UNIT 9

SYSTEMATIC SAMPLING

Structure

9.1 INTRODUCTION

As far as the theory of sampling is concerned, until so far, we discussed two sampling schemes, namely, Simple Random Sampling (SRS) and Stratified Random Sampling (STRS) which basically belong to the category of ‘Equal Probability Selection Method (EPSEM)’ schemes. However, we have seen that Stratified Random Sampling is, in fact, nothing but the repetitive use of Simple

Random Sampling scheme in all the sub-populations (strata) of the original population. Besides Equal Probability Selection Method schemes, we also described and discussed the Varying Probability Sampling Schemes (VPSS). All these sampling schemes have their own significance in the literature of sampling theory and are, therefore, found useful for different types of situations, either due to their theoretical and conceptual strengths or due to their convenience in operational procedures.

Besides these sampling schemes, there is a sampling scheme in the literature of sampling theory, which, in true sense, neither belongs to the category of Equal Probability Selection Method nor to the category of Varying Probability Sampling Schemes (VPSS). This sampling scheme is popularly known as ***"Systematic Sampling (SYS)"*** which has been observed to have its applications in many specific types of populations. More elaborately it, instead of selecting all the units randomly from the population, as we do in Simple Random Sampling scheme; selects only the first unit randomly and according to some specific rule of selection, other units are automatically included in the sample. Owing to this fact, Systematic Sampling (SYS) scheme has been proved to be very much convenient in practice which consumes less time and resource for selecting the sample of required size as compared to SRS scheme.

The present unit focusses on the Systematic Sampling scheme with its related theories / concepts and necessity for applying it in some specific type of populations, its operational procedure of selecting a sample of fixed size, say, n, from the population and estimation technique(s) which is/are used to estimate some of the population parameters. Section 9.2 deals with the concept of Systematic Sampling scheme, its definition, meaning of Systematic Sampling and different types of it, like, Linear Systematic Sampling (LSYS) Scheme and Circular Systematic Sampling (CSYS) scheme. Section 9.3 presents the details of Linear Systematic Sampling scheme and the arrangement of units of the population in the form of an array consisting of some rows and columns. This Section also explains how the Linear Systematic Sampling scheme resolved a problem of selecting a random sample through SRS scheme in some special types of populations. It also discusses the advantages and disadvantages of the Systematic Sampling (SYS) scheme as compared to Simple Random Sampling scheme. Section 9.4 deals with the problem of estimation of the population mean with the help of Linear Systematic Sampling sample. The sampling variance of the suggested estimator has also been derived. Section 9.5 presents the comparison of the Linear Systematic Sampling scheme with respect to Simple Random Sampling and Stratified Random Sampling schemes in respect of their efficiencies and derives the necessary conditions for it. It is seen that the performance of the Linear Systematic Sampling scheme as compared to Stratified Random Sampling and Stratified Random Sampling schemes mainly depend on the nature of the population and on the positions of units in it. In this respect, it becomes necessary to investigate the performance of the Linear Systematic Sampling scheme in different types of populations. Such a comparison of the above-mentioned three sampling schemes has been made in the Section 9.6, for a special type of population in which values of the study variable exhibit an increasing linear trend.

It has been explained in the early discussions made in this unit that Linear Systematic Sampling scheme is sometimes not suitable for populations in which the population size is not an integer multiple of the sample size. In such situations, Circular Systematic Sampling (CSYS) Scheme is found to be helpful in getting unbiased estimator of the population mean. Section 9.7 has been, therefore, devoted to the study of Circular Systematic Sampling scheme and explains what we mean by Circular Systematic Sampling and what would be the pattern of the arrangement of units. The problem of estimation of the population mean has also been taken up and two different types of estimators have been defined. The properties of these estimators have been mentioned.

Expected Learning Outcomes

After studying this unit, you should be able to:

- ❖ explain the Systematic Sampling (SYS) scheme as an alternative to Simple Random Sampling (SRS) scheme for using it to select a random sample from some special types of populations;
- ❖ define the two different types of Systematic Sampling (SYS) scheme, namely, Linear Systematic Sampling (LSYS) scheme and Circular Systematic Sampling (CSYS) scheme and describe how the units are arranged for selecting samples from these schemes;
- ❖ describe the methods of selection of units of the population in both the schemes for including them in a sample of fixed size;
- ❖ describe the method of obtaining an estimator for the population mean under both the schemes and discuss their properties;
- ❖ derive the sampling variance expressions for the corresponding estimators under both the schemes;
- ❖ compare the Linear Systematic Sampling and Circular Systematic Sampling schemes with Simple Random Sampling and Stratified Random Sampling schemes for their respective efficiencies;
- ❖ derive the necessary conditions under which Systematic Sampling scheme is more efficient than Simple Random Sampling and Stratified Random Sampling schemes; and
- ❖ discuss the performance of Linear Systematic Sampling scheme and compare it with Simple Random Sampling and Stratified Random Sampling schemes in such populations which follow a specific pattern in respect of the variate-values of the study variable.

9.2 SYSTEMATIC RANDOM SAMPLING SCHEME

Although, Systematic Sampling scheme is recognized as a probability sampling scheme, which is found to have a lot of applications in the estimation of population parameters in such populations which exhibit some peculiar types of nature; it is only one of the sampling schemes which does not follow the theory of Simple Random Sampling scheme in strict sense and is found to be practically handy and more efficient too than Simple Random Sampling for these types of populations. However, this deviation in Systematic Sampling scheme from Simple Random Sampling scheme, in selecting the units in the sample, should not be thought of as an arbitrarily taken decision but, in fact, is genuinely designed with some intuitive reasons to deal with some specific type of populations.

In other words, it can be stated that Systematic Sampling was initially invented by survey scientists for collecting a sample from the population where the direct application of Simple Random Sampling for this purpose, created practical difficulties in the operational procedure of sample selection. You should not see the Systematic Sampling scheme to be the only example of deviation from Simple Random Sampling scheme with respect to the process of selecting a random sample. We have seen earlier that Stratified Random Sampling scheme is also an example of such deviation from Simple Random Sampling scheme. Virtually, Stratified Random Sampling scheme, in comparison to Simple Random Sampling scheme, also deviates in the process of selecting a random sample of fixed size from the population. Since, the population is necessarily available in the form of divided sub-populations (strata), Simple Random Sampling samples are selected independently from each of the sub-populations using some specific rules of selection of units from each sub-population and then combined them together to form the fixed size sample. This is an indication that, depending upon the nature of the concerned population, wherever necessary the most fundamental scheme, that is, Simple Random Sampling scheme might undergo to suitable changes so as to generate a new sampling scheme which might be observed to be suitable for that population. Systematic Sampling scheme is an example of such change in Simple Random Sampling scheme.

9.2.1 Systematic Sampling Scheme

As mentioned in the Introductory Section 9.1, Systematic sampling scheme is quite different from the Simple Random Sampling and Stratified Random Sampling schemes in respect of selection procedure of a random sample from the population. In both the Simple Random Sampling and Stratified Random Sampling schemes, units of the population are selected in the sample of pre-determined size one by one in a number of draws using some chance mechanism; ensuring the selection probability to be equal in each draw. In Simple Random Sampling only one random sample is selected whereas in Stratified Random Sampling the number of such random samples are equal to the number of strata.

Contrast to this, in Systematic Sampling scheme in order to select a random sample of fixed size, units are thought to form an array consisting of some rows and columns or lying on the circumference of a circle. Following the specified rule of selection of units, only one unit is selected randomly using a chance mechanism either from the first row or the first column in the first case where the units are arranged in the form of an array and then all the units lying on that particular row or column are supposed to be automatically selected in the sample. In the second case, where units are supposed to form a circle, only one unit out of all the units of the population is selected randomly using a chance mechanism and then following the specified rule, other units are supposed to be selected in the sample so that the pre-determined size of the sample is achieved.

From the brief description of Systematic Sampling scheme made above, it is clear that the described method of selection of sample applies the concept of randomness during the selection process only partially. However, on the basis of the discussions made so far in this sub-section, we can define it as follows:

Given a finite population of units, if the sample of a fixed size is selected from the population in such a way that only the unit drawn at the first draw is selected randomly using a chance mechanism and rest of the units are assumed to be selected automatically in the sample according to some pre-designed pattern; then it is called ***"Systematic Random Sampling"*** *or simply* ***"Systematic Sampling"*** *(SYS) scheme.*

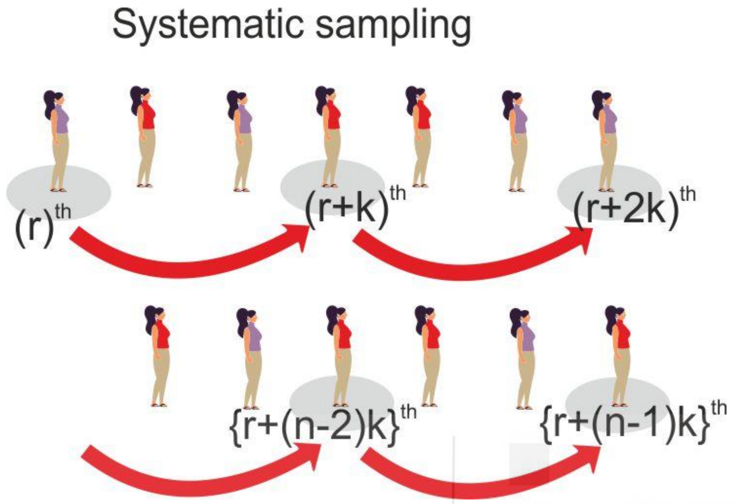

Fig. 9.1: Systematic Sampling Scheme

Fig. 9.1 shows how the Systematic Sampling scheme works. If the first unit selected is the r^{th} unit of the population, then every k^{th} unit after the first selection id assumed to be selected automatically in the sample.

Remark 9.1: At first sight, it seems that Systematic Sampling does not provide us a random sample as Simple Random Sampling and Stratified Random Sampling schemes do, but this is not true. We shall discuss under Remark 9.2 in sub-section 9.3.1 about the probability structure of the units while selecting units in the sample using the rule specified in the Systematic Sampling scheme. In fact, initially Systematic Sampling scheme came into existence as an alternative method of sample selection to Simple Random Sampling scheme in a specific type of survey which created operational difficulty in applying the method of selection used in Simple Random Sampling scheme. The detailed description of the invention of Systematic Sampling scheme as an alternative to Simple Random Sampling scheme will be made in Sub-section 9.3.2.

9.2.2 Types of Systematic Sampling Scheme

Virtually, Systematic Sampling scheme is comprised of two different methods of selection of a random sample of fixed size. These are:

(i) **Linear Systematic Sampling** and

(ii) **Circular Systematic Sampling**.

Basically, the difference between these two types of Systematic Sampling scheme occurs only at the stage of sample selection, due to the arrangement of the population units in two different ways because of some genuine reasons. Theoretically, there is no difference between these two types as far as the estimation stage of the sampling scheme is concerned.

However, in order to give more emphasis to the selection stage in Systematic Sampling scheme, which is the sole theme in this scheme, we shall present the theories and concepts of the two types in two different sections. Whereas the next section starts the discussion of the Linear Systematic Sampling Scheme, the concept and theories related to Circular Systematic Sampling shall be presented in the Section 9.7.

Now, you may try to answer the following Self-Assessment Question:

SAQ 1

What is the basic difference between Systematic Sampling and Simple Random Sampling schemes?

9.3 LINEAR SYSTEMATIC SAMPLING SCHEME

The "Systematic Sampling" scheme is most frequently known as "Linear Systematic Sampling" (LSYS) scheme in order to differentiate it with the "Circular Systematic Sampling" scheme. Linear Systematic Sampling scheme is generally useful for application when the condition $\mathbf{N = n.k}$ is fulfilled, where N and n, respectively, stand for the population size and the size of the sample which is supposed to be selected from the population and k is an integer. In other words, for applying the Linear Systematic Sampling scheme in a sample survey, the population size must be an integer multiple of the sample size.

9.3.1 Selection Procedure of a Sample in Systematic Sampling Scheme

Let a complete and exhaustive list (sampling frame) of the units of the given population be either available from some previous record or can be prepared by the survey scientist himself/herself. Let the relation, $N = n.k$ be fulfilled for the sizes of the given population and the sample size. In this case, first of all, the population units are linearly ordered, that is, units of the population are assigned labels 1, 2, 3, …, N as usual, without considering the order in which they appeared in the population. Let the label i be assigned to the i^{th} unit of the population occurred in the sampling frame where i = 1, 2, …, N. Then, the N labels of the units of the population are arranged in the form of an array consisting of n rows and k columns, as follows:

Array of N (= n.k) Labels assigned to Units of the Population

Rows	Columns						
	1	**2**	**3**	**….**	**r**	**….**	**k**
1	1	2	3	….	r	….	k
2	k+1	k+2	k+3	….	k+r	….	2k
3	2k+1	2k+2	2k+3	….	2k+r	….	3k
….	….	….	….	….	….	….	….
….	….	….	….	….	….	….	….
J	(j-1) k+1	(j-1) k+2	(j-1) k+3	….	(j-1) k+r	….	Jk
….	….	….	….	….	….	….	….
….	….	….	….	….	….	….	….
n	(n-1) k+1	(n-1) k+2	(n-1) k+3	….	(n-1) k+r	….	nk

In the above array, remember that each row consists of k number of labels and each column consists of n number of labels, where k is the multiplier of n. We shall now describe the method of selecting a sample of size n in Systematic Random Sampling. As described, the first unit in the process of selecting a sample of size n has to be selected randomly using some chance mechanism and rest of the units are automatically selected in the sample following some specific rule. How to complete the selection process then?

In fact, we observe that the arrangement of units in the array provides us k columns each of size n for each row. Therefore, in order to select a random sample of size n such that the first unit is selected at random using some chance mechanism and rest units are selected automatically, it would be sufficient to select a column randomly out of k columns. It suggests to use random number tables for selecting a random number, say r, where $1 \le r \le k$. Then, the r^{th} column of the array is said to be selected completely, that is, n labels; {r, k+ r, 2k+ r, …. , (j-1)k+ r, …., (n-1)k+ r} are selected for the sample of size n, where the first label r is selected randomly using a chance mechanism and other labels are selected automatically under a specified rule. In fact, the specified rule in Systematic Random Sampling scheme is that after the first draw, every k^{th} unit will be selected in the sample of size n. This method of selection of units from the population, thus, provides us a systematic sample of size n. Similarly, let the random number selected be 3; where the condition $3 \le k$ is satisfied, then, the labels selected in the sample of size n would be {3, k +3, 2k +3, …, (j-1) k+ 3, …, (n-1) k+ r}.

Fig. 9.2: Linear Systematic Sampling Scheme

Remark 9.2: In the array, we have k columns, each consisting of n labels and as per pre-specified rule, any one of the columns can be selected randomly. This means that all the columns have an equal probability of being selected which is 1/k. Thus, the technique will generate k systematic samples of size n, each having the probability of selection 1/k. Hence, with reference to the Section 5.5 of Unit 5 of Block 1, the inclusion probability of any particular unit in the r^{th} systematic sample is

$$\frac{1}{k} = \frac{1}{N/n} = \frac{n}{N} = \pi_j,$$

which is the result obtained in Probability Proportional to Size without Replacement scheme and also in SRS scheme.

Further, we observe that the chance of two different units, say, U_{rj} and $U_{rj'}$, both belonging to the r^{th} systematic sample, are included in the sample is obviously given by $\pi_{jj'} = \dfrac{1}{k}$.

However, if two units U_{rj} and $U_{sj'}$ belong to two different samples, say, r^{th} and s^{th} systematic samples, then $\pi_{jj'} = 0$.

So, virtually, Systematic Random Sampling is also a probability sampling scheme which ensures equal probability of inclusion of each unit in the sample and at the same time, as you see, is operationally more convenient in practice than Simple Random Sampling scheme.

Remark 9.3: In the selection process in linear systematic sampling, as described above, the number r, which is selected randomly in the first draw, is said to be **"Random Start"** and the number k is called the **"Sampling Interval"**.

Let us use the following examples to clarify the process of linear systematic sampling:

Example 1: (a) Let from a population of size 300, a systematic sample of size 50 is to be selected. Using linear systematic sampling scheme, select the sample if the first label selected is 3.

(b) Using linear systematic sampling scheme, select a sample of size 25 from the population of size 150 given that the first label selected is 5.

Solution:

(a) We have N = 300 and n = 50. Therefore, we have N = 6xn, that is, k, the sampling interval, is 6.

Here, given that random start, r = 3. Then other labels selected are

r +k = 9, 2k+r = 15, 3k+r = 21 and so on.

Thus, the sample of size 50 will include labels:

3, 9, 15, 21, 27, 33, 39, 45, 51, 57, 63, 69, 75, 81, 87, 93, 99, 105, 111, 117, 123, 129, 135, 141, 147, 153, 159, 165, 171, 177, 183, 189, 195, 201, 207, 213, 219, 225, 231, 237, 243, 249, 255, 261, 267, 273, 279, 285, 291, 297.

(b) Here, N = 150, n = 25 and random start, r = 5; therefore, the sampling interval will be 6.

The selected labels are, therefore,

5, 11, 17, 23, 29, 35, 41, 47, 53, 59, 65, 71, 77, 83, 89, 95, 101, 107, 113, 119, 125, 131, 137, 143, 149.

Example 2: Let from a population of size N, a linear systematic sample of size n be selected. The selected labels are arranged in a haphazard way which are given as:

54, 144, 74, 104, 34, 4, 164, 84, 64, 14, 24, 94, 174, 44, 114, 154, 124, 134.

Find the population size, sample size, random start and sampling interval.

Solution: Since the sample is a linear systematic sample, we have the relation N = nk.

In the sample there are 18 labels, therefore, we observe that n = 18. The smallest label selected is 4 so the random start r is 4. This indicates that the fourth column is randomly selected and if the labels are arranged in ascending order of magnitude, the selected labels will be

4, 14, 24, 34, 44, 54, 64, 74, 84, 94, 104, 114, 124, 134, 144, 154, 164, 174

which are 18 in number. This indicates that the sampling interval k would be 10.

Therefore, the population size

$$N = nk = 18 \times 10 = 180.$$

Remark 9.4: Resemblance with Stratified Sampling scheme

It can be shown that linear systematic sampling scheme resembles with Stratified Random Sampling scheme. If the rows in the array are considered to be strata, systematic sampling is equivalent to choosing one unit at random from the first stratum and then choosing units having the same relative positions from subsequent strata. So, the Systematic Sampling scheme is comparable with Stratified Random Sampling scheme with one unit per stratum, considering rows of the array as strata. However, this resemblance of Systematic Sampling scheme with Stratified Random Sampling scheme is only casual, since, in Stratified Random Sampling scheme, units from each stratum are selected randomly whereas here in Systematic Sampling scheme its position relative to the unit in the first stratum is pre-determined.

9.3.2 Systematic Sampling Scheme as an Alternative to Simple Random Sampling Scheme

The Systematic Sampling scheme initially came into existence as an alternative to Simple Random Sampling (SRS) in some specific types of surveys, particularly in “Forest Surveys” and “Land Use Surveys”. There are a number of references on the use of systematic sampling made by the statisticians during its early years of development such as, Madow and Madow (1944), Madow (1946, 1949, 1953) and Finney (1948); who illustrated the use of Systematic Sampling (SYS) scheme either in forest surveys for estimating the volume of timber in a dense forest or in land use surveys for estimating the volume of cultivation. In both of these types of surveys, if the use of Simple Random Sampling (SRS) scheme would have been made, the main difficulty faced by the surveyors was the inability to approach the selected units for their at the spot enumeration. In forest surveys, where trees were found to be so thickly populated that to reach the interior parts of the forest to locate the trees, chosen in the sample using SRS scheme, was a very difficult task due to the close surrounding of other trees. Similar problem was found in land use surveys also where to approach the cultivated lands selected in the sample was quite a difficult task due to non- existence of approaching path to these lands. In this sense, Systematic Sampling was an invention in its earlier days which was thought to be an alternative to the Simple Random Sampling (SRS) scheme.

In reference to Forest Surveys for estimating volume of timber, we can explain what operational advantage was obtained in applying Systematic Sampling in such cases. Consider the following diagram of a dense forest:

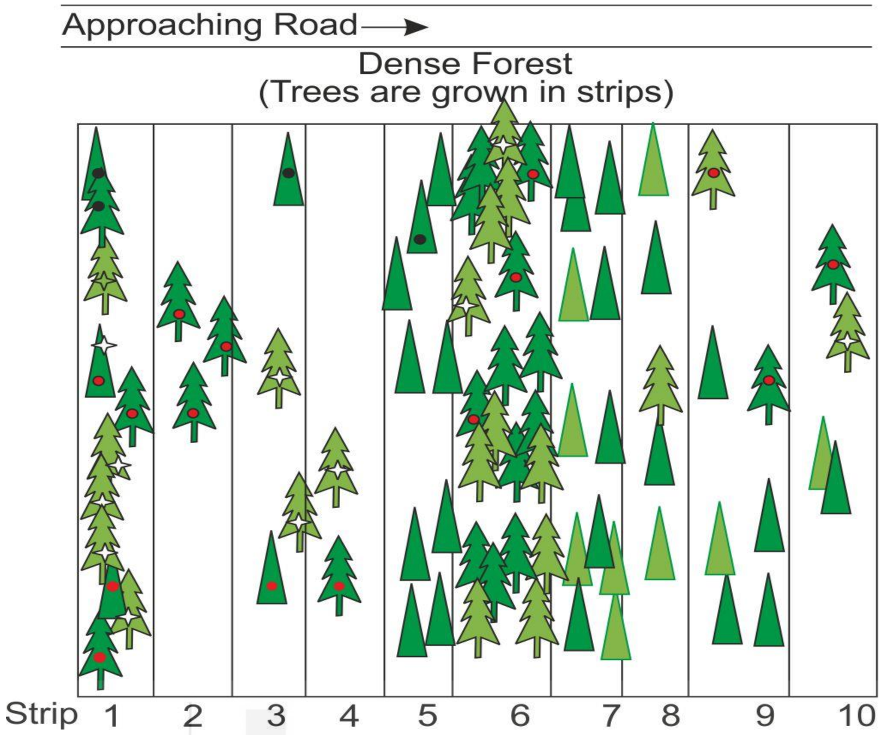

Fig. 9.3: Diagrammatic Representation of a Forest Survey

Consider the above diagram of a thickly populated forest in which trees, shown by the figures grew too close to each other in strips. There is no crossing path through which one can enter into the forest except only one approaching road parallel to one side of the forest. Let the problem be to select a random sample of trees in order to estimate the volume of timber in the forest. Let a Simple Random Sampling (SRS) sample be selected randomly, and the selected units are denoted by dots (•). We see that sampled trees are scattered over the entire forest and, therefore, not approachable easily due to the thickness of the forest and non-availability of any path inside the strips. On the other hand, even if trees which are deeply situated in the forest, like trees denoted by (✧) are approached by any means, it is not simple to cut down these trees into pieces for weighing them at its own position due to the reason that it is closely surrounded by other trees from all the directions and, therefore, it will not fall down completely to the ground. This is a genuine operational problem, which has to be resolved. Therefore, some alternative method to Simple Random Sampling (SRS) scheme has to be developed.

Now, let us think to sample the trees in another way. Since, as per the diagram, there are 10 strips of trees, one can approach to the start of any of these 10 strips through the approaching road. To select any of the 10 strips randomly, select a random number less than equal to 10 from the random number table. Let it be 6, so that strip 6 is randomly selected and all the trees belonging to strip 6 are then supposed to be selected automatically in the sample. Now, we can see what benefit we gain by selecting all the trees of the selected strip in the sample. The very first tree in this strip just to the roadside can be allowed to fall on the roadside as there would be ample space for it and then to cut down easily into small pieces for taking their weights. After removing the small parts of this tree, the next tree behind it in the strip can easily be cut down and be allowed to fall on the same side. Following this process, the sampler gets the approach into the strip for cutting down all the trees of the strip one by one till the end.

Thus, in this type of process of sampling, we see that only the first unit is required to be selected randomly and others are automatically selected in the sample following a pre-designed rule. In this sense, Systematic Sampling scheme can be thought of as a practical solution of the genuine operational problem which was mentioned above. Also, it is expected to be less time-consuming and less costly as compared to Simple Random Sampling (SRS) scheme, since all the sampled units are very near to each other and easy to enumerate.

The linear Systematic Sampling (SYS) scheme has a lot of applications in many other types of populations for selecting a random sample in less time with less effort.

9.3.3 Advantages and Disadvantages of Systematic Sampling Scheme

Since, Systematic Sampling scheme is seen as an alternative to Simple Random Sampling scheme, it is not out of place to compare Systematic Sampling scheme with Simple Random Sampling scheme for its advantages and disadvantages. The advantages of this scheme are as follows:

(i) The main advantage of Systematic Sampling scheme over Simple Random Sampling scheme is its simplicity in sample selection, more speedy than Simple Random Sampling scheme in all respects, operational convenience and almost evenly spread of the sampled units over the entire population due to a suitable arrangement of units of the population before starting the selection process.

(ii) Systematic Sampling scheme provides no scope for personal bias or mistakes in the sample selection process.

(iii) It can even be used when a sampling frame of the population units is not readily available, since a Systematic Sampling sample can be selected by serially numbering the units while progressively listing them.

(iv) Systematic Sampling scheme is sometimes seen more advantageous than Simple Random Sampling scheme in terms of producing efficient estimates in some populations which exhibit some special types of nature. Examples of its use are: (i) in fisheries for estimating the total catch of marine fish in India, (ii) in geological mine sampling, in spatial sampling, (iii) in industrial sampling, (iv) in selecting a sample of books in a library using catalogue cards and others.

The disadvantages of Systematic Sampling scheme are as follows:

(i) Systematic Sampling scheme is the worst kind of sampling scheme for selecting a random sample in a population which shows a periodicity trend for the appearance of its units, since, then if the care is not taken in fixing the sampling interval in a suitable manner, the same types of units might be appearing in the sample.

(ii) Another disadvantage of this scheme is that the estimator of the sampling variance is obtained with a single sample.

Now, you may try to answer the following Self-Assessment Question:

SAQ 2

Define the Linear Systematic Sampling (LSYS) scheme.

9.4 ESTIMATION OF POPULATION MEAN AND VARIANCE OF SAMPLE MEAN ESTIMATOR

We shall now discuss and describe the problem of estimation of the population mean when a random sample is selected through Linear Systematic Sampling scheme.

9.4.1 Estimation of Population Mean

In Linear Systematic Sampling scheme, we know that $N = n * k$. In order to estimate the population mean on the basis of a sample selected through Linear Systematic Sampling scheme, we shall again consider here the array of $n.k$ units shown in the sub-section 9.3.1 by replacing the labels of units by their values on the study variable Y. Thus, we have the following array showing the values of the variable under study:

Array of n.k(=N) Units of the Population showing the Values of Study Variable Y

Rows			**Columns**				
	1	2	3	….	r	….	k
1	Y_{11}	Y_{21}	Y_{31}		Y_{r1}		Y_{k1}
2	Y_{12}	Y_{22}	Y_{32}		Y_{r2}		Y_{k2}
3	Y_{13}	Y_{23}	Y_{33}		Y_{r3}		Y_{k3}
….	….	….	….		….		….
….	….	….	….		….		….
j	Y_{1j}	Y_{2j}	Y_{3j}		Y_{rj}		Y_{kj}
….	….	….	….		….		….
….	….	….	….		….		….
n	Y_{1n}	Y_{2n}	Y_{3n}		Y_{rn}		Y_{kn}

In the above array, obviously, Y_{rj} denotes the value of the variable Y on the r^{th} unit belonging to the j^{th} row, where $r = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n$. Since, columns 1 to k are, in fact, showing all the possible systematic samples of size n, which can be selected from the population, r^{th} column actually shows that we are talking about the r^{th} systematic sample for $r = 1, 2, \ldots, k$. Let us find the sample mean denoted by $\bar{y}_{ro}$, of the r^{th} systematic sample. Clearly, it will be given by

$$\bar{y}_{ro} = \frac{1}{n}\sum_{j=1}^{n} Y_{rj}; \quad \text{for } r = 1, 2, \ldots, k. \qquad \ldots (9.1)$$

Since, the r^{th} column is randomly chosen, as usual the sample mean of the r^{th} systematic sample will be a random variable with probability of selection equal to $1/k$.

As usual, the sample mean $\bar{y}_{ro}$ is considered to be the immediate estimator of population mean, given by $\bar{Y} = \frac{1}{nk}\sum_{r=1}^{k}\sum_{j=1}^{n} Y_{rj}$

Theorem 1: In the Linear Systematic Sampling with sampling interval k, the sample mean $\bar{y}_{ro}$ is an unbiased estimator of population mean $\bar{Y}$.

Proof: Since the probability of selection of the r^{th} systematic sample is 1/k, we have

$$E(\bar{y}_{ro}) = \frac{1}{k}\sum_{r=1}^{k}\bar{y}_{ro} = \frac{1}{k}\sum_{r=1}^{k}\left(\frac{1}{n}\sum_{j=1}^{n} Y_{rj}\right) = \frac{1}{nk}\sum_{r=1}^{k}\sum_{j=1}^{n} Y_{rj} = \bar{Y} \qquad \dots (9.2)$$

which means that the sample mean in Linear Systematic Sampling is an unbiased estimator of the population mean.

9.4.2 Sampling Variance of the Estimator

The sampling variance of the Linear Systematic Sampling estimator of population mean, $\bar{y}_{ro}$ is obtained in the following theorem:

Theorem 2: The sampling variance of the estimator $\bar{y}_{ro}$ is given by

$$V(\bar{y}_{ro}) = \frac{(k-1)}{k} S_C^2; \qquad \dots (9.3)$$

where $S_C^2 = \frac{1}{k-1}\sum_{r=1}^{k}\left(\bar{y}_{ro} - \bar{Y}\right)^2$ denotes the population mean square between the column means in the population.

Proof: As per definition of the variance of a random variable, we have

$$V(\bar{y}_{ro}) = E(\bar{y}_{ro} - \bar{Y})^2 = \frac{1}{k}\sum_{r=1}^{k}(\bar{y}_{ro} - \bar{Y})^2. \qquad \dots (9.4)$$

Let us denote the mean square between the column means by S_C^2, that is,

$$S_C^2 = \frac{1}{k-1}\sum_{r=1}^{k}(\bar{y}_{ro} - \bar{Y})^2 \qquad \dots (9.5)$$

Then, (9.4) reduces to

$$V(\bar{y}_{ro}) = \frac{(k-1)}{k} S_C^2. \qquad \dots (9.6)$$

This establishes the theorem.

An Alternative Expression of $V(\bar{y}_{ro})$:

The expression of the sampling variance of the sample mean estimator can be obtained in a different form. The following theorem shows how it can be derived:

Theorem 3: The variance of the estimator $\bar{y}_{ro}$ can also be expressed as

$$V(\bar{y}_{ro}) = \frac{1}{N}\left[(N-1)S^2 - k(n-1)S_{wsy}^2\right]; \qquad \dots (9.7)$$

where $S_{wsy}^2 = \frac{1}{k(n-1)}\sum_{r=1}^{k}\sum_{j=1}^{n}(Y_{rj} - \bar{y}_{ro})^2,$

stands for the variance among units belonging to the same systematic sample or mean square within systematic sample.

Proof: We know that by definition, the mean square of the population is given by

$$S^2 = \frac{1}{N-1}\sum_{r=1}^{k}\sum_{j=1}^{n}(Y_{rj} - \bar{Y})^2$$

Hence, we have

$$(N-1)S^2 = \sum_{r=1}^{k}\sum_{j=1}^{n}(Y_{rj} - \bar{Y})^2$$

$$= \sum_{r=1}^{k}\sum_{j=1}^{n}(Y_{rj} - \bar{y}_{ro} + \bar{y}_{ro} - \bar{Y})^2$$

$$= \sum_{r=1}^{k}\sum_{j=1}^{n}(Y_{rj} - \bar{y}_{ro})^2 + \sum_{r=1}^{k}\sum_{j=1}^{n}(\bar{y}_{ro} - \bar{Y})^2$$

$$+2\sum_{r=1}^{k}\sum_{j=1}^{n}(Y_{rj} - \bar{y}_{ro})\left(\bar{y}_{ro} - \bar{Y}\right). \qquad \dots (9.8)$$

But, we see that the first term is,

$$\sum_{r=1}^{k}\sum_{j=1}^{n}(Y_{rj} - \bar{y}_{ro})^2 = k(n-1)S_{wsy}^2 \;;$$

and the second term is

$$\sum_{r=1}^{k}\sum_{j=1}^{n}(\bar{y}_{ro} - \bar{Y})^2 = \sum_{r=1}^{k} n(\bar{y}_{ro} - \bar{Y})^2;$$

being independent of j and the third term

$$2\sum_{r=1}^{k}\sum_{j=1}^{n}(Y_{rj} - \bar{y}_{ro})\left(\bar{y}_{ro} - \bar{Y}\right) = 2\sum_{r=1}^{k}\left(\bar{y}_{ro} - \bar{Y}\right)\sum_{j=1}^{n}(Y_{rj} - \bar{y}_{ro}) = 0,$$

as $\sum_{j=1}^{n}\left(Y_{rj} - \bar{y}_{ro}\right)$; being the sum of deviations of the values of the variable Y_{rj} in the r^{th} column from the respective mean $\bar{y}_{ro}$, would be zero.

Therefore, (9.8) reduces to

$$(N-1)S^2 = k(n-1)S_{wsy}^2 + n\sum_{r=1}^{k}(\bar{y}_{ro} - \bar{Y})^2$$

$$= k(n-1)S_{wsy}^2 + nkV\left(\bar{y}_{ro}\right); \qquad \text{due to (9.4).}$$

Therefore, from this expression, we get

$$V\left(\bar{y}_{ro}\right) = \frac{1}{N}\left[(N-1)S^2 - k(n-1)S_{wsy}^2\right]. \qquad \dots (9.9)$$

This establishes the theorem.

Try to answer the following Self-Assessment Question:

SAQ 3

Let the mean of a population of size N is to be estimated with the help of a systematic sample of size n, such the N = n.k, where k is an integer multiple of n. Show how would you proceed to select the sample of desired size and what estimator would you use for this purpose. Show that the estimator used is unbiased for the population mean.

9.5 EFFICIENCY COMPARISON WITH SIMPLE RANDOM SAMPLING AND STRATIFIED RANDOM SAMPLING SCHEMES

Since, Systematic Sampling scheme is considered to be an alternative to Simple Random Sampling scheme and also have a resemblance with the Stratified Random Samling scheme; it seems necessary to compare the Linear Systematic Sampling scheme with Simple Random Sampling and Stratified Random Samling schemes in respect of their efficiency. We shall derive the conditions under which Linear Systematic Sampling scheme would be preferable over Simple Random Sampling and Stratified Random Sampling schemes.

9.5.1 Comparison with Simple Random Sampling Scheme

The required condition is derived in the following theorem:

Theorem 4: The estimator of population mean in Linear Systematic Sampling scheme is more precise than compared to Simple Random Sampling scheme if $S^2 < S^2_{wsy}$.

Proof: From the expression (9.9), we have

$$V(\bar{y}_{ro}) = \frac{1}{N}\left[(N\text{-}1)S^2 - k(n-1)S^2_{wsy}\right]$$

Also, we have

$$V(\bar{y}) = \frac{(N-n)}{Nn}S^2$$

Therefore, $V(\bar{y}_{ro}) < V(\bar{y})$, if

$$\frac{1}{N}\left[(N-1)S^2 - k(n-1)S^2_{wsy}\right] < \frac{(N-n)}{Nn}S^2$$

or, when

$$\left(1-\frac{1}{N}\right)S^2 - \left(\frac{1}{n}-\frac{1}{N}\right)S^2 < \frac{k(n-1)}{nk}S^2_{wsy}$$

$$\Rightarrow \left(1-\frac{1}{n}\right)S^2 < \left(1-\frac{1}{n}\right)S^2_{wsy}$$

$$\Rightarrow S^2 < S^2_{wsy}.$$

This establishes the theorem.

Remark 9.5: The above condition states that the Linear Systematic Sampling scheme would be more efficient than Simple Random Sampling scheme when the variability within the systematic sample would be larger than the overall population variability. This means that the units within the same systematic sample should be as much heterogeneous as possible. In other words, the systematic sample should not consist of similar types of values but must contain values of the variable which are very much heterogeneous among themselves.

The two schemes, that is, Linear Systematic Sampling and Simple Random Sampling schemes may be compared for their efficiency in an alternative way. This method assumes that the intra-class correlation coefficient between the units of a column (systematic sample) is known. The following theorem provides the comparison of Linear Systematic Sampling and Simple Random Sampling schemes under this assumption:

Theorem 5: The Linear Systematic Sampling scheme is more efficient than Simple Random Sampling scheme when

$$\rho_{wsy} < -1/(nk-1)$$

where ρ_{wsy} stands for the intra-class correlation coefficient between the units of a column in the array of the variate-values.

Proof: We know that the variance of the sample mean estimator in Simple Random Sampling without Replacement scheme is given by

$$V(\bar{y}) = \frac{(N-n)}{Nn} S^2; \qquad \dots (9.10)$$

where S^2 is the mean square between the population units. Further, we know that the variance of the sample mean estimator in Linear Systematic Sampling is given by

$$V(\bar{y}_{ro}) = \frac{(k-1)}{k} S_C^2$$

In order to compare the two variances, it is first necessary to establish a relation between S^2 and S_C^2. Using the definition of the variance of a variable, we re-write $V(\bar{y}_{ro})$ as

$$V(\bar{y}_{ro}) = \frac{1}{k}\sum_{r=1}^{k}(\bar{y}_{ro} - \bar{Y})^2$$

$$= \frac{1}{k}\sum_{r=1}^{k}\left\{\frac{1}{n}\sum_{j=1}^{n} Y_{rj} - \bar{Y}\right\}^2$$

$$= \frac{1}{kn^2}\sum_{r=1}^{k}\left\{\sum_{j=1}^{n}\left(Y_{rj} - \bar{Y}\right)\right\}^2$$

$$= \frac{1}{kn^2}\sum_{r=1}^{k}\left\{\sum_{j=1}^{n}(Y_{rj} - \bar{Y})^2 + \sum_{j\neq j'=1}^{n}(Y_{rj} - \bar{Y})\left(Y_{rj'} - \bar{Y}\right)\right\}; \qquad \dots (9.11)$$

since, Square of Sums = Sum of Squares + Product terms.

We know that, by definition, population mean square, S^2, is defined as

$$S^2 = \frac{1}{nk-1}\sum_{r=1}^{k}\sum_{j=1}^{n}(Y_{rj} - \bar{Y})^2; \qquad \dots (9.12)$$

therefore, from (9.11), we have

$$V(\bar{y}_{ro}) = \frac{1}{kn^2}\left\{(nk-1)S^2 + \sum_{r=1}^{k}\sum_{j\neq j'=1}^{n}(Y_{rj} - \bar{Y})\left(Y_{rj'} - \bar{Y}\right)\right\} \qquad \dots (9.13)$$

In order to evaluate the product term inside the bracket of the expression (9.13), we use the concept of intra-class correlation coefficient between the units of the same systematic sample. As per definition, it is given by

$$\rho_{wsy} = \frac{E\left(Y_{rj} - \bar{Y}\right)\left(Y_{rj'} - \bar{Y}\right)}{E(Y_{rj} - \bar{Y})^2} \qquad \dots (9.14)$$

Since, by definition of expectation,

$$E(Y_{rj} - \bar{Y})^2 = \frac{1}{nk}\sum_{r=1}^{k}\sum_{j=1}^{n}(Y_{rj} - \bar{Y})^2 = \frac{(nk-1)S^2}{nk}$$

using (9.12), and

$$E\left(Y_{rj} - \bar{Y}\right)\left(Y_{rj'} - \bar{Y}\right) = \frac{\sum_{r=1}^{k}\sum_{j\neq j'=1}^{n}(Y_{rj} - \bar{Y})\left(Y_{rj'} - \bar{Y}\right)}{nk(n-1)};$$

we have

$$\sum_{r=1}^{k}\sum_{j\neq j'=1}^{n}(Y_{rj} - \bar{Y})\left(Y_{rj'} - \bar{Y}\right) = (n-1)(nk-1)\rho_{wsy}S^2 \qquad \dots (9.15)$$

Now, substituting from (9.15) in (9.13), we have

$$V\left(\bar{y}_{ro}\right) = \frac{(nk-1)}{nk}\frac{S^2}{n}\left[1 + (n-1)\rho_{wsy}\right]. \qquad \dots (9.16)$$

Therefore, using (9.10) and (9.16), we have

$$\frac{V\left(\bar{y}_{ro}\right)}{V\left(\bar{y}\right)} = \frac{(nk-1)\left[1 + (n-1)\rho_{wsy}\right]}{n(k-1)}. \qquad \dots (9.17)$$

This expression is useful for comparing the performance of the two sampling schemes. Since, N = nk is always pre-fixed, we can say that relative precision of the schemes depends on the value of the intra-class correlation coefficient ρ_{wsy}.

It can be observed that

(i) the Linear Systematic Sampling scheme and SRS scheme are equally efficient, if $\rho_{wsy} = -1/(nk-1)$;

(ii) the Linear Systematic Sampling scheme is less efficient than Simple Random Sampling scheme if $\rho_{wsy} > -1/(nk-1)$; and

(iii) the Linear Systematic Sampling scheme is more efficient than Simple Random Sampling scheme if $\rho_{wsy} < -1/(nk-1)$.

Moreover, variance of the estimator $\bar{y}_{ro}$ would be zero when ρ_{wsy} assumes its minimum value $-1/(n-1)$; therefore, in this situation Linear Systematic Sampling scheme is 100% efficient than Simple Random Sampling scheme. Similarly, for the highest value of ρ_{wsy}, which is 1; the relative precision of Linear Systematic Sampling scheme with respect to Simple Random Sampling scheme is $(k-1)/(N-1)$. This proves the theorem.

9.5.2 Comparison with Stratified Random Sampling Scheme

Let us now compare the relative efficiency of Linear Systematic Sampling scheme with Stratified Random Sampling (STRS) scheme.

We explained the resemblance between the Systematic Sampling and Stratified Random Sampling schemes in Remark 9.4 under Sub-section 9.3.1. We have seen that Linear Systematic Sampling scheme is equivalent to Stratified Random Sampling scheme if rows, in the array arrangement, are considered to be strata and from each of the n strata, one unit is randomly selected to constitute a stratified sample of size n.

Obviously, then size of the sample selected from each stratum would be, $n_j = 1$ for j = 1, 2,…, n. Further, we observe that in such situation the strata sizes would be, $N_j = k$ for all j and, hence, $W_j = N_j/N$ for j = 1, 2, …, n. Then, the sampling variance of the sample mean estimator, $\bar{y}_{STRS}$ would be

$$V(\bar{y}_{STRS}) = \sum_{j=1}^{n}(\frac{1}{n_j} - \frac{1}{N_j})W_j^2 S_j^2 = \sum_{j=1}^{n}(\frac{1}{1} - \frac{1}{k})\frac{N_j^2}{N^2}S_j^2 = \sum_{j=1}^{n}(1 - \frac{1}{k})\frac{k^2}{n^2k^2}S_j^2$$

or $$V(\bar{y}_{STRS}) = \frac{1}{n^2}\sum_{j=1}^{n}(1 - \frac{1}{k})\, S_j^2 . \qquad \ldots (9.18)$$

If we compare S_j^2, the mean square within the j^{th} stratum, appeared in the above expression, with S_i^2; the mean square within the i^{th} stratum, as defined in the Sub-section 6.4.1 of Unit 6, we see that it can be written as

$$S_j^2 = \frac{1}{k-1}\sum_{r=1}^{k}(Y_{rj} - \bar{Y}_{oj})^2,$$

where $\bar{Y}_{oj} = \frac{1}{k}\sum_{r=1}^{k}Y_{rj}$ represents the mean of the j^{th} stratum.

Substituting the expression of S_j^2 in (9.18), we have

$$V(\bar{y}_{STRS}) = \frac{1}{n}\left(1 - \frac{1}{k}\right)\frac{1}{n(k-1)}\sum_{r=1}^{k}\sum_{j=1}^{n}(Y_{rj} - \bar{Y}_{oj})^2 = \frac{1}{n}\left(1 - \frac{1}{k}\right)S_w^2;$$

where $S_w^2 = \frac{1}{n(k-1)}\sum_{r=1}^{k}\sum_{j=1}^{n}(Y_{rj} - \bar{Y}_{oj})^2$ represents the pooled mean square between units within rows.

From (9.3), we see that the variance of sample mean estimator under Linear Systematic Sampling scheme is

$$V(\bar{y}_{ro}) = \frac{(k-1)S_C^2}{k} = \frac{1}{k}\sum_{r=1}^{k}\{\sum_{j=1}^{n}(y_{rj} - \bar{Y})\}^2.$$

Further, we know that $\bar{Y} = \frac{1}{n}\sum_{j=1}^{n}\bar{Y}_{oj}$, therefore,

$$V(\bar{y}_{ro}) = \frac{(k-1)S_C^2}{k} = \frac{1}{kn^2}\sum_{r=1}^{k}\left[\sum_{j=1}^{n}(y_{rj} - \bar{Y}_{oj})\right]^2$$

$$= \frac{1}{kn^2}\left[\sum_{r=1}^{k}\sum_{j=1}^{n}(y_{rj} - \bar{Y}_{oj})^2 + \sum_{r=1}^{k}\sum_{j\neq j'=1}^{n}(y_{rj} - \bar{Y}_{oj})(y_{rj'} - \bar{Y}_{oj'})\right] \ldots (9.19)$$

Now, we define the non-circular serial correlation coefficient ρ_{wst}. It is, in fact, the correlation between deviations from strata means of pairs of units that are in the same systematic sample. It is given by the formula

$$\rho_{wst} = \frac{E\left(y_{rj} - \bar{Y}_{oj}\right)\left(y_{rj'} - \bar{Y}_{oj'}\right)}{E(y_{rj} - \bar{Y}_{oj})^2}$$

$$= \frac{\sum_{r=1}^{k}\sum_{j\neq j'=1}^{n}\left(y_{rj} - \bar{Y}_{oj}\right)\left(y_{rj'} - \bar{Y}_{oj'}\right)}{(n-1)n(k-1)S_w^2}; \qquad \ldots (9.20)$$

since, by definition

$$E\left(y_{rj} - \bar{Y}_{oj}\right)\left(y_{rj'} - \bar{Y}_{oj'}\right) = \frac{\sum_{r=1}^{k}\sum_{j\neq j'=1}^{n}\left(y_{rj} - \bar{Y}_{oj}\right)\left(y_{rj'} - \bar{Y}_{oj'}\right)}{(n-1)n(k-1)}$$

and

$$E(y_{rj} - \bar{Y}_{oj})^2 = \frac{1}{n(k-1)}\sum_{r=1}^{k}\sum_{j=1}^{n}(Y_{rj} - \bar{Y}_{oj})^2 = S_w^2$$

Therefore, from (9.19), we have

$$V(\bar{y}_{ro}) = \frac{1}{kn^2}\left[n(k-1)S_w^2 + n(n-1)(k-1)S_w^2\rho_{wst}\right]$$

or,

$$V(\bar{y}_{ro}) = \frac{(k-1)}{nk}S_w^2\left[1 + (n-1)\rho_{wst}\right]. \qquad \ldots (9.21)$$

Thus, we have

$$\frac{V(\bar{y}_{STRS})}{V(\bar{y}_{ro})} = 1/\left[1 + (n-1)\rho_{wst}\right].$$

This states that the relative precision of Linear Systematic Sampling scheme over Stratified Sampling scheme depends upon the value of ρ_{wst}. For a positive value of it, Stratified Sampling scheme will provide an estimate with more precision as compared to Linear Systematic Sampling scheme. If it assumes value zero, then both schemes are equally efficient.

Remark 9.6: From the variance comparison, it is evident that the results depend upon the non-circular serial correlation coefficient, which is not easy to obtain. The dependency of results on this correlation does not lead us to general Conclusions.

On the basis of the contents of the above section, provide the answer of the following Self-Assessment Question:

SAQ 4

If ρ_{wsy} stands for the intra-class correlation coefficient between the units of a column in the array of the variate-values, then obtain the condition under which the Linear Systematic Sampling scheme is more efficient as compared to Simple Random Sampling scheme.

9.6 EFFICIENCY COMPARISON WITH SIMPLE RANDOM SAMPLING AND STRATIFIED RANDOM SAMPLING SCHEMES FOR SPECIFIC TYPE OF POPULATION

By now, perhaps you might be familiar with the fact that the performance of Systematic Sampling scheme, as compared to Simple Random Sampling and Stratified Random Sampling schemes, mainly depend upon the order of arrangement of population units, because Systematic Sampling scheme needs specific arrangements of population units before selecting the units from the population in order to constitute a random sample. We have seen that in Linear Systematic Sampling scheme, since, only one of the columns in the array arrangement, is randomly selected as the sample of size n; this sample might be worse if, by chance, the column includes units with almost similar values of the study variable so that the heterogeneity of the population is not reflected in the selected sample; or might be comparatively better if it consists of units with larger variability among themselves so that the sample is a true representative of the population variability. Clearly, this might be possible theoretically only if the survey statistician is having complete knowledge of the variate values for all the units; which is never in any sample survey.

It seems necessary, therefore, to have an idea about the nature of the population values, as much as possible, before ordering the units in the form of an array or in a circular form. Although, we come across with many types of populations with different nature, but we shall discuss here only a particular type of population for which we shall prove the conditions under which the Linear Systematic Sampling, Simple Random Sampling and Stratified Random Sampling schemes would be useful for application.

However, before this, we have presented a brief mention of some other types of populations below which can also be taken into consideration for obtaining similar conditions for the three sampling schemes:

(a) Population in Random Order:

In most of the cases of use of Systematic Sampling scheme, nothing is known in advance to the surveyor about the distribution of units in the population or about the nature of the study variable. In such cases, therefore, we can assume that the population is in random order and there is no sign that it exhibits any specific trend in respect of the variate values. Therefore, it is very difficult to guess what the result of Linear Systematic Sampling scheme would be if it is applied for sample selection and for estimation purpose.

(b) Population with Periodic Variation:

Sometimes, we face populations in which sampling units with high and low values follow one another according to a regular pattern. Such populations are called “Population with periodic variation”. For example, let variate-values follow the following periodic rule:

$$Y_r = \sin\left\{\theta + \frac{r\,\pi}{n}\right\}$$

where r varies from 0 to an integral multiple of 2n, the periodicity of the curve. In such a case, the values of the variable show a sin curve which has ups and downs at regular intervals. If the points showing exactly the same trend on this sin curve, are selected in the systematic sample, then such a sample would be as good as a single value. On the other hand, if points having just opposite trend that the nature of the previous unit, this sample would produce more precise estimates. Thus, in a population showing a periodic variation, it is most important to focus on the choice of suitable sampling interval keeping in view the periodicity of the variation of the values. However, in practice, we rarely come across with such populations.

(c) Auto-Correlated Population:

It is a general consensus that in a population, units which are very closer to each other exhibit a high degree of correlation in comparison to the units which are quite apart from each other. In practice, such populations are frequently observed. As for example, prices of commodities in the market in two or three consecutive years are almost alike to each other, thus, exhibiting a high correlation among them; while those which are at a distance of 5 to 10 years or more exhibit a less degree of correlation. A population in which closer units shows a high degree of correlation between them are popularly called *"Auto-Correlated Populations".* While applying Linear Systematic Sampling scheme for selecting a random sample in such populations, the efficiency of this scheme depends upon the autocorrelation of units and, therefore, conditions can be mathematically obtained for the choice of a suitable sampling scheme.

(d) Populations with Linear Trend:

Let us assume that in a population the variate- values show an increasing linear trend which can be approximated by the following linear model:

$$Y_i = \alpha + i\theta; \quad \text{for } i = 1, 2, \ldots, N, \qquad \ldots (9.22)$$

where Y_i expresses the value of the i^{th} unit in the population and α and θ are two positive constants. Clearly then, the population mean $\bar{Y}$ would be

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N}(\alpha + i\theta).$$

$$= \frac{1}{N}.N\alpha + \frac{\theta}{N}\sum_{i=1}^{N} i$$

$$= \alpha + \frac{\theta}{N} \times \frac{N(N+1)}{2}$$

$$= \alpha + \frac{N+1}{2}\theta \qquad \ldots (9.23)$$

Further, we have the population mean square as,

$$S^2 = \frac{1}{N-1}\left\{\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right\}. \qquad \ldots (9.24)$$

But, we know that

$$\sum_{i=1}^{N} Y_i^2 = N\alpha^2 + \theta^2 \sum_{i=1}^{N} i^2 + 2\alpha\theta \sum_{i=1}^{N} i$$

$$= N\alpha^2 + \theta^2 \frac{N(N+1)(2N+1)}{6} + 2\alpha\theta \frac{N(N+1)}{2}$$

$$= N\alpha^2 + \theta^2 \frac{N(N+1)(2N+1)}{6} + \alpha\theta N(N+1).$$

Therefore, substituting this value of $\sum_{i=1}^{N} Y_i^2$ and using (9.23) in (9.24), we can see that

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

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

So that,

$$S^2 = \frac{N(N+1)}{12}\theta^2. \qquad \dots (9.25)$$

Therefore, the sampling variance of the sample mean estimator in Simple Random Sampling without Replacement scheme, given by,

$$V(\bar{y}) = \frac{(N-n)}{Nn} S^2$$

reduces to

$$V(\bar{y}) = \frac{(N-n)}{Nn} \cdot \frac{N(N+1)}{12}\theta^2 = \frac{(k-1)(nk+1)}{12}\theta^2 \qquad \dots (9.26)$$

Similarly, for obtaining the expressions of

$$V(\bar{y}_{STRS}) = \frac{1}{n}\left(1 - \frac{1}{k}\right) S_w^2,$$

as obtained in the Sub-section 9.5.2 and

$$V(\bar{y}_{ro}) = \frac{(k-1)}{k} S_C^2 ,$$

as given in (9.6); we have to obtain the expressions of

$$S_w^2 = \frac{1}{n(k-1)} \sum_{r=1}^{k} \sum_{j=1}^{n} (Y_{rj} - \bar{Y}_{oj})^2$$

and

$$S_C^2 = \frac{1}{k-1} \sum_{r}^{k} (\bar{y}_{ro} - \bar{Y})^2$$

for this population.

Since, the observations within each row increase by the same amount, we have

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

and for the same reason, since the column means corresponding to k different systematic samples also increase by the same amount, we have

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

Therefore, we have

$$V(\bar{y}_{STRS}) = \frac{1}{n}\left(1-\frac{1}{k}\right)S_w^2 = \frac{k^2-1}{12n}\theta^2 \qquad \dots (9.27)$$

and

$$V(\bar{y}_{ro}) = \frac{(k-1)}{k}S_C^2 = \frac{k^2-1}{12}\theta^2 \qquad \dots (9.28)$$

Therefore, we see that

$$V(\bar{y}_{STRS}) : V(\bar{y}_{ro}) : V(\bar{y}) = \frac{(k^2-1)}{12n}\theta^2 : \frac{(k^2-1)}{12}\theta^2 : \frac{(k-1)(nk+1)}{12}\theta^2$$

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

or approximately,

$$V(\bar{y}_{STRS}) : V(\bar{y}_{ro}) : V(\bar{y}) \approx \frac{1}{n} : 1 : n.$$

This shows that the variance in a stratified population is only $\left(\frac{1}{n}\right)^{th}$ of the variance in a systematic sample in the population which shows a linear trend. Similarly, the variance of a systematic sample is approximately $\left(\frac{1}{n}\right)^{th}$ of the variance of a simple random sample. Thus, Stratified Sampling scheme is the most efficient of all the methods of sample selection.

Now, you may try to answer the following Self-Assessment Question:

SAQ 5

Considering that a population shows a linear trend in variate-values as given by

$$Y_i = \delta + i\beta; \quad \text{for } i = 1, 2, \dots, N;$$

Y_i being the value of the i^{th} unit, compare the Linear Systematic Sampling, SRS and Stratified Random Sampling schemes for their efficiency.

9.7 CIRCULAR SYSTEMATIC SAMPLING

While dealing with the Linear Systematic Sampling (LSYS) scheme under the Section 9.3, we mentioned that Linear Systematic Sampling (LSYS) scheme is not suitable always due to the reason that if population size, N is not an integer multiple of the desired sample size, n. If the relation N = n.k is not satisfied with integer k, then, in arranging the units in the form of an array, some of the columns (which are, in fact, systematic samples) would not be of size n and, hence, the sample mean estimator for these systematic samples would not be an unbiased estimator of the population mean. Therefore, Linear Systematic Sampling (LSYS) scheme is not applicable in such cases. We shall discuss here the Circular Systematic Sampling (CSYS) scheme as a remedy of this type of situations.

Let us assume that for the given population, the condition N = n.k is not satisfied for an integer value of k. As an example, let N = 300 and a random sample of size 40 be selected from the population using Systematic Sampling (SYS) scheme, then k = 300/40 = 7.5, a fractional number. If in this case, the units are arranged in the form an array with number of columns 7 and number of rows 40 (= n) we would have sample size 43 in the first 6 systematic samples (columns) and 42 in the last sample (seventh column). On the other hand, if k=7.5 is approximated to the nearest integer 8, then the sample size in the first 4 columns would be 38 whereas the sample size of the last 4 columns would be 37. Thus, if $N \neq nk,$ we observe that the desired sample size varies from sample to sample, which creates a great difficulty at the estimation stage due to reason that the sample mean estimator would not be unbiased for the population mean.

In order to overcome this difficulty in Systematic Sampling scheme, an alternative arrangement of units is suggested which is ***"Circular Systematic Sampling"*** (CSYS) scheme.

9.7.1 Selection of Random Sample in Circular Systematic Sampling Scheme

In Circular Systematic Sampling (CSYS) scheme, population units are supposed to be arranged in circular form where units are arranged serially in clockwise direction on the circumference of the circle. In this scheme, the sampling interval, k is chosen to be [N/n], where [N/n] stands for the greatest integer nearest to N/n. In order to decide the random start, a random number r such that $1 \leq r \leq N$, is selected from the Random Number Table. As soon as the random start point is decided, every k^{th} unit situated on the circumference from the random start point is assumed to be included in the sample. This process is continued till n such numbers on the circumference are selected in the sample. This process is generally known as 'Circular Systematic Sampling' (CSYS). It can be seen that if the random start point is r, then the systematic sample consists of units corresponding the serial numbers:

$$\{r + jk\} \quad \text{if } r + jk \leq N \quad \text{and}$$

$$\{r + jk - N\} \quad \text{if } r + jk > N, \quad \text{where } j = 0, 1, 2, \ldots, (n-1). \qquad \ldots (9.29)$$

Since, a number r; $(1 \le r \le N)$ is randomly selected using the chance mechanism, all the units of the population have an equal probability, 1/N, of selection in the sample. This means that the inclusion probability of each unit in the sample is n/N.

It is to be mentioned here, while selecting a sample in Circular Systematic Sampling scheme, there is no need to remember the rule (9.29). If every k^{th} unit after the random start is chosen serially in the clockwise direction till n units are selected the rule (9.29) will automatically be followed.

We shall now explain the selection method in Circular Systematic Sampling scheme with the help of some practical examples:

Example 3: Select a systematic sample of size 3 from the population of size 25.

Solution: We have given N = 25; n = 3. Then, we see that [N/n] = [25/3] = 8.

Let the random number selected from the random number table be 17. Then 17^{th} unit of the population is selected randomly in the sample of size 3. Since k = 8, we shall select 25^{th} and then 8^{th} units of the population in the sample. Thus, the sample consists of labels as (8, 17, 25).

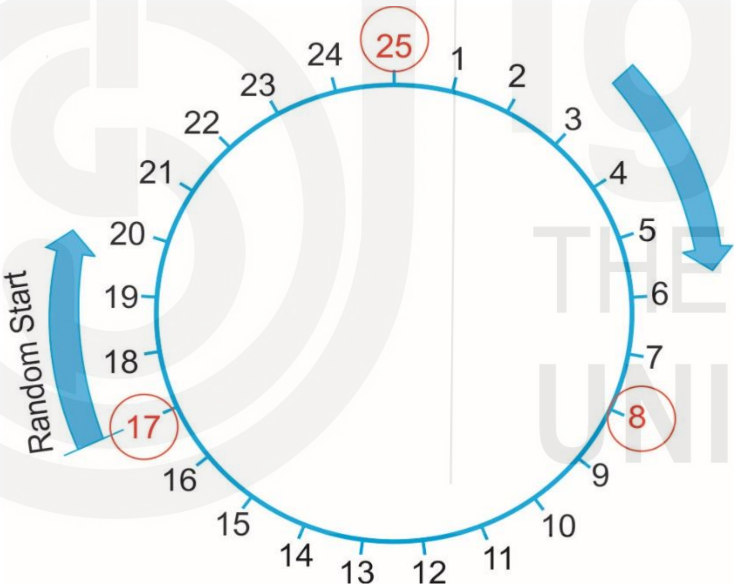

Fig. 9.4: Circular Systematic Sampling Scheme

Since, we have N = 25, n =3, k = 8 and r = 17 here, therefore, j = 0, 1, 2 in the expression (9.29). Therefore,

$$r + jk = 17 + 0x8 = 17 < N \text{ for } j = 0;$$

$$17 + 1x8 = 25 < N \text{ for } j = 1 \text{ and}$$

$$17 + 2x8 = 33 > N \text{ for } j = 2.$$

This shows that labels 17 and 25 are selected in the sample and for j = 2, we select the label r + jk – N = 33 – 25 = 8.

So, labels (8, 17, 25) constitute the sample of size 3.

Example 4: Select a systematic sample of size n from the following**:**

(a) Let N = 15 and n = 4. **(b)** Let N = 15, n = 4 and r = 6.

Solution: (a) We have to select a systematic sample of size 4. Obviously, $N \neq nk$ for any integer k.

We have N/n = 15/4 = 3.75, which is near to integer 4. So, $k = [N/n] \cong 4$.

Now, let r = 1, such that $1 \le r \le N$ then label 1 is selected.

Other selected labels will be 1 + 4 = 5; 5 + 4 = 9 and 9 + 4 = 13.

You can verify the formula (9.29) which also yields labels (1, 5, 9, 13).

(b) We have given, N = 15, n = 4 and r = 6.

We have $k = [15/4] \cong 4$ and r = 6.

Then we can verify the formula (9.29) and get the labels

For j = 0; 6 + 0 × 4 = 6

For j = 1; 6 + 1× 4 = 10

For j = 2; 6 + 2 × 4 = 14

and

For j = 3; 6 + 3 × 4 – 15 = 3

(6; 6 + 4 = 10; 10 + 4 = 14; 14 + 4 –15 = 3)

Hence, the units (3, 6, 10, 14) selected in the sample.

9.7.2 Estimation of Population Mean

Let us consider the problem of estimating the population mean under Circular Systematic Sampling scheme.

Since, $N \neq n.k$, we can write N = nk + t, where t < k. We know that in this case, sample size will vary, it will be either n or n +1 depending upon the random start. Let the total of variate-values belonging in the sample be y_{ro}; then the sample mean would be

$$\bar{y}_{ro} = \frac{y_{ro}}{n} = \frac{\sum_{j=1}^{n} Y_{rj}}{n}, \qquad \text{if } r > t$$

$$= \frac{y_{ro}}{n+1} = \frac{\sum_{j=1}^{n+1} Y_{rj}}{n+1}, \qquad \text{if } r \le t. \qquad \ldots (9.30)$$

Therefore,

$$E(\bar{y}_{ro}) = \frac{1}{k}\sum_{r=1}^{k} \bar{y}_{ro} \neq \bar{Y}.$$

Thus, the sample mean $\bar{y}_{ro}$ in this case is not an unbiased estimator of the population mean, $\bar{Y}$.

However, we can obtain an unbiased estimator of population mean by adjusting the weight assigned to the sample mean $\bar{y}_{ro}$. Let us define an estimator as follows:

$$\bar{y}^{*}_{ro} = \frac{k}{N} y_{ro}; \qquad \ldots (9.31)$$

where y_{ro} denotes the total of all the units belonging to the r^{th} row. Then, we have

$$E\left(\bar{y}_{ro}^{*}\right)=E\left(\frac{k}{N}y_{ro}\right)=\frac{k}{N}E\left(y_{ro}\right)$$

$$=\frac{k}{N}E\left(n\bar{y}_{ro}\right)=\frac{nk}{N}E\left(\bar{y}_{ro}\right)=\frac{nk}{N}\frac{1}{k}\sum_{r=1}^{k}E\left(\bar{y}_{ro}\right)$$

$$=\frac{nk}{N}\frac{1}{k}\sum_{r=1}^{k}\frac{1}{n}\sum_{j=1}^{n}Y_{rj}$$

$$=\frac{1}{N}\sum_{r=1}^{k}\sum_{j=1}^{n}Y_{rj}=\bar{Y}$$

showing that $\bar{y}_{ro}^{*}$ is an unbiased estimator of population mean.

9.7.3 Sampling Variance of the Sample Mean Estimator

Since the estimator $\bar{y}_{ro}^{*}$ is an unbiased estimator of the population mean $\bar{Y}$, we can find its sampling variance.

We have the following theorem:

Theorem 6: The sampling variance of the estimator $\bar{y}_{ro}^{*}$ is given by

$$V\left(\bar{y}_{ro}^{*}\right)=\frac{k(k-1)}{N^{2}}S_{C}^{*2}; \qquad \text{… (9.32)}$$

where $S_{C}^{*2}=\frac{\sum_{r=1}^{k}(y_{ro}-\bar{y}_{ko})^{2}}{k-1}$ with $\bar{y}_{ko}=\frac{1}{k}\sum_{r=1}^{k}y_{ro}=\frac{N}{k}\bar{Y}.$

Proof: We have

$$V\left(\bar{y}_{ro.}^{*}\right)=\frac{k^{2}}{N^{2}}V\left(y_{ro}\right).$$

Now, from (9.5) we know that

$$S_{C}^{2}=\frac{1}{k-1}\sum_{r=1}^{k}(\bar{y}_{ro}-\bar{Y})^{2};$$

which is the mean square between the column means. It can alternatively be written as

$$S_{C}^{2}=\frac{1}{k-1}\sum_{r=1}^{k}\left\{\bar{y}_{ro}-E\left(\bar{y}_{ro}\right)\right\}^{2}; \qquad \text{since } E\left(\bar{y}_{ro}\right)=\bar{Y}$$

In a similar way, we can define the mean square between the column totals as

$$S_{C}^{*2}=\frac{1}{k-1}\sum_{r=1}^{k}\left\{y_{ro}-E\left(y_{ro}\right)\right\}^{2}; \qquad \text{… (9.33)}$$

where

$$E\left(y_{ro}\right)=\frac{1}{k}\sum_{r=1}^{k}y_{ro}=\frac{N}{k}\bar{Y}; \qquad \text{since } \sum_{r}^{k}y_{ro}=N\bar{Y}.$$

Therefore, similar to expression (9.6) we can write

$$V\left(y_{ro}\right)=\frac{k-1}{k}S_{C}^{*2}$$

Hence, we have

$$V\left(\bar{y}_{ro}^{*}\right)=\frac{k^2}{N^2}V\left(y_{ro}\right)$$

$$=\frac{k^2}{N^2}\frac{k-1}{k}S_C^{*2}=\frac{k\left(k-1\right)}{N^2}S_C^{*2}$$

Now you may try to answer the following Self-Assessment Question:

SAQ 6

Let it is desired to select a systematic random sample of size n = 9 from the population of size 55. Show the process of selecting it. What would be the units selected in the sample if the first unit is selected randomly using the random number table is 12.

Now, let us summarize in the next section, the contents presented in different sections of this unit.

9.8 SUMMARY

In this unit, we have discussed:

- The concept of Systematic Sampling (SYS) scheme as an alternative to Simple Random Sampling, which was proved to be operationally simple in selecting a random sample of fixed size from the population and sometimes more efficient as compared to SRS scheme.
- The reason of selecting only one unit of the population at random in the sample and other units automatically, following a pre-defined pattern in Systematic Sampling (SYS) scheme.
- Two different types of method of selection of a sample from the population, namely, Linear Systematic Sampling (LSYS) and Circular Systematic Sampling (CSYS) and, accordingly, the arrangement of population units in both of these methods for selecting a sample with the cases under which these are applied.
- The method of developing appropriate estimator of population mean under the case of Linear Systematic Sampling scheme.
- The unbiased estimator of the population mean and derivation of sampling variance of sample mean estimator.
- Comparative study of the Linear Systematic Sampling scheme with the Simple Random Sampling and Stratified Random Sampling schemes for their efficiencies and conditions under which Linear Systematic Sampling scheme is found to be better than Simple Random Sampling and Stratified Random Sampling.
- Some of the populations which show a peculiar type of nature and, hence, might be more suitable for the application of Linear Systematic Sampling scheme.
- The comparison of the three schemes, namely, Linear Systematic Sampling, Simple Random Sampling and Stratified Random Sampling, for their efficiency in populations showing an increasing linear trend of variate-values.

- Lastly, the description of Circular Systematic Sampling scheme was given and the way in which the units of the population should be arranged before selecting a sample from it.
- Similar to Linear Systematic Sampling scheme, the appropriate estimators of population mean in Circular Systematic Sampling scheme one biased and another an unbiased estimator along with the sampling variance of the unbiased estimator.

9.9 TERMINAL QUESTIONS

1. Define the systematic sampling scheme.

2. What are the types of Systematic Sampling Scheme?

3. Describe how the units of the population must be arranged in order to select a Systematic sample of a fixed size using the Linear Systematic Sampling.

4. Show that the Systematic Sampling scheme provides a probability sample with equal probability selection method. Define the terms "Random start" and "Sampling interval" in reference to Systematic Sampling scheme.

5. The following data show the cultivated area in acres for 60 villages in a particular state of India:

111	210	370	70	249	644	970	156	111	275
91	102	132	482	72	203	62	655	70	149
439	68	732	206	97	530	131	312	127	85
125	466	755	600	131	291	82	144	456	333
761	535	155	110	65	220	180	90	245	138
38	145	230	390	583	333	80	146	210	258

In order to estimate the mean cultivated area per village, it is decided to apply systematic sampling scheme. Select three different samples of sizes 12, 10 and 20 respectively and find the corresponding sample means.

6. Let a population be of size 250. In order to estimate a population parameter, it is desired to select two systematic samples of sizes 10 and 25 respectively. Let the random starts in selecting these samples be respectively 17 and 5, write the units which will come under the two samples.

7. If S^2_{wsy} denotes the variance among units belonging to the same systematic sample or mean square within systematic sample, obtain an expression of the sampling variance of the systematic sample mean estimator in terms of S^2_{wsy}.

8. Compare the Linear Systematic Sampling scheme with Stratified Random Sampling scheme and show that Stratified Random Sampling scheme is more efficient than Linear Systematic Sampling scheme and is equally efficient to Linear Systematic Sampling scheme respectively when the non-circular serial correlation coefficient ρ_{wst} given by

$$\rho_{wst} = \frac{E\left(y_{rj} - \bar{Y}_{oj}\right)\left(y_{rj'} - \bar{Y}_{oj'}\right)}{E(y_{rj} - \bar{Y}_{oj})^2}$$

is positive or zero.

9. Under what condition, the circular systematic sampling used for sample selection? Explain with an example what do you mean by Circular Systematic Sampling scheme.

10. Suggest an unbiased estimator of population mean under Circular Systematic Sampling scheme and obtain its sampling variance.

9.10 ANSWERS / SOLUTIONS

Self-Assessment Questions (SAQ)

1. **Hint:** For answering this exercise, you are advised to go through the Section 9.2.

2. **Hint:** You are advised to consult the Section 9.3 for answering the question.

3. **Hint:** You are referred to go through the sub-section 9.4.1 and the **Theorem 1** under the Sub-section 9.4.1 for getting answer of both the parts of this question.

4. **Hint:** See Theorem **5** under the Sub-section 9.5.1 for answering this question.

5. **Hint:** You are referred to Section 9.6(d) for the answer of this exercise.

6. Given that N = 55 and n = 9. Then, we see that the multiple of n is not an integer multiple. So, we use circular systematic sampling scheme here. Clearly, then k = [55/9] =6.11 which is taken to be greatest integer equal to 6. Now, we arrange 55 population units in clockwise direction in a circular form as follows:

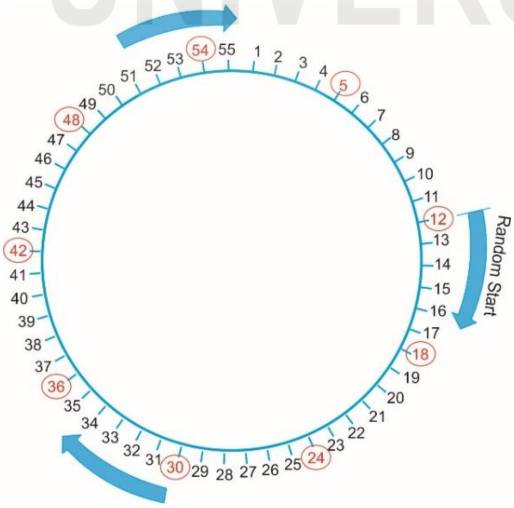

Fig. 9.5: Circular Systematic Sampling Scheme

Doing this, we select a random number r from the Random Number Table such that $1 \le r \le 55$. Let it be 12. Then we start with the first selected label as 12. Since k = 6, we select the $(12 + 6)^{th}$, i.e. 18^{th} label as the second selected label. Similarly, every 6^{th} unit from the previously selected unit would be selected in the sample by moving the circle in clockwise direction until the sample size is fulfilled. So, the sample of size six would be {12, 18, 24, 30, 36, 42, 48, 54 and (60 – 55 = 5)}.

Terminal Questions

1. **Hint:** For answering this exercise, you are advised to go through the Sub-section 9.2.1.

2. **Hint:** See the Sub-section 9.2.2 for getting your answer.

3. **Hint:** You are advised to consult Sub-section 9.3.1 for answering the two parts of this exercise.

4. **Hint:** You can answer the two parts of this exercise on the basis of the **Remark 9.2** and **Remark 9.3** under Sub-section 9.3.1.

5. We have N = 60. If the sample size be 12, then the condition N = n.k is satisfied such the k is an integer multiple of n and therefore, we have k = 5. We then place the 60 units of the population in the form of an array consisting of 12 rows and 5 columns. The array will consist of 12 rows and 5 columns as:

111	210	370	70	249
644	970	156	111	275
91	102	132	482	72
----	----	----	----	----
----	----	----	----	----
38	145	230	390	583
333	80	146	210	258

In order to select a sample of size 12, therefore, we select a random number r such that $1 \le r \le 5$ from the Random Number Table.

Let r = 2. Then the second column is selected as a systematic sample of size 12. From the given set of data, you can see that the sample of size 12 will consist of values (210, 970, 102, 62, 68, 131, 466, 82, 535, 180, 145, 80).

The sum of these 12 values is 3031, therefore, the sample mean would be 3031/12 = 252.58.

Further, if the sample size be 10, then clearly, n = 10 and hence, we have k = 6. Obviously, then the 60 values would be arranged in an array with 10 rows and 6 columns. Arranging in a similar fashion as we did above, we have the array as:

111	210	370	70	249	644
970	156	111	275	91	102
132	482	72	203	62	655
----	----	----	----	----	----
----	----	----	----	----	----
245	138	38	145	230	390
583	333	80	146	210	258

If the selected random number be r where $1 \leq r \leq 6$ and it is selected from the Random Number Table and found to be 5, then our sample of size 10 would be (249, 91, 62, 732, 127, 131, 761, 180, 230, 210) as per the values given in the data set.

The mean of sample would be 2773/10 = 277.3.

If sample size be 20, then we have n = 20 and k =3. Therefore, we arrange 60 values in the form of an array consisting of 20 rows and 3 columns. So, the random start would be a number r such that $1 \leq r \leq 3$.

Let it be 3. Then, the first value which is randomly selected in the sample would be 370. Then, we select one by one every 3rd unit in the sequence as the unit selected in the sample. From the given set of data, we can see that the sample of size 20 would be

(370, 644, 111, 102, 72, 655, 439, 206, 131, 85, 755, 291, 456, 535, 65, 90, 38, 390, 80, 258).

The Mean of this sample would be 5773/20 = 288.65.

6. Given that N = 250 and two systematic random samples of sizes 10 and 25 are to be selected from this sample.
For the first sample, random start is 17 whereas for the second sample, it is 5. For both the cases, we see that condition N = n.k is satisfied, so we use Linear Systematic Sampling scheme for the selection of samples.
In the first case where n = 10 we observe that the array will consist of 10 rows and 25 columns. Since, in this case random start r = 17, the sample would consist of

(**17**th, 17+25 = **42**nd, 17+2 x25 = **67**th, 17+75 = **92**nd, 17+ 100 = **117**th, 17+125 = **142**nd, 17 + 150 = **167**th, 17+175 = **192**nd, 17+200=**217**th and 17+225 = **242**nd) units of the population.

If n = 25, then we have an array consisting of 25 rows and 10 columns. Since the random start is 5, the first unit in the sample would be 5 and then every 10th unit would be included in the sample. Thus, we have the sample of size 25 consisting labels as

(5, 15, 25, 35, 45, 55, 65, 75, 85, 95, 105, 115, 125, 135, 145, 155, 165, 175, 185, 195, 205, 215, 225, 235 and 245).

7. **Hint:** See the **Theorem 3** under the Sub-section 9.4.2.

8. **Hint:** You are advised to consult the Sub-section 9.5.2 for answering this exercise.

9. **Hint:** For the first part of the exercise go through the introductory part of the Section 9.7. For the second part you are referred to Sub-section 9.7.1 where Circular Systematic Sampling scheme is explained with the help of an example.

10. **Hint:** You are advised to go through the Sub-sections 9.7.2 and 9.7.3 for the answer of this exercise.