UNIT 18 STATISTICAL ESTIMATION

Structure

18.0 OBJECTIVES

After going through this Unit you will be in a position to:

- explain the concept of estimation;
- distinguish between point estimate and interval estimate;
- estimate confidence interval for a parameter; and
- explain the concept of confidence level.

18.1 INTRODUCTION

Many times due to certain constraints such as inadequate funds or manpower or time we are not in a position to survey all the units in a population. In such situations we take resort to sampling, that is, we survey only a part of the population. On the basis of the information contained in the sample we try to draw conclusions about the population. This process is called statistical inference. We must emphasise that statistical inference is widely applied in economics as well as in many other fields such as sociology, psychology, political science, medicine, etc. For example, before election process starts or just before declaration of election results many newspapers and television channels conduct exit polls. The purpose is to predict election results before the actual results are declared. At that point of time, it is not possible for the surveyors to ask all the voters about their preferences for electoral candidates — the time is too short, resources are scarce, manpower is not available, and a complete census before election defeats the very purpose of election!

In the above example the surveyor actually does not know the result, which is the outcome of votes cast by all the voters. Here all the voters taken together comprise the population. The surveyor has collected data from a representative sample of the population, not all the voters. On the basis of the information contained in the sample, (s)he is making forecast about the entire population.

In this Unit we deal with the concept of statistical inference and methods of statistical estimation. Parameter, as you know, is a function of population units while statistic is a function of sampling units. There could be a number of *parameters* and corresponding *statistics*. However, in order to keep our presentation simple, we will confine ourselves mostly to arithmetic mean.

18.2 STATISTICAL BACKGROUND

In the previous two blocks we have discussed two important aspects: theoretical probability distributions and sampling techniques. These two aspects form the basis of statistical inference.

In Unit 14, Block 5 we explained the concept of a random variable. We learnt that *X* is a random variable if it assumes values $x_1, x_2, ..., x_n$ with corresponding probabilities $p_1, p_2, ..., p_n$ attached to it. Here the probability of occurrence of x_1 is p_1, the probability of occurrence of x_2 is p_2, and so on. If the values $x_1, x_2, ..., x_n$ are discrete we call *X* a discrete random variable and find out the probability for isolated values of *X*. On the other hand, if *X* is a continuous random variable we can find out the probability of *X* within certain range such that $P(a \le X \le b) = p_1$.

In Units 14 and 15 of Block 5 we discussed theoretical discrete probability distributions (such as binomial and Poisson) and continuous probability distributions (such as normal and *t*). We learnt that if the range of *X* increases infinitely then these probability distributions approach normal distribution. Thus normal distribution is a limiting case of these probability distributions and is considered as a sort of ideal among probability distributions.

The normal distribution is defined by two parameters: mean (μ) and standard deviation (σ). If the probabilities associated with a random variable are distributed according to normal distribution (that means, if *X* follows normal distribution), we can find out the probability of $P(a \le X \le b) = p_1$ by using the equation for its probability distribution function.

A problem encountered here is that μ and σ can take any values and finding out corresponding probability is time consuming. This problem is tackled by subtracting μ from the normal variable and dividing it by σ. This way we obtain the 'standard normal variate', $z = \frac{x-\mu}{\sigma}$, which has mean = 0 and standard deviation = 1. By plotting the probabilities for different values of *z* on a graph paper we obtain 'standard normal curve' which is symmetrical and area under the curve is = 1. Remember that in the case of standard normal curve we measure $z = \frac{x-\mu}{\sigma}$ on the x-axis and probability of occurrence of *z*, that is $p(z)$, on the y-axis. Thus if we consider a particular segment of the normal curve (bounded by two values of *z*, say, z_1 and z_2) the area under the curve gives its probability. Remember that normal curve is different from the frequency curve considered in Block 1 of this course. Area under the normal curve does not give frequencies; it gives probabilities.

In Unit 16 of Block 6 we leant that very often it is not possible to study the entire population and we undertake a sample survey. If the sample is drawn in a random manner through appropriate probability attached to each population unit and the sample size is not very small, the sample can be a representative one of the population. Recall that we can draw a number of samples from a given population and each sample provides us with a sample mean. Thus the sample means can be arranged in the form of a frequency distribution, called the 'sampling distribution'.

We know from Unit 16, Block 6 that sample mean ($\bar{x}$) assumes different values and for each value we can attach a probability. Thus sample mean can be considered as a random variable. In real life situations we have a finite population and the number of samples (and therefore the number of sample means) is finite. In this case $\bar{x}$ is a discrete random variable but when there are infinite number of samples, $\bar{x}$ could be a continuous random variable.

Now let us consider another important concept discussed in Unit 16: the central limit theorem. It says that sampling distribution of $\bar{x}$ is normal if the population from which the sample is drawn is normal. However, sampling distribution of $\bar{x}$ is approximately normal if sample size (n) is large, even if the parent population (that is, population from which it is drawn) is not normal. If the parent population is approximately normal then sampling distribution of sample means is approximately normal even when sample size is small.

We know that dispersion of sample means is smaller in value than dispersion of the parent population from which the sample is drawn. Recall that the standard deviation of the sampling distribution is called standard error. Thus if the population has a standard deviation σ then the standard error of sample mean is $\frac{\sigma}{\sqrt{n}}$.

From the above we learn that sample mean can be considered as a random variable and it approximates normal distribution when sample size is large. Usually we consider a sample to be large in size if $n > 30$. For small samples ($n \leq 30$), sampling distribution of sample means is similar to student's t distribution. Recall that in the case of t distribution the shape of the probability curve changes according to its 'degrees of freedom'.

18.3 CONCEPT OF STATISTICAL INFERENCE

As mentioned earlier, statistical inference deals with the methods of drawing conclusions about the population characteristics on the basis of information contained in a sample drawn from the population. Remember that population mean is not known to us, but we know the sample mean. In statistical inference we would be interested in answering two types of questions. First, what would be the value of the population mean? The answer lies in making an informed guess about the population mean. This aspect of statistical inference is called 'estimation'. The second question pertains to certain assertion made about the population mean. Suppose a manufacturer of electric bulbs claims that the mean life of electric bulbs is equal to 2000 hours. On the basis of the sample information, can we say that the assertion is not correct? This aspect of statistical inference is called hypothesis testing.

Thus statistical inference has two aspects: estimation and hypothesis testing. We will discuss about statistical estimation in the present Unit while testing of hypothesis will be taken up for discussion in the next Unit.

Fig. 18.1 below summarises different aspects of statistical inference. A crucial factor before us is whether we know the population variance or not. Of course when we do not know the population mean, how do we know the population variance? We begin with the case where population variance is known, because it will help us in explaining the concepts. Later on we will take up the more realistic case of unknown population variance.

Estimation could be of two types: point estimation and interval estimation. In point estimation we estimate the value of the population parameter as a single point. On the other hand, in the case of interval estimation we estimate lower and upper bounds around sample mean within which population mean is likely to remain.

The assertion or claim made about the population mean would be in the form of a null hypothesis and its counterpart, alternative hypothesis. We will explain these concepts and the methods of testing of hypothesis in the next Unit.

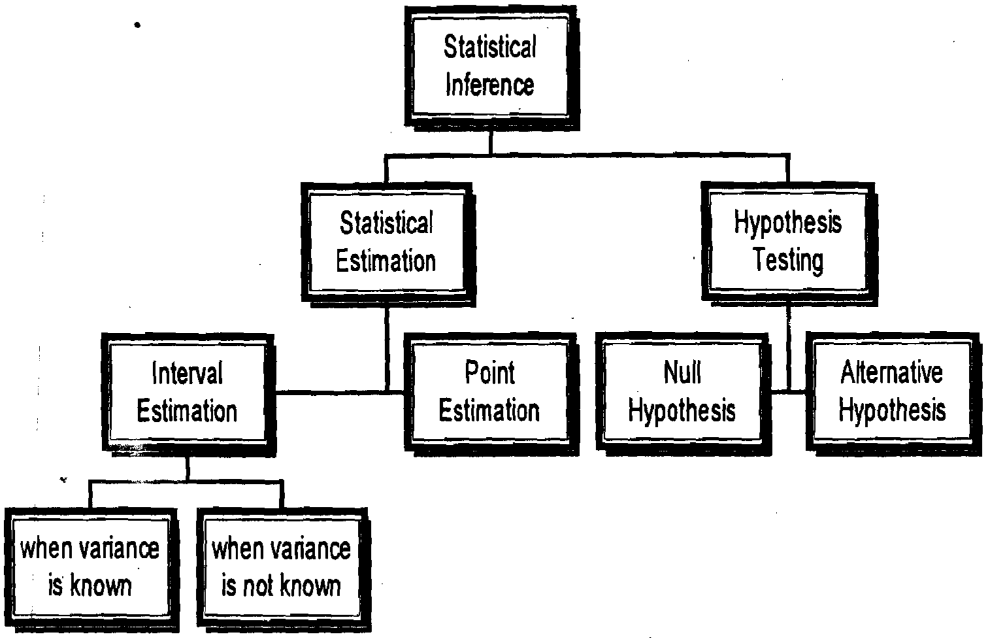

Fig. 18.1: Statistical Inference

Check Your Progress 1

1) Explain the following concepts.
   a) standard normal variate
   b) random variable
   c) sampling distribution
   d) central limit theorem

....

....

....

....

....

....

2) State whether the following statements are true or false.
   a) Normal distribution is a limiting case of binomial distribution.
   b) Standard deviation of sampling distribution of a statistic is termed as standard error.
   c) Poisson distribution is an example of continuous distribution.
   d) Statistical estimation is a part of statistical inference.

....

....

....

....

....

18.4 POINT ESTIMATION

As mentioned earlier we do not know the parameter value and want to guess it by using sample information. Obviously the best guess will be the value of the sample statistic. For example, if we do not know the population mean the best guess would be the sample mean. Here in this case we use a single value or point as 'estimate' of the parameter.

In Unit 16 we have explained the concepts of estimate and estimator. Also we have pointed out the distinction between the two. Recall that estimator is the formula and estimate is the particular value obtained by using the formula. For example, if we use sample mean for estimation of population mean, then $\frac{1}{n}\sum_i x_i$ is the estimator. Suppose I collect data on a sample, and put the sampling units to this formula and obtain a particular value for sample mean, say 120. Then 120 is an estimate of population mean. It is possible that you draw another sample from the same population, use the formula for sample mean, that is $\frac{1}{n}\sum_i x_i$, and obtain a different value, say 123. Here both 120 and 123 are estimates of population mean. But in both the cases the estimator is the same, which is $\frac{1}{n}\sum_i x_i$. Remember that the term statistic, which is used to mean a function of sample values, is a synonym for estimator.

There may be situations when you would find more than one potential estimator (alternative formulae) for a parameter. In order to choose the best among these estimators, we need to follow certain criteria. Based on these criteria an estimator should fulfill certain desirable properties. There are quite a few desirable properties for an estimator, but the most important is its unbiasedness.

Unbiasedness means that an estimate may be higher or lower than the unknown value of the parameter. But the expected value of the estimate should be equal to the parameter. For example, sample mean ($\bar{x}$) may fluctuate from sample to sample but on an average it would be equal to population mean. In other words, $E(\bar{x}) = \mu$.

However, $\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is not an unbiased estimator of the population variance $\sigma^2 = \frac{1}{N}\sum(X_i - \bar{X})^2$. In fact, if we define $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$, then s^2 is an unbiased estimator of σ^2. Usually a sample is less dispersed than the population from which it is drawn. Therefore, there is a tendency for the sample standard deviation s to be little less than population standard deviation σ. In order to rectify this condition we artificially inflate s by dividing by a smaller number $(n-1)$, instead of n.

The point estimate is quite important for testing of hypothesis, as we will see in Unit 19.

18.5 CONFIDENCE INTERVAL FOR KNOWN VARIANCE

We have seen above that in point estimation, we estimate the parameter by a single value, usually the corresponding sample statistic. The point estimate may not be realistic in the sense that the parameter value may not exactly be equal to it. An alternative procedure is to give an interval, which would hold the parameter with certain probability.

Here we specify a lower limit and an upper limit within which the parameter value is likely to remain. Also we specify the probability of the parameter remaining in the interval. We call the *interval* as 'confidence interval' and the *probability* of the parameter remaining within this interval as 'confidence level' or 'confidence coefficient'.

Let us take an example. Suppose you are asked to estimate the average income of people in Raigarh district of Chhattisgarh state. You collected data from a sample of 500 households and found the average income (say, $\bar{x}$) to be Rs. 18250 per annum. This sample average may not be equal to the actual average income of Raigarh district of Chhattisgarh (μ) because of sampling error. Thus we are not sure whether average income of the above district is Rs. 18250 or not. On the other hand, it will be more sensible if we say that average income of Raigarh district of Chhattisgarh is between Rs. 17900 and Rs. 18600 per annum. Also we may specify that the probability that average income will remain within these limits is 95 per cent. Thus our confidence interval in this case is Rs. 17900-18600 and the confidence level or confidence coefficient is 95 per cent.

Here a question may be shaping up in your mind, 'How do we find out the confidence interval and confidence coefficient?' Let us begin with confidence coefficient. We know that the sampling distribution of $\bar{x}$ for large samples is normally distributed with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$, where n is the size of the sample. By transforming the sample mean ($z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$) we obtain *standard normal variate*, which has zero mean and unit variance. The standard normal curve is symmetrical and therefore, the area under the curve for $0 \le z \le \infty$ is 0.5 which is presented in the form of a table (See Table 15.1 in Unit 15 of Block 5). Let us assume that we want our confidence coefficient to be 95 per cent (that is, 0.95). Thus we should find out a range for z which will cover 0.95 area of the standard normal curve. Since distribution of z is symmetrical, 0.475 area should remain to the right and 0.475 area should remain to the left of $z = 0$. If look into normal area table (Table 15.1) we find that 0.475 area is covered when $z = 1.96$. Thus the probability that z ranges between −1.96 to 1.96 is 0.95. From this information let us work out backward and find the range within which μ will remain.

We find that

$$P(-1.96 \le z \le 1.96) = 0.95 \qquad \ldots(18.1)$$

$$\text{or } P\left(-1.96 \le \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \le 1.96\right) = 0.95$$

$$\text{or } P\left(-1.96\frac{\sigma}{\sqrt{n}} \le \bar{x}-\mu \le 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$\text{or } P\left(\bar{x}-1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95 \qquad \ldots(18.2)$$

Let us interpret the above. Recall that each sample would provide us with a different value of $\bar{x}$. Accordingly, the confidence interval would be different. In each case the confidence interval would contain the unknown parameter or it would not. Equation (18.2) means that if a large number of random samples, each of size n, are drawn from the given population and if for each such sample the interval $\left(\bar{x}-1.96\frac{\sigma}{\sqrt{n}}, \leq \bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right)$ is determined, then in about 95% of the cases, the interval will include the population mean μ.

The confidence coefficient is denoted by $(1-\alpha)$ where α is the level of significance (we will discuss the concept of 'level of significance' in Unit 19). Confidence coefficient could take any value. We can ask for a confidence level of say 81 per cent or 97 per cent depending upon how precise our conclusions should be. However, conventionally two confidence levels are frequently used, namely, 95 per cent and 99 per cent. Of course at times we take 90 per cent confidence level also, though not frequently.

Let us find out the confidence interval when confidence coefficient $(1-\alpha)=0.99$. In this case 0.495 area should remain on either side of the standard normal curve. If we look into the normal area table (Table 15.1) we find that 0.495 area is covered when z = 2.58.

Thus

$$P(-2.58. \leq z \leq 2.58) = 0.99 \qquad \ldots(18.3)$$

By rearranging the terms in the above we find that

$$P\left(\bar{x}-2.58\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x}+2.58\frac{\sigma}{\sqrt{n}}\right) = 0.99 \qquad \ldots(18.4)$$

Equation (18.4) implies that 99 per cent confidence interval for μ is given by $\bar{x} \pm 2.58\frac{\sigma}{\sqrt{n}}$.

By looking into the normal area table you can work out the confidence interval for confidence coefficient of 0.90 and find that

$$P\left(\bar{x}-1.65\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x}+1.65\frac{\sigma}{\sqrt{n}}\right) = 0.90 \qquad \ldots(18.5)$$

We observe from (18.2), (18.4) and (18.5) that as the interval widens, the chance for the interval holding a population parameter (in this case μ) increases.

The two limits of the confidence interval are called *confidence limits*. For example, for 95 per cent confidence level we have the lower confidence limit as $\bar{x}-1.96\frac{\sigma}{\sqrt{n}}$ and the upper confidence limit as $\bar{x}+1.96\frac{\sigma}{\sqrt{n}}$. The confidence coefficient can be interpreted as the confidence or trust that we place in these limits for actually holding μ.

Example 18.1

A paper company wants to estimate the average time required for a new machine to produce a ream of paper. A random sample of 36 reams shows an average production time of 1.5 minutes per ream of paper. The population standard deviation is known to be 0.30 minute. Construct an interval estimate with 95% confidence limits.

The information given is

$\bar{x} = 1.5$, $\sigma = 0.30$ and $n = 36$

Since $n = 36\ (> 30)$, we can take the sample as a large sample and accordingly $\bar{x}$ is normally distributed with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$. Now, the standard error is

$$\frac{\sigma}{\sqrt{n}} = \frac{0.30}{\sqrt{36}} = 0.05$$

The 95% confidence interval is given by

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

or $1.5 - 1.96 \times 0.05 \le \mu \le 1.5 + 1.96 \times 0.05$

i.e., $1.402 \le \mu \le 1.598$

Thus with 95% confidence, we can state that the average production time for the new machine will be between 1.402 minutes and 1.598 minutes. Here, 1.402 is the lower confidence limit and 1.598 is the upper confidence limit.

18.6 CONFIDENCE INTERVAL FOR UNKNOWN VARIANCE

In the previous Section we estimated confidence interval for population mean on the assumption that population variance is known. It is a bit unrealistic that we do not know population mean (we want to estimate it) but know population variance. A realistic case would be the assumption that both population mean and variance are unknown. On the basis of sample mean and variance we want to find out confidence interval for population mean.

Since the population standard deviation (σ) is not known we use the sample standard deviation (s) in its place. However, in such a case the sampling distribution of $\bar{x}$ is not normal, rather it follows student's *t* distribution. The standard error of the sample means would be $\frac{s}{\sqrt{n}}$.

Like the standard normal variate z, the *t*-distribution has a mean of zero, is symmetrical about mean and ranges between $-\infty$ to ∞. But its variance is greater than 1. Actually its variance changes according to degrees of freedom. However, when $n > 30$ the *t*-distribution has a variance very close to 1 and thus resembles *z*-distribution.

The *t*-statistic, like the *z*-statistic, is calculated as

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

By looking into the area table for *t*-distribution (see Table 15.3 in Unit 15) we find the probability values for the confidence level that we require. The degrees of freedom is $(n-1)$. Thus the confidence interval would be

$$\bar{x} - t \cdot \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t \cdot \frac{s}{\sqrt{n}} \qquad \ldots(18.6)$$

Example 18.2

The mean weight (in kilogram) of 20 children are found to be 15 with a standard deviation of 4. On the basis of the above information estimate 95 per cent confidence interval for mean weight of the population from which the sample is drawn. Assume that population is normally distributed.

Since population is normal and sample size is small we apply *t*-distribution for estimation of confidence interval. Since $n = 20$ we have degrees of freedom (d.f.) = 19. We move down the first column of Table 15.3 till we reach the row corresponding to 19. Since we need 95 per cent confidence interval we should leave 0.025 area on each side of $t = 0$ as we did in the previous Section. Thus for 19 degrees of freedom and $\alpha = 0.025$ we find that *t*-value is 2.093.

Hence the confidence interval is

$$15 - 2.093 \times \frac{4}{\sqrt{20}} \leq \mu \leq 15 + 2.093 \times \frac{4}{\sqrt{20}}$$

or $$15 - 1.87 \leq \mu \leq 15 + 1.87$$

or $$13.13 \leq \mu \leq 16.87$$

Similarly you can find out confidence intervals for different sample sizes and confidence coefficients.

Let us summarise the rules for application of z or *t* statistic for estimation of confidence interval.

1) If sample size is large ($n>30$) apply z-statistic — it does not matter whether i) parent population is normal or not, and ii) variance is known or not.
2) If sample size is small ($n \leq 30$) check whether i) parent population is normal, and ii) variance is known.
   a) If parent population is not normal apply nonparametric tests.
   b) If parent population is normal and variance is known apply *z*.
   c) If parent population is normal and variance is not known apply *t*.

In Fig. 18.2 we present the above in the form of a chart.

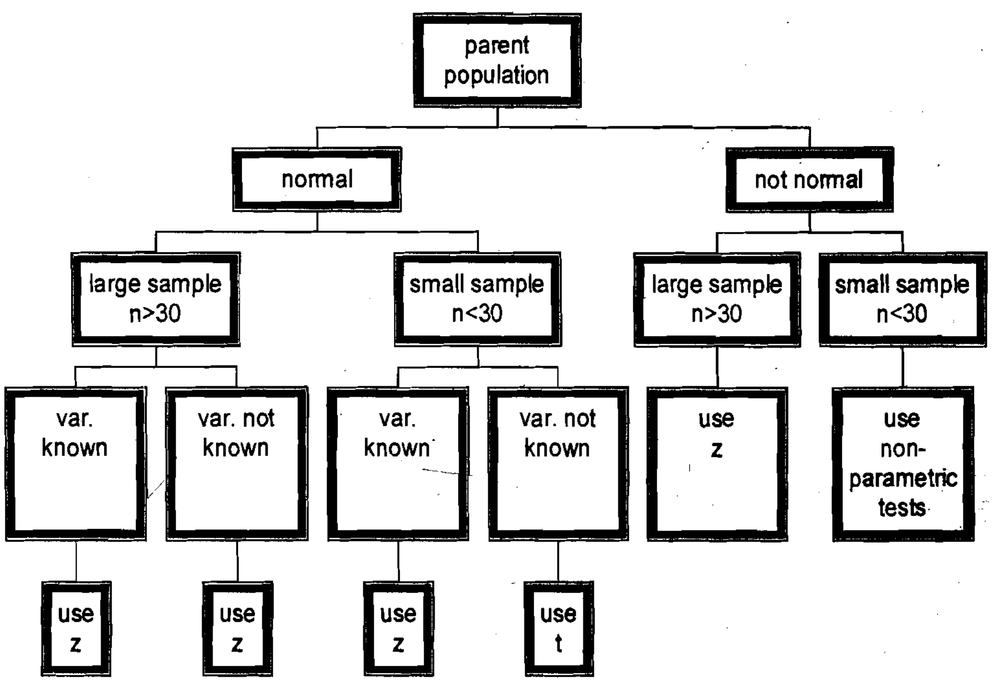

Fig. 18.2: Selection of Proper Test Statistic

Check Your Progress 2

1) A sample of 50 employees were asked to provide the distance commuted by them to reach office. If sample mean was found to be 4.5 km. find 95 percent confidence interval for the population. Assume that population is normally distributed with a variance of 0.36.

....

....

....

2) For a sample of 25 students in school the mean height was found to be 95 cm. with a standard deviation of 4 cm. Find the 99 percent confidence interval.

....

....

....

3) State whether the following statements are true or false.

   a) When parent population is not normal and sample size is small we use t-distribution to estimate confidence interval.

   b) The range of t-distribution is 0 to infinity.

   c) When confidence level is 90 per cent, level of significance is 10 per cent.

....

....

....

18.7 LET US SUM UP

Drawing conclusions about a population on the basis of sample information is called statistical inference. Here we have basically two things to do: estimation and hypothesis testing. In this unit we took up the first issue while the second one will be discussed in the remaining units of the block.

An estimate of an unknown parameter could be either a point or an interval. Sample mean is usually taken as a point estimate of population mean. On the other hand, in interval estimation we construct two limits (upper and lower) around the sample mean. We can say with stipulated level of confidence that the population mean, which we do not know, is likely to remain within the confidence interval. In order to construct confidence interval we need to know the population variance or its estimate. When we know population variance, we apply normal distribution to construct the confidence interval. In cases where population variance is not known, we use student's t for the above purpose. Remember that when sample size is large (n>30) t-distribution approximates normal distribution. Thus for large samples, even if population variance is not known, we can use normal distribution for construction of confidence interval on the basis of sample mean and sample variance.

18.8 KEY WORDS

Confidence Level : It gives the percentage (probability) of samples where the population mean would remain within the confidence interval around the sample mean. If α is the significance level the confidence level is $(1-\alpha)$.

Estimation : It is the method of prediction about parameter value on the basis of statistic.

Estimator : It is another name given to statistic in the theory of estimation.

Parameter : It is a measure of some characteristic of the population.

Population : It is the entire collection of units of a specified type in a given place and at a particular point of time.

Random Sampling : It is a procedure where every member of the population has a definite chance or probability of being selected in the sample. It is also called probability sampling.

Sample : It is a sub-set of the population. It can be drawn from the population in a scientific manner by applying the rules of probability so that personal bias is eliminated. Many samples can be drawn from a population and there are many methods of drawing a sample.

Sampling Error : In the sampling method, we try to approximate some feature of a given population from a sample drawn from it. Now, since in the sample all the members of the population are not included, howsoever close the approximation is, it is not identical to the required population feature and some error is committed. This error is called the sampling error.

Significance Level : There may be certain samples where population mean would not remain within the confidence interval around sample mean. The percentage (probability) of such cases is called significance level. It is usually denoted by α. When $\alpha = 0.05$ (that is, 5 percent) we can say that in 5 per cent cases we are likely to reach an incorrect decision or commit Type I error. Level of significance could be at any level but it is usually taken at 5 percent or 1 percent level.

Statistic : It is a function of the values of the units that are included in the sample. The basic purpose of a statistic is to estimate some population parameter.

Sampling Distribution : It is the relative frequency or probability distribution of the values of a statistic when the number of samples tends to infinity.

Standard Error : It is the standard deviation of the sampling distribution of a statistic.

Statistical Inference : It is the process of concluding about an unknown population from a known sample drawn from it.

Problem of Estimation : We may be interested in some feature of the population that is *completely* unknown to us and we want to make some intelligent guess about it on the basis of a random sample drawn from the population. This problem of statistical inference is known as the problem of estimation.

18.9 SOME USEFUL BOOKS

Nagar, A. L. and Das, R. K., 1989, *Basic Statistics:* Oxford University Press, Delhi, Chapter 9.

Newbold, P., 1991, *Statistics for Business and Economics* (Third Edition): Prentice Hall, New Jersey, Chapters 6, 7, 8 and 9.

Keller, G, and B. Warrack, 1991, *Essentials of Business Statistics*, Wordsworth Publishing Co., California, Chapters 7 and 8.

18.10 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Go through Section 18.2 and answer.
2) a) true b) true c) true d) true

Check Your Progress 2

1) Since it is large sample we apply *z*-statistic. The confidence interval is

   $4.40 \le \mu \le 4.60$

2) Since it is small sample and population variance is not given we apply *t*-statistic with degrees of freedom 24. The tabulated value of *t* at 99 per cent confidence level is 2.49. The confidence interval is $93.01 \le \mu \le 96.99$.

3) a) false b) false c) true