UNIT 4 SOME OTHER SAMPLING SCHEMES

Structure

4.1 INTRODUCTION

When one has to make an inference about a large lot and it is not practically possible to examine each individual unit.Then a few units of the lot are examined and on the basis of the information of those units, one makes decisions about whole lot. In previous unit, we have discussed about stratified random sampling. As we stated in Unit 3, a stratified random sample is selected randomly from the groups of population units named strata, where the strata are formed of the homogeneous units. But, if the formation of the strata is not done in proper manner, then the results would be biased. So, to overcome this drawback, it is suggested to adopt another sampling method known as Systematic sampling. The systematic random sampling is one of among the mixed sampling schemes, which is partly probabilistic and partly non-probabilistic.

A brief introduction of systematic sampling is given in Section 4.2, whereas the methods of selection of systematic random samples are explained with examples in Section 4.3. Properties of the systematic random sampling in terms of mean and variance are derived in Section 4.4. In Section 4.5 the notations and the methods of cluster sampling are explained whereas the properties of the cluster sampling are described in Section 4.6. Similarly, the brief introduction about the two stage sampling is given in Section 4.7 and the properties of the two-stage sampling are discussed in Section 4.8.

Objectives

After studying this unit, you would be able to

- define the systematic random sampling;
- draw a systematic random sample and estimate the population mean;
- obtain the variance of the estimate of the population mean;
- define the cluster sampling and obtain the estimate of population mean; and
- define and explain the two-stage sampling scheme.

4.2 INTRODUCTION TO SYSETMATIC SAMPLING

In previous units, we have discussed those sampling techniques where units were selected randomly. In this unit, we shall discuss a sampling technique which has a nice feature of selecting the whole sample with just one random start. Systematic random sampling is commonly employed if the complete and up-to-date list of population units is available. In systematic random sampling only the first unit is selected with the help of random method and the rest being automatically selected according to some predetermined pattern. The systematic random sampling is a kind of mixed sampling, which is partly probabilistic and partly non-probabilistic. This is random since the first unit of the sample is selected at random and non-random or purposive since the rest of units in the sample are selected by predesigned pattern.

4.3 METHODS OF SYSTEMATIC SAMPLING

Generally, in day to day life, we have to obtain the information from cards or register which are full of information arranged in serial order. For example, the books in library, a telephone directory, etc. In such case systematic sampling often works better than simple random or stratified sampling. There are two ways of selection of a systematic random sample and these are (a) Linear systematic sampling and (b) Circular systematic sampling.

4.3.1 Linear Systematic Sampling

In order to draw a systematic random sample of size n from a population of N units, Let us suppose that N sampling units are serially numbered from 1 to N in some order.

Let N = nk, where n is the sample size and k is an integer. Therefore, $k = N/n$. In other words, in order to draw a sample of size n from N we divide the total number of units into n equal parts. Suppose each part consists of k units. From the serially arranged 1 to k units, draw a unit with random method. Let the selected unit is i^{th} unit where $i \leq k$. The selected unit would be the first sample unit. Then select every k^{th} unit after the i^{th} unit in order to select rest of $(n-1)$ units. Thus, the systematic sample of size n will consists of $i^{th}, (i+k)^{th}, (i+2k)^{th}, ..., (i+(n-1)k)^{th}$ unit. The random sample unit i.e. the i^{th} unit is called the random start. For example, suppose there are 100 units in a population serially numbered 1 to 100 units. We will divide the whole

population into 10 equal parts of 10 units each if we have to draw a sample of size10:

$$\frac{N}{n}=\frac{100}{10}=10=k$$

Then we draw a unit randomly from 1 to 10 units and let the selected number is 7 i.e. i = 7. Then we select rest of 9 units in a systematic way i.e.

$(i+k)^{th}$, $(i+2k)^{th}$, $(i+3k)^{th}$, ..., $(i+9k)^{th}$ unit from the serially ordered 100 units. Then the systematic sample consists of 7^{th}, 17^{th}, 27^{th}, 37^{th}, 47^{th}, 57^{th}, 67^{th}, 77^{th}, 87^{th} and 97^{th} units.

4.3.2 Circular Systematic Sampling

Suppose a population consists of N units and from this we have to select a systematic sample of n units. Also, assume that N is a multiple of n i.e. N = nk. The procedure is to select a random number, let it be i such that $1 \leq i \leq k$ and then we have to select i^{th} unit from first k unit and then every k^{th} unit i.e. $(i + k)^{th}$, $(i+2k)^{th}$,…,$(i + (n-1) k)^{th}$ positional units. This sampling technique is known as linear systematic sampling. But in general N does not be always a multiple of n. For example N = 17 and n = 4, then $k=\frac{N}{n}=\frac{17}{4}=4.25$ has to be taken as 5. Now we select a random number between 1 and 5 and suppose it is 4. Then the remaining three units to be selected are at positions 9, 14, 19. There is no unit in the population at serial number 19. Hence in this situation we can select a sample of 3 only instead of 4. In this situation, the circular systematic sampling is used.

The circular systematic sampling is used when size of the population N is not a multiple of sample n. In this situation we take N/n as k by rounding off N/n to the nearest integer. With regard to selection of a systematic random sample from N units, we have to select random number from 1 to N. Let this number is i. Now we select every $(i+jk-N)^{th}$ unit, when (i + jk) >N putting j = 1, 2, 3, … till n unit are selected. By using the circular systematic sampling we always get a sample of size n. For the same example discussed above with N = 17, n = 4 and k = 5, let the randomly selected number from 1 to 17 is 8 and latter the 13^{th}, 1^{th} and 6^{th} units are selected. When N = nk, the linear and circular systematic sampling plans become identical.

4.3.3 Advantages and Disadvantages of Systematic Sampling

Advantages

Systematic sampling has some advantages over other sampling schemes which are given as:

1. The systematic sampling is very simple and is not very expensive;
2. The systematic sample is uniformly distributed over the whole population and therefore, all sections of the population are represented in the sample; and
3. The managerial control of field work provides an advantage over other sampling methods.

Disadvantages

Systematic sampling has some disadvantages also along with the advantages which are:

1. The main disadvantage of systematic sampling is that samples are not generally random samples;
2. The sample size is different from that required if N is not a multiple of n;
3. In systematic random sampling the sample mean would not be an unbiased estimate of population mean if N is not a multiple of n;
4. We cannot obtain an unbiased estimate of the variance of the estimate of the population mean since it does not provide sampling error; and
5. It may provide highly biased estimate if the sampling frame has a periodic feature.

4.4 PROPERTIES OF SYSTEMATIC SAMPLING

Let x_{ij} denote the j^{th} member of the i^{th} systematic sample (where, i =1, 2, … k; j=1, 2, … n). $\bar{x}_i$ may be denoted as mean of the i^{th} sample, i.e.

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij} \qquad (i = 1, 2, ..., k)$$

and $$\bar{x}_{sys} = \frac{1}{k}\sum_{i=1}^{k} \bar{x}_i$$

Then $\bar{X}_i$, $\bar{X}_{..}$ and S^2 may be denoted as population mean and population mean square as follows:

$$\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij} \qquad (i = 1, 2, ..., k)$$

$$\bar{X}_{..} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij} = \frac{1}{k}\sum_{i=1}^{k} \bar{X}_i$$

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

Theorem 1: In systematic sampling with interval k sample mean $\bar{x}_{sys}$ is an unbiased estimator of population mean $\bar{X}_{..}$ and its variance is given by

$$\text{Var}\left(\bar{x}_{sys}\right) = \frac{N-1}{N}S^2 - \frac{(n-1)k}{N}S_{sys}^2$$

where, $$S_{sys}^2 = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right)^2$$

is the mean square among the units which lie within the same systematic sample.

Proof: We have

$$E(\bar{x}_{sys}) = E\left(\frac{1}{k}\sum_{i=1}^{k}\bar{x}_i\right)$$

$$= E\left[\frac{1}{k}\sum_{i=1}^{k}\frac{1}{n}\sum_{j=1}^{n}x_{ij}\right]$$

$$= \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n}\sum_{j=1}^{n}E(x_{ij})$$

$$= \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n}X_{ij}$$

$$= \frac{1}{k}\sum_{i=1}^{k}\bar{X}_i = \bar{X}_{..}$$

Now, we have

$$(N-1)S^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X}_{..})^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i+\bar{X}_i-\bar{X}_{..})^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{X}_i-\bar{X}_{..})^2$$

Covariance term vanishes, since

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)(\bar{X}_i-\bar{X}_{..}) = \sum_{i=1}^{k}\left[(\bar{X}_i-\bar{X}_{..})\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)\right] = 0$$

Therefore,

$$(N-1)S^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X}_{..})^2 + n\sum_{i=1}^{k}(\bar{X}_i-\bar{X}_{..})^2$$

$$(N-1)S^2 = k(n-1)S_{sys}^2 + nk\,Var(\bar{x}_{sys})$$

$$Var(\bar{x}_{sys}) = \frac{N-1}{N}S^2 - \frac{k(n-1)}{N}S_{sys}^2$$

4.4.1 Comparison between Simple Ranodm, Stratified and Systematic Sampling

If the population consists of linear trend and given by

$X_i = i$; $i = 1, 2, 3, ..., N$

Then, $\sum_{i=1}^{N}X_i = \sum_{i=1}^{N}i = \frac{N(N+1)}{2}$

$$\sum_{i=1}^{N}X_i^2 = \sum_{i=1}^{N}i^2 = \frac{N(N+1)(2N+1)}{6}$$

Therefore,

$$\overline{X} = \frac{N+1}{2}$$

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}_{..}\right) = \frac{1}{(N-1)}\left[\sum_{i=1}^{N} X_i^2 - N\overline{X}^2\right]$$

$$S^2 = \frac{1}{N-1}\left[\frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{4}\right]$$

$$S^2 = \frac{N(N+1)}{2(N-1)}\left[\frac{(2N+1)}{3} - \frac{(N+1)}{2}\right]$$

$$= \frac{N(N+1)}{12}$$

$$\text{Var}\left(\overline{x}_{srswor}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2$$

$$= \frac{N-n}{N.n} \cdot \frac{N(N+1)}{12}$$

$$= \frac{nk-n}{n^2k} \cdot \frac{nk(nk+1)}{12}$$

$$\text{Var}\left(\overline{x}_{srswor}\right) = \frac{(k-1)(nk+1)}{12} \quad \ldots (1)$$

We have $S^2 = \frac{N(N+1)}{12}$ for population of N units.

In stratified random sampling we have

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

In our case, there are n strata of size k and we draw one unit from each stratum so we put $N_i = k$, $n_i = 1$, $k = n$ and $W_i = N_i / N = 1/n$.

$$\text{Var}\left(\overline{x}_{st}\right) = \sum_{j=1}^{n} \frac{1}{n^2}\left(1 - \frac{1}{k}\right) S_i^2$$

Since i^{th} stratum consists of k units, we have

$$S_i^2 = \frac{k(k+1)}{12}$$

Therefore, $$\text{Var}\left(\overline{x}_{st}\right) = \frac{(k-1)}{n^2k} \cdot \frac{nk(k+1)}{12}$$

$$\text{Var}\left(\overline{x}_{st}\right) = \frac{\left(k^2-1\right)}{12n} \quad \ldots (2)$$

For finding out $\mathrm{Var}(\bar{x}_{sys})$, we have

$$\bar{x}_i = \text{mean of the values of } i^{th} \text{ sample} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$$

$$= \frac{1}{n}\left[i + (i+k) + (i+2k) + \ldots + (i+(n-1)k)\right]$$

$$= \frac{1}{n}\left[ni + \{1+2+3+\ldots+(n-1)\}k\right]$$

$$= \frac{1}{n}\left[ni + \frac{(n-1)n}{2}.k\right] = i + \frac{(n-1)}{2}k$$

Also $\bar{x}_{..} = \bar{x} = \frac{N+1}{2} = \frac{nk+1}{2}$

$$\bar{x}_i - \bar{x}_{..} = i + \frac{(n-1)k}{2} - \frac{(nk+1)}{2}$$

$$= i - \frac{(k+1)}{2}$$

$$\mathrm{Var}(\bar{x}_{sys}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{x}_i - \bar{x}_{..})^2$$

$$= \frac{1}{k}\sum_{i=1}^{k}\left(i - \frac{(k+1)}{2}\right)^2$$

$$= \frac{1}{k}\sum_{i=1}^{k}\left(i^2 + \left(\frac{k+1}{2}\right)^2 - 2i\left(\frac{k+1}{2}\right)\right)$$

$$= \frac{1}{k}\sum_{i=1}^{k} i^2 + \left(\frac{k+1}{2}\right)^2 - \left(\frac{k+1}{k}\right)\sum_{i=1}^{k} i$$

$$= \frac{(k+1)(2k+1)}{6} + \frac{(k+1)^2}{4} - \frac{(k+1)^2}{2}$$

$$\mathrm{Var}(\bar{x})_{sys} = \frac{k^2-1}{12} \qquad \ldots (3)$$

From equations (1), (2) and (3), we get

$$\mathrm{Var}(\bar{x}_{st}) : \mathrm{Var}(\bar{x}_{sys}) : \mathrm{Var}(\bar{x}_{srswor})$$

$$\frac{(k^2-1)}{12n} : \frac{(k^2-1)}{12} : \frac{(k-1)(nk+1)}{12}$$

$$\frac{(k+1)}{n} : (k+1) : nk+1$$

$$\frac{1}{n} : 1 : n \text{ (approx)}$$

$$\mathrm{Var}(\bar{x}_{st}) \le \mathrm{Var}(\bar{x}_{sys}) \le \mathrm{Var}(\bar{x}_{srswor})$$

Example 1: In a class of Statistics, total number of students is 30. Select a systematic random sample of 10 students.The age of 30 students is given below:

Age:	22	25	22	21	22	25	24	23	22	21
	20	21	22	23	25	23	24	22	24	24
	21	20	23	21	22	20	20	21	22	25

Solution: We have given a population of size N = 30 values of the age of 30 students. Now, first of all we arrange all the values together with their serial numbers. Therefore,

Sr No:	1	2	3	4	5	6	7	8	9	10
Age:	22	25	22	21	22	25	24	23	22	21
Sr No:	11	12	13	14	15	16	17	18	19	20
Age:	20	21	22	23	25	23	24	22	24	24
Sr No.	21	22	23	24	25	26	27	28	29	30
Age:	21	20	23	21	22	20	20	21	22	25

After that we obtain value k as

$$k = \frac{N}{n} = \frac{30}{10} = 3$$

From first k values i.e. 1 to 3, we select a value randomly. Let we select the age 25 which is in 2^{nd} place in data. Therefore, our 1^{st} unit which selected in the sample is i = 25.

Now rest of the 9 units we will select systematically which are at the position (i+1k), (i+2k), (i+3k), …, (i+(n-1)k) in the given data. So according to the given data rest of the 9 units in our case would be the age given in 5^{th}, 8^{th}, 11^{th}, 14^{th} , 17^{th}, 20^{th}, 23^{rd}, 26^{th}, 29^{th} position.

Therefore, all the 10 units which has been selected in the sample are {25, 22, 23, 20, 23, 24, 24, 23, 20, 22}.

E1) The information regarding production of wheat (in Thousand kg) in 25 districts are collected, for a particular season. Select a systematic random sample of 7 units from the data given below:

23, 20 ,30, 37, 76, 36, 13, 36, 16, 58, 53, 83, 10, 15,13, 17, 12, 16, 17, 21, 20, 18, 61, 31, 71.

E2) A data of 50 values of heights (in cm) according to the roll no. of the students is given as follows:

146,	156,	152,	167,	178,	180,	172,	162,	148,
153,	161,	173,	163,	164,	175,	168,	161,	180,
173,	185,	169,	167,	168,	173,	145,	153,	154,
162,	164,	170,	172,	160,	161,	158,	152,	163,
165,	170,	168,	158,	149,	155,	160,	150,	149,
167,	176,	169,	159,	160.				

Select a systematic random sample of size 10.

4.5 INTRODUCTION TO CLUSTER SAMPLING

The population has been considered as a group of a finite number of distinct and identified units defined as sampling units. The smallest identity content in a population is known as element or elementary unit of the population. A group of such elementary units is known as cluster. Clusters are generally made up of which all the elements tend to have similar characteristics. When these clusters are treated as sampling units and few of them are selected either by equal or unequal probabilities then this procedure is known as cluster sampling. All the elements in selected clusters are to be observed, measured and interviewed. The number of elements in the cluster should be small and the number of clusters in the population should be large. For example, if we are interested in obtaining the information or data for monthly average income in a colony, then the whole colony may be divided into N numbers of block known as clusters and a simple random sample of n blocks is to be drawn. The individuals living in the selected clusters would be determined for interviewing to collect the information.

4.5.1 Significance of Cluster Sampling

Following are the various reasons, which cause problems in the selection of a sample of elementary units and cluster sampling enables us to overcome those problems:

1. When the sampling frame is unavailable, so the identifying and interviewing of sampling units is costly in terms of money, time consuming and need much labor. For example a list of households in metro city, list of farmhouse owners in a state, etc.;
2. The location of the identified sampling units may be situated far apart from one another and consume a lot of time and money to survey them;
3. It may not be possible to find well identifiable and easily locatable elementary units.

Thus, to overcome the above problems, cluster sampling yields satisfactory results in sampling of elementary units. The elementary units are formed in groups on the basis of location, class or area in cluster sampling.

4.5.2 Formation of Clusters

Certain precautions should be taken necessarily while dealing with the clusters sampling, which are given as follows:

1. The clusters should be made like that each elementary units should belong to one and only one cluster;
2. All units of similar characteristics should belong to the same cluster;
3. Each and every unit of the population should be included in any of the clusters constituting the population. In other words, there should neither be overlapping clusters nor omission of units;
4. All clusters should be heterogeneous themselves; and
5. Clusters should be as small as possible.

4.6 PROPERTIES OF CLUSTER SAMPLING

Notations

N = Number of clusters in the population

M = Number of elements in the clusters

n = Number of clusters in the sample

X_{ij} be the value of characteristic under study for the j^{th} element (j=1, 2,…, M) in the i^{th} cluster (i = 1, 2,…, N) of a population.

$\overline{X}_i = \sum_{j=1}^{M} X_{ij}/M$ is the mean of the i^{th} cluster in the population

$\overline{X} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} X_{ij}$ is the mean per element in the population

Similarly, x_{ij} be the value of characteristic under study for the j^{th} element (j =1, 2, …, M) in the i^{th} cluster (i = 1, 2,…, n) in the sample

Let $\overline{x}_i = \frac{1}{M}\sum_{j=1}^{M} x_{ij}$ is the mean of the i^{th} cluster in sample

and $\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} \overline{x}_i$ is the mean of cluster means in the sample of size n

We have

$S_i^2 = \frac{1}{(M-1)}\sum_{j=1}^{M}(X_{ij} - \overline{X}_i)^2$ is the mean squar between elements within the i^{th} cluster

$S_w^2 = \frac{1}{N}\sum_{i=1}^{N} S_i^2$ the mean square within clusters

and $S_b^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(\overline{X}_i - \overline{X})^2$ is the mean squar between the cluster means in the population.

Therefore,

$S^2 = \frac{1}{(NM-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(X_{ij} - \overline{X})^2$ denotes the mean square between elements in the population and

$$\rho = \frac{E(X_{ij} - \overline{X})(X_{ik} - \overline{X})}{E(X_{ij} - \overline{X})^2} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j=1}^{M}(X_{ij} - \overline{X})(X_{ik} - \overline{X})}{(M-1)(NM-1)S^2}$$

denote the intra-cluster correlation coefficient.

Theorem 2: In simple random sample without replacement of n clusters each containing M elements drawn form a population of N clusters, the sample mean $\overline{x}_n$ is an unbiased estimator $\overline{X}$ and its variance is given by

$$Var(\overline{x}_n) = \left(\frac{1-f}{n}\right)S_b^2 \cong \left(\frac{1-f}{n}\right)S_b\left[1+(M-1)\rho\right]$$

where, ρ is the intra-cluster correlation coefficient.

Proof: We have

$$E(\bar{x}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n}\bar{x}_i\right)$$

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

$$= \frac{1}{n}\sum_{i=1}^{N}\frac{n}{N}\bar{X}_i = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}X_{ij}$$

$$= \bar{X}$$

and now we have

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

$$= \left(\frac{1-f}{n}\right)\sum_{i=1}^{N}\frac{(\bar{X}_i - \bar{X})^2}{(N-1)}$$

where f = n/N.

Using

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

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

$$\sum_{i=1}^{N}(\bar{X}_i - \bar{X})^2 = \frac{1}{M^2}\sum_{i=1}^{N}\sum_{j=1}^{M}(X_{ij} - \bar{X})^2 + \frac{1}{M^2}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j=1}^{M}(X_{ij} - \bar{X})(X_{jk} - \bar{X})$$

$$= \frac{(NM-1)}{M^2}S^2[1 + (M-1)\rho]$$

Therefore,

$$Var(\bar{x}_n) = \left(\frac{1-f}{n}\right)\left[\frac{(NM-1)}{M^2(N-1)}S^2\{1 + (M-1)\rho\}\right]$$

$$= \left(\frac{1-f}{nM}\right)S^2[1 + (M-1)\rho] \text{ for large N}$$

Variance in cluster sampling depends on the number of clusters in the sample, the size of the cluster, the intra-cluster correlation coefficient ρ and the mean square between the elements in the population. The variance of cluster sampling reduces to the variance of simple random sampling if M = 1.

Example 2: To determine the yield rate of wheat in a district of Punjab, 6 groups were constructed of 6 plots each. The data is given in the following table:

Plot No.	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
1	8	6	18	13	17	12
2	13	5	8	7	15	15
3	11	16	6	13	10	11
4	26	5	10	6	21	17
5	13	16	16	7	20	8
6	31	5	20	2	25	10

Select a cluster sample of 3 clusters from the given data and find sample mean.

Solution: In the given data, 6 groups have been formed and we have to draw a sample of 3 groups. Therefore, there will be 20 possible samples of size 3 which may be drawn from the population of size 6. Let we consider one sample of 3 groups drawn from population of 6 groups is {Group 1, Group 3 and Group 5}. Therefore, the sample mean will be

$$\bar{x} = \frac{n_1\bar{x}_1 + n_3\bar{x}_3 + n_5\bar{x}_5}{n_1 + n_3 + n_5}$$

where,

$$\text{Mean of Group 1} \qquad \bar{x}_1 = \frac{8+13+11+26+13+31}{6} = \frac{102}{6} = 17$$

$$\text{Mean of Group 3} \qquad \bar{x}_3 = \frac{18+8+6+10+16+20}{6} = \frac{78}{6} = 13$$

$$\text{Mean of Group 5} \qquad \bar{x}_5 = \frac{17+15+10+21+20+25}{6} = \frac{108}{6} = 18$$

Therefore,

$$\text{Grand Mean } \bar{x} = \frac{6\times17 + 6\times13 + 6\times18}{18}$$

$$= \frac{6\times(17+13+18)}{18} = \frac{6\times48}{18}$$

$$= 16$$

Hence, the sample mean is 16.

E3) A Housing board is constructing the 10 Duplex on each of 5 different locations. The plot areas (in square fit) as given in the following table:

S. No.	location 1	location 2	location 3	location 4	location 5
1	800	800	1300	2100	700
2	1300	700	1000	1600	2000
3	600	1500	1100	1300	1700
4	1800	1000	1600	600	1300
5	1300	1600	500	700	1600
6	1700	1300	1000	800	1300
7	1200	1800	1300	1200	1000
8	1300	600	1000	1500	500
9	500	700	1700	1600	1500
10	1000	1000	1000	1000	1000

Select a sample of 2 clusters with the help of cluster sampling scheme and calculate the sample mean.

4.7 INTRODUCTION TO TWO-STAGE SAMPLING

A sample survey as pointed out in Unit 1 of this block has certain limitations, mainly regarding the budget and time availability. Hence in survey too many elementary units is often not possible. In stratified random sampling, a sample is selected of optimum size from each stratum and then each and every unit selected from different stratum is to be observed. In Section 4.5 of this unit we have discussed cluster sampling in which the population is divided into some number of clusters and clusters were considered as sampling units. All the units in the selected clusters are enumerated completely. It has been pointed out there that cluster sampling is economically better than other samplings but the method restricts the spread of sample over the population which increased the variance of the estimator.

Instead of enumerating all the sampling units in the selected clusters, one can select a subsample of identified and specific units from the selected clusters by the same or different sampling methods.

The sampling which consists of selected clusters and then select the specified number of units from each selected cluster is known as two stage sampling. In this sampling technique, clusters being termed as primary stage units and units within clusters as secondary stage units. This method can be generalized upto three or more stages and is termed as multi stage sampling.

When large scale surveys on district, state or national level are to be conducted, it is one of the most suitable sampling schemes. For example, suppose we would like to draw the information about the monthly income of a household in a colony, it is better to select a sample of some blocks or wards and then households from the selected blocks or wards. In this procedure, the wards or blocks are the first stage units and the households are the second stage units.

4.7.1 Terminologies

N = Total number of first stage units.

n = Sample size of first stage units.

M = Number of second stage units in each first stage unit.

m = Number of second stage units in the sample from each first stage unit.

X_{ij} = Observation on the j^{th} second stage unit belonging to i^{th} first stage unit.

$\overline{X}_i$ = Mean of the i^{th} first stage unit $= \frac{1}{M}\sum_{j=1}^{M} X_{ij}$

$\bar{x}_i$ = Sample mean of the i^{th} first stage unit $= \frac{1}{m}\sum_{j=1}^{m} x_{ij}$

$\bar{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} \bar{x}_i$ = Overall sample mean on the basis of each second stage unit when sub – sampling has been done

$\bar{\bar{X}} = \frac{1}{N}\sum_{i=1}^{N} \bar{X}_i$ = Overall population Mean

4.8 PROPERTIES OF TWO-STAGE SAMPLING

Theorem 3: If the n first stage units and m second stage units from each selected first stage unit are selected by simple random sampling without replacement. There sample mean ($\bar{\bar{x}}$) is an unbiased estimator of population mean ($\bar{\bar{X}}$) and having the variance

$$Var(\bar{\bar{x}}) = \frac{(N-n)}{N}\frac{S_b^2}{n} + \frac{M-m}{M}\frac{S_w^2}{mn}$$

where, $S_b^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(\bar{X}_i - \bar{\bar{X}})^2$ and $S_w^2 = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(X_{ij} - \bar{X}_i)^2$

Proof: Since the units are selected in two stages by considering a probability sampling in each stage in two stage sampling. At both stages, selection procedures are to be considered in deriving expected value and variance of the sample statistic based on the number of units selected in second stage.

For getting the expected value and variance we have to follow:

$$E(\bar{\bar{x}}) = E_1[E_2[\bar{x}_i / i]] \quad \ldots (4)$$

$$Var(\bar{\bar{x}}) = Var_1[E_2(\bar{x}/i)] + E_1[Var_2(\bar{x}/i)] \quad \ldots (5)$$

Where E_1 and Var_1 are expectation and variance over the first stage units and E_2 and Var_2 are the conditional expectation and variance over the second stage unit for a given sample of first stage units.

Now since we draw second stage units from the first stage units by SRSWOR so

$$E_2[\bar{x}/i] = \bar{X}_i$$

Therefore, from equation (4), we have

$$E[\bar{\bar{x}}] = E_1[\bar{X}_i] = \bar{\bar{X}}$$

It shows that sample mean of all elements in the sample is an unbiased estimate of the population mean.

To obtain the variance

$$Var(\bar{\bar{x}}) = Var_1[E_2(\bar{x}/i)] + E_1[Var_2(\bar{x}/i)]$$

$$= \text{Var}_1\left(\overline{X}_i\right) + E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{m}-\frac{1}{M}\right)S_i^2\right]$$

$$= \frac{N-n}{Nn}S_b^2 + \frac{(M-m)}{mM}\frac{S_w^2}{n}$$

where, $S_w^2 = \frac{1}{N}\sum_{i=1}^{N}S_i^2$

Example 3: Select a sample of size 6 from given population of 36 units. The data given in 'Example 2' is divided in 6 clusters or groups each of them having 6 units.

Solution: Let us select 3 groups as first stage units from the given 6 groups. Let the selected units are Group-1, Group-3 and Group-6. Therefore the first stage sample is:

S. No.	Group-1	Group-3	Group-6
1	8	18	12
2	13	8	15
3	11	6	11
4	26	10	17
5	13	16	18
6	31	20	10

Now, from the selected first stage units, we shall select the 6 second stage units. These units are selected on the basis of their importance. Let we select 2 units each from these three selected first stage units of size 6.

Let us select 2nd & 4th second stage units from Group-1, 1st & 4th unit from Group-3 and 3rd & 6th unit from Group-6.

Therefore, the second stage sample units selected from 3 first stage units of size 6 are {13, 26, 18, 10, 11, 10 }.

Let us answer the given exercise.

E4) Select a first stage sample of size 2 then the second stage sample of size 10, from the data given in E3) by two-stage sampling method.

4.9 SUMMARY

In this unit, we have discussed:

1. Systematic random sampling;
2. How to draw a systematic random sample and estimate the population mean;
3. The variance of the estimate of the population mean;
4. The cluster sampling and the estimate of the variance of the sample mean; and
5. Method of drawing a two-stage sample; and
6. Method of solving the numerical examples

4.10 SOLUTIONS / ANSWERS

E1) First of all we arrange the data of production of wheat (in Thousand kg) with their serial order:

Sr.No. :	1	2	3	4	5	6	7	8	9	10	11	12	13
Production:	23	20	30	37	76	36	13	36	16	58	53	83	10
Sr. No.:	14	15	16	17	18	19	20	21	22	23	24	25	
Production:	15	13	17	12	16	17	21	20	18	61	31	71	

Now we obtain a number $k = \frac{N}{n} = \frac{25}{7} = 3.5$

We have N = 25 and n = 7

So we have to take k = 4

Now, from first 4 values in serially arranged data let we select 3rd value (i =30), so this will be our 1st sample unit selected in the sample by random method. Now remaining 6 values will be selected in systematic way i.e. $(i+1k)^{th}$, $(i+2k)^{th}$, …, $(i+6k)^{th}$ order value in the data. So in this way, we have to select the values which are at 7^{th}, 11^{th}, 15^{th}, 19^{th}, 23^{rd} and 27^{th}position. But in the data, only 25 values are available. So, we have to adopt the circular systematic sampling method for the selection of all 7 units. By following the circular systematic sample after selecting the first unit ($i = 3^{rd}$) in between 1 to 25 the remaining units would be 7^{th}, 11^{th}, 15^{th}, 19^{th}, 23^{rd} and 2^{nd} positioned units.

Therefore, from the population of 25 units the systematic random sample of size 7 would be {30, 13, 53, 13, 17, 61, 20}.

E2) We have N = 50 and we have to draw a systematic sample of size 10. We arrange the values of the heights of 50 students corresponding to their roll numbers from 1 to 50.

Roll No:	1	2	3	4	5	6	7	8	9
Height:	146	156	152	167	178	180	172	162	148
Roll No:	10	11	12	13	14	15	16	17	18
Height:	153	161	173	163	164	175	168	161	180
Roll No:	19	20	21	22	23	24	25	26	27
Height:	173	185	169	167	168	173	145	153	154
Roll No:	28	29	30	31	32	33	34	35	36
Height:	162	164	170	172	160	161	158	152	163
Roll No:	37	38	39	40	41	42	43	44	45
Height:	165	170	168	158	149	155	160	150	149
Roll No:	46	47	48	49	50				
Height:	167	176	169	159	160				

Now we shall obtain a number $k = \frac{N}{n} = \frac{50}{10} = 5$ that is k = 5. So we have to select the first unit of the sample of size 10 by random selection method from serially 1 to 5. Let we have selected the 4^{th} unit in order. So the 1^{st}unit selected in the sample is 167.

Now we shall select the remaining 9 unit in a systematic way. We have
i = 4 and k = 5 so the remaining 9 sample unit to be selected in the sample would be 9^{th},14^{th}, 19^{th}, 24^{th} , 29^{th} , 34^{th}, 39^{th}, 44^{th} and 49^{th} order in the data.

Therefore, the values of systematic random sample of size 10 selected from the population of size 50 would be 167, 148, 164,173, 173, 164, 158, 168, 150, 159.

E3) In cluster sampling, we know that if population is divided in several N homogeneous groups (known as clusters) then we have to select the clusters as sample units for selection of sample of size n.

Here, in the given data N = 5 and n =2. Therefore, there will be ${}^{N}C_{n} = {}^{5}C_{2} = 10$ number of possible sample of size 2 clusters.

Let us select the Location 2 and Location 4 as the sample units. Therefore, the sample of 2 locations selected from population of 5 locations is

Location	Size of Plots										Total
Location 2	800	700	1500	1000	1600	1300	1800	600	700	1000	11000
Location 4	2100	1600	1300	800	1000	900	1200	1500	1600	1000	13000

Therefore, sample mean of first sample unit

$$\bar{x}_1 = \frac{11000}{10} = 1100$$

and sample mean of second sample unit

$$\bar{x}_2 = \frac{13000}{10} = 1300$$

Therefore, the sample mean

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 +}{n_1 + n_2} = \frac{10\times 1100 + 10\times 1300}{10+10}$$

$$= \frac{10(1100+1300)}{20} = 1200 \text{ square fit}$$

E4) The data given in E3) are having 50 number of sizes of plots, which are divided into 5 groups on the basis of the different locations, i.e. Location 1, Location 2, …, Location 5. Now, we select a first stage sample of two locations. Let it be Location 2 and Location 5 as first stage sample units.

Therefore, we have

Plot No.	Location 2	Location 5
1	800	700
2	700	2000
3	1500	1700
4	1000	1300
5	1600	1600
6	1300	1300
7	1800	1000
8	600	500
9	700	1500
10	1000	1400
Total	11,000	13,000

Now, from the selected first stage units we shall select a second stage sample of 10 units by selecting 5 units each from both locations. Let the 2^{nd}, 3^{rd}, 6^{th}, 7^{th} and 9^{th} are selected from the Location 2 and 3^{rd}, 5^{th}, 7^{th}, 8^{th} and 10^{th} unit are selected from the Location 5.

Therefore, the sample selected by the two-stage sampling method is {700, 1500, 1300, 1800, 700, 1700, 1600, 1000, 500, 1400}.