UNIT 15 STATISTICAL DATA SAMPLING

Structure

15.1 INTRODUCTION

It is a common practice for all of us to draw some conclusions about a large bulk by examining a small part of it. For example suppose a person wants to make judgement about the quality of oil in a tinful of oil. Then he will not check all the oil in the tin. He examines only a small part of the oil from the tin. Similar is the situation when a doctor takes a few drops of blood from a patient to decide if the patient has malarial infection. In such cases it is not practical to examine the entire data. So we adopt a method called sampling by which we select a sample of the whole mass and study it for valid information. These two examples have the special phenomenon that the tinful of oil and the total blood in the patient, are known to be perfectly homogeneous material so that every part of the material represents the material exactly. Often, however, we are not in this simple situation. For example, suppose we are interested in knowing what the average height of an adult Indian male is. Obviously we will not consider it satisfactory to measure just a few adult males, as not all adult Indian males are of the same height or weight. They show considerable heterogeneity. So, how do we take a part of a large heterogeneous mass to draw valid conclusions about it? This is the central question in statistical sampling.

In this unit we will discuss the basic ideas of sampling. We will then describe the most common method of collecting samples, namely the random sampling method. We use samples to draw conclusion about entire populations. But it would be unreasonable to expect a sample result to have exactly the same value as population characteristic. In this unit we shall acquaint you with some measures of accuracy of sample results. For this we will use the concepts of mean and standard deviation which you have studied in Unit 11. In the final stage, we shall briefly explain different sampling procedures.

Objectives

After reading this unit, you should be able to

- explain the need for sample selection
- define the meaning of a good sample
- use random sampling procedure for drawing random samples
- obtain the sampling distribution of means and proportions and compute the standard deviation for sampling distributions
- define an 'unbiased estimator' and 'accuracy and precision of an estimator', and
- describe and distinguish between two sampling methods—Stratified sampling and Cluster sampling.

15.2 SAMPLE SELECTION

In Unit 11 you have seen the formal definition of 'population'. We recall that a population is the totality of any kind of individuals about which we wish to have knowledge in a given context. In the examples given in the introduction, the population were, respectively,

a) a tinful of oil
b) all the blood in a patient's body
c) all adult Indian males.

Note that the individuals which constitute the population may not be human always. They may be non-human items such as drops of oil that go to make a tinful of oil as in (a).

Also recall from Unit 11 that a sample is a part of the population selected for study to draw conclusions about the population. In the examples considered earlier,

a) the drops of oil taken, and
b) the adult Indian males actually measured for their heights,

are the samples.

Before proceeding further you may try the following exercise in order to make sure that you are able to distinguish between population and sample.

E 1) Suppose you are interested in knowing the average height of female students enrolled at Madras University. Which of the following groups would be the population or a sample for this problem.

a) All female students enrolled in a psychology course.
b) All female students enrolled at Madras University.
c) All students enrolled in a business school.

Note that information obtained from a study of a sample is of interest to us only if we can use it to draw conclusions about the entire population. For example, the response to a drug of a sample of rats or monkeys is of interest to us only if on the basis of this sample, we can say what is likely to be the response of all rats or all monkeys to that drug. This point should always be kept in mind while dealing with samples.

Perhaps you are wondering why we are talking of studying a sample, rather than the entire population. In the next sub-section we list some advantages of studying samples.

15.2.1 Why Use a Sample?

Here we will explain to you some situations where we can save our time, cost and energy by adopting sampling methods.

(i) **Studying a sample may be the only approach** to throw light on the population characteristic of interest. Suppose a manufacturer wishes to make a statement about the expected length of life of the bulbs manufactured in 1987. Then he has to burn a bulb till it gets fused to determine its life. This process is destructive and obviously, he can't burn all the bulbs! Only a sample can be used in this context.

(ii) **Studying a sample obviously saves money and time.** You will agree that measuring the heights of only a sample of adult Indian males to draw a conclusion about the average height of the population of adult Indian males, can be done in a shorter time and involves less expenditure.

(iii) **Often information of high quality can be collected from only a sample** rather than the population. This could be because of lack of monetary resources or more importantly, because of lack of technical resources. In medical surveys, for example, complex laboratory investigations may have to be carried out by highly trained staff for the exact diagnosis of a disease. Now, such highly trained staff are likely to be very limited in number and so the investigations can be done only on a sample

You can now try this exercise.

E 2) Give an example where the study of a population is a necessity, and cannot be substituted by the study of a sample.

Next we will pass onto certain characteristics of a good sample.

15.2.2 Criteria of A Good Sample

We have seen that many times it is not possible to study the entire population; only a sample can be used for study. Does that mean that we can select any part of the population as we like, and draw valid conclusions about the population? No. We have to be careful in our choice of a good sample. But what are the criteria of a good sample? Let us consider an example to find an answer to this question.

Suppose we wish to make a valid statement about the health condition of a class of 100 students by taking a sample of 10 students. Let us assume that the students sit in the class in 10 rows of 10 students each. An easy way to select 10 students would be to select the 10 students sitting in the first row in the class. However, it may happen that those students whose eyesight is not good enough, or whose hearing is impaired, or who are short in height, may prefer to sit in the first row. If that happens, the sample of 10 students in the first row will give a poorer picture of the health of the class as a whole. In other words, the sample will have a **systematic error** in the direction of a poorer representation of the true health condition of the class. If, to avoid this possibility, we eliminate the first row while selecting the sample, we will tend to commit a systematic error in the opposite direction.

In statistical terminology, **systematic error** is called **'bias'**. So, **we want to select samples free of bias** or, in other words, we want samples to be unbiased. That is, we want a sample to be representative of the population from which it is selected. We shall elaborate this point later in Sec. 15.2.3.

In Sec. 15.2.1, we stated that the purpose of studying a sample is to enable us to say something about one or more unknown characteristics of the population. **The unknown characteristic in the population is called a parameter,** and the quantity we compute from the sample to make a statement about the unknown parameter is called **a statistic.** In general, **a statistic is any quantity computed from a sample.**

For example, in E 1, the average height is the parameter and the average height measured from group (a) is a statistic. If we want to compute the mean of a population then, the mean of a sample is a statistic. The median, range, standard deviation computed from a sample is each a statistic.

The following exercise will help you get a clearer idea about these terms.

E 3) A workers union has a membership of 300 persons. Data were collected from 25 of them, and their average age was found to be 39. The average age of the entire union membership was therefore estimated to be approximately 39. A subsequent polling of all members indicated that the true average age was 42.

a) What is the parameter?
b) What is the statistic?

Can you see that a parameter is a fixed quantity? For example, the average height of the population of adult Indian males (that is, all individuals who are adult Indian males) at a given time, has a single fixed value.

The **statistic,** on the other hand is a variable quantity. Let us see why it is a variable. We shall take the examples of the heights of adult Indian males to explain our statements. The population is, in general, heterogeneous. We know that the heights of adult Indian males vary. A sample also will reflect this variability in some measure. If two samples of the same size are drawn from this population, they are unlikely to be identical. One sample may have a few more taller people than the other. Hence, the average height computed from one sample is likely to be different from that computed from the other. That is, the average height computed from a sample is likely to show variability from sample to sample. Thus, the statistic is a variable quantity. This variability of the statistic is called **sampling variability or standard error** of the mean.

All statistics computed from samples show standard error if the population is heterogeneous. You can note that if the population is homogeneous, then any sample selected will represent the population exactly. So, we try to choose a sample such that the statistic computed from this sample has **an acceptably small sampling error. Finally, we also want to select a sample in such a way that we can compute the standard error of the statistic from the sample itself.** This gives us some idea of the accuracy of the sample.

Let us summarise our discussion in this section. There are three criteria which determine a good sample.

a) It should be unbiased.
b) It should provide the desired statistic with an acceptably small sampling error.
c) It should permit the computation of the sampling error from the sample itself.

So, we know what to look for, in a good sample. In the next sub-section we shall see how to select a good sample.

15.2.3 Random Sampling Procedure

In the last sub-section we have listed three criteria of a good sample. Here we shall first discuss how to select a good sample that satisfies the first and the third criteria. That is, we shall see how to select a sample which is unbiased and, which permits the computation of the standard error of the statistic of interest to us.

In the example, we considered at the beginning of Sec. 15.2.2 regarding the health condition of a class of 100 students, we saw that selecting the first row of 10 students may result in our describing the health of the whole class as poorer than what it really is. We also noted that if we decided to exclude the first row from selection, we would commit an error in the opposite direction. This has a general lesson for us in the context of selecting an unbiased sample from any population. We want a method of choosing a sample, which is independent of the characteristics of the individuals in the population. In other words, **we do not want the selection of a sample to be influenced, consciously or unconsciously, by a knowledge of the characteristics of the individuals in the population.**

To achieve this, we adopt a procedure called **random sampling,** which **ensures that each individual** in the population is **chosen, or not chosen,** as a member of a sample **strictly** by a chance mechanism. In other words, random sampling ensures that no one individual in the population is preferred to any other individual, for any reason for selection into the sample. This, in turn, ensures that the selection of the sample is unbiased. Since the process of selection is random, we call such samples, random samples. The defined chance or probability of inclusion into the sample of the units in the population may or may not be the same. We shall see more of this in Sec. 15.4. Let us first understand the method of drawing a random sample..

The process of drawing a random sample is essentially equivalent to a fair lottery procedure. For example, suppose we have a population of 500 individuals and we wish to draw a random sample of 50 individuals. We can number the population serially from 1 to 500. Then we can write numbers from 1 to 500 on 500 identical slips of paper, place them in a box, mix them thoroughly and pick out 50 slips, one by one without looking. That should give us a random sample of 50 individuals. But this procedure is not easily manageable. Firstly, because the numbering of the slips become inconvenient, when the population size increases. Secondly, we have to be very careful to see that the **bowl** is thoroughly mixed after a piece of paper is selected. Statisticians have devised an elegant method of drawing random samples, using what are called **random sampling numbers,** or **tables of random digits.** These tables can be produced by a computer giving each of the digits 0 to 9 an equal chance or probability of appearing at each draw. Tables of such random numbers are published, and we have appended a page of such random digits at the end of this unit for your use. (Appendix).

Let us work out some examples of drawing a random sample using this table of random numbers. We must note that the numbers in the random number table in the appendix can be used as single digited numbers, or two digited numbers or three or

more digited numbers depending on the size of the population you are sampling from. We will illustrate this with few examples.

Example 1 : Let us first consider the case of two digited numbers. Suppose, we wish to take a random sample of 10 students from a class of 100 students.

Solution : Let us assume that the students are given serial numbers from 00 to 99, 00 actually standing for the 100th student and 01, 02, ..., 99 standing for the 1st, 2nd ..., 99th student, respectively. In this case, since the population size is 100, the 100th student being given the number 00, we can use the table as one of two digited numbers. While using the table, it is advisable to take a blind start on the table to avoid repeatedly using the same starting point for sampling. Let the blind starting point—by placing your finger on the table with closed eyes—be the number 26, corresponding to the fourth row and fifth column of two digited numbers. The two-digited numbers can be read either along the columns or the rows. Let us read along the columns. The first 10 numbers in serial order, starting from 26, are 26, 64, 39, 31, 59, 29, 44, 46, 75, 74. Then we select individuals corresponding to these numbers from the population as our random sample.

Here you can note one thing. It may happen that the same two digited number occurs more than once. In such a case, we include this number only when it occurs for the first time, and reject it when it occurs for the second time. We then choose the next number occurring serially in the table. We do not wish to include the same individual twice. This is called **sampling without replacement,** that is, a person is not chosen into the sample more than once. In this case, we can see that each possible sample of 10 **different** students from the class of 100 has the same chance of being selected as our sample. As we said earlier, in the random sampling number tables, digits 0 to 9 occur equally frequently. Because of this, all single digited, two digited, three or more digited numbers occur equally frequently.

Let us consider a slight variation of the above example.

Example 2 : Suppose we have a population of 32 students and we wish to draw a random sample of 10 from this.

Solution : We could follow the same procedure as above by numbering the 32 students from 01 to 32 and selecting two-digited random numbers from the table by a blind-folded start, till we get 10 different students as our sample. However, if we do this, since there are only 32 students in the population, we will have to reject numbers from 33 to 99 and also 00. It is not desirable to reject too many numbers like this since that may affect the randomness of our sample. To avoid rejecting too many numbers, we do the following. We divide 100 (total two-digited numbers from 00 to 99) by the population size 32, and take the quotient which is 3 in this case. We allot 3 two-digited numbers to each student as below :

Sl. No. of student	**Two-digited numbers allotted**
1	01; 33; 65
2	02; 34; 66
3	03; 35; 67
⋮	⋮
32	32; 64; 96

97 and numbers above are rejected. Then, student 1 will be selected into the sample if any one of the three numbers 01, 33 or 65 occurs while selecting random numbers from the table. By this process, we reject only numbers 97, 98, 99 or 00, that is, only 4 out of the 100 two-digited numbers.

Let us consider another example which illustrates that we can use random number table for three-digited numbers.

Example 3 : Suppose, we have a population of 430 students in a school, and we wish to select a random sample of 30 students from this population.

Solution : You would readily agree that we now have to use the random number table as consisting of 3-digited numbers. As before, take a blind-folded start on the table. Suppose it is the 6th column and the 5th row starting with the three-digited number

385. Now, let us read row wise the successive three-digited numbers. The numbers are 385, 462, 482, 231, 624, etc. Since our population consists of only 430 students, we have to drop numbers exceeding 430 in this procedure, that is, in the above list, we will have to drop 462, 482, 624 etc. To minimise the rejection of numbers, we follow a procedure similar to the one in Example 2. We divide 1000 (i.e., the number of all three-digited numbers from 000 to 999) by 430 getting a quotient of 2. Hence we can give two numbers to each of the 430 students as shown below:

Student Serial No.	Random Number
1	001 and 431
2	002 and 432
3	003 and 433
⋮	⋮
430	430 and 860

We will reject numbers above 860, and select students with serial numbers 385, 482, 462, 169, etc. till we reach the sample size of 30 different students.

In fact, you don't have to actually allocate numbers to each student, and then select. You can dispense with this tedious step by doing as follows. When you get a random number, divide it by 430 (= the population size) and see what the remainder is. The remainder stands for the serial number of the student who is to be selected. For example, the first random number is 385. Dividing this by 430, we get the remainder 385. So, the 385th student is selected. The next random number is 482. Dividing by 430, we get the remainder 52. So, select the 52nd student and continue this procedure till you have got the sample of 30 different students you want.

Now here is an exercise for you.

E 4) Select a random sample of 5 children from a class of 80 using the random number tables given in Appendix at the end of this unit. Explain your procedure clearly.

Any sample selected from a population provides only partial information about the population. Therefore the statements we make about the population may be subjected to error. In the next section we will study this error in detail.

15.3 MEASURES OF VARIATION AND ACCURACY

We have seen that, in most situations, it is extremely difficult to have a sample completely representative of the population. It would be unreasonable to expect a sample result to have exactly the same value as some population characteristic because sampling error is always present. In the next sub-section we shall discuss some measures of sampling error.

15.3.1 Sampling Distribution

In Sec. 15.2.2, we mentioned that 'Statistics' show variability from sample to sample when the population is heterogeneous. We need to study this variability, if we want to use a statistic from a sample as an estimate of the unknown parameter. We have already noted that the unknown parameter is a fixed quantity. Let us consider an example.

Example 4 : Suppose we have a population of 8 individuals with heights 5′6″, 5′4″, 6′, 5′8″, 5′4″, 5′6″, 5′10″ and 5′6″.

a) What is their mean height?
b) Calculate the sample means by selecting samples of 2 individuals.

Solution : a) You can easily calculate that their mean height is 5′7″.
(b) Using the random sampling procedure that we described in Sec. 15.2.3 of this unit, we can draw random samples of size 2 from this population of 8 individuals. We can draw a total of C (8,2) = 28 different samples of 2 individuals, as we have shown in Table 1, listed below. For convenience, the 8 individuals in the pupulation are listed as A, B, C, D, E, F, G, and H.

All possible samples of size 2 from a population of 8 individuals

Sample No.	Individual selected in the sample	Height of the sampled individuals	Mean height in sample
1	A,B	5′6″ , 5′4″	5′5″
2	A,C	5′6″ , 6′0″	5′9″
3	A,D	5′6″ , 5′8″	5′7″
4	A,E	5′6″ , 5′4″	5′5″
5	A,F	5′6″ , 5′6″	5′6″
6	A,G	5′6″ , 5′10″	5′8″
7	A,H	5′6″ , 5′9″	5′6″
8	B,C	5′4″ , 6′0″	5′8″
9	B,D	5′4″ , 5′8″	5′6″
10	B,E	5′4″ , 5′4″	5′4″
11	B,F	5′4″ , 5′6″	5′5″
12	B,G	5′4″ , 5′10″	5′7″
13	B,H	5′4″ , 5′6″	5′5″
14	C,D	6′0″ , 5′8″	5′10″
15	C,E	6′0″ , 5′4″	5′8″
16	C,F	6′0″ , 5′6″	5′9″
17	C,G	6′0″ , 5′10″	5′11″
18	C,H	6′0″ , 5′6″	5′9″
19	D,E	5′8″ , 5′4″	5′6″
20	D,F	5′8″ , 5′6″	5′7″
21	D,G	5′8″ , 5′10″	5′9″
22	D,H	5′8″ , 5′6″	5′7″
23	E,F	5′4″ , 5′10″	5′7″
24	E,G	5′4″ , 5′10″	5′7″
25	E,H	5′4″ , 5′6″	5′5″
26	F,G	5′6″ , 5′10″	5′8″
27	F,H	5′6″ , 5′6″	5′6″
28	G,H	5′10″ , 5′6″	5′8″
All samples	—	—	5′7″

Note : This example may appear artificial to you. You are right because in this case the mean height can be directly computed from the population. This example is actually taken to illustrate the ideas in a simple way.

If you examine the average or mean of each of the samples given in the last column of Table 1, you see that it differs from the population mean of 5′7″ in a number of samples. In this example, sample numbers 3, 12, 20, 22, 23 and 24 have a mean of 5′7″ which is the population mean value. But there might be some situations where none of the sample means is equal to the mean of the population. Coming back to our example, you see that the sample means vary between 5′4″ and 5′11″ from a low value of 5′4″ to a high value of 5′11″. This clearly brings out one point. When we select random samples, that is, unbiased samples, there is some chance that some of these samples might give a statistic that differs considerably from the parameter. From Table 1, let us identify the samples that give a mean value which differs from the population mean value by a specified amount. We write them in another table.

Table 2

Deviation of the means of the 28 samples from the population mean

Deviation (+ or —) of Sample mean from Population mean (″)	Serial number of sample	Number of samples
0	3, 12, 20, 22, 24	5
1	5, 6, 7, 8, 9, 15, 19, 26, 27, 28	10
2	1, 2, 4, 11, 13, 16, 18, 21, 23, 25	10
3	10, 14	2
4	17	1

From Table 2, you can see that the proportion of samples whose means do not differ by more than 1″ from the population mean is 15/28. This is obtained by dividing the

sum of the first two numbers in the last column of Table 2 i.e. 5 and 10, by the total number of samples i.e. 28. You can say the same thing using the concept of probability. You would then say that the probability is 15/28 that the sample mean does not differ from the population mean by more than 1″. Similarly, the probabilities that the sample mean does not differ from the population mean by more than 2″, more than 3″ and more than 4″, are, respectively, 25/28, 27/28 and 28/28. You can see from this, that even when we select our sample by a random process—which is the best way of ensuring an unbiased sample—the sample statistic often differs from the population parameter. Also, the closer we want the sample statistic to be to the population parameter, the smaller becomes the chance or probability of it being so.

Now, we make a frequency distribution for the value of the mean height in samples that we obtained from all the 28 possible samples of size 2, given in Table 1. We get the following Table 3.

Table 3

Distribution of sample mean in samples of size 2

Sample mean value $\bar{x}$	Number of samples giving this mean (frequency) f	Relative frequency (= probability)
5′4″	1	1/28
5′5″	6	6/28
5′6″	5	5/28
5′7″	5	5/28
5′8″	5	5/28
5′9″	4	4/28
5′10″	1	1/28
5′11″	1	1/28
Total	28	1

Table 3 gives the **sampling distribution** of the sample mean. It shows the different values that the sample mean can take in repeated sampling and the relative frequency or probability with which it can take each of them. Note that the sum of the probabilities is equal to 1. This is so because the 28 samples can take only the mean values listed in column 1. That is, these sample means listed in column 1 are exhaustive. Thus we got a probability distribution corresponding to the sample statistic mean. The sampling distribution of the sample mean could be written down by us only because we had used a random sampling procedure that gave an equal chance of selection to every possible sample of 2 individuals from the population of 8 individuals. Such a random sampling procedure which gives equal probability of selection to every possible sample of a given size, is called **simple random sampling.**

There are other types of random sampling where all samples may not get the same probability of selection. We shall learn about this and other types of sampling procedures in Sec. 15.4.

The statistic used to throw light on the unknown parameter is called an **'estimator'**. In our example of the heights of 8 individuals, the sample mean is the estimator. The particular value the sample mean takes in a given sample is called the **estimate.**

Now you will be able to do the following exercise easily.

E 5) Suppose we have a population of 5 students enrolled in a statistics course and an instructor wants to find the average amount of time spent by each student in preparing for classes each week. The amount of time (in hours) each student spends per week is given by 7, 3, 6, 10 and 4? If the instructor takes a sample of three students, obtain the sampling distribution of the sample mean. Compute the population mean and the mean of the sampling distribution.

Recall the formula for computing the mean of a frequency distribution from Unit 11.

So far we have been discussing about the sample statistic: the mean. Now corresponding to each sample statistic, we consider its probability distribution called

the sampling distribution of the statistic. In the next section we will measure the standard error in sampling using the most common measure of central tendency, namely, mean and standard deviation of a distribution which you have seen in Unit 11.

15.3.2 Standard Error

We have already seen how random samples taken from a population show variability from sample to sample in the estimator of interest to us. We wish to measure this variation by calculating the standard deviation of the sampling distribution of the statistic. For example, let us go back to Example 4 of measuring the mean height of the population of 8 individuals. Table 1 shows that the sample mean differs from sample to sample. The distribution of the sample mean in samples of size 2 is given in Table 3. Let us now calculate the standard deviation of the sampling distribution given in Table 3. For that we need to calculate the mean of the distribution, say $\bar{\bar{x}}$. The mean $\bar{\bar{x}}$ is given by,

$\bar{\bar{x}} = \frac{\Sigma f \bar{x}}{\Sigma f} = \frac{156'4''}{28} = 5'7''$ which is equal to the population mean. This shows that the **mean of the sampling distribution is the same as the population mean.** The details of the calculations of the standard deviation are given in Table 4.

Table 4

Computation of sampling standard deviation of the sample mean

Sample mean value ('') ($\bar{x}$)	Frequency (f)	Deviation of $\bar{\bar{x}}$ from sample mean x ($\bar{x} - \bar{\bar{x}}$)	$(\bar{x} - \bar{\bar{x}})^2$	f. $(\bar{x} - \bar{\bar{x}})^2$
5'4''	1	−3''	9	9
5'5''	6	−2''	4	24
5'6''	5	−1''	1	5
5'7''	5	0''	0	0
5'8''	5	+ 1''	1	5
5'9''	4	+ 2''	4	16
5'10''	1	+ 3''	9	9
5'11''	1	+ 4''	16	16
	$\Sigma f = 28$		$\Sigma f (\bar{x} - \bar{\bar{x}})^2 = 84$	

Recall the formula for computing the standard deviation σ of a frequency distribution

$\sigma = \sqrt{\frac{\Sigma f (x - \mu)^2}{\Sigma f}}$

where
μ—the mean of the distribution
x—observed values
f—frequency of x

The sample standard deviation of the sample means

$$S_{\bar{x}} = \sqrt{\frac{\Sigma f.(\bar{x} - \bar{\bar{x}})^2}{\Sigma f}} = \sqrt{\frac{84}{28}} = \sqrt{3}$$

The standard deviation of the sampling distribution is called the **standard error.** In the present case, since the sampling distribution is that of the mean, we call the standard error of this distribution the **standard error of the mean** which is equal to $\sqrt{3}$. The standard error of the mean gives us a quantitative measure of the average variability of the sample mean due to sampling variability.

You will study later in sub-section 15.3.4 that the standard error of any statistic gives us an idea of how good a statistic is in estimating the parameters.

You may have realised that the computation of the standard deviation from the sampling distribution is a tedious process. There is an alternative method to compute standard error of the means, SE ($\bar{x}$), from a single sample if we know the poulation standard deviation. By this method, we have the following formula for obtaining the standard error of the mean.

$$SE(\bar{x}) = \sqrt{\frac{N-n}{N-1} - \frac{\sigma^2}{n}}$$

where N = population size = the total number of individuals
n = sample size = number of individuals selected in the random sample.

$\sigma =$ standard deviation of the individuals in the population.

We will not derive the formula here since the process is too technical for the scope of this course. The factor $\sqrt{\frac{N-n}{N-1}}$ is called the finite population correction factor. As a rule of thumb, when $\frac{n}{N}$ is less than 0.1, this correction factor can be ignored. We use the above formula for computing SE $(\bar{x})$ when N—n is not very large. When N is large, relative to n, we use the formula,

$$SE(\bar{x}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Now you can try this exercise.

E 6) Compute the standard error of the data given in E 5.

Thus we have seen that we can compute the standard error of the sample means if we know the population standard deviation. Here you can note one thing. Usually, we do not know σ . But it is possible to estimate σ from the sample. We will talk about this later in the next section. Using that estimate we compute the standard error by the shortcut formula given earlier.

Next we willl discuss the sampling distribution of another sample statistic, namely, sample proportion.

Standard Error of a Proportion

Very often we are interested in studying a parameter in proportion form rather than in measurement form. For example, let us again go back to Example 4 in sub-section 15.3.1. In Example 4, suppose, our interest is not in the mean height of the population of 8 individuals, but in finding the proportion of individuals exceeding a height of 5′6″. In the population of 8 individuals, the height of C, D, and G is more than 5′6″. That is, the population proportion of interest to us is 3/8. In general for a finite population we define a population proportion as $\pi = \frac{k}{N}$ where k is the number of elements that possess a certain characteristics and N is the total no. of items in the population. Then as in the case of population mean, we estimate the population proportion by taking random samples. Let us see how we can do this in the case of the above example. As before, we take simple random samples of size 2. For convenience in computation, let us adopt the following procedure. Whenever an individual selected in the sample has a height greater than 5′6″, give him a score of 1 and, otherwise give him a score of 0. Now the mean of these scores will give the proportion of individuals exceeding 5′6″. Let us verify this for the population of 8 individuals. The scores of A, B, C, D, E, F, G and H will be, respectively, 0, 0, 1, 1, 0, 0, 1, 0 and the mean of this is $(0 + 0 + 1 + 1 + 0 + 0 + 1 + 0) \div 8 = 3/8$ which is the population proportion exceeding a height of 5′6″

Let us now work out the sampling distribution of the sample proportion by getting the sample proportion of individuals exceeding 5′6″ from each of 28 possible samples listed in Table 1. In each sample, we score an individual as 0, if his height is 5′6″ or lower and as 1 if his height exceeds 5′6″. Then the mean score in each sample shown in Table 5 (a) below gives the sample proportions.

Table 5 (a)

Computation of the sample proportions

Sample No.	Score	Mean score = sample proportion
1	0,0	0
2	0,1	1/2
3	0,1	0
4	0,0	0
5	0,1	0
6	0,0	1/2
7	0,1	0
8	0,1	1/2
9	0,1	1/2
10	0,0	0
11	0,0	0
12	0,1	1/2
13	0,0	0
14	1,1	1
15	1,0	1/2
16	1,0	1/2
17	1,1	1
18	1,0	1/2
19	1,0	1/2
20	1,0	1/2
21	1,1	1
22	1,0	1/2
23	0,0	0
24	0,1	1/2
25	0,0	0
26	0,1	1/2
27	0,0	0
28	1,0	1/2

Then we form the following table (Table 5 (b)) and get the sampling distribution of sample proportion.

Table 5 (b)

Sampling distribution of the sample proportion

Sample proportion (p)	Frequency (f)	f.p
0	10	0
1/2	15	7.5
1	3	3.0
Total	28	10.5

$$\text{Mean of Sampling Distribution} = \frac{\Sigma fp}{\Sigma f} = \frac{10.5}{28} = \frac{3}{8}$$

Here also you can note that the mean of the sampling distribution of sample proportion is the same as the population proportion exceeding 5'6". Now the standard error of sample proportions, by definition, is the standard deviation of this sampling distribution. As mentioned earlier, the computation of the standard deviation from the table is a tedious process. So we make use of a shortcut formula by which we can compute the standard error if we know the population proportion, population size and sample size.

The formula is given by $SE(p) = \sqrt{\frac{N-n}{N-1} \quad \frac{\pi(1-\pi)}{n}}$,

where π is the population proportion and N and n are the population size and sample size, respectively. When n is small compared to N, that is, if the size of the population, relative to the sample size is extremely large, then

$$SE(p) = \sqrt{\frac{\pi(1-\pi)}{n}},$$

Let us consider an example.

Example 5 : Compute the standard error of the proportion of individuals exceeding a height 5′6″ in the population given in Example 4.

Solution : We have calculated earlier that the population proportion $\pi = \frac{3}{8}$.

Also N = 8 and n = 2. Substituting these values in the formula for SE (p), we get,

$$SE(p) = \sqrt{\frac{8-2}{8-1} \times \frac{3}{8} \left(1-\frac{3}{8}\right) \times \frac{1}{2}}$$

$$= \sqrt{\frac{6}{7} \times \frac{3}{8} \times \frac{5}{8} \times \frac{1}{2}}$$

$$= \frac{3}{8}\sqrt{\frac{5}{7}}$$

$$= .32$$

Why don't you try this exercise!

E 7) An organisation has a total of eight members whose ages in years are 27, 32, 33, 26, 43, 52, 28 and 25. The organisation has a rule which requires a minimum age of 33 for member to be a President. Assume a simple random sample of size 4 is selected to provide an estimate of the population proportion eligible for presidentship. What would be the mean and standard deviation of the sampling distribution?

From our discussions about sampling distribution, we observe one fact. Even though the statistics computed from different samples of a population vary from population parameter, we can expect that the average of the sampling distribution of the statistic is equal to the population parameter. In the next sub-section we will discuss this nature of the statistic.

15.3.3 Unbiased Estimator of Population Parameter

We have seen that the sample mean is selected as an estimator of the population mean and sample proportion is selected as an estimator of the population proportion.

What are the reasons for selecting a particular statistic to be an estimator? If you look back at Table 4, you will find that the mean of the sampling distribution, $\bar{\bar{x}}$ is 5′7″ which is the population mean, that is, the mean of the height of the 8 individuals in the population we have considered. Similarly, if you look at Table 5, you will find that the mean of the sampling distribution of the sample proportion is the same as the population proportion. Whenever the mean of the sampling distribution of an estimator equals the corresponding unknown population parameter, then the estimator is said to be **unbiased.** In other words, sample mean and sample proportion are unbiased estimators of the population mean and population proportion, respectively. So, one very important criterion for the selection of a statistic as an estimator is 'unbiasedness'.

We noted earlier that the standard error of the sample mean is expressed in terms of the population standard deviation σ, by the formula

$$SE(\bar{x}) = \sqrt{\frac{N-n}{N-1}} \times \frac{\sigma}{\sqrt{n}}$$

Here σ is given by the formula

$$\sigma = \sqrt{\sum \frac{(x-\mu)^2}{N}}$$

where μ is the population mean. We also noted that σ is usually unknown and therefore we have to estimate it from the sample itself. Intuitively, it appears that the sample standard deviation S given by the formula,

$$S = \sqrt{\sum \frac{(x-\bar{x})^2}{n}}$$ where $\bar{x}$ — sample mean

n — sample size

is an estimator of the population standard deviation because of their similarity in computation. But is S an unbiased estimator of σ? We know that, by definition, S will be an unbiased estimator of σ, if the sampling distribution of S has a mean value exactly equal to σ.

The sampling distribution of the sample standard deviation in Example 4 of a population of 8 individuals from which we draw samples of size 2, is given below in Table 6.

$$\text{Here, } S = \sqrt{\sum \frac{(x - \bar{x})^2}{n}}$$

Table 6

Sampling distribution of sample standard deviation when n = 2

Sample standard deviation (S)	Frequency (f)	f.s
0	4	0
1	11	11
2	8	16
3	3	9
4	2	8
Total	28	44

The mean of the sampling distribution $= \frac{44}{28} = \frac{11}{7} = 1.59$. (Let us call this '$\mu_\sigma$')

The population standard deviation of the 8 individuals,

$$\sigma = \sqrt{\sum \frac{(x - \mu)^2}{8}} \text{ is } \sqrt{7},$$

which is quite different from the mean of the sampling distribution of sample standard deviation μ_σ which is 1.59. It is, therefore clear that the sample standard deviation S is not an unbiased estimator of the population standard derivation σ, though S and σ appear to be similar.

This example teaches us an important lesson. The statistic or estimator in a sample exactly corresponding to the population parameter is not necessarily an unbiased estimator.

Actually, in the example above, we can use $\sqrt{\sum \frac{(x-\bar{x})^2}{n-1}}$ [instead of $\sqrt{\sum \frac{(x - \bar{x})^2}{n}}$]

as an estimator of σ. But this estimator is not unbiased. (We are not proving this result here since it is beyond the scope of this course.) So, whenever you wish to estimate the population standard deviation from a sample, you can compute the standard deviation as

$\sqrt{\sum \frac{(x - \bar{x})^2}{n - 1}}$ where 'n' is the sample size.

Let us consider an example.

Example 5 : A Psychologist measures the reaction times of sample of 6 individuals to certain stimula. The measures are given by 0.53, 0.46, 0.50, 0.49, 0.52, 0.53 seconds. Determine (i) an unbiased estimate of the population mean, (ii) an estimate of the population standard deviation.

Solution : (i) An unbiased estimate of the population mean is

$$\bar{x} = \frac{\Sigma x}{n}$$

$$= \frac{0.53 + 0.46 + 0.50 + 0.49 + 0.52 + 0.53}{6}$$

$= 0.51$ seconds

(ii) An estimate of the population standard deviation is given by

$$S = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}}$$

$$\frac{\Sigma (x - \bar{x})^2}{n - 1} = \frac{1}{n - 1} \left[(0.53 - 0.51)^2 + (0.46 - 0.51)^2 + (0.50 - 0.51)^2 + \right.$$

$$\left. (0.49 - 0.51)^2 + (0.52 - 0.51)^2 + (0.53 - 0.51)^2 \right]$$

$= 0.0006$

Therefore $S = \sqrt{.0006}$ seconds.

Try this exercise.

E 8) Measurements of a sample of six weights were determined as 8.3, 10.6, 9.7, 8.8, 10.2 and 9.4 kilograms respectively. Determine (i) an unbiased estimate of the population mean and (ii) compare the sample standard deviation with the estimated population standard deviation.

Next we will talk about some measure of the degree to which a sample statistic differs from the true parameter.

15.3.4 Accuracy and Precision of Sample Estimator

In Table 3, we listed the sample means of 28 samples and noted that 23 of them gave a mean that was different from the population mean of 5′7″. Only 5 of the 28 samples gave a sample mean of 5′7″. If the sample mean coincides with the population mean then we can say that the sample mean is an accurate estimate of the population mean. That is, we can define accuracy in terms of the agreement between the sample mean and the population mean. If the sample mean differs from the population mean, then it is an inaccurate estimate. The degree of inaccuracy depends on the size of the difference between the sample mean and population mean. In general, the accuracy of an estimate can be defined as follows.

Accuracy of an estimate = Estimate − Parameter value.

In Example 4 of the population of 8 individuals, we knew the parameter value and used that knowledge to study the behaviour of the sample estimator. But in most cases the parameter value is not known. Since the parameter value is unknown in real life, we cannot measure accuracy as defined above. Therefore, we have to find some other way of assessing accuracy. We do that by computing the **precision** or **probable accuracy** of the estimate, using the standard error.

The standard error is a measure of the variability or the spread of the sampling distribution of the estimator. The smaller the standard error, the closer is the sample estimate to the population parameter. We know that the standard error of the sample mean is given by $\frac{\sigma}{\sqrt{n}}$. For a given population 'σ' is fixed. Hence the standard error will decrease with increasing n. But, the decrease is proportional to $\sqrt{n}$, and not n. Therefore, for cutting the standard error of mean by 50% in a given situation, the sample size will have to be increased by 4 times. Thus, we can use the standard error of an estimator to measure the precision or probable accuracy of an estimate. Let us examine what happens to the sampling distribution of the sample mean when we increase the sample size. Table 7 presents the sampling distribution in our illustrative Example 4 of a population of 8 individuals, when the sample size is increased from 2 to 6. Again we will get C (8,6) = 28 different samples as we got when the sample size was 2.

Sampling distribution of sample mean in samples of size 2 and size 6

Sample mean value (Class interval of 1″)	Number of samples giving mean in this class interval	
	when n = 2	when n = 6
5′4″	1	0
5′5″	6	1
5′6″	5	10
5′7″	5	16
5′8″	5	1
5′9″	4	0
5′10″	1	0
5′11″	1	0
	28	28

You can see that when n has increased to 6, the sampling distribution of the mean clearly has become less variable. In other words, the sample standard deviation (and therefore the standard error) has decreased, and the estimator has become more precise.

You can now easily do this exercise.

E 9) Compute the standard error of the data given in E 5 for sample size 4. Compare this with the standard error obtained in E 6.

Next we state a theorem which gives a very interesting and extremely useful observation on the behaviour of the sampling distribution of the sample means

Theorem 1 (The Central Limit Theorem) : If large random samples of size n ($n \geq 30$) are taken from a population with mean μ and standard deviation σ, and if a sample mean $\bar{x}$ is computed for each sample, then the following three things will be true about the distribution of sample means.

a) The sampling distribution of the sample means will be approximately a normal distribution.
b) The mean of the sampling distribution will be equal to the mean of the population.
c) The standard deviation of the sampling distribution will be equal to the standard deviation of the population divided by the square root of the number of items in each sample.

Recall that you have seen normal distribution in Unit 14.

According to Theorem 1 (a), sample means $\bar{x}$'s are approximately normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. Then we make use of the normal distribution chart given in the Appendix of Unit 14 and conclude that we can expect 95% of the $\bar{x}$'s to fall between $\mu - 1.96 \frac{\sigma}{\sqrt{n}}$ and $\mu + 1.96 \frac{\sigma}{\sqrt{n}}$.

From the normal distribution chart you note that 0.95 probability implies that Z = 1.96 or − 1.96.

Thus, if all possible samples of size n are selected, and the intervals $\bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$ are established for each sample, then 95% of all such intervals are expected to contain the population parameter μ. Such intervals are called confidence intervals because these intervals give some confidence that the estimated value is close to the parameter value.

We are not going into the details of confidence intervals as it is beyond the scope of your course. In the next section, we describe two sampling methods which are widely used. We have already discussed one basic method of sampling called simple random sampling. There you can see that simple random sampling is certainly a practical procedure if the population is not large and if it is relatively easy and inexpensive to find the sampling units. Now, suppose the population size is large, then there is difficulty in numbering the population if random procedure has to be adopted. For

ignou
THE PEOPLE'S
UNIVERSITY

ignou
THE PEOPLE'S
UNIVERSITY

15.4.2 Cluster Sampling

Let us try to understand this type of sample design through a simple example. Suppose we want to estimate the prevalence of heart disease in children and adults in Delhi. Let us assume that there are 12 lakh households in all. If we wish to draw a simple random sample of 20000 individuals or 4000 households (if each household has approximately 5 members), we need a complete list of the 12 lakh households which is not easily available. Even if that sort of a list is available, if we select a sample of 4000 households, they may be distributed so widely over Delhi that a good deal of effort and money will have to be spent on travel from one household to another. In such situations, we can adopt a cluster sampling approach. Delhi is divided into urban blocks each containing about 200 households. A list of such urban blocks is more readily available. From a total of 6000 urban blocks, we can select 20 urban blocks randomly and survey the 4000 households contained in these 20 blocks. What we have done now is to select 20 clusters of households, which has simplified and lowered the costs of the field work. Thus, distinct benefit of cluster sampling is savings in cost, time and energy.

We would like to make some general points about clustering.

i) A cluster sample is a sample in which the individual units are groups or clusters of single items. These clusters may be natural, like urban blocks or schools, or artificially marked areas on maps.

ii) In cluster sampling, we can have different stages to select the final sample. Such sampling is called multi-stage cluster sampling. In this case clusters may be defined differently at each stage of sampling. For example suppose a statistical group wants to study voter opinion on the construction of additional nuclear power plants in India. They can select all districts within each state as clusters. But then they cannot possibly interview all the people living in the district. So we go to the next stage where we take a part of the original samples which will now be the municipalities in these districts. Thus, in the second stage our clusters have changed to municipal divisions.

Now we will discuss the essential difference between cluster sampling and stratified sampling. In the stratified sample we choose a sample from each stratum to be representative of that stratum. The samples from each stratum when put together become a full sample from the population. In cluster sampling, however, we select a sample of clusters from among all the clusters in the population. But in most population, similar units tend to cluster together. For instance, rich people tend to live in the same neighbourhood in the city, while the poor families are concentrated in another area. This difficulty causes cluster sampling to have greater standard error in practical work than stratified sampling will have. Accordingly the precision will be less in this case. This doesn't mean that cluster sampling is less efficient. It is usually the case that cost in cluster sampling is lower than the stratified sampling as we must spend money to find out on what basis we make the strata (lower cost is most evident in area sampling, when interviewer's time and travelling costs are cut down). Then we can include more points in the sample and it is possible to have a larger cluster sample that can yield the same precision with lower cost than a stratified random sample.

You can now easily do these exercises.

E 12) The State Government of Kerala wants to determine the level of unemployment in the state. Explain how he can do this task by cluster sampling.

E 13) Determine the sample design most appropriate for the following situations.

i) A labour union is to select a sample among its members to determine the attitude of the members towards recently introduced social activities.

ii) The Food Corporation of India wants to know the average household expenditure on food in the rural areas of India.

iii) Suppose we want to estimate the average annual expenditures of a college student for non-school items.

15.5 SUMMARY

In this unit, we have

1) explained the purpose of sampling and the advantages of sampling,
2) defined the terms bias, random sample, statistic and parameter,
3) described the method of drawing random sample,
4) shown how to obtain sampling distributions of sample means and sample proportions,
5) shown how to compute the standard errors of sample means and sample proportions,
6) discussed the concepts of an unbiased estimator, accuracy and precision of an estimate, and
7) discussed two major types of sampling design commonly used in sample surveys, namely, stratified sampling, and cluster sampling.

15.6 SOLUTIONS/ANSWERS

E 1) a) sample b) population c) none

E 2) If we want to know the number of children in a district of India who have physical disability so that they can be provided needed treatment and rehabilitation, we have to study all children in the district. Every child has to be identified who needs treatment. Sampling will not help in this situation.

E 3) The average age, the population mean, 42 is the parameter. The sample mean 39 is the statistic.

E 4) Since there are 80 children in the class, we use two digited numbers from the random numbers table. We place our finger blindfolded on this table to decide which number should be the starting point. Suppose this happens to be the number 35 in the fourth column and fifth row of the table. If the 80 children are numbered from 1 to 80, then the first child selected is the 35th child. Then we proceed down the column to select the remaining four children. The next number below 35 is 94. This cannot be used since there are only 80 children in the class. So we go to the next number below. This is 53. So, the 53rd child is selected. Next number 78, then number 34 and finally number 32 is selected completing the sample 5. The sample selected is, therefore, numbers 35, 53, 78, 34 and 32.

E 5) Using the random sampling procedure, we can draw $C(5, 3) = 10$ different samples of 2 individuals as listed in Table 10. For convenience, the 5 individuals in the population are listed as A, B, C, D, E.

Table 10

Sample No.	Individual selected in the sample	Sample data	Sample means
1	A, B, C	7, 3, 6	5.33
2	A, B, D	7, 3, 10	6.67
3	A, B, E	7, 3, 4	4.67
4	A, C, D	7, 6, 10	7.67
5	A, C, E	7, 6, 4	5.67
6	A, D, E	7, 10, 4	7.0
7	B, C, D	3, 6, 10	6.33
8	B, C, E	3, 6, 4	4.33
9	B, D, E	3, 10, 4	5.67
10	C, D, E	6, 10, 4	6.67

The sampling distribution for the value of mean time in samples that we obtained from all the 10 possible samples of size 3 is given by the following table.

Table 11

Distribution of sample mean in samples of size 3

Sample Mean value	Number of samples giving this mean (frequency)	Relative frequency
4.33	1	1/10
4.67	1	1/10
5.33	1	1/10
5.67	2	2/10
6.33	1	1/10
6.67	2	2/10
7.0	1	1/10
7.67	1	1/10
Total	10	1

Population mean $= \dfrac{7+3+6+10+4}{5} = \dfrac{30}{5} = 6$

The mean of the above distribution

$$= \frac{4.33\times1 + 4.67\times1 + 5.33\times1 + 5.67\times2 + 6.33\times1 + 6.67\times2 + 7\times1 + 7.67\times1}{10}$$

$$= \frac{60}{10} = 6.$$

E 6) The formula for the standard error is

$$\sigma_x = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

where σ is the population standard deviation and N and n are population size and sample size respectively. Firstly we have to calculate 'σ' from the data given by

$$\sigma = \sqrt{\frac{\sum_{i=1}^{5} (x_i - \mu)^2}{N}}$$

where x_is are the observed samples given in the data and μ is the population. Here $\mu = 6$,

$$\text{Then, } \sigma = \sqrt{\frac{(7-6)^2 + (3-6)^2 + (6-6)^2 + (10-6)^2 + (4-6)^2}{5}}$$

$$= \sqrt{\frac{1^2 + 3^2 + 4^2 + 2^2}{5}}$$

$$= \sqrt{\frac{1+9+16+4}{5}}$$

$$= 2.45$$

and n = 3 and N = 5

The standard error σ_x is then calculated by

$$\sigma_x = \frac{2.45}{\sqrt{3}} \sqrt{\frac{5-3}{5-1}}$$

$$= 1.415 \times .707$$

$$= .899 = 1.00 \text{ (approximately)}$$

E 7) Mean of the sampling distribution $\bar{\bar{x}}_p$ is Population proportion π.

$$\pi = \frac{\text{the number possessing the age qualification}}{\text{the population size}}$$

Therefore $\bar{\bar{x}}_p = \frac{3}{8}$

The standard deviation for the sampling distribution σ_p = the standard

$$\text{error} = \sqrt{\frac{\pi(1-\pi)}{n}} \sqrt{\frac{N-n}{N-1}}$$

$$= \sqrt{\frac{\frac{3}{8}(1-\frac{3}{8})}{4}} \sqrt{\frac{8-4}{8-1}}$$

$$= \sqrt{\frac{3}{8} \times \frac{5}{8} \times \frac{1}{4}} \sqrt{\frac{4}{7}} = 0.183$$

E 8) i) $x = \frac{\Sigma x}{n} = 9.5$ kg.

ii) Sample standard deviation $S = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n}}$

$$= \sqrt{\frac{3.68}{6}}$$

$$= .78$$

An unbiased estimate of the population

$$\text{standard deviation} = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{3.68}{5}} = .86$$

The sample standard deviation is less than the unbiased estimate of the population standard deviation.

E 9) Here the population is the students in the academic institution. One way by which he can stratify the population is to classify the students as Ist year's, 2nd year's, 3rd year's etc. Another way of stratification is by selecting each department in the institution as different strata. In each case a sample can be taken from each stratum. The result obtained from these samples are then weighed to provide an overall estimate of the average grade.

E 10) a)

Table 12

Items selected in the stratified sample	Mean
ABE	13/3
ABF	15/3
BCE	14/3
BCF	16/3
CDE	15/3
CDF	17/3
ADE	14/3
ADF	16/3
ACE	13/8
ACF	15/3
BDE	15/3
BDF	17/3

b)

Table 13

Sampling distribution of sample means

Sample mean value $\bar{x}$	Frequency f	Relative frequency
13/3	2	2/12
14/3	2	2/12
15/3	4	4/12
16/3	2	2/12
17/3	2	2/12

c) To compute the sample standard deviation—first we have to calculate the mean of the sampling distribution. From Table 13

$$\bar{\bar{x}} = \frac{\Sigma f\bar{x}}{\Sigma f} = \frac{15}{3}$$

Table 14

Computation of sample standard deviation

Sample mean value $\bar{x}$	Frequency f	$(\bar{x}-\bar{\bar{x}})$	$(\bar{\bar{x}}-\bar{\bar{x}})^2$	$f(\bar{x}-\bar{\bar{x}})^2$
13/3	2	−2/3	4/9	8/9
14/3	2	−1/3	1/9	2/9
15/3	4	0	0	0
16/3	2	1/3	1/9	2/9
17/3	2	2/3	2/3	8/9
	$\Sigma f = 12$			$\Sigma f(\bar{x}-\bar{\bar{x}})^2 = \frac{20}{9}$

$$\text{The sample standard deviation} = \sqrt{\frac{\Sigma f(\bar{x}-\bar{\bar{x}})^2}{\Sigma f}}$$

$$= \sqrt{0.185185}$$

E 11) He might do this by multi-stage cluster sampling. He first selects a set of districts randomly. From these selected districts he can select a set of villages. Finally from the selected villages he can select a required number of households. Then study the situation in these households.

E 12) a) Stratified sampling
b) Multi-stage cluster sampling
c) Stratified sampling

Appendix

Random Numbers

03	47	43	73	86	36	96	47	36	61	46	98	63	71	62
97	74	24	67	62	42	81	14	57	20	42	53	32	37	32
16	76	62	27	60	56	50	26	71	07	32	90	79	78	53
12	56	85	99	26	96	96	68	27	31	05	03	72	93	15
55	59	56	35	64	38	54	82	46	22	31	62	43	09	90
16	22	77	94	39	49	54	43	54	82	17	37	93	23	78
84	42	17	53	31	57	24	55	06	88	77	04	74	47	67
63	01	63	78	59	16	95	55	67	19	98	10	50	71	75
33	21	12	34	29	78	64	56	07	82	52	42	07	44	38
57	60	86	32	44	09	47	27	96	54	49	17	46	09	62
18	18	07	92	46	44	17	16	58	09	79	83	86	19	62
26	62	38	97	75	84	16	07	44	99	83	11	46	32	24
23	42	40	64	74	82	97	77	77	81	07	45	32	14	08
52	36	28	19	95	50	92	26	11	97	00	56	76	31	38
37	85	94	35	12	83	39	50	08	30	42	34	07	96	88
70	29	17	12	13	40	33	20	38	26	13	89	51	03	74
56	62	18	37	35	96	83	50	87	75	97	12	25	93	47
99	49	57	22	77	88	42	95	45	72	16	64	36	16	00
16	08	15	04	72	33	27	14	34	09	45	59	34	68	49
31	16	93	32	43	50	27	89	87	19	20	15	37	00	49
68	34	30	13	70	55	74	30	77	40	44	22	78	84	26
74	57	25	65	76	59	29	97	68	60	71	91	38	67	54
27	42	37	86	53	48	55	90	65	72	96	57	69	36	10
00	39	68	29	61	66	37	32	20	30	77	84	57	03	29
29	94	98	94	24	68	49	69	10	82	53	75	91	93	30

UNIT 16 HYPOTHESIS TESTS

Structure

16.1 INTRODUCTION

Many television commercials often make various performance statements. For example, consider the following statements :

(i) A particular brand of tyre will last an average of 40,000 miles before replacement is needed.

(ii) A certain detergent produces the cleanest wash.

(iii) Brand X of disposable nappies are stronger and are more absorbent.

How much confidence can one have in such statements? Can they be verified statistically? Fortunately, in many cases the answer is yes. In Unit 15, you have seen that samples are taken and partial bits of informations are gathered from the sample about such statements. In this unit we will study how to make a decision on whether to accept or to reject a statement on the basis of sample information. This process is called **hypothesis testing.**

In Unit 15, we have studied that population parameter can be estimated by using samples. These estimates are in turn used in arriving at a decision to either accept or reject the hypothesis. By a **hypothesis** we mean an assumption about one or more of the population parameters that will either be accepted or rejected on the basis of the information obtained from a sample.

In this unit we will discuss three types of tests : chi-square test, t-test and analysis of variance for testing a hypothesis. These tests are widely used in making decisions concerning problems in biology and in other fields. We shall not talk about the theories of these tests, but shall consider only their applications under simple situations. To apply these tests you should know what is meant by null hypothesis, level of significance and degrees of freedom. We shall begin with discussing these concepts in brief.

Objectives

After you have completed this unit, you should be able to

- select an appropriate test to analyse a given problem;
- apply the chi-square test, t-test and analysis of variance;
- substitute numerical data into selected formulas and solve the corresponding equations;
- draw conclusions from your statistical analysis.