UNIT 13 STRATIFIED SAMPLING

Structure **Page No.**

13.1 INTRODUCTION

As seen in the previous unit, *sampling* is a process by which we choose a *representative sample* from a given *population*. The main *objective of sampling* is to make an optimal use of the budget and other resources for a study to obtain as precise an estimate of a population parameter as possible.

On the basis of what you read in Unit 12, it is plausible to accept that if the *target population* is homogeneous with respect to the study variable, then the sample size requirement is likely to be small. However, as we know, a population is always associated with certain amount of variability. And so, if the population is heterogeneous, a *larger sample* is required to increase the precision of the population estimates. But, as you will agree, it is not possible always to adopt a *sampling design* needing *large* samples.

Many modifications have evolved from the central concept of *simple random sampling* that permit more precise inferences to be attained for different types of populations, especially wherein the *variability* within *small subpopulations* is small. One of the most practically useful designs is **stratified sampling**, in which we first divide the *target population* into *homogenous segments* and then, following any method of sample selection, choose samples from each of these *segments* (or *subpopulations*) independently in such a way that the total number of units pooled over all the selected groups is the desired sample size. The *small groups* thus formed are called **strata** and the process of forming strata, is called **stratification**.

The term **strata** is the plural of the word *stratum*.

In Sec.13.2, we shall discuss *preliminaries about stratification, broad principles adopted while using stratified sampling*, and some *advantages of stratified random sampling*. In Sec.13.3, we shall illustrate the use of certain relations commonly employed in the estimation of the population characteristics. In Sec.13.4, we shall discuss *equal allocation, proportional allocation*, and *optimum allocation*, which are some of the commonly used *methods of sample size allocation*. In Sec.13.5, some aspects consideration involved in the construction of strata are discussed. Finally, the unit is concluded with a discussion on the principle of *post-stratification*.

Objectives

After reading this unit, you should be able to

- discuss the basics, principles and advantages of stratified sampling;
- estimate the population mean, population total, and proportion alongwith their efficiency in stratified random sampling;

- allocate the total sample size using various methods of allocation;
- construct strata using a simple procedure (*Dalenius-Hodges rule*);
- discuss the principle of post-stratification.

13.2 STRATIFIED SAMPLING

As said in the introduction of the unit, *stratification is the process of partitioning the entire population into small groups, called* **strata**, *each containing units homogenous with reference to study variables under consideration.* That is, homogeneity within a stratum is based on the characteristic under study.

Quite often, strata are available in natural forms. For example, in Agricultural Surveys, geographically contiguous units form a stratum under the assumption that nearby units are likely to be homogenous due to similar agro-climatic conditions as well as similar cultivation practices.

However, in other situations strata are formed on the basis of related variables. Listed below are some of the practical situations where the use of stratified sampling is a common practice.

The **stratified sampling**, with all its advantages of convenience, flexibility, efficiency with respect to sampling variance as well as cost, has become an essential component in all sample surveys of practical importance.

a) In crop estimation surveys, geographically contiguous areas such as tehsils/talukas or groups thereof are taken as strata.

b) In many socio-economic surveys, (within villages) small, medium, and large cultivators are taken as strata.

c) In National Sample Surveys, which are conducted continuosely with multi-subject surveys being conducted in successive rounds, strata are formed by grouping contiguous tehsils which are homogeneous with respect to population density, altitude above sea level and cultivation of food crops.

d) A wide variety of maps from Indian National Atlas showing population density, food crops etc., are used for stratifying the population.

e) In many other situations, the geographical and topographical considerations are taken into account for resorting to stratification.

We shall use the following definition for our discussion in this unit.

Definition. *The procedure of partitioning a given population into homogeneous groups, called* **strata**, *and then selecting samples independently from each stratum is known as* **stratified sampling**. *If a sample from each stratum is selected by random sampling, the procedure will be called* **stratified random sampling**.

The following is the list of some of the broad principles that one has to keep in mind while adopting stratified sampling.

1. For obvious reasons, the strata should be non-overlapping and should together comprise the whole population.
2. To minimise variance within a stratum, the units forming any stratum should be similar with respect to the study variable.
3. Sometimes administrative convenience may be considered as the basis for stratification. However, if such strata are not necessarily homogeneous, sub-stratification within each geographical stratum may be adopted based on homogeneity criteria using some ancillary information.

Try the following exercise now.

E1) With the help of *two* examples, explain the concept of stratified sampling clearly stating the principles adopted in the process.

13.2.1 Preliminaries

The essence of stratification is that it capitalizes on the known homogeneity of the *subpopulations*, so that only relatively small samples are required to estimate the characteristic for each subpopulation. These individual estimates are then combined to produce an estimate for the entire population.

Practical considerations like *cost, administrative convenience, simplicity of the methods*, etc., are kept in mind while stratifying a population.

To illustrate this point, let us consider the case of a city in which the northern districts are predominantly *low-income* areas and the southern districts are primarily *high-income* areas. So, to estimate the *average income* for the whole city, it is intuitively apparent that *relatively small simple random samples taken separately from the northern and southern districts* are likely to provide more accurate information than a single random sample taken from the entire city.

Also, it is clear from above that in situations when *variability within population is wide ranging* more precise inferences can be made using stratified sampling rather by using simple random sampling.

In general, the following four basic questions are important in the process of stratified sampling.

a) *How to form the strata*?

b) *How many strata to be formed*?

c) *How to select the samples in each stratum*?

d) *How to allocate the sample size to different strata*?

Of course, these fundamental questions are addressed with a purpose to minimise the sampling variance.

In the following illustration, we shall analyse a practical situation in order to understand the *importance* of above listed four questions in the context of stratified sampling.

Example 1: The case we are discussing is that of a *transport company*, which we shall refer to as A&B in our subsequent discussion. Common understanding is that if a shipment travels over several roads, the total freight charge is divided among all the transport companies sharing the responsibility of shipment. Also, it is acceptable that the computations involved in determining each transport company's revenue are cumbersome and expensive.

Hence, A&B decided to conduct a study to determine if the division of total revenue among several companies could be made accurately on the basis of a *sample survey* and at a substantial savings in clerical exercises. So, the purpose of such a study was to determine *how much of this total revenue belongs to A&B.*

A *waybill*, which is a document issued with every shipment of freight, gives details about the goods, route, and charges.

In one of the experiment during a *six-month* period, A&B studied the division of revenue for all shipments travelling over more than twenty districts. *The* **waybills**, *from which the amounts due each transport company can be computed, of these shipments constituted the* **population** *under examination.* The total number of waybills in the population, as well as the total freight revenue accounted for by the population of waybills, was known.

For the six-month period under study, there were nearly 23,000 waybills in the population. The amounts of the freight charges on these waybills vary greatly (some freight charges were as low as Rs.200 and others as high as Rs.2000) and, so, it was decided to follow the stratified sampling procedure.

Since the amount due the A&B on a waybill tended to be larger when the total amount on that waybill was large, the strata in this case were set up according to the amount of the total freight charge. Specifically, the strata formed were as given in Table 1 below.

Table 1. Formation of Strata.

Stratum	Waybills with freight charges (in Rs.)
1	0 to 200
2	201 to 400
3	401 to 700
4	701 to 1400
5	over 1401

Table 2. Proportion to be sampled.

Stratum	Proportion to be sampled (in %)
1	1
2	10
3	20
4	50
5	100

We assume that each stratum contains waybills with total freight charges of roughly the same order of magnitude.

The next problem before A&B was to decide *how large a sample from a stratum must be selected* so that the amount of the revenue due them could be estimated with a required amount of precision from as small a sample as possible. One piece of information needed for this task was the number of waybills in each stratum. The final *sample size allocation* decided on for the starta were as given in Table 2 above. As is clear from this table, more sample units are taken from the strata containing wider ranges of freight charges and smaller number of sample units are taken from the strata containing narrow ranges of freight charges.

You will read about the three methods of *sample size allocation* in Sec.13.4

To understand the method of sample size allocation adopted by A&B, consider Stratum-1. Here, the stratum contains waybills with charges less than Rs.200. As the variation between the waybills amounts is small, so, a small sample will provide adequate information about the amounts of all of the waybills in this stratum. On the other hand, Stratum-4, containing waybills with charges between Rs.701 and Rs.1400, has much greater variation. And, hence a larger sample is required from Stratum-4 to obtain adequate information about the amounts of all waybills.

Once the sample sizes allocation to different strata is determined, the next problem with A&B was to *select the samples from each stratum* i.e., construction of strata. At this stage, it is important to select the samples according to a procedure which facilitate the evaluation of the *sample statistics* as precisely as possible. That is, we should be able to judge how close the sample results are to the relevant population characterstics. Simple random sampling, you read in the previous unit, is one such procedure that can be applied to select samples from each stratum.

You will read about the method of *construction of strata* in Sec.13.5.

* * *

The A&B actually used a slightly different method of selecting waybills from each stratum, called *serial number sampling*. In this procedure, the sample from each stratum were selected according to certain *digits in the serial numbers* of the waybill. Since the serial numbers appear prominantly on the waybills, so, in comparison to other methods, this procedure for selecting the sample was found simple.

Try the following exercise.

E2) Why A&B choose to go ahead with a stratified sampling? Also, give reasons why each stratum is relatively homogenous with respect to the amount of freight charges due the A&B company.

More generally, as is clear from above discussion, there could be certain population related *constraints* because of which we are led to consider the stratified sampling. And then, in the next step, the population characterstics guide us in *strata formation process*. Finally, on the basis of our practice with some of the basic sampling methods, we adopt an appropirate *procedure* for selecting samples from each stratum.

Once through with these three stages, we next make sample statistics and check the consistency of results with population estimates. About *sample statistics* and *population estimates* we shall talk in the next section.

Let us wind up this section with a discussion on some of the *advantages* of the stratified sampling.

13.2.2 Advantages

Some of the advantages of stratified samplingse are briefly described in the following list.

1. It is clear that exclusion of a proportion of the population under study may lead to a wrong estimation of the population parameters. And, as we divide the population into various strata and then draw samples from each stratum so formed, there is very little possibility that a part of the population remains completely excluded. Hence, *stratification* ensures that a better cross section of the population is represented in a sampling design. Certainly, this is not the case with any *unstratified sampling* procedure.
2. It is desirable in some situations to use different sampling designs in different strata. *Stratification* allow us to do so, thereby, enabling effective utilization of the available auxiliary information. It is particularly true, when the extent and nature of the available information vary from stratum to stratum. A separate estimates obtained for different strata are combined to get a precise estimate for the whole population.
3. By using a proper strata formation procedure, the variability within strata can be considerably reduced. So, the stratification normally provides more efficient estimates than any unstratified sampling. For example, when there are several extreme values for the study variable in a population, they are grouped into a separate stratum thereby reducing the variability within strata.
4. In case of stratified sampling, the cost of conducting the survey is expected to be less. This is particularly true when strata are formed keeping administrative convenience in mind. This facilitate the supervision and organisation of field work involved in the sampling process.
5. 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 these areas.

Try the following exercise.

E3) Discuss the above stated four advantages of stratified sampling, giving one example in each case.

In this section, we have discussed certain preliminaries related to the process of stratified sampling. Also, we talked about certain advantages of stratified sampling that it has over other sampling methods. Let us now discuss the method of *estimation of population parameters* in the next section.

13.3 ESTIMATION OF POPULATION PARAMETERS

As said above, stratified sampling has got the flexibility that *any method of sampling* can be used independently within a stratum. However, in this unit, we consider only those situations wherein *simple random sampling without replacement* (**SRS-wor**) is used for selecting a sample from each stratum.

Let a (finite) population (under study) contain N units and suppose this population is divided into L number of strata (by any method), each containing units homogenous with respect to certain characterstics in question. Also, in the following table, the suffix h stands for the *hth* stratum ($h = 1, 2, ..., L$) and the suffix i will indicate the *ith* unit within a stratum. The *notations* defined in Table 3 are fixed for our convenience and future use.

Since samples in stratified sampling are drawn independently from each stratum, the estimates of strata *means, totals* and *proportions* are obtained on the basis of sampling

Table 3. Meaning of the notations used in this unit.

N_h (size of the *hth* stratum)	total number of units in the *hth* stratum;
n_h	number of units selected in the sample from the *hth* stratum;
$W_h = \frac{N_h}{N}$ (*hth* stratum weight)	proportion of the population units falling in the *hth* stratum;
$f_h = \frac{n_h}{N_h}$	sampling fraction for the *hth* stratum;
Y_{hi} (y_{hi})	the value of study variable for the *ith* unit (or sample unit) in the *hth* stratum, $1 \leq i \leq N_h$;
$Y_h = \sum_{i=1}^{N_h} Y_{hi}$	*hth* 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 of the study variable in the *hth* stratum;
$\overline{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} Y_{hi}$	*hth* stratum sample mean for the study variable;
$\sigma_h^2 = \frac{1}{N_h} \left((\sum_{i=1}^{N_h} Y_{hi}^2) - N_h \overline{Y}_h^2 \right)$	*hth* stratum variance based on N_h units;
$S_h^2 = \frac{N_h}{N_h - 1} \sigma_h^2$	*hth* 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 *hth* stratum;

schemes followed in each individual stratum. These estimates when pooled over all the strata give an overall estimate of the respective population parameters.

Try the following exercise to get familiar with notations defined in above table.

E4) Let a population of 100 units is divided into *four* strata of size $N_1 = 10$, $N_2 = 15, N_3 = 50, N_4 = 25$, and let the corresponding sample sizes allocation to these four strata be $n_1 = 3$, $n_2 = 4, n_3 = 15, n_4 = 8$, respectively. Also, let the value Y_{hi} of study variable for the *i-th* unit in the *hth* stratum ($1 \leq h \leq 4$) be given by $Y_{hi} = h, \forall$ i. Calculate the value of the terms $\overline{Y}_h, \overline{y}_h, \sigma_h^2, S_h^2$ and s_h^2.

Next, we shall discuss certain relations commonly used for calculating *sample statistics* and *population estimates*. Throughtout, notations will have meaning as explained in Table 3 above.

Firstly, recall that the *population mean* and the *population total* are given by relations

$$\overline{Y} = \sum_{h=1}^{L} W_h \overline{Y}_h, \quad \text{and} \quad Y = \sum_{h=1}^{L} Y_h, \text{ repectively.}$$

The following are some important relations which are used for an unbiased estimator of *population mean*, its *variance*, and the *estimated variance*.

Unbiased estimator of population mean: An estimator $\overline{y}_{st}$ for the population mean $\overline{Y}$ is given by

$$\overline{y}_{st} = \sum_{h=1}^{L} W_h \overline{y}_h.$$

Throughout, the *suffix* st refers to *stratification*.

Variance of estimator $\overline{y}_{st}$: For stratified random sampling, without replacement, it is known that the sample estimator $\overline{y}_{st}$ is unbiased and its variance is given by

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

$$= \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}$$

$$= \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h - 1}{N_h - 1}\right) \frac{\sigma_h^2}{n_h}.$$

Estimator of variance $v(\bar{y}_{st})$: With stratified random sampling (without replacement) an unbaised estimator of the variance of $\bar{y}_{st}$ is given by

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

Unbiased estimation of population total: Since the population total is

$$Y = N\bar{Y} = \sum_{h=1}^{L} N_h \bar{Y}_h,$$

so, the (unbaised) *estimator of population total* Y is given by

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

The *variance* of this estimator is given by

$$V(\hat{Y}_{st}) = \sum_{h=1}^{L} \left(1 - \frac{n_h - 1}{N_h - 1}\right) \frac{\sigma_h^2}{n_h}.$$

Unbiased estimator of proportion: As in case of simple random sampling (see Unit 12), estimation of proportion of units having an attribute is tackled by defining a variable given by

$$y_{hi} = \begin{cases} 1, & \text{if the } \textit{ith} \text{ unit in the } \textit{hth} \text{ stratum possesses the attribute} \\ 0, & \text{otherwise} \end{cases}$$

So, if N_h' denotes the number of units having the attribute in *hth* stratum and $P_h = \frac{N_h'}{N_h}$, then

$$N' = \sum_{h=1}^{L} N_h' \quad \text{and} \quad P = \frac{N'}{N} = \sum_{h=1}^{L} W_h P_h.$$

Also, it is clear that the population mean $\bar{Y}$ in this case is P i.e., it equals proportion of units having the attribute. An unbiased estimator of P is given by

$$p_{st} = \sum_{h=1}^{L} W_h p_h, \quad \text{where } p_h = \frac{n_h'}{n_h},$$

n_h' being the number of units having the attribute in the sample in the *hth* stratum. Expressions for variance and estimator of variance of p_{st} are given by

$$v(p_{st}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h - 1}\right) \left(\frac{P_h(1 - P_h)}{n_h}\right); \text{ and}$$

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

In order to understand the use of some of the above given relations, let us discuss a practical problem.

Problem 1. A sample for estimating the *number of orchards of apple* is conducted in some district of Himachal Pradesh. And, *Four strata* A, B, C and D of villages are formed according to the acreage of temperate fruit trees as per records avaiable with the revenue records. The *sizes of strata* (in acres) were 0-3, 3-6, 6-15, and 15 and above, respectively. A simple random sample of villages in each stratum was selected and the number of apple orchards was noted in selected villages. The collected data for various strata are as given below.

Stratum	Total number of villages	villages selected	No of orchards in the selected villages
A	275	15	2, 5, 1, 9, 6, 7, 0, 4, 7, 0, 5, 0, 0, 3, 0
B	146	10	21, 11, 7, 5, 6, 19, 5, 24, 30, 24
C	93	12	3, 10, 4, 11, 38, 11, 4, 46, 4, 18, 1, 19
D	62	11	30, 42, 20, 38, 29, 22, 31, 28, 66, 41, 15

Estimate the number of orchards in the district. Also estimate the variance of estimated number of orchards.

Solution. For our use in the subsequent computations, we prepare the following table of values for different quantities required in the final calculations.

Stratum-I	Stratum-II	Stratum-III	Stratum-IV
$n_1 = 15$	$n_2 = 10$	$n_3 = 12$	$n_4 = 11$
$N_1 = 275$	$N_2 = 146$	$N_3 = 93$	$N_4 = 62$
$W_1 = 0.4774$	$W_2 = 0.2535$	$W_3 = 0.1615$	$W_4 = 0.1076$
$\bar{y}_1 = 3.27$	$\bar{y}_2 = 15.2$	$\bar{y}_3 = 14.08$	$\bar{y}_4 = 32.91$
$s_1^2 = 9.6381$	$s_2^2 = 88.84$	$s_3^2 = 205.91$	$s_4^2 = 192.69$

Using values from above table, we have

$$\begin{aligned}\bar{y}_{st} &= W_1\bar{y}_1 + W_2\bar{y}_2 + W_3\bar{y}_3 + W_4\bar{y}_4 \\ &= \frac{1}{N}(N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3 + N_4\bar{y}_4) \\ &= \frac{1}{576}[275(3.27) + 146(15.2) + 93(14.08) + 62(32.91)] = 11.23.\end{aligned}$$

And, the estimate of variance is

$$\begin{aligned}V(\bar{y}_{st}) &= \sum_{h=1}^{4} \frac{W_h^2(N_h - n_h)s_h^2}{N_h n_h} \\ &= \frac{(.4774)^2(275-15)}{275\times 15}\times 9.6381 + \frac{(0.2535)^2(146-10)}{146\times 10}\times 88.84 \\ &\quad + \frac{(.115)^2(93-12)}{93\times 12}\times 205.91 + \frac{(.1076)^2(62-11)}{62\times 11}\times 192.69 \\ &= 0.134 + 0.5321 + 0.3901 + 0.1671 = 1.2277.\end{aligned}$$

In this case, since total number of orchards in the district are to be estimated,

$$\hat{Y}_{st} = N\bar{y}_{st} = 576 \times 11.23 = 6468.48,$$

and an estimate of variance is

$$V(\hat{Y}_{st}) = (576)^2 \times 1.2277 = 407321.39.$$

———— × ————

Try the following exercise now.

E5) Verify the values given in the last two rows of the table given in the solution of Problem 1.

As stated before, in stratified sampling, we need to address the problems of *construction of strata*, deciding the *proportion of sample* for each stratum, and calculating *sample statistics* and *population estimates*.

We start discussion with the methods of sample size allocations.

13.4 ALLOCATION OF SAMPLE SIZE

In a practical situation, the (total) sample size is normally decided by a single consideration viz., the budget available for a survey. However, the allocation of sample

size to different strata is made by a statistician. Here, it is important to remember that the precision of estimators largely depends on the *allocation plans*.

In fact, in order to increase the efficiency of estimators, it is imperative to choose a proper allocation plan. In this process, (1) the *strata sizes* i.e., the values of N_h $(1 \leq h \leq L)$, (2) *variability within a stratum*, and (3) the *cost of observing a sampling unit* within various strata are three considerations which can affect the choice of allocation.

(1) *Equal allocation*, (2) *Proportional allocation*, and (3) *Optimum allocation* are the three methods of sample size allocation that are commonly used in practice. Let us discuss them one by one. In what follows, n stands for the total sample size i.e., total number of units in a stratified sample.

Equal allocation: Here, *the number of sampling units selected from each stratum is equal.* Thus, in this case,

$$n_h = \frac{n}{L}, \text{ for } h = 1, 2, ...L.$$

This method is preferred when strata sizes do not differ too much from each other and the information about the variability within the strata is not available. Sometimes equal allocation is also used for equal allocation of work to different strata. One of the advantage of using this method of allocation is that it is convenient for adminstration and field work. However, from the efficiency point of view, this is not a desirable method of sample size allocation.

Proportional allocation: This method was proposed by **Bowley** (1926) and has its motivation in the argument that *samples are distributed to different strata in proportion to strata sizes*. That is, larger strata should get a larger share of allocation while the smaller strata are allocated smaller number of units. Hence, here the sample allocation to the *hth-stratum* is made by

$$n_h = \frac{n\, N_h}{N}, \text{ where notations are used from Table 3.}$$

This method is simple to use and numerous estimates can be made with greater degree of precision by this method. However, it does not take into account an important aspect associated with stratified sampling, namely, the variability within strata.

Optimum allocation: This allocation method is given by **Neyman** (1934). Here, the basic idea is that, for population with larger variability, sample sizes have to be large. That is, we should take larger allocation sample sizes for strata with higher variability. Also, as we want that larger strata should have a higher allocation, so, to improve the precision of estimates (i.e., to reduce the variance), an important criterion of allocating the sample sizes should be to minimise the variance $v(\bar{y}_{st})$ of stratified sample mean estimator for a fixed total sample size, n.

The minimisation of variance $v(\bar{y}_{st})$, subject to the constraint $\sum_{h=1}^{L} n_h = n$, leads to the allocation

$$n_h = n \frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h}$$

$$= n \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h}.$$

Clearly, here we have taken into account both the strata sizes as well as strata variability.

However, in Neyman allocation as described above, it is assumed that the sampling cost per unit among different strata is same and the size of the sample is fixed.

Alternatively, if we want to minimise the cost for a specified value of the variance of stratified sample mean $\bar{y}_{st}$, then the simplest *cost function*, we referred to above, is

given by

$$C = C_0 + \sum_{h=1}^{L} C_h n_h,$$

where C stands for the overall budget, C_0 for the (fixed) overhead cost, and C_h is the average cost of observing the study variable for each unit selected in the sample from the *hth stratum*. Then, an **optimum allocation** is given by that value of n_h for which C is minimum. And, using standard techniques from calculus, one can see that such a value of n_h is given by the relation

$$n_h = \frac{\frac{(C - C_0) W_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{C_h}}}$$

The corresponding relations for the variance, in above discussed three types of allocations methods, are given by

$$V(\overline{Y}_{st})_{eq} = L \sum_{h=1}^{L} W_h^2 \left(\frac{1}{n} - \frac{1}{N_h}\right) S_h^2,$$

$$V(\overline{y}_{st})_{prop} = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{L} W_h S_h^2,$$

$$V(\overline{y}_{st})_{opt} = \frac{1}{n} \left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2, \text{ respectively.}$$

A comparative study between unstratified, proportional and optimum allocations shows that if we ignore *finite population correction* then

$$V(\overline{y}_{st}) \geq V(\overline{y}_{st})_{prop} \geq V(\overline{y}_{st})_{opt}.$$

Thus, in terms of efficiency, optimum allocation is better than proportional allocation, which, in turn, is better than unstratified simple random sampling.

In the following problem, a practical situation is discussed to illustrate the above explained methods of sample size allocation.

Problem 2. Suppose three small towns are under study, having population $N_1 = 50000, N_2 = 30000$ and $N_3 = 40000$, respectively. A stratified random sample is to be taken with a total sample size of $n = 500$. Determine the sample size to be taken from each town individually using the method of (a) proportional, and (b) optimal allocation. It is (roughly) known from a previous survey that $S_1 = 30, S_2 = 15$ and $S_3 = 20$.(*Notations here have same meaning as given in Table 3.*)

Solution. (a) Under proportional allocation:

$$n_1 = n\left(\frac{N_1}{N}\right) = 500 \times \frac{5}{12} = 208;$$

$$n_2 = n\left(\frac{N_2}{N}\right) = 500 \times \frac{3}{12} = 125;$$

$$n_3 = n\left(\frac{N_3}{N}\right) = 500 \times \frac{4}{12} = 167.$$

(b) Under optimal allocation:

$$n_1 = n\left(\frac{W_1 S_1}{\sum_{h=1}^{3} W_h S_h}\right) = 500 \times \frac{150}{23 \times 12} = 272;$$

$$n_2 = n\left(\frac{W_2S_2}{\sum_{h=1}^{3} W_hS_h}\right) = 500 \times \frac{45}{23 \times 12} = 82;$$

$$n_3 = n\left(\frac{W_3S_3}{\sum_{h=1}^{3} W_hS_h}\right) = 500 \times \frac{80}{23 \times 12} = 145.$$

Now, try the following exercise for your practice.

E6) In Problem 1, let the mean square errors in different strata be $S_1^2 = 8, S_2^2 = 2$, $S_3^2 = 15$ and $S_4^2 = 20$. Obtain the allocations using equal, proportional and optimum allocations. Also, work out the variances of the estimated mean corresponding to these allocations.

In this section, we discussed how to allocate the sample sizes to various strata. So, you know how many units per stratum should be selected so that the survey is cost effective and the variance of the estimate is minimum. Recall, the variability within a stratum can be minimised provided we could manage to take extra care while *constructing strata.*

Let us discuss this aspect of stratified sampling in the next section.

13.5 CONSTRUCTION OF STRATA

In the process of *construction of strata*, the basic consideration involved is that *the strata should be internally homogeneous*. For the construction of strata in which the distribution of a (single) study variable y is available, L strata could then be formed by cutting this distribution at $(L - 1)$ suitable points.

However, in practice, the distribution of y is not available always. So, in the absence of this information, the next best alternative for the formation of strata is to look at the frequency distribution of some other variable which is highly correlated with the relevant study variable y.

Here, we have to remember that the construction of strata on such an auxiliary variable will not yield exactly optimum strata, but atleast it can provide a good approximation.

Working on these lines, **Dalenius** and **Hodges** (1957) gave a procedure called **cumulative square root rule**. This rule is used on the frequency distribution of a highly positively correlated auxiliary variable x (also called **stratification variable**). This rule uses the argument that the distribution of y within strata can be assumed to be rectangular if the number of strata is large.

Cumulative square root rule, though proposed for the optimum allocation method, is found to yield approximately optimum strata for equal and proportional allocation methods as well. This rule can, therefore, be used for the construction of strata for all allocation methods.

The various steps involved in the construction of strata for this method are listed as under:

1. Obtain a frequency table with K classes for the *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, use linear interpolation method on the class intervals and cumulative square root frequency column to obtain the value of $x = x_1$, which corresponds to the value $\frac{T}{L}$ in the cumulative square root frequency column.
5. Repeat the process in above step to obtain $x = x_i$ corresponding to the value $\frac{iT}{L}, i = 2, 3, ..., L-1$, 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 $(\geq x_{L-1})$.

Let us try to understand the steps involved in this rule with the help of a particular situation.

Problem 3. 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. Optimum 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 the table on the next page.

Lectation length	No. of cows (f)	$\sqrt{f}$	Cummulative $\sqrt{f}$
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. Here, we are given the frequency for the stratification variable, namely, *the first lactation length.* In the next step, we obtain the square roots of the frequencies (f) as given in the second column of above table. The third column of the gives the square root values $(\sqrt{f})$ and the cumulative totals of $\sqrt{f}$ constitute the fourth column of this table.

Now, here 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. By above stated Step-4 and Step-5 of the rule, x_1 and x_2 correspond to the values

$$\frac{T}{3} = \frac{35.97}{3} = 11.99 \quad \text{and} \quad \frac{2T}{3} = \frac{2(35.97)}{3} = 23.98, \text{respectively,}$$

in the fourth column of the table. You can see that value 9.01 in the fourth column correspond to value 190 in the first column, whereas, the value 13.48 in the fourth column corresponds to the value 230 in the first column of the table. Thus, an increase of 4.47 in cumulative $\sqrt{f}$ value takes place over the interval 190 – 230. The first lactation length x_1 corresponding to the cumulative value of 11.99, therefore, lies in the interval 190 – 230.

Hence, by the method of *linear intrapolation,*

$$11.99 = \frac{230 \times 9.01 - 190 \times 13.48}{230 - 190} + \frac{13.48 - 9.01}{230 - 190} x_1$$

Recall that, if $f(a_1) = b_1$ and $f(a_2) = b_2$, then the value $f(x)$ of the function f at any point x on the line joining points (a_1, b_1) and (a_2, b_2), by the **method of linear intrapolation**, is given by the relation $f(x) = \frac{a_2b_1 - a_1b_2}{a_2 - a_1} + \frac{(b_2 - b_1)}{a_2 - a_1} x$.

$\Rightarrow x_1 = 216.67.$

Similarly, it can be seen that $x_2 = 313.21$. This shows that the cows with the first lactation length in the range $[30, 216.67]$ will constitute the first stratum, whereas those having lactation length in the ranges $[216.67, 313.21]$ and $[313.21, 510]$ will form second and third strata, respectively.

——— × ———

Now, you try the following exercise.

E7) It is proposed to estimate total wool yield in a certain region of Rajasthan, using stratified simple random sampling. An overall sample of 20 villages is to be selected employing Optimum method of sample allocation. The stationary sheep population data for 141 villages of this region is as in the frequency table given below.

No. of Sheep	No. of Villages (f)
0 – 100	46
100 – 200	36
200 – 300	23
300 – 400	11
400 – 500	6
500 – 600	4
600 – 700	4
700 – 800	1
800 – 900	4
900 – 1000	4
1000 – 1100	1
1100 – 1200	1

Construct three approximately optimum strata taking stationary sheep population as the stratification variable.

Next, we discuss the quantitative part of the strata formation. That is, about the method by which we decide the *number* of strata to be taken in a stratified sampling. Certainly, in any stratifying sampling, the *minimum number of strata* is **two** and *maximum number of strata* could be, say n, with one unit selected from each stratum.

However, to obtain an estimate of variance, one should have at least two units per stratum. Thus, the maximum number of strata may be (n/2). As expected, increasing the number of strata adds towards increase in efficiency but the fact is that this *gain goes on decreasing with increase in number of strata.* Also, in situations where estimation of variance is not possible due to increase in the number of strata collapsing of strata is done.

So, as a general rule, number of strata beyond 6 to 8 is seldom profitable. It is possible to use more general and complex methods of sampling in each stratum separately. And, estimation of population parameters can be done accordingly.

Finally, in the concluding part of the unit, we want to discuss with you the concept of *post-startification.*

13.6 POST-STRATIFICATION

Recall that, in stratified sampling, we *presuppose* that the strata sizes and the sampling frame for each stratum are available. However, there do exist situations when it is difficult to obtain these two informations.

For instance, this is a case when details about classification of farmers' population by farm size (i.e., small, medium, large) is required. Though the size of the population here can be obtained from the census records, but the *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 corresponding farm size.

This technique is typically useful when published journals/reports may provide clear indication of strata sizes and, due to non-availability of strata frames, it is difficult to sample the units from different strata.

This simply means that we can assign the units to different strata once the sample units are contacted and observed. The underlying procedure is called **post-stratification**.

The following are some useful formulations required for calculating the estimator of population mean, approximate variance, and estimator of variance when the sample units are stratified after they have been selected as a *single without replacement simple random sampling* from the entire unstratified population.

Estimator of population mean $\bar{Y}$: $\bar{y}_{ps} = \sum_{h=1}^{L} W_h \bar{y}_h$.

Approximate variance of $\bar{y}_{ps}$:

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

The suffix **ps** here refers to *post-stratification.*

Note that the first term in the last equation is the value of $V(\bar{y}_{st})_{prop}$. In fact, in a random allocation, units get distributed to different strata in proportion to strata sizes. Thus, the first term belonging to proportional allocation is as expected. And, the second term is due to deviations in proportionality and, so, is a contribution due to *post-stratification*. For large sample sizes, second term is likely to be small and post-stratification is nearly as good as proportional allocation.

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

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

With this we end our discussion on stratified sampling. Let us summarize what we have discussed in this unit.

13.7 SUMMARY

In this unit, we have discussed the following points.

1. Basic principles of stratified sampling with the help of various type of illustrative examples. A case study of a *transport company* is analysed to facilitate the understanding of the following four questions in the context of stratified sampling.
   a) How to form strata?
   b) How many strata to be formed?
   c) How to select a sampling technique for various strata?
   d) How to allocate the sample size to different strata?

   Also, some advantages of stratified sampling method, that it has over other sampling methods have been breifly discussed.
2. With the assumption that SRS-wor is being used for sampling within a stratum, the use of the following relations has been illustrated.

   (unbaised estimator of population mean) $\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h$;

   (Variance of estimator $\bar{y}_{st}$) $v(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h} \right) S_h^2$;

   (Estimator of variance $v(\bar{y}_{st})$)$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h} \right) s_h^2$;

(Unbiased estimation of population total Y) $\hat{Y}_{st} = \sum_{h=1}^{L} N_h \bar{y}_h$;

(variance of $\hat{Y}_{st}$) $V(\hat{Y}_{st}) = \sum_{h=1}^{L} \left(1 - \frac{n_h - 1}{N_h - 1}\right) \frac{\sigma_h^2}{n_h}$,

where the notations have their meaning as defined in Table 3.

3. The following three method of sample size allocation are explained.
   a) Equal allocation $\left(n_h = \frac{n}{L}\right)$;
   b) Proportional allocation $\left(n_h = \frac{nN_h}{N}\right)$;
   c) optimum allocation $\left(n_h = n\frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h}\right)$;

   Also, the optimum allocation

$$n_h = \frac{\frac{(C-C_0)W_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{C_h}}}.$$

   is discussed with reference to the cost function

$$C = C_0 + \sum_{h=1}^{L} C - hn_h.$$

4. Construction of a strata, using cummulative square root rule, is discussed. The method is illustrated with help of examples.
5. The concept of post stratification is discussed briefly.

13.8 ANSWERS/SOLUTIONS

E1) Do it yourself.

E2) From the discussion held before this exercise.

E3) Use some examples given in this unit.

E4) Substitute values in relevant formula given in Table 3.

E5) Use relevent formula.

E6) Here $S_1^2 = 8, S_2^2 = 12, S_3^2 = 15$ and $S_4^2 = 20$. According to data given in Problem 1, we have

$N_1 = 275, N_2 = 146, N_3 = 93, N_4 = 62$, and total sample size $n = 48$.

(a) **Equal allocation** : $n_h = \frac{n}{4} = 12$, for all h, $1 \leq h \leq 4$. So, the variance

$$\begin{aligned} V(\bar{y}_{st})_{eq} &= 4\sum_{h=1}^{4} W_h^2 \left(\frac{1}{48} - \frac{1}{N_h}\right) S_h^2 \\ &= 1.0034 + 0.5176 + 0.2366 + 0.0871 \\ &= 1.845. \end{aligned}$$

(b) **Propoprtional allocation** : Here $n_1 = 23$, $n_2 = 12$, $n_3 = 8$ and $n_4 = 5$. And so, the variance in this case

$$V(\bar{y}_{st})_{prop} = \left(\frac{1}{48} - \frac{1}{576}\right) \sum_{h=1}^{4} W_h S_h^2$$

$$= 0.0191[0.4774 \times 64 + 0.2535 \times 144 + 0.1615 \times 225 + 0.1076 \times 400]$$

$$= 2.797.$$

(c) **Optimum allocation** : Here, we first compute

$$\sum_{h=1}^{4} W_h S_h = [3.8192 + 3.042 + 2.4225 + 2.152]$$

$$= 11.4357.$$

Then,

$$n_1 = 48 \times \frac{3.8192}{11.4357} = 16;$$

$$n_2 = 48 \times \frac{3.042}{11.4357} = 13;$$

$$n_3 = 48 \times \frac{2.4225}{11.4357} = 10;$$

$$n_4 = 48 \times \frac{2.152}{11.4357} = 9.$$

And, the variance in this case is

$$V(\bar{y}_{st})_{opt} = \frac{1}{48}\left(\sum_{h=1}^{4} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{4} W_h S_h^2.$$

$$= 2.7245 - 0.2542$$

$$= 2.4703.$$

E7) Observe that in given table we already have the frequency for the stratification variable, namely, ***stationary sheep population.*** For rest of the calculations follow the steps as in Problem 3.