) = 718.1 \approx 718$$

$$n_2 = \left(\frac{n}{N}\right) N_2 = \left(\frac{1016}{103280}\right)(30280) = 297.9 \approx 298$$

The estimate of proportion of urban readers who are in favor of continuing the 'Children Column' is

$$p_1 = \frac{570}{718} = .7939$$

Similarly, the estimate of proportion of rural readers, who wish the continuance of the said column, is

$$p_2 = \frac{143}{298} = .4799$$

The estimate of the required overall population proportion is now worked out from (5.23) as

$$\begin{aligned} p_{st} &= \frac{1}{N}(N_1p_1 + N_2p_2) \\ &= \frac{1}{103280}[(73000)(.7939) + (30280)(.4799)] \\ &= .7018 \end{aligned}$$

Next, we compute estimate of variance from (5.25). It yields

$$\begin{aligned} v(p_{st}) &= W_1^2\left(\frac{N_1 - n_1}{N_1}\right)\frac{p_1q_1}{n_1 - 1} + W_2^2\left(\frac{N_2 - n_2}{N_2}\right)\frac{p_2q_2}{n_2 - 1} \\ &= (.7068)^2\left(\frac{73000 - 718}{73000}\right)\left(\frac{.7939\ (1 - .7939)}{718 - 1}\right) \\ &\quad + (.2932)^2\left(\frac{30280 - 298}{30280}\right)\left(\frac{.4799\ (1 - .4799)}{298 - 1}\right) \\ &= .0001129 + .0000715 \\ &= .0001844 \end{aligned}$$

The confidence interval for population proportion is worked out by using (2.8) as

$$\begin{aligned} & p_{st} \pm 2\sqrt{v\,(p_{st})} \\ &= .7018 \pm 2\sqrt{.0001844} \\ &= .6746,\ .7290 \end{aligned}$$

To summarize, the sample estimate of proportion indicates that about 70 percent of reader population is in favor of continuing the column in question. The confidence limits obtained above assure the management of the paper that the population proportion favoring the continuance of the column will lie in the closed interval [.6746, .7290] with probability approximately equal to .95. ■

Example 5.13

After 6 months of the survey considered in example 5.12, the management of that newspaper again wishes to estimate the proportion of readers who would favor continuance of 'Children Column'. The budget at disposal, overhead cost, and per unit cost for collecting information for urban and rural readers remain same as in the survey considered in example 5.12. The estimates of strata variances obtained in example 5.12 are to be used as known strata variances for the present survey for the purpose of sample allocation. Using all this information, allocate the sample to different strata following optimum allocation method.

Solution

From example 5.12, we have $N_1 = 73000$, $N_2 = 30280$, $N = 103280$, $C = \$3000$, $c_0 = \$820$, $c_1 = \$2$, $c_2 = \$2.5$, $W_1 = .7068$, and $W_2 = .2932$. Also,

$$A_1 = \frac{n_1p_1q_1}{n_1 - 1} = \frac{718\,(.7939)\,(1-.7939)}{718-1} = .1639$$

$$A_2 = \frac{n_2p_2q_2}{n_2 - 1} = \frac{298\,(.4799)\,(1-.4799)}{298-1} = .2504$$

The optimum allocation for quantitative data is given by (5.11). By using the estimate s_h in place of S_h, the expression is written as

$$n_h = \frac{(C - c_0)W_h s_h/\sqrt{c_h}}{\sum\limits_{h=1}^{L} W_h s_h \sqrt{c_h}}$$

In case of attributes, corresponding to s_h^2 we have $\frac{n_h p_h q_h}{n_h - 1}$. The above allocation formula, therefore, reduces to

$$n_h = \frac{(C - c_0)W_h\sqrt{A_h}/\sqrt{c_h}}{\sum\limits_{h=1}^{L} W_h\sqrt{A_h}\sqrt{c_h}}, \; h = 1, 2$$

where A_1 and A_2 have been defined and calculated above. Now, we work out

$$\sum_{h=1}^{L} W_h\sqrt{A_h}\sqrt{c_h} = .7068\,(\sqrt{.1639})\,(\sqrt{2}) + .2932\,(\sqrt{.2504})\,(\sqrt{2.5})$$
$$= .6367$$

The substitution of different values in the above formula for n_h gives

$$n_1 = \frac{(3000 - 820)\,(.7068)\,(\sqrt{.1639})}{(\sqrt{2})\,(.6367)} = 692.8 \approx 693$$

$$n_2 = \frac{(3000 - 820)\,(.2932)\,(\sqrt{.2504})}{(\sqrt{2.5})\,(.6367)} = 317.7 \approx 318$$

Therefore, under optimum allocation that minimizes the variance, 693 respondents will be selected from urban readers and 318 from the rural stratum. ■

5.7 CONSTRUCTION OF STRATA

As pointed out earlier, the basic consideration involved in the formation of strata is that the strata should be internally homogeneous. For a single study variable y, the best characteristic for construction of strata is the distribution of study variable y itself. L strata could then be formed by cutting this distribution at (L-1) suitable points. This distribution of y is generally not available in practice, and in absence of this information, the next best alternative is the frequency distribution of some other variable which is highly correlated with the study variable y. Construction of strata on such an auxiliary variable will not yield exactly optimum strata, but these will be approximately optimum. In what follows, the procedures of constructing strata for different allocation methods are discussed.

5.7.1 *Neyman and Equal Allocation*

Dalenius and Hodges (1957) gave cumulative square root rule to obtain approximately optimum strata for Neyman allocation by using the frequency distribution for the study variable y. As this distribution is not available in practice, the rule is used on the frequency distribution of a highly positively correlated auxiliary variable x (also called stratification variable) to obtain approximately optimum stratification on x. Cumulative square root rule, though proposed for the Neyman allocation method, is also found to yield approximately optimum strata for equal allocation method. This rule can, therefore, be used for the construction of strata for both Neyman and equal allocation methods. Steps involved in the construction of strata, through the above rule, are listed as under :

Steps involved in cumulative square root rule :

1. Obtain a frequency table for stratification variable x.
2. In the frequency table for x, obtain square roots of the frequencies for each of the K classes.
3. Obtain the cumulative totals of the square roots of frequencies for each of the K classes. Let T denote the cumulative total for the K-th class.
4. If L strata are to be constructed, then using linear interpolation method on the class intervals and the cumulative square root frequency column, obtain the value of $x = x_1$, which corresponds to the value T/L in the cumulative square root frequency column.
5. Repeat the process in step (4) to obtain $x = x_i$ corresponding to the value i.T/L, $i = 2,3,\ldots,L-1$, in the cumulative square root frequency column.
6. The values $(x_1, x_2,\ldots, x_{L-1})$ so obtained define L strata with boundaries $(< x_1)$, $(x_1$ to $x_2)$, $(x_2$ to $x_3),\ldots,(x_{L-2}$ to $x_{L-1})$, and $(\geq x_{L-1})$.

Example 5.14

It is desired to estimate average annual milk yield per cow for a *tharparkar* herd of 127 cows at a certain government cattle farm using stratified simple random sampling. Cows in the herd are to be grouped into three strata on the basis of first lactation length in days. Neyman method of sample allocation is to be used for selecting the overall sample of 25 cows from the three strata. Determine approximately optimum strata boundaries using the information on first lactation length given in table 5.11.

Table 5.11 First lactation length (in days) and other related computations

Lactation length	No. of cows (f)	$\sqrt{f}$	Cumulative $\sqrt{f}$
30 - 70	4	2.00	2.00
70 - 110	6	2.45	4.45
110-150	3	1.73	6.18
150-190	8	2.83	9.01
190-230	20	4.47	13.48
230-270	27	5.20	18.68
270-310	25	5.00	23.68
310-350	14	3.74	27.42
350-390	7	2.65	30.07
390-430	6	2.45	32.52
430-470	6	2.45	34.97
470-510	1	1.00	35.97

Solution

In this example, we are already given the frequency table for the stratification variable, first lactation length. As the next step, we find square roots of the frequencies (f) given in column (2) of the table 5.11. These square root values ($\sqrt{f}$) are presented in column (3). The cumulative totals of $\sqrt{f}$ are then obtained. These totals constitute column (4) of the table.

For this illustration, we have L=3, K=12, and T=35.97. For constructing three strata, we need to determine only two boundaries, x_1 and x_2 in days, using linear interpolation between the class intervals and the cumulative $\sqrt{f}$ values. As stated earlier, x_1 and x_2 are to correspond to T/3 = 35.97/3 = 11.99 and 2T/3 = 2(35.97)/3 = 23.98 in column (4). From table 5.11, we find that a value of 9.01 in column (4) corresponds to the value 190 in column (1), whereas a value of 13.48 in column (4) corresponds to the value 230 in column (1). Thus, an increase of 4.47 in cumulative $\sqrt{f}$ value takes place over the interval 190-230. First lactation length x_1 corresponding to the cumulative $\sqrt{f}$ value of 11.99, therefore, lies in this interval. Hence,

$$\begin{aligned} x_1 &= 190 + \frac{(40)\,(11.99-9.01)}{4.47} \\ &= 216.67 \end{aligned}$$

Similarly,

$$\begin{aligned} x_2 &= 310 + \frac{(40)\,(23.98-23.68)}{3.74} \\ &= 313.21 \end{aligned}$$

It shows that the cows with the first lactation length in the range [30, 216.67] will constitute the first stratum, whereas those having lactation lengths in the ranges [216.67, 313.21] and [313.21, 510] will form second and third strata respectively. ■

5.7.2 *Proportional Allocation*

Singh (1975) proposed a cumulative cube root rule to obtain approximately optimum strata boundaries for the proportional allocation method. The rule is to be applied on the frequency table for the stratification variable x in exactly the same way as the cumulative square root rule of Dalenius and Hodges (1957), except for the difference that in this method we shall operate with the cube roots of the class frequencies in place of their square roots.

Example 5.15

It is proposed to estimate total wool yield in a certain region of Rajasthan state in India, using stratified simple random sampling. An overall sample of 20 villages is to be selected employing proportional allocation method. The stationary sheep population data, for 141 villages of this region, is given in the frequency table below. Construct three approximately optimum strata taking stationary sheep population as the stratification variable.

Table 5.12 Stationary sheep population with other related computations

No. of sheep	No. of villages (f)	$\sqrt[3]{f}$	Cumulative $\sqrt[3]{f}$
0-100	46	3.583	3.583
100-200	36	3.302	6.885
200-300	23	2.844	9.729
300-400	11	2.224	11.953
400-500	6	1.817	13.770
500-600	4	1.587	15.357
600-700	4	1.587	16.944
700-800	1	1.000	17.944
800-900	4	1.587	19.531
900-1000	4	1.587	21.118
1000-1100	1	1.000	22.118
1100-1200	1	1.000	23.118

Solution

For this case, the cube roots of frequencies in column (2) are given in column (3), and their cumulated values are presented in column (4) of table 5.12. Here we have L=3, K=12, and T= 23.118. The two strata boundaries, x_1 and x_2, needed to form three strata will now correspond to T/3 = 23.118/3 = 7.706 and 2T/3 = 2(23.118)/3 =15.412 respectively. Thus, on proceeding as in example 5.14, we get from columns (1) and (4)

$$x_1 = 200 + \frac{(100)\,(7.706 - 6.885)}{2.844} = 228.87$$

$$x_2 = 600 + \frac{(100)\,(15.412 - 15.357)}{1.587} = 603.47$$

Hence, the villages having stationary sheep population in the range [0, 228.87] constitute first stratum, and those having sheep population in the ranges [228.87, 603.47] and [603.47, 1200] form second and third strata respectively. ■

5.8 POSTSTRATIFICATION

In stratified sampling, it is presupposed that the strata sizes and the sampling frame for each stratum are available. However, the situations do exist where the latter is difficult to obtain. For instance, the details about classification of farmers' population by farm size (small, medium, large) can be had from the census records, but list of farmers falling in each of the three classes may not be available. Consequently, it is not possible to determine in advance as to which stratum a farmer belongs until he is observed for the farm size. This means, the units can be assigned to different strata only after the sample units are contacted and observed. This whole procedure is termed as *poststratification*. This technique is useful where published journals/ reports may provide clear indication of strata sizes, but due to nonavailability of strata frames it is difficult to sample the units from different strata. Below we give the estimator of population mean, its variance, and estimator of this variance when the sample units are stratified after they have been selected as a single WOR simple random sample from the entire unstratified population.

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

$$\overline{y}_{ps} = \sum_{h=1}^{L} W_h \overline{y}_h \tag{5.26}$$

Approximate variance of estimator $\overline{y}_{ps}$:

$$V(\overline{y}_{ps}) = \frac{N-n}{Nn} \sum_{h=1}^{L} W_h S_h^2 + \frac{1}{n^2} \sum_{h=1}^{L} (1-W_h)\, S_h^2 \tag{5.27}$$

Estimator of variance $V(\overline{y}_{ps})$:

$$v(\overline{y}_{ps}) = \sum_{h=1}^{L} \left(\frac{N_h - n_h}{N_h n_h}\right) W_h^2\, s_h^2 \tag{5.28}$$

The first term of (5.27) is the value of $V(\overline{y}_{st})$ for proportional allocation. The second term is due to the fact that $\{n_h\}$ do not distribute themselves exactly proportionally because of poststratification. However, for sufficiently large n, this second term will be small in comparison to first. It means that for reasonably large n, the poststratification is almost as precise as stratified sampling with proportional allocation.

Example 5.16

The list of all the 800 farmers in a development block can be obtained but the information about their farm sizes is not available. Thus the farmers can not be classified as large, medium, or small, and the usual stratified sampling procedure is not applicable. However, from census records the proportion of large, medium, and small farmers in the above said population is known to be .2, .5, and .3 respectively. The State Bank of India is interested in estimating the total amount of loan expected to be taken by the farmers of the development block during the next financial year. For this purpose, a WOR simple random sample of 40 farmers was selected and then classified according to their farm sizes. The information regarding the expected amount of loan to be taken by the selected farmers is given below :

Table 5.13 Expected amount of loan (in '000 rupees) to be taken by the sample farmers

Large farmers	Medium farmers			Small farmers	
44	30	29	35	18	16
50	42	30	33	22	20
60	28	29	36	17	14
38	20	34	27	26	23
52	30	25	31	21	28
43	31	27	24	19	24
				16	15
				10	16

Estimate the total amount of loan expected to be taken by all the farmers in the block. Also, place confidence limits on this total.

Solution

Here, we have $W_1 = .2$, $W_2 = .5$, $W_3 = .3$, and $N = 800$. Also, $n_1 = 6$, $n_2 = 18$, and $n_3 = 16$. Now proceeding as in examples 5.1 and 5.3, we compute sample means and sample mean squares for each stratum. These are given below along with values of N_h, W_h, and n_h. The N_h values for h=1, 2, 3, are computed from the equation $N_h = NW_h$.

Table 5.14 Certain sample and population values

Large farmers	Medium farmers	Small farmers
$n_1 = 6$	$n_2 = 18$	$n_3 = 16$
$W_1 = .2$	$W_2 = .5$	$W_3 = .3$
$N_1 = 160$	$N_2 = 400$	$N_3 = 240$
$\bar{y}_1 = 47.833$	$\bar{y}_2 = 30.056$	$\bar{y}_3 = 19.063$
$s_1^2 = 60.967$	$s_2^2 = 24.526$	$s_3^2 = 22.596$

From (5.26), the estimate of the total amount of loan, expected to be taken by all the farmers of the block, will be

$$\begin{aligned}\hat{Y} &= N\bar{y}_{ps} = \sum_{h=1}^{L} N_h\bar{y}_h \\ &= 160\,(47.833) + 400\,(30.056) + 240\,(19.063) \\ &= 24250.80\end{aligned}$$

We now compute estimate for the variance of $\hat{Y}$. For this, we use (5.28). Thus,

$$\begin{aligned}v\,(\hat{Y}_{ps}) &= N^2\, v(\bar{y}_{ps}) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{n_h}\, s_h^2 \\ &= \frac{160\,(160-6)}{6}\,(60.967) + \frac{400\,(400-18)}{18}\,(24.526) \\ &\quad + \frac{240\,(240-16)}{16}\,(22.596) \\ &= 250371.14 + 208198.48 + 75922.56 \\ &= 534492.18\end{aligned}$$

The required confidence limits are worked out following (2.8). These are given by

$$\begin{aligned}&\hat{Y}_{ps} \pm 2\sqrt{v(\hat{Y}_{ps})} \\ &= 24250.80 \pm 2\sqrt{534492.18} \\ &= 22788.62,\ 25712.98\end{aligned}$$

The investigator could, therefore, be reasonably confident that the total amount of loan, expected to be taken by all the 800 farmers of the block, is in the range of 22788.62 to 25712.98 thousand rupees. ■

While dealing with poststratification, we have assumed that strata sizes are known exactly. But in certain situations, it may not be the case. In such a situation, one is either to use guess values of $\{N_h\}$ from some past survey or census record, or one can estimate them by drawing a large preliminary sample and then use a subsample to obtain information on the study variable. In both the cases, the values of $\{N_h\}$ used may differ from the actual $\{N_h\}$. These inaccuracies affect the precision of the required estimate. Detailed discussion of the problem is given in Sukhatme *et al.* (1984).

5.9 SOME FURTHER REMARKS

5.1 Suppose a sample of n_h units is selected from N_h units of the h-th stratum with PPS with replacement using x as the size variable. Let Y_{hi} and $P_{hi} = (X_{hi}/X_h)$ respectively, denote the study variable value and the probability of selection for the i-th unit of the h-th stratum. Also, let y_{hi} and p_{hi} be the corresponding sample values. Then we have the following results :

Unbiased estimator of population total Y:

$$\hat{Y}_{pst} = \sum_{h=1}^{L} \hat{Y}_h = \sum_{h=1}^{L} \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{y_{hi}}{p_{hi}} \qquad (5.29)$$

Variance of estimator $\hat{Y}_{pst}$:

$$V(\hat{Y}_{pst}) = \sum_{h=1}^{L} V(\hat{Y}_h) = \sum_{h=1}^{L} \frac{1}{n_h} \sum_{i=1}^{N_h} \left(\frac{Y_{hi}}{P_{hi}} - Y_h\right)^2 P_{hi} \qquad (5.30).$$

Estimator of variance $V(\hat{Y}_{pst})$:

$$v(\hat{Y}_{pst}) = \sum_{h=1}^{L} v(\hat{Y}_h) = \sum_{h=1}^{L} \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} \left(\frac{y_{hi}^2}{p_{hi}^2} - n_h\hat{Y}_h^2\right) \qquad (5.31)$$

where $\hat{Y}_h$ is defined in (5.29).

5.2 In certain situations, the population is highly heterogeneous and several effective criteria are available for stratification. In such cases, it may be desirable to carry stratification to the extent that only one unit is selected from each stratum. In this event, it is not possible to estimate $V(\bar{y}_{st})$ and $V(\hat{Y}_{st})$. However, approximate estimate of variance is possible by using method of *collapsed strata*. With L even, this method consists in grouping the adjoining strata in pairs to form collapsed strata, and then estimating the sampling variance as if two units had been sampled from each collapsed stratum. Hansen, Hurwitz, and Madow (1953) have given a more general procedure by grouping the strata into g groups.

5.3 Goodman and Kish (1950) proposed that it is possible to enhance the control of error further. The method they suggested is termed *controlled selection* procedure. It increases the probability of selection of preferred samples, and, consequently, reduces the probability of selection of nonpreferred samples without altering the probabilities of selection to the units in a stratified sampling design.

5.4 Sometimes, the investigator is interested in estimating population characteristics for several variables. If all the variables of interest are closely related to a single auxiliary variable, say x, and information for all the population units on x is available, then stratification and allocation is done as discussed in this chapter. If all the study variables are not related to a single auxiliary variable but are related to more than one auxiliary variable, the procedure of stratification and allocation is little modified and termed *multiple stratification* or *deep stratification*. In this procedure, the population units are first stratified using the most important auxiliary variable. The strata thus formed are called *primary strata*. Then each primary strata is further stratified using another auxiliary variable. For details, the reader may refer to Sukhatme *et al.* (1984).

5.5 Certain other methods of determining approximately optimum strata boundaries on the study variables have been proposed by Aoyama (1954), Dalenius and Gurney (1951), Dalenius and Hodges (1959), and Mahalanobis (1952). These methods are, however, not preferred to cumulative square root rule. Methods of approximately optimum stratification on an auxiliary variable for certain other allocation procedures have been given by Singh (1971), Singh and Parkash (1975), and Mehta *et al.* (1995).

LET US DO

5.1 Explain, why should one use stratified simple random sampling ? What are the various points one should keep in mind while stratifying a population ?

5.2 Discuss various problems that are to be resolved before one could start selecting a stratified simple random sample.

5.3 How does stratification increase efficiency of the estimator of mean/total and proportion ?

5.4 Do you think it appropriate to use stratified sampling to estimate

a. per head travelling allowance (TA) of teaching and nonteaching staff of a university,
b. average calories taken per day by boys and girls,
c. proportion of nonexistent voters in a voter list, and
d. mean weight of adult men and women in a city block.

5.5 What do you understand by sample allocation? Describe merits and demerits of various sample allocation methods.

5.6 A stratified sample of size 80 is to be drawn from a population of 6400 units, divided into 3 strata of sizes 2400, 3200, and 800 units. If the allocation is to be equal, or proportional, how many units should be selected from individual stratum in each case?

5.7 An insurance company's records show that out of the total of 500 claims, 280 are major claims (from Rs 1000 to Rs 2500) and 220 are minor (below Rs 1000). A WOR simple random sample of 10 claims was drawn from each category (stratum), and claim amounts were recorded as :

Stratum I : 1200, 1600, 1800, 1400, 1980, 2110, 2440, 1660, 1790, 1910
Stratum II : 720, 880, 760, 660, 790, 840, 550, 960, 640, 800

Estimate the total amount of all the 500 claims, and construct the confidence interval for it.

5.8 The adult population in a colony consists of 400 Sikhs, 260 Muslims, 200 Hindus, and 140 Christians. An investigator selected 40 Sikhs, 26 Muslims, 18 Hindus, and 16 Christians so as to draw a total sample of 100 adults. Do you think the allocation is proportional ?

5.9 During 1990-91 session, a student doing M.Sc. (Statistics) was given a project to estimate average time taken by the university employees to get ready for office in the morning. The population was grouped into 3 strata. The first stratum consisted of women. The males were divided into 2 strata – teachers and the other staff. A WOR random sample of 400 employees was drawn using proportional allocation. The information on time (in hours) taken by selected respondents to be ready for office was collected. Below are given the sample average and sample mean square for each stratum along with the values of N_h and n_h.

Strata	N_h	n_h	$\bar{y}_h$	s^2_h
Women	1250	125	2.0	.1587
Teachers	2390	239	1.6	.2667
Others	360	36	1.2	.0811
Total	4000	400		

Estimate the average time taken to get ready for office, and place the required confidence limits on it.

5.10 Discuss the method of sample allocation to different strata when (a) total cost of survey is fixed and the aim is to minimize variance, (b) precision of the estimator $\bar{y}_{st}$ is fixed and the objective is to minimize survey cost.

5.11 Take the estimates of strata mean squares obtained in exercise 5.9 as known. Using this information along with strata sizes, determine Neyman allocation when the total budget at disposal is Rs 1000. Assume that the overhead cost, and the cost of eliciting and processing information per respondent, are Rs 100 and Rs 3 respectively.

5.12 In March 1992, another student of M.Sc., majoring in statistics, was assigned the job to estimate the parameters of exercise 5.9 using Neyman allocation. For this, $\{s^2_h\}$ values in exercise 5.9 were treated as actual strata mean squares. One thousand rupees were allotted to him for the completion of this job. The overhead cost was expected to be Rs 200, and the cost of collecting and processing the information per respondent was known as Rs 2. This way, the number of units to be selected from each stratum came out to be $n_1 = 109$, $n_2 = 269$, and $n_3 = 22$. The allocated number of individuals were drawn from each stratum by using SRS without replacement, and the requisite information on time (in hours) spent by each sample individual to get ready for office work was obtained. In table below are presented the sample average and sample mean square for each stratum, along with the values of N_h and n_h.

Stratum	N_h	n_h	$\bar{y}_h$	s^2_h
Women	1250	109	2.3	.1308
Teachers	2390	269	1.6	.2111
Others	360	22	1.4	.0905

Estimate the parameter of exercise 5.9, and construct confidence interval for it.

5.13 Below are given 50 population values divided into three strata :

Stratum I : 78, 87, 71, 88, 76, 99, 98, 86, 75, 92

Stratum II : 40, 46, 42, 58, 60, 55, 49, 54, 61, 48, 44, 52,
54, 46, 58, 57, 50, 48, 44, 50, 47, 50, 41, 43

Stratum III : 30, 28, 28, 24, 26, 21, 34, 28, 36, 27, 26, 32,
24, 25, 29, 23

Compute the overall population mean $\overline{Y}$ and the population mean square S^2. If one is to select WOR random sample of size 10, determine the appropriate stratum sample sizes using proportional allocation and Neyman allocation. Work out the relative efficiency of the stratified sample mean $\overline{y}_{st}$ for proportional and Neyman allocations, with respect to the usual SRS based mean estimator $\overline{y}$, for the same total sample size of 10.

5.14 The list of all the 50,000 adults in a town was available. In order to estimate proportion of literate adults (educated up to at least 8th grade), the population was stratified into 3 strata with respect to age. A WOR random sample of size 500 persons was drawn using proportional allocation. The sample size allocation to each stratum and the number of literate persons recorded in sample of size n_h, h = 1, 2, 3, are given below :

Age group (years)	Persons	n_h	Literate persons
20-40	25600	256	243
40-60	18100	181	144
60 and over	6300	63	31

Compute the estimate of proportion of literate persons in the town, and construct confidence interval for it.

5.15 The Electricity Department has 4 offices in a town. There are complaints of receiving faulty electricity bills by the consumers. The 4 offices were treated as strata and the bills issued during the last six months were to be scrutinized. An amount of Rs 1720 was allotted to carry out the survey. The overhead cost was expected to be Rs 100, and the cost of scrutinizing a bill, and processing the collected information, is Rs 3. In all, 540 units could be observed with these funds. For determining the sample size for each stratum, proportional allocation was used. The results obtained from the survey are reported as follows :

Office	Total bills issued	n_h	Faulty bills
1	7000	140	13
2	4280	86	7
3	9560	191	9
4	6160	123	10
Total	27000	540	

Estimate total number of faulty bills issued by all the 4 offices, and place confidence limits on it.

5.16 Describe how will you obtain approximately optimum strata boundaries for Neyman allocation ?

5.17 The objective of a survey is to estimate the total wheat production in a certain region having 220 farmers. The area (in hectares) under wheat in respect of each farmer is available from revenue records. Its distribution is given below :

Area	Farmers	Area	Farmers
0-2	6	10-12	41
2-4	13	12-14	30
4-6	14	14-16	16
6-8	36	16-18	11
8-10	50	18-20	3

Neyman allocation method is to be employed for selecting the overall sample of 30 farmers after dividing the population into 3 strata. Determine the approximately optimum strata boundaries using the cumulative square root method.

5.18 Describe cumulative cube root rule for constructing approximately optimum strata. For which allocation method, this rule is appropriate ?

5.19 An overall sample of 30 farmers is to be selected from 3 strata using proportional allocation for the estimation of parameter of exercise 5.17. Determine approximately optimum strata boundaries by using Singh's (1975) cumulative cube root method on the frequency table of exercise 5.17.

5.20 What do you understand by poststratification? Identify 4 situations where this technique could be useful. Give expressions for the estimator of total, its variance, and estimator of variance.

5.21 The objective of a study was to estimate mean fibre length of three newly developed strains of cotton (*Gossypium arboreum* L.). Incidentally, the seeds of these strains got mixed at the time of packing and transporting from the headquarters. Since no spare seed was available, the plants were raised using the mixed seed. However, the strains could be distinguished if the plant characteristics were examined carefully by using laboratory equipment. On maturity, the total number of plants were counted as 1200. The proportion of plants of each variety was guessed from the number of each type of seed known before mixing. These were $W_1 = .60$, $W_2 = .25$, and $W_3 = .15$. An overall WOR random sample of 60 plants was selected. The selected plants were observed critically and assigned to the appropriate strain category. The fibre length recorded (in mm) for the sample plants is given below :

Strain I				Strain II		Strain III
18.5	17.0	16.9	18.3	19.6	20.4	21.6
17.4	18.3	17.4	17.9	19.4	20.6	20.7
19.3	18.1	16.8	17.6	20.3	19.3	21.4
18.0	19.6	17.0	17.1	19.8	18.4	21.4
17.6	17.5	18.5	18.0	18.6	19.1	20.5
17.6	17.4	18.8	19.4	18.9	17.9	22.3
18.2	18.4	18.1	19.2	19.5	20.5	21.7
17.5	19.1	17.1	18.3	18.4	18.3	22.2
17.0	17.3	16.8	17.3			

Estimate the mean fibre length, and place confidence limits on this average.