UNIT 3 STRATIFIED RANDOM SAMPLING

Structure

3.1 INTRODUCTION

When the units of the population are scattered and not completely homogeneous in nature, then simple random sample does not give proper representation of the population. So if the population is heterogeneous the simple random sampling is not found suitable. In simple random sampling the variance of the sample mean is proportional to the variability of the sampling units in the population. So, in spite of increasing the sample size n or sampling fraction n/N, the only other way of increasing the precision is to device a sampling which will effectively reduce the variability of the sample units, the population heterogeneity. One such method is stratified sampling method.

In stratified sampling the whole population is to be divided in some homogeneous groups or classes with respect to the characteristic under study which are known as strata. That means, we have to do the stratification of the population. Stratification means division into layers. The auxiliary information related to the character under study may be used to divide the population into various groups or strata in such a way that units within each stratum are as homogeneous as possible and the strata are as widely different as possible.

Thus, all strata would comprise the population. Then from each stratum sample would be drawn and lastly all samples would be combined to get the ultimate sample. For example, let us consider that population consists of N units and these are distributed in a heterogeneous structure. Now first of all we divide the population into 'k' non overlapping strata of sizes N_1, N_2, N_3, ..., N_k such that each stratum becomes homogeneous. Evidently $N = N_1 + N_2 + N_3 + ... + N_k$. Then from first stratum a sample of size n_1 would be drawn by simple random sampling method. Similarly, from the second stratum a sample of n_2 units would be drawn and so on, up to k^{th} stratum. Now all these k samples would be combined to get the ultimate sample. So, the ultimate size of sample would be $n = n_1 + n_2 + n_3 + ... + n_k$. This method of sampling is known as Stratified random sampling because here stratification is done first to make population homogeneous and then samples are drawn randomly by simple random sampling from each stratum.

The principles of stratification are explained in Section 3.2. The properties of stratified random sampling are described in Section 3.3, whereas Section 3.4 provides the derivation of the mean and variance of proportions in stratified random sampling. The allocation of sample size with the help of different techniques is described in Section 3.5. The comparative study between stratified random sampling and simple random sampling is given in Section 3.6.

Objectives

After studying this unit, you would be able to

- define the stratified random sampling;
- explain the principles of stratification;
- describe the properties of stratified random sampling;
- derive the mean and variance of proportions in stratified random sampling;
- describe the allocation of sample size with the help of different techniques; and
- calculate the estimate of population mean and variance of sample mean.

3.2 PRINCIPLES OF STRATIFICATION

The principles to be kept in mind while stratifying a population are given below:

1. The strata should not be overlapping and should together comprise the whole population.

2. The strata should be homogeneous within themselves and heterogeneous between themselves with respect to characteristic under study.

3. If a investigator is facing difficulties in stratifying a population with respect to the characteristic under study, then he/she has to consider the administrative convenience as the basis for stratification.

4. If the limit of precision is given for certain sub-population then it should be treated as stratum.

3.2.1 Notations and Terminology

N = Population size

n = Sample size

k = Number of strata

N_i = Size of i^{th} stratum

Then $N = \sum_{i=1}^{k} N_i$

n_i = Size of sample drawn from i^{th} stratum

Then $n = \sum_{i=1}^{k} n_i$

X_{ij} = Value of character under study for j^{th} unit of i^{th} stratum

$\overline{X}_i$ = Population mean of i^{th} stratum $= \dfrac{1}{N_i}\sum_{j=1}^{N_i} X_{ij}$

$\overline{X}$ = Population Mean $= \dfrac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i} X_{ij}$

$$= \frac{1}{N}\sum_{i=1}^{k} N_i\,\overline{X}_i = \sum_{i=1}^{k} W_i\,\overline{X}_i$$

where, $W_i = \dfrac{N_i}{N}$ is called the weight of i^{th} stratum

S_i^2 = Population mean square of the i^{th} stratum

$$= \frac{1}{N_i - 1}\sum_{j=1}^{N_i}\left(X_{ij} - \overline{X}_i\right)^2, \ \left(j = 1, 2, \ldots, N_i \ \& \ i = 1, 2, \ldots, k\right)$$

x_{ij} = Value of j^{th} sample unit taken from i^{th} stratum

$\bar{x}_i = \dfrac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$ = Mean of sample units selected from i^{th} stratum

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_i\right)^2 \ , \quad \left(i = 1, 2, \ldots, k\right)$$

= Sample mean square of the i^{th} stratum

Let us consider the following sample means to estimate the populations mean of i^{th} stratum $\overline{X}$ which are:

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{k} n_i\,\bar{x}_i$$

and $$\bar{x}_{st} = \frac{1}{N}\sum_{i=1}^{k} N_i\,\bar{x}_i = \sum_{i=1}^{k} W_i\,\bar{x}_i$$

where, $\bar{x}_{st}$ is the weighted mean of the strata sample means, weights being equal to strata sizes. These two will be identical if $n_i \propto N_i$

3.3 PROPERTIES OF STRATIFIED RANDOM SAMPLING

Theorem 1: $\bar{x}_{st}$ is an unbiased estimate of the population mean $\bar{X}$ i.e.

$$E(\bar{x}_{st}) = \bar{X}$$

Proof: We have

$$\bar{x}_{st} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{x}_i$$

Therefore,

$$E(\bar{x}_{st}) = E\left[\frac{1}{N}\sum_{i=1}^{k} N_i \bar{x}_i\right]$$

$$\frac{1}{N}\sum_{i=1}^{k} N_i E(\bar{x}_i)$$

Since the sample units selected from each of stratum are simple random sample, then we have

$$E(\bar{x}_i) = \bar{X}_i$$

Therefore,

$$E(\bar{x}_{st}) = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{X}_i$$

$$= \frac{1}{N}\sum_{i=1}^{k} N_i \frac{1}{N_i}\sum_{j=1}^{N_i} X_{ij}$$

$$= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i} X_{ij} = \bar{X}$$

Hence proved

Theorem 2: Prove that

$$\text{Var}(\bar{x}_{st}) = \frac{1}{N^2}\sum_{i=1}^{k} N_i (N_i - n_i)\frac{S_i}{n_i} = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 S_i^2$$

Proof: We have

$$\text{Var}(\bar{x}_{st}) = \text{Var}\left(\sum_{i=1}^{k} W_i \bar{x}_i\right)$$

$$= \sum_{i=1}^{k} W_i^2 \text{Var}(\bar{x}_i)$$

The covariance term vanish since the samples from different strata are independent and the sample units in each stratum are the simple random sample without replacement, we have

$$\operatorname{Var}(\bar{x}) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$$

or $$\operatorname{Var}(\bar{x}_i) = \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_i^{\,2}$$

Therefore,

$$\operatorname{Var}(\bar{x}_{st}) = \sum_{i=1}^{k} W_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_i^2$$

$$= \frac{1}{N^2} \sum_{i=1}^{k} N_i \; (N_i - n_i) \frac{S_i^2}{n_i}$$

From the above result the variance depends on $S_i^{\,2}$ the heterogeneity within the strata. Thus, if S_i^2 are small i.e. strata are homogeneous then stratified sampling schemes provides estimates with greater precision.

Theorem 3: If $S_i^{\,2}$ is not known then prove that estimate of the variance of the sample mean of the stratified random sample is given by

$$E\left[\operatorname{Var}(\bar{x}_{st})\right] = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 S_i^2$$

Proof: In general $S_i^{\,2}$ are not known. A simple random sample is drawn from each stratum. If we assume a individual stratum as a population then the sample, drawn from it, would be a simple random sample. If the sample is drawn from i^{th} stratum, the sample mean square $s_i^{\,2}$ would be an estimate of population mean square $S_i^{\,2}$

$$\text{i.e.} \quad E\left(s_i^2\right) = S_i^2 \qquad i = 1, 2, ..., k \qquad \ldots (1)$$

Accordingly an unbiased estimate of the variance is given by

$$\operatorname{Var}(\bar{x}_{st}) = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 \, s_i^2$$

Therefore,

$$E\left[\operatorname{Var}(\bar{x}_{st})\right] = E\left[\sum_{i=1}^{n} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 \, s_i^2\right]$$

$$= \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 \, E\left(s_i^2\right)$$

Substituting from equation (1), we get

$$E\left[\operatorname{Var}(\bar{x}_{st})\right] = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 \, S_i^2$$

3.4 MEAN AND VARIANCE FOR PROPORTIONS

As in simple random sampling, we can divide a population into two classes with respect to a attribute. Hence the units in the population are classified in these two classes accordingly as it possesses or does not possess the given attribute. After taking a sample of size n, we may be interested in estimating the population proportion of the defined attribute.

If a unit possesses the attribute, it receives the code value 1 and if an unit does not possesses the attribute, it receives the value 0. Let the number of units belonging to A in the i^{th} stratum of size N_i be M_i and if the sample of size n_i taken from i^{th} stratum, the number of units belonging to A be m_i. Denoting the proportion of units belonging to A in the population, in the i^{th} stratum and sample from the i^{th} stratum by π, π_i and p_i respectively, various formula for mean and variance are as follows:

$$\pi_i = \frac{M_i}{N_i} \text{ and } p_i = \frac{m_i}{n_i}$$

and

$$\pi = \sum_{i=1}^{k} \frac{N_i}{N} \pi_i = \sum_{i=1}^{k} W_i \pi_i \qquad \text{for } i = 1, 2, \ldots, k$$

The estimated proportion p_{st} under stratified sampling for the units belonging to A is

$$p_{st} = \sum_{i=1}^{k} W_i\, p_i$$

$$\text{Mean}(p_{st}) = E\left(\sum_{i=1}^{k} W_i\, p_i\right) = \sum_{i=1}^{k} W_i E(p_i)$$

since we draw SRS from each stratum so by Theorem 10 of Section 2.4 of Unit 2 we have

$$E(p_i) = \pi_i$$

By putting this value in above formula, we have

$$\text{Mean}(p_{st}) = \sum_{i=1}^{k} W_i \pi_i = \pi$$

Hence sample proportion under stratified sampling is unbiased estimate for population proportion.

Now variance of p_{st} is given by

$$\text{Var}(p_{st}) = \text{Var}\left(\sum_{i=1}^{k} W_i\, p_i\right) = \frac{1}{N^2} \sum_{i=1}^{k} N_i^2\, V(p_i)$$

since we draw SRS from each stratum so by Theorem 11 of Section 2.4 of Unit 2 we have

$$\text{Var}(p_i) = \frac{N_i - n_i}{N_i - 1} \cdot \frac{\pi_i(1-\pi_i)}{n_i}$$

By putting this value in above formula, we have

$$\text{Var}(p_{st}) = \frac{1}{N^2} \sum_{i=1}^{k} \frac{N_i^{\,2}(N_i - n_i)\pi_i(1-\pi_i)}{(N_i - 1)n_i}$$

If N_i is large enough, consider $1/N_i$ as negligible and $N_i - 1 \sim N_i$, formula for Var (p_{st}) reduces to

$$\text{Var}(p_{st}) = \frac{1}{N^2} \sum_{i=1}^{k} N_i(N_i - n_i)\frac{\pi_i(1-\pi_i)}{n_i}$$

and if n_i / N_i is negligible, therefore

$$\text{Var}(p_{st}) = \sum_{i=1}^{k} W_i^2 \frac{\pi_i(1-\pi_i)}{n_i}$$

An unbiased estimate of Var (p_{st}) is given by

$$E\ [\text{Var}(p_{st})] = \frac{1}{N^2} \sum_{i=1}^{k} N_i \frac{N_i - n_i}{(n_i - 1)} p_i\, q_i$$

where, $q_i = 1 - p_i$

3.5 ALLOCATION OF SAMPLE SIZE

In stratified sampling, the allocation of the sample to different strata is done by considering the following factors:

1. The total number of units in the stratum, i.e. stratum size;
2. The variability within the stratum; and
3. The cost in taking observations per sampling unit in the stratum.

A good allocation is one where maximum precision is obtained with minimum resources. In other words, the criterion for allocation is to minimize the cost for a given variation or minimize the variance for a fixed cost, thus making the most effective use of the available resources.

Types of Allocation of Sample Size

It is evident from the formula for variance of $\bar{x}_{st}$ that it depends on n_i the number of units selected at random from i^{th} stratum. Hence, the problem arises, what optimum value of $n_i\ (i = 1, 2, ..., k)$ can be chosen out of n, so that, the variance is as small as possible. Four types of allocations are considered here:

3.5.1 Equal Number of Units from Each Stratum

This is a situation of considerable practical interest for reasons of administrative convenience. In this allocation method, the total sample size n is divided equally among all the strata i.e. if the population is divided in k strata then the size of sample for each stratum would be

$$n_i = \frac{n}{k} \qquad \text{for all } i = 1, 2, \ldots, k$$

3.5.2 Proportional Allocation

This allocation was originally proposed by Bowley in 1926. This procedure of allocation is very common in practice because of its simplicity. As its name indicates, proportional allocation means that we select a small sample from a small stratum and a large sample from a large stratum. The sample size in each stratum is fixed in such a way that for all the strata, the ratio n_i/N_i is equal to n/N i.e.

$$\frac{n_i}{N_i} = \frac{n}{N}$$

$$\text{or} \quad n_i = \frac{n}{N} N_i$$

$$n_i = n W_i$$

$$\text{or} \quad n_i \propto N_i \qquad \ldots (2)$$

In other words, the allocation of a sample of size n to different strata is to be done in proportional to their sizes. We have variance of $\bar{x}_{st}$

$$\text{Var}(\bar{x}_{st}) = \sum_{i=1}^{k} \left(1 - \frac{n_i}{N_i}\right) \frac{W_i^2}{n_i} S_i^2$$

Thus, substituting the values of n_i and n_i/N_i as $n W_i$ and n/N respectively in variance formula then we get variance under proportional allocation

$$\text{Var}(\bar{x}_{st})_{PROP} = \left(1 - \frac{n}{N}\right) \sum_{i=1}^{k} \frac{W_i S_i^2}{n}$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} W_i S_i^2 \qquad \ldots (3)$$

In case, the sampling fraction (n/ N) is negligible

$$\text{Var}(\bar{x}_{st})_{PROP} = \sum_{i=1}^{k} \frac{W_i S_i^2}{n} \qquad \ldots (4)$$

3.5.3 Neyman's Allocation

This allocation of the total sample size n to the different stratum is called minimum variance allocation and is due to Neyman (1934). This result was first discovered by Tchuprow (1923) but remained unknown until it was rediscovered independently by Neyman. This allocation of samples among different strata is based on a joint consideration of the stratum size and the stratum variance. In this allocation, it is assumed that the sampling cost per unit among different strata is same and the size of the sample is fixed. Sample sizes are allocated by

$$n_i = n \frac{W_i S_i}{\sum_{i=1}^{k} W_i S_i} = n \frac{N_i S_i}{\sum_{i=1}^{k} N_i S_i} \quad \ldots (5)$$

A formula for the minimum variance with fixed n is obtained by substituting the value of n_i in variance formula, then we get

$$\operatorname{Var}(\bar{x}_{st})_{NEY} = \frac{\left(\sum_{i=1}^{k} W_i S_i\right)^2}{n} - \frac{\sum_{i=1}^{k} W_i S_i^2}{N} \quad \ldots (6)$$

3.5.4 Optimum Allocation

The variance of estimated mean depends on n_i which can arbitrarily be fixed. One more factor, which is none the less important, also influences the variance of estimated mean. The allocation problem is two fold:

1. We attain maximum precision for the fixed cost of the survey; and
2. We attain the desired degree of precision for the minimum cost.

Thus, the allocation of the sample size in various strata, in accordance with these two objectives, is known as optimum allocation.

In any stratum the cost of survey per sampling unit cannot be the same. That is, in one stratum the cost of transportation may be different from the other. Hence, it would not be wrong to allocate the cost of the survey in each stratum differently.

Let c_i be the cost per unit of survey in the i^{th} stratum from which a sample of size n_i is stipulated. Also suppose c_0 as the over head fixed cost of the survey. In this way the total cost C of the survey comes out to be

$$C = c_0 + \sum_{i=1}^{k} c_i n_i \quad \ldots (7)$$

c_0 and c_i are beyond our control. Hence we will determine the optimum value of n_i which minimizes the variance of stratified sample mean.

To determine the optimum value of n_i, we consider the function

$$\begin{aligned} \varphi &= \operatorname{Var}(\bar{x}_{st}) + \lambda C \\ &= \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 S_i^2 + \lambda \left(c_0 + \sum_{i=1}^{k} c_i n_i\right) \end{aligned} \quad \ldots (8)$$

where λ is constant and known as Lagrange's multiplier.

Using the method of Lagrange's multiplier we select n_i and the constant λ to minimize φ. Differentiating φ with respect to n_i, we have

$$-\frac{W_i^2 S_i^2}{n_i^2} + \lambda c_i = 0 \qquad \ldots (9)$$

or

$$n_i = \frac{W_i \; S_i}{\sqrt{\lambda c_i}} \qquad \ldots (10)$$

Since λ is an unknown quantity, it has to be determined interms of known values. So, we take the sum over all in equation (10) and thus obtain

$$\sum_{i=1}^{k} n_i = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{k} \frac{W_i S_i}{\sqrt{c_i}}$$

or $$n = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{k} \frac{W_i S_i}{\sqrt{c_i}}$$

$$\Rightarrow \sqrt{\lambda} = \frac{1}{n} \sum_{i=1}^{k} \frac{W_i S_i}{\sqrt{c_i}}$$

Substituting the value of $\sqrt{\lambda}$ in equation (10), we get the value of n_i

$$n_i = n \frac{\left(W_i S_i / \sqrt{c_i}\right)}{\sum_{i=1}^{k} \left(W_i S_i / \sqrt{c_i}\right)} = n \frac{\left(N_i S_i / \sqrt{c_i}\right)}{\sum_{i=1}^{k} \left(N_i S_i / \sqrt{c_i}\right)} \qquad \ldots (11)$$

Thus, the relation (11) leads to the following important conclusions that we have to take a larger sample in a given stratum if

1. The stratum size N_i is larger;
2. The stratum has larger variability (S_i); and
3. The cost per unit is lower in the stratum.

3.6 STRATIFIED SAMPLING VERSUS SIMPLE RANDOM SAMPLING

Now, we shall make a comparative study of simple random sampling without replacement and stratified random sampling under different kinds of allocations i.e. Proportional allocation and Neyman's allocation.

3.6.1 Proportional Allocation versus Simple Random Sampling

The variance of the estimate of stratified sample mean with proportional allocation and variance of the sample mean of simple random sampling is given respectively by

$$\text{Var}\left(\overline{x}_{st}\right)_{PROP} = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} W_i S_i^2 \qquad \ldots (12)$$

where, $S_i^2 = \frac{1}{(N_i - 1)} \sum_{j=1}^{N_i} \left(X_{ij} - \overline{X}_i\right)^2$

and $\quad \text{Var}(\overline{x})_{SRSWOR} = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 \quad \ldots (13)$

where, $S^2 = \frac{1}{N-1} \sum_{i=1}^{k} \sum_{j=1}^{N_i} \left(X_{ij} - \overline{X}\right)^2$

In order to comparing (12) and (13) we shall first express S^2 in terms of S_i^2 we have

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{k} \sum_{j=1}^{N_i} \left(X_{ij} - \overline{X}_i + \overline{X}_i - \overline{X}\right)^2$$

$$(N-1)S^2 = \sum_{i=1}^{k} \left[\sum_{j=1}^{N_i} \left(X_{ij} - \overline{X}_i\right)^2\right] + \sum_{i=1}^{k} \sum_{j=1}^{N_i} \left(\overline{X}_i - \overline{X}\right)^2 + 2\sum_{i=1}^{k} \left[\left(\overline{X}_i - \overline{X}\right) \sum_{j=1}^{N_i} \left(X_{ij} - \overline{X}_i\right)\right]$$

$$(N-1)S^2 = \sum_{i=1}^{k} (N_i - 1) S_i^2 + \sum_{i=1}^{k} N_i \left(\overline{X}_i - \overline{X}\right)^2$$

The product term vanishes since

$$\sum_{j=1}^{N_i} \left(X_{ij} - \overline{X}_i\right) = 0$$

being the sum of square of deviation from the stratum mean. If we assume that N_i and consequently N are sufficiently large so that we can put $N_i - 1 = N_i$ and $N - 1 = N$, then we get

$$NS^2 = \sum_{i=1}^{k} N_i S_i^2 + \sum_{i=1}^{k} N_i \left(\overline{X}_i - \overline{X}\right)^2$$

$$S^2 = \sum_{i=1}^{k} W_i S_i^2 + \sum_{i=1}^{k} W_i \left(\overline{X}_i - \overline{X}\right)^2 \quad \ldots (14)$$

Substituting in equation (13), we get

$$\text{Var}(\overline{x})_{SRSWOR} = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} W_i S_i^2 + \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} W_i \left(\overline{X}_i - \overline{X}\right)^2$$

$$\text{Var}(\overline{x})_{SRSWOR} = \text{Var}(\overline{x}_{st})_{PROP} + \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} W_i \left(\overline{X}_i - \overline{X}\right)^2$$

$$\text{Var}(\overline{x})_{SRSWOR} \geq \text{Var}(\overline{x}_{st})_{PROP} \quad \ldots (15)$$

3.6.2 Neyman's Allocation versus Proportional Allocation

Considering the variances of estimated sample mean in stratified random sampling with proportional allocation and Neyman's allocation respectively we have

$$\text{Var}(\bar{x}_{st})_{PROP} = \left(\frac{1}{n}-\frac{1}{N}\right)\sum_{i=1}^{k} W_i S^2 \quad \ldots (16)$$

and

$$\text{Var}(\bar{x}_{st})_{NEY} = \frac{1}{n}\left(\sum_{i=1}^{k} W_i S_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} W_i S_i^2 \quad \ldots (17)$$

By subtracting equation (17) from equation (16) we get

$$\text{Var}(\bar{x}_{st})_{PROP} - \text{Var}(\bar{x}_{st})_{NEY} = \left(\frac{1}{n}-\frac{1}{N}\right)\sum_{i=1}^{k} W_i S_i^2 - \left\{\frac{1}{n}\left(\sum_{i=1}^{k} W_i S_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} W_i S_i^2\right\}$$

$$= \frac{1}{n}\left[\sum_{i=1}^{k} W_i S_i^2 - \left(\sum_{i=1}^{k} W_i S_i\right)^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{k} W_i (S_i - \bar{S})^2 \quad \ldots (18)$$

where, $\bar{S} = \sum_{i=1}^{k} W_i S_i = \frac{1}{N}\sum_{i=1}^{k} N_i S_i$ is the weighted mean of the stratum sizes N_i

Hence from equation (18) we can say

$$\text{Var}(\bar{x}_{st})_{PROP} \geq \text{Var}(\bar{x}_{st})_{NEY}$$

because, R.H.S. of equation (18) is non-negative.

3.6.3 Neyman's Allocation versus Simple Random Sampling

From the relationship between the proportional allocation and simple random sampling and the relation between proportional and Neyman allocation we have

$$\text{Var}(\bar{x})_{SRSWOR} = \text{Var}(\bar{x}_{st})_{PROP} + \left(\frac{1}{n}-\frac{1}{N}\right)\sum_{i=1}^{k} W_i \left(\bar{X}_i - \bar{X}\right)^2 \quad \ldots (19)$$

and $$\text{Var}(\bar{x}_{st})_{PROP} = \text{Var}(\bar{x}_{st})_{NEY} + \frac{1}{n}\sum_{i=1}^{k} W_i \left(S_i - \bar{S}\right)^2 \quad \ldots (20)$$

By substituting the value of the variance under proportional allocation in equation (19) from equation (20), we have

$$\text{Var}(\bar{x})_{SRSWOR} = \text{Var}(\bar{x}_{st})_{NEY} + \frac{1}{n}\sum_{i=1}^{k} W_i \left(S_i - \bar{S}\right)^2 + \left(\frac{1}{n}-\frac{1}{N}\right)\sum_{i=1}^{k} W_i \left(\bar{X}_i - \bar{X}\right)^2 \quad \ldots (21)$$

That means

$$\text{Var}(\bar{x})_{SRSWOR} \geq \text{Var}(\bar{x}_{st})_{NEY}$$

because both the terms in R.H.S. of equation (21) are positive. From the results of the relations of variance of simple random sample mean and the variance of stratified sample means with proportional and Neyman allocations, we can reach on the conclusion that

$$\text{Var}(\bar{x})_{SRSWOR} \geq \text{Var}(\bar{x}_{st})_{PROP} \geq \text{Var}(\bar{x}_{st})_{NEY}$$

3.6.4 Merits and Demerits of Stratified Random Sampling

Merits

1. **More Representative**
   Stratified random sampling ensures any desired representation in the sample of the various strata in the population. It overruled the probability of any essential group of the population being completely excluded in the sample.

2. **Greater Accuracy**
   Stratified random sampling provides estimate of parameters with increased precision in comparison to simple random sampling. Stratified random sampling also enables us to obtain the results of known precision for each of the stratum.

3. **Administrative Convenience**
   The stratified random samples would be more concentrated geographically in comparison to simple random samples. Therefore, this method needs less time and money involved in interviewing the supervision of the field work can be done with greater case and convenience.

Demerits

However, stratified random sampling has some demerits too, which are:

1. **May Contain Error due to Subjectiveness**
   In stratified random sampling the main objective is to stratify the population in homogeneous strata. But stratification is a subjective issue and so it may contain error.

2. **Lower Efficiency**
   If the sizes of samples from different stratum are not properly determined then stratified random sampling may yield a larger variance that means lower efficiency.

Example 1: A sample of 60 persons is to be drawn from a population consisting of 600 belonging to two villages A and B. The means and standard deviations of their marks are given below:

Villages	Strata Sizes (N_i)	Means $(\bar{X}_i)$	Standard Deviation (σ_i)
Village A	400	60	20
Village B	200	120	80

Draw a sample using proportional allocation techniques.

Solution: If we regard the villages A and B as representing two different strata then the problem is to draw a stratified random sample of size 30 using technique of proportional allocation. In proportional allocation, we have

$$n_i = \frac{n}{N} N_i$$

Therefore,

$$n_1 = \frac{60}{600} \times 400 \;= 40$$

$$n_2 = \frac{60}{600} \times 200 \;= 20$$

Thus, the required sample sizes for the villages A and B are 40 and 20 respectively.

E1) A sample of 100 employees is to be drawn from a population of collages A and B. The population means and population mean squares of their monthly wages are given below:

Village	N_i	$\bar{X}_i$	S_i^2
Collage A	300	25	25
Collage B	200	50	100

Draw the samples using proportional and Neyman allocation technique and compare.

E2) Obtain the sample mean and estimate of the population mean for the given information in Example 1 discussed above.

3.7 SUMMARY

In this unit, we have discussed:

1. The definition and procedure of stratified random sampling;
2. The principles of stratification;
3. The properties of stratified random sampling;
4. The mean and variance of proportions in stratified random sampling;
5. The allocation of sample size with the help of different techniques; and
6. Calculation of the estimate of population mean and variance of sample mean.

3.8 SOLUTIONS /ANSWERS

E1) If we regard the collages A and B representing two different strata then the problem is to draw as stratified sample of 100 employees using technique of proportional allocation and Neyman's allocation.

In proportional allocation we have

$$n_i = \frac{n}{N} N_i = \frac{100}{500} \times N_i$$

$$n_1 = \frac{100}{500} \times 300 = 60$$

$$n_2 = \frac{100}{500} \times 200 = 40$$

In Neyman's allocation, we have

$$n_i = n \times \frac{N_i S_i}{\sum_{i=1}^{k} N_i S_i}$$

$$\sum_{i=1}^{2} N_i S_i = 300 \times 5 + 200 \times 10$$

$$= 1500 + 2000 = 3500$$

$$n_i = 100 \times \frac{N_i S_i}{3500}$$

$$n_1 = 100 \times \frac{300 \times 5}{3500} = 42.85 \cong 43$$

$$n_2 = 100 \times \frac{200 \times 5}{3500} = 57.14 \cong 57$$

Therefore, the samples regarding the colleges A and B for both allocations are obtained as:

	Proportional	**Neyman**
Collage A	60	43
Collage B	40	57
Total	100	100

E2) We have the following data:

Village	$\mathbf{N_i}$	$\mathbf{\overline{X}_i}$	σ_i	$\mathbf{S_i^2 = \frac{N}{N-1}\sigma_i^2}$	$\mathbf{N_i\, S_i^2}$	$\mathbf{N_i\, \sigma_i^2}$	$\mathbf{\overline{X}_i^2}$	$\mathbf{N_i\, \overline{X}_i^2}$
A	400	60	20	400.67	160267.11	160000	3600	3440000
B	200	120	80	6432.16	1286432.16	1280000	14400	880000
Totat	600				1446699.27	1440000	180000	4320000

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{k} N_i \overline{X}_i$$

$$= \frac{1}{600}\left[400 \times 60 + 200 \times 120\right]$$

$$= \frac{1}{600}\left[24000 + 24000\right] = \frac{48000}{600} = 80$$

$$\mathrm{Var}\left(\bar{x}\ \right)_{PROP} = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{\sum_{i=1}^{k} N_i\, S_i^2}{N}$$

$$= \left(\frac{600-60}{60 \times 600}\right) \times \frac{1446699.27}{600} = \frac{1446699.27}{40000}$$

$$= 36.1675$$

$$\mathrm{Var}\left(\bar{x}\right)_{SRSWOR} = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$$

where,

$$S^2 = \frac{1}{(N-1)}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}\right)^2$$

$$= \frac{1}{(N-1)}\left[\sum_{i=1}^{k} N_i \sigma_i^2 + \sum_{i=1}^{k} N_i\left(\overline{X_i} - \bar{X}\right)^2\right]$$

$$= \frac{1}{(N-1)}\left[\sum_{i=1}^{k} N_i \sigma_i^2 + \sum_{i=1}^{k} N_i \bar{X}_i^2 - N\bar{X}^2\right]$$

$$= \frac{1}{599}\left(1440000 + 4320000 - 600 \times 80 \times 80\right)$$

$$= \frac{1}{599}\left[5760000 - 3840000\right] = \frac{1920000}{599}$$

$$\mathrm{Var}\left(\bar{x}\right)_{SRSWOR} = \frac{600-60}{600 \times 60} \times \frac{1920000}{599}$$

Then the conclusion is

$$\mathrm{Var}\left(\bar{x}_{st}\right)_{PROP} = 36.1675$$

$$\mathrm{Var}\left(\bar{x}\right)_{SRSWOR} = 48.08$$

Therefore, precision of $\bar{x}_{st}$ can be obtained by

$$\text{Gain in precision} = \frac{\mathrm{Var}\left(\bar{x}\right)_{SRS} - \mathrm{Var}\left(\bar{x}_{st}\right)_{PROP}}{\mathrm{Var}\left(\bar{x}_{st}\right)_{PROP}} \times 100$$

$$= \frac{48.04 - 36.1675}{36.1675} \times 100$$

$$= 32.8\ \%$$