Stratified Sampling

The basic idea behind the stratified sampling is to

- divide the whole heterogeneous population into smaller groups or subpopulations, such that the sampling units are homogeneous with respect to the characteristic under study within the subpopulation and
- heterogeneous with respect to the characteristic under study between/among the subpopulations.Such subpopulations are termed as strata.
- Treat each subpopulation as a separate population and draw a sample by SRS from each stratum.

Stratified Sampling

- 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*.

The major advantages of stratification are (1) estimates for each stratum can be obtained separately, (2) differences among the strata can be evaluated, (3) the total, mean, and proportion of the entire population can be estimated with high precision by suitably weighting the estimates obtained from each stratum.

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.

Notation:

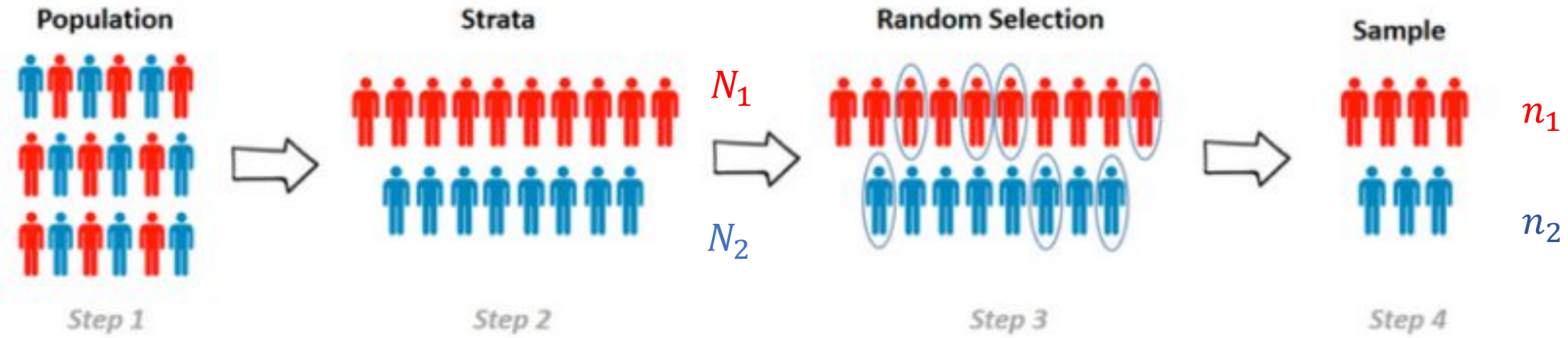
N_h = total number of units in the stratum

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

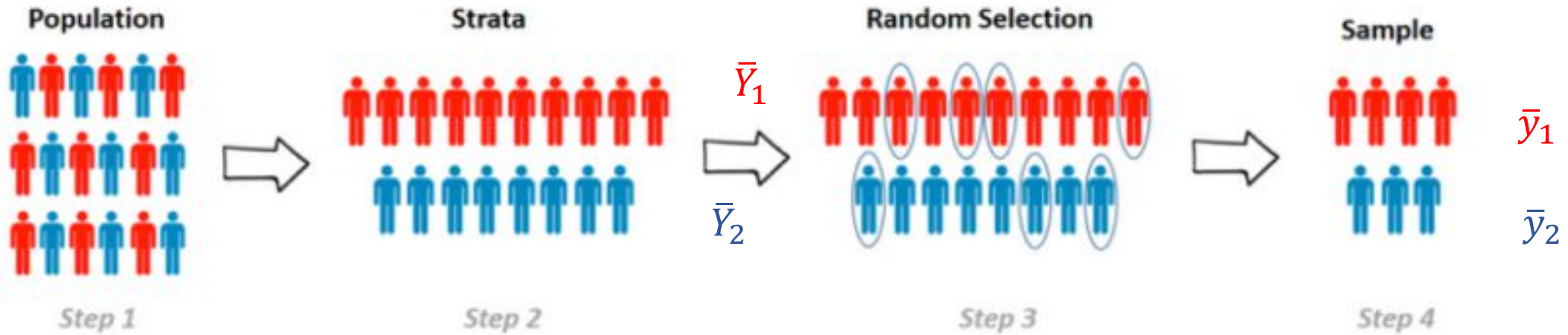
Y_{hi} = the value of study variable for the i-th unit in the stratum, i=1,2,...,N_h



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

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

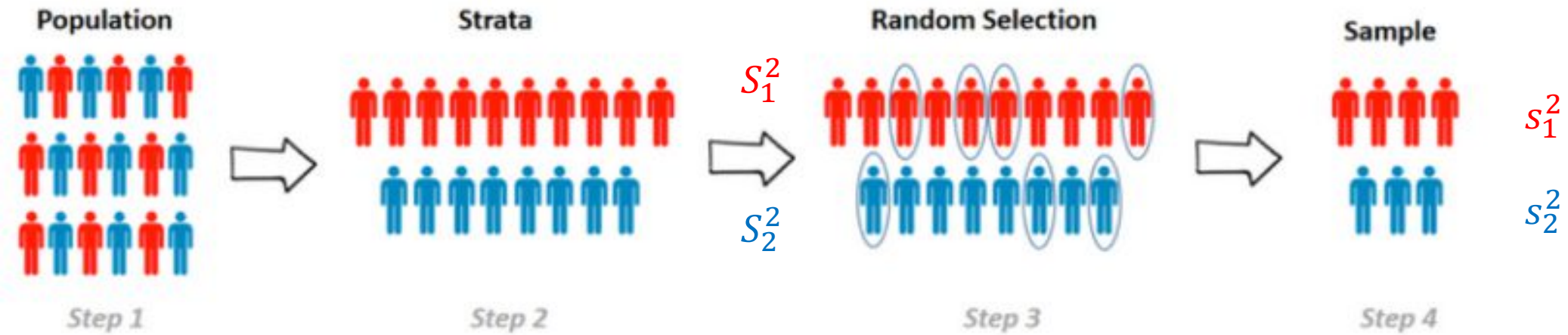
$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} = \text{stratum sample mean for the estimation variable}$$



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

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

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



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

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

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

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

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

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

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.

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

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

$$= \frac{1}{2000}[1300\,(3.800) + 450\,(9.600) + 250\,(16.167)]$$

$$= 6.651$$

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

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

$$= \frac{(.650)^2\ (1300 - 20)\ (7.958)}{(1300)\ (20)} + \frac{(.225)^2\ (450 - 10)\ (4.933)}{(450)\ (10)}$$

$$+ \frac{(.125)^2\ (250 - 12)\ (17.049)}{(250)\ (12)}$$

$$= .16553 + .02442 + .02113$$

$$= .21108$$

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

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

$$= 6.651 \pm 2\sqrt{.21108}$$

$$= 5.732,\ 7.570$$

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.

Methods of sample allocation to different strata :

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

Equal Allocation:

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 :

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

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,

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

$$= \frac{3\,(150 - 24)}{3+3+3}$$

$$= 42 \blacksquare$$

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 :

Estimate the parameters of example 5.1 from the **below** data.

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

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

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

$$= \frac{1}{2000}[1300\,(4.643) + 450\,(8.786) + 250\,(17.286)]$$

$$= 7.156$$

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

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

$$= \frac{(.650)^2\,(1300 - 14)\,(10.707)}{(1300)\,(14)} + \frac{(.225)^2\,(450 - 14)\,(10.484)}{(450)\,(14)}$$

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

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

$$= .3870$$

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. ■

HW

Page 140 question 5.7, 5.9, 5.13 by assuming an equal allocation only.