
UNIT 9 ESTIMATING PARAMETERS

Structure

- 9.0 Introduction
- 9.1 Objectives
- 9.2 Basic Terminology
- 9.3 Characteristics of Estimators
- 9.4 Point Estimation
- 9.5 Method of Maximum Likelihood
- 9.6 Interval Estimation
- 9.7 Confidence Interval for Population Mean
 - 9.7.1 Confidence Interval for Population Mean when Population Variance is Known
 - 9.7.2 Confidence Interval for Population Mean when Population Variance is Unknown
 - 9.7.3 Confidence Interval for Difference of Two Population Means when Samples are Independent and Population Variances are known
 - 9.7.4 Confidence Interval for Difference of Two Population Means when Population Variances are Unknown.
- 9.8 Confidence Interval for Population Proportion
- 9.9 Determination of Sample Size
- 9.10 Summary
- 9.11 Solutions/ Answers
- 9.12 References

9.0 INTRODUCTION

In many real-life problems, population parameter(s) is (are) unknown and someone is interested to obtain the value(s) of the parameter(s). But if the whole population is too large to study or the units of the population are destructive in nature or there are limited resources and manpower available, then it is not practically convenient to examine each unit of the population to find the value(s) of the parameter(s). In such situations, one can draw a sample from the population under study and utilise sample observations to estimate the parameter(s). The technique of finding an estimator to produce an estimate of the unknown parameter on the basis of a sample is called estimation. There are two methods of estimation:

1. Point Estimation
2. Interval Estimation

In point estimation, we determine an appropriate single statistic whose value is used to estimate the unknown parameter, whereas in interval estimation, we determine an interval that contains the true value of the unknown parameter with a certain confidence. The unit also elaborates on confidence intervals for population means for one and two variables under different conditions and the confidence interval for population proportions.

9.1 OBJECTIVES

After studying this unit, you should be able to:

- define the parameter space and estimator;

- explain the properties of an estimator;
- explore the different methods of point estimation;
- explain the method of maximum likelihood;
- describe the properties of maximum likelihood estimators;
- define the interval estimation;
- obtain the confidence interval for population mean of a normal population when population variance is known and unknown;
- describe the method of obtaining the confidence interval for difference of means of two normal populations when variances are known and unknown;
- obtain the confidence interval for population proportion; and
- determination of sample size.

9.2 BASIC TERMINOLOGY

Before discussing the properties of a good estimator, we discuss basic definitions of some important terms. These terms are very useful in understanding the fundamentals of the theory of estimation discussed in this block.

Discrete and Continuous Distributions

In the earlier Units, we have discussed standard discrete and continuous distributions as binomial, Poisson, normal, etc. We know that the populations can be described with the help of distributions, therefore, standard discrete and continuous distributions are used in statistical inference. Here, we present some standard discrete and continuous distributions in brief as in tabular form:

S. No.	Distribution	Parameter (s)	Mean	Variance
1	Bernoulli (discrete) $P[X = x] = p^x (1 - p)^{1-x}; x = 0, 1$	p	p	pq
2	Binomial (discrete) $P[X = x] = {}^n C_x p^x q^{n-x}; x = 0, 1, \dots, n$	$n \text{ \& } P$	np	npq
3	Poisson (discrete) $P[X = x] = \frac{e^{-\lambda} \lambda^x}{x!}; x = 0, 1, \dots \text{ \& } \lambda > 0$	λ	λ	λ
4	Uniform (discrete) $P[X = x] = \frac{1}{n}; x = 0, 1, \dots, n$	n	$\frac{n+1}{2}$	$\frac{n^2 - 1}{12}$
5	Hypergeometric (discrete) $P[X = x] = \frac{{}^M C_x {}^{N-M} C_{n-x}}{{}^N C_n}; x = 0, 1, \dots, \min\{M, n\}$	$N, M \text{ \& } n$	$\frac{nM}{N}$	$\frac{NM(N-M)(N-n)}{N^2(N-1)}$

6	Normal(continuous) $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}; -\infty < x < \infty$ $\& \sigma > 0, -\infty < \mu < \infty$	$\mu \& \sigma^2$	μ	σ^2
7	Standard Normal(continuous) $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}; -\infty < x < \infty$	--	0	1
8	Uniform (continuous) $f(x) = \frac{1}{b-a}; a < x < b, b > a$	a & b	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$

Parameter Space

The set of all possible values that the parameter θ or parameters $\theta_1, \theta_2, \dots, \theta_k$ can assume is called the parameter space. It is denoted by Θ and is read as “**big theta**”. For example, if parameter θ represents the average life of electric bulbs manufactured by a company, then parameter space of θ is $\Theta = \{\theta : \theta \geq 0\}$, that is, the parameter average life θ can take all possible values greater than or equal to 0. Similarly, in a normal distribution (μ, σ^2) , the parameter space of parameters μ and σ^2 is $\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty; 0 < \sigma < \infty\}$.

9.3 CHARACTERISTICS OF ESTIMATORS

It is to be noted that a large number of estimators can be proposed for an unknown parameter. For example, if we want to estimate the average income of the persons living in a city then the sample mean, sample median, sample mode, etc. can be used to estimate the average income. Now, the question arises, “Are some of the possible estimators better, in some sense, than the others?” Generally, an estimator can be called good for two different situations:

- (i) **When the true value of the parameter is being estimated is known**— An estimator might be called good if its value is close to the true value of the parameter to be estimated. In other words, the estimator whose sampling distribution concentrates as closely as possible near the true value of the parameter may be regarded as the good estimator.
- (ii) **When the true value of the parameter is unknown**— An estimator may be called good if the data give good reason to believe that the estimate will be close to the true value.

In the whole estimation, we estimate the parameter when the true value of the parameter is unknown. Hence, we must choose estimates not because they are certainly close to the true value, but because there is a good reason to believe that the estimated value will be close to the true value of the parameter. Let us describe certain properties, which help us in deciding whether an estimator is better than others.

Prof. Ronald A. Fisher was the man who pushed ahead the theory of estimation and introduced these concepts and gave some properties of a good estimator as follows:

1. Unbiasedness
2. Consistency



*Prof. Ronald A. Fisher,
a Great English
Mathematical
Statistician.
(1890-1962)*

3. Efficiency
4. Sufficiency

We shall discuss these properties one by one in this section.

1. Unbiasedness

Generally, population parameter(s) is (are) unknown and if the whole population is too large to study to find the value of an unknown parameter(s) then one can estimate the population parameter(s) with the help of estimator(s) which is (are) always a function of sample values.

An estimator is said to be unbiased if the expected value of the estimator is equal to the true value of the parameter being estimated.

An estimator is said to be unbiased for the population parameter such as population mean, population variance, population proportion, etc. if and only if the average or mean of the sampling distribution of the estimator is equal to the true value of the parameter.

Now, we explain the procedure to show that a statistic is unbiased or not for a parameter with the help of some examples:

Example 1: A random sample of 10 cadets of a centre is selected and their weights (in kg) are measured, which are given below:

48, 50, 62, 75, 80, 60, 70, 56, 52, 78

Determine an unbiased estimate of the average weight of cadets of the centre.

Solution: We know that sample mean ($\bar{X}$) is an unbiased estimator of the population mean and its particular value is the unbiased estimate of the population mean, therefore,

$$\begin{aligned}\bar{X} &= \frac{1}{n} \sum_{i=1}^n X_i = \frac{X_1 + X_2 + \dots + X_n}{n} \\ &= \frac{48 + 50 + 62 + 75 + 80 + 60 + 70 + 56 + 52 + 78}{10} = 63.10\end{aligned}$$

Hence, an unbiased estimate of the average weight of cadets of the centre is 63.10 kg.

Note 1: If $X_1, X_2, \dots, X_n$ is a random sample taken from a population with mean μ and variance σ^2 , then

$$S'^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \text{ is a biased estimator of } \sigma^2.$$

whereas, $S^2 = \frac{n}{n-1} S'^2$ i.e. $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ is an unbiased estimator of σ^2 .

The proof of the above result is beyond the scope of this course.

Remark:

1. Unbiased estimators may not be unique. For example, the sample mean and sample median are unbiased estimators of the population mean of a normal population.
2. Unbiased estimators do not always exist for all the parameters. For example, for a Bernoulli distribution (θ), there is no unbiased estimator for θ^2 .

Similarly, for a Poisson distribution (λ), there exists no unbiased estimator for $1/\lambda$.

3. If an estimator is unbiased for all types of distribution, then it is called an absolutely unbiased estimator. For example, sample mean is an absolutely unbiased estimator of the population mean, if population mean exists.

One weakness of unbiasedness is that it requires only the average value of the estimator equals the true value of the population parameter. It does not require those values of the estimator to be reasonably close to the population parameter. For this reason, we require some other properties of a good estimator - consistency, efficiency and sufficiency which are described next.

2. Consistency

As defined earlier, an estimator T is said to be an unbiased estimator of a parameter, say, θ if the mean of the sampling distribution of estimator T is equal to the true value of the parameter θ . This concept was defined for a fixed sample size. Now, we will learn about consistency which is defined for increasing sample size.

If $X_1, X_2, \dots, X_n$ is a random sample of size n taken from a population whose probability density (mass) function is $f(x, \theta)$ where, θ is the population parameter then consider a sequence of estimators, say, $T_1 = t_1(X_1)$, $T_2 = t_2(X_1, X_2)$, $T_3 = t_3(X_1, X_2, X_3), \dots, T_n = t_n(X_1, X_2, \dots, X_n)$. A sequence of estimators is said to be consistent for a parameter θ if the deviation of the values of an estimator from the parameter tends to zero as the sample size increases. That means values of the estimator tend to get closer to the parameter θ as sample size increases.

Remark:

1. Consistent estimators may not be unique. For example, sample mean and sample median both are consistent estimators of the population mean of a normal population.
2. An unbiased estimator may or may not be consistent.
3. A consistent estimator may or may not be unbiased.

3. Efficiency

In some situations, we see that there are more than one estimators of a parameter which are unbiased as well as consistent. For example, sample mean and sample median both are unbiased and consistent for the parameter μ when sampling is done from a normal population with mean μ and known variance σ^2 . In such situations, there arises a necessity of some other criterion which will help us to choose 'best estimator' among them. A criterion which is based on the concept of the variance of the sampling distribution of the estimator is termed as efficiency.

If T_1 and T_2 are two estimators of a parameter θ , then T_1 is said to be more efficient than T_2 for all sample sizes if

$$\text{Var}(T_1) < \text{Var}(T_2) \quad \text{for all } n$$

4. Sufficiency

In statistical inference, the aim of the investigator or statistician may be to decide the value of the unknown parameter (θ). The information that guides the investigator in making a decision is supplied by the random sample

$X_1, X_2, \dots, X_n$. However, in most of the cases, the observations would be too numerous and too complicated. Direct use of these observations is complicated or cumbersome, therefore, a simplification or condensation would be desirable. The technique of condensing or reducing the random sample $X_1, X_2, \dots, X_n$ into a statistic such that it contains all the information about parameter θ that is contained in the sample is known as sufficiency. So before continuing our search of finding the best estimator, we introduce the concept of sufficiency.

A sufficient statistic is a particular kind of statistic that condenses random sample $X_1, X_2, \dots, X_n$ in a statistic $T = t(X_1, X_2, \dots, X_n)$ in such a way that no information about parameter θ is lost. That means, it contains all the information about θ that is contained in the sample and if we know the value of sufficient statistic, then the sample values themselves are not needed and can tell you nothing more about θ . In other words:

A statistic T is said to be sufficient statistic for estimating a parameter θ if it contains all the information about θ which are available in the sample.

This property of an estimator is called sufficiency.

9.4 POINT ESTIMATION

The technique of estimating the unknown parameter with a single value is known as point estimation.

There are so many situations in our day to day life where we need to estimate some unknown parameter(s) of the population on the basis of the sample observations. For example, a housewife may want to estimate the monthly expenditure, a sweet shopkeeper may want to estimate the sales of sweets on a day, a student may want to estimate the study hours for the reading of a particular unit of this course, etc. This need is fulfilled by the technique of estimation. So the technique of finding an estimator to produce an estimate of the unknown parameter is called estimation.

Estimation is broadly divided into two categories namely:

- Point estimation and
- Interval estimation

If we find a single value with the help of sample observations which is taken as the estimated value of unknown parameter then this value is known as a point estimate and the technique of estimating the unknown parameter with a single value is known as “**point estimation**”.

If instead of finding a single value to estimate the unknown parameter if we find two values between which the parameter may be considered to lie with a certain probability(confidence) is known as interval estimate of the parameter and this technique of estimating is known as “**interval estimation**”.

For example, if we estimate the average weight of men living in a colony on the basis of sample mean, say, 62 kg then 62 kg is called point estimate of the average weight of men in the colony and this procedure is called as point estimation. If we estimate the average weight of men by an interval, say, [40,110] with 90% confidence that true value of the weight lies in this interval then this interval is called interval estimate and this procedure is called as interval estimation.

Now, the question may arise in your mind that “how the point and interval estimates are obtained?” So we will describe one of the important and frequently used methods of point estimation, next.

Methods of Point Estimation

Some of the important and frequently used methods of point estimation are:

1. Method of maximum likelihood
2. Method of moments
3. Method of least squares
4. Method of minimum chi-square
5. Method of minimum variance

In the next section, we will explain the method of maximum likelihood. Other methods are beyond the scope of this course.

Check Your Progress 1

- 1) Write the four properties of a good estimator.

.....

- 2) Find which technique of estimation (point estimation or interval estimation) is used in each case given below:

- (i) An investigator estimates average income Rs. 1.5 lakh per annum of the people living in a particular geographical area, on the basis of a sample of 50 people taken from that geographical area.
- (ii) A product manager of a company estimates the average life of electric bulbs in the range of 800 hours and 1000 hours, with a certain confidence, on the basis of a sample of 20 bulbs.
- (iii) A pathologist estimates the mean time required to complete a certain analysis in the range 30 minutes to 45 minutes, with a certain confidence, on the basis of a random sample of 25.

.....

- 3) List any two methods of point estimation.

.....

.....

9.5 METHOD OF MAXIMUM LIKELIHOOD

For describing the method of maximum likelihood, first, we have to define likelihood function.

Likelihood Function

If $X_1, X_2, \dots, X_n$ is a random sample of size n taken from a population with joint probability density (mass) function $f(x_1, x_2, \dots, x_n, \theta)$ of sample values then likelihood function is denoted by $L(\theta)$ and is defined as follows:

$$L(\theta) = f(x_1, x_2, \dots, x_n, \theta)$$

For discrete case,

$$L(\theta) = P[X_1 = x_1]P[X_2 = x_2] \dots P[X_n = x_n]$$

For continuous case,

$$L(\theta) = f(x_1, \theta) \cdot f(x_2, \theta) \dots f(x_n, \theta)$$

From the theoretical point of view, one of the most important methods of point estimation is method of maximum likelihood because it generally gives very good estimators as judged from various criteria. It was initially given by Prof. C.F. Gauss but later on, it was used as a general method of estimation by Prof. Ronald A. Fisher in 1912. The principle of maximum likelihood estimation is to find /estimate /choose the value of the unknown parameter which would most likely generate the observed data. We know that the likelihood function gives the relative likelihoods for different values of the parameters for the observed data. Therefore, we search the value of the unknown parameter for which the likelihood function is maximum corresponding to the observed data. The concept of maximum likelihood estimation is explained with a simple example given below:

Suppose we toss a coin 5 times and we observe 3 heads and 2 tails. Instead of assuming that the probability of getting head is 0.5, we want to find/ estimate the value of p that makes the observed data most likely. Since the number of heads follows the binomial distribution, therefore, the probability (likelihood function) of getting 3 heads in 5 tosses is given by

$$P[X = 3] = {}^5C_3(p)^3(1-p)^2$$

Imagine that p was 0.1 then

$$P[X = 3] = {}^5C_3(0.1)^3(0.1)^2 = 0.0081$$

Similarly, for different values of p the probability of getting 3 heads in 5 tosses is given in Table 9.1 given below:

Table 9.1: Probability/Likelihood Function Corresponding to Different Values of p

S. No.	p	Probability/ Likelihood Function
1	0.1	0.0081
2	0.2	0.0512
3	0.3	0.1323
4	0.4	0.2304
5	0.5	0.3125
6	0.6	0.3456
7	0.7	0.3087
8	0.8	0.2048
9	0.9	0.0729

From Table 9.1, we can conclude that p is more likely to be 0.6 because at $p = 0.6$ the probability is maximum or the likelihood function is maximum.

Therefore, the principle of maximum likelihood (ML) consists in finding an estimate for the unknown parameter θ within the admissible range of θ , i.e.

within parameter space Θ , which makes the likelihood function as large as possible, that is, maximise the likelihood function. Such an estimate is known as maximum likelihood estimate for unknown parameter θ . Thus, if there exists an estimate, say, $\hat{\theta}(x_1, x_2, \dots, x_n)$ of the sample values which maximises likelihood function $L(\theta)$, is known as “**maximum likelihood estimate**”. That is,

$$L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta)$$

Example 2: If the number of weekly accidents occurring on a mile stretch of a particular road follows a Poisson distribution with parameter λ then find the maximum likelihood estimate of parameter λ on the basis of the following data:

Number of Accidents	0	1	2	3	4	5	6
Frequency	10	12	12	9	5	3	1

Solution: Here, the number of weekly accidents occurring on a mile stretch of a particular road follows a Poisson distribution with parameter λ so probability mass function (pmf) of the Poisson distribution is given by

$$P[X = x] = \frac{e^{-\lambda} \lambda^x}{x!}; \quad x = 0, 1, 2, \dots \text{ \& } \lambda > 0$$

We find the theoretical maximum likelihood estimate of the parameter λ (refer to references). It was found that maximum likelihood estimate of parameter λ of Poisson distribution is sample mean.

Since we observed that the sample mean is the maximum likelihood estimate of parameter λ so we calculate the sample mean of the given data as

S. No.	Number of Accidents(X)	Frequency(f)	fX
1	0	10	0
2	1	12	12
3	2	12	24
4	3	9	27
5	4	5	20
6	5	3	15
7	6	1	6
		N = 52	$\sum fX = 104$

The formula for calculating mean is

$$\begin{aligned} \bar{X} &= \frac{1}{N} \sum fX \text{ where, N is the total number of accidents} \\ &= \frac{1}{52} \times 104 = 2 \end{aligned}$$

Hence, maximum likelihood estimate of λ is 2.

Example 3: For random sampling from normal population $N(\mu, \sigma^2)$, find the maximum likelihood estimators for μ and σ^2 .

Estimation Parameters

$\hat{\theta}$ - is used to represent the estimate / estimator of parameter and read as cap i.e.

$\hat{\theta}$ - means it is the estimate of parameter θ and read as theta cap.

Solution: Let $X_1, X_2, \dots, X_n$ be a random sample of size n taken from normal population $N(\mu, \sigma^2)$, whose probability density function is given by

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}; -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$$

The likelihood function for parameters μ and σ^2 are computed in the references.

The maximum likelihood estimates for μ and σ^2 are $\bar{x}$ and s'^2 respectively. Hence, ML estimators for μ and σ^2 are $\bar{X}$ and S'^2 respectively.

Note 2: Since throughout the course we are using capital letters for estimators, therefore in the last line of the above example we use capital letters for maximum likelihood estimators for μ and σ^2 .

Properties of Maximum Likelihood (ML) Estimators

The following are the properties of maximum likelihood estimators:

1. A ML estimator is not necessarily unique.
2. A ML estimator is not necessarily unbiased.
3. A ML estimator may not be consistent in a rare case.
4. If a sufficient statistic exists, it is a function of the ML estimators.
5. If $T = t(X_1, X_2, \dots, X_n)$ is a ML estimator of θ and $\gamma(\theta)$ is a one to one function of θ , then $\gamma(T)$ is a ML estimator of $\gamma(\theta)$. This is known as invariance property of ML estimator.
6. When ML estimator exists, then it is most efficient in the group of such estimators.

9.6 INTERVAL ESTIMATION

When we find two values with the help of sample observations and constitute an interval such that it contains the true value of the parameter with a certain probability, then it is known as an interval estimate of the parameter. This technique of estimation is known as “Interval Estimation”.

In this section, we will define:

- Confidence Interval and Confidence Coefficient
- One-sided Confidence Intervals

in the following two sub-sections.

Confidence Interval and Confidence Coefficient

Let $X_1, X_2, \dots, X_n$ be a random sample of size n taken from a population whose probability density (mass) function is $f(x, \theta)$. Let $T_1 = t_1(X_1, X_2, \dots, X_n)$ and $T_2 = t_2(X_1, X_2, \dots, X_n)$ (where $T_1 \leq T_2$) be two statistics such that the probability that the random interval $[T_1, T_2]$ includes the true value of population parameter θ is $(1 - \alpha)$, that is,

$$P[T_1 \leq \theta \leq T_2] = 1 - \alpha$$

as shown in the Fig.9.1, where, α does not depend on θ .

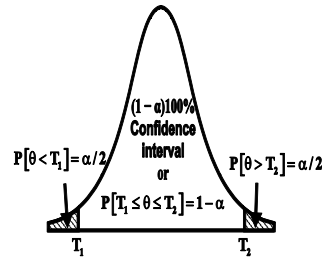


Fig. 9.1: Confidence Interval

Then the random interval $[T_1, T_2]$ is known as $(1-\alpha)$ 100% confidence interval for unknown population parameter θ and $(1-\alpha)$ is known as confidence coefficient or confidence level. The above probability statement may be explained with the help of an example:

Suppose we say that the probability that the random interval contains the true value of parameter θ is 0.95 then by this statement we simply mean that if 100 samples of the same size, say, 'n' are drawn from the given population $f(x, \theta)$ and the random interval $[T_1, T_2]$ is computed for each sample, then 95 times out of 100 intervals, the random interval $[T_1, T_2]$ contains the true value of the parameter θ . Hence, the higher the probability $(1-\alpha)$, more the confidence we will have that the random interval $[T_1, T_2]$ will include the true value of θ . The statistics T_1 and T_2 are known as lower and upper confidence limits or confidence bounds or fiducial limits, respectively for θ . And the interval is known as two-sided confidence interval. The length of the confidence interval is defined as

$$L = \text{Upper confidence limit} - \text{Lower confidence limit}$$

$$\text{i.e. } L = T_2 - T_1$$

The confidence interval may also be one-sided so we define one-sided confidence intervals next.

One-Sided Confidence Intervals

In some situations, we may be interested in finding an upper bound or lower bound but not both for population parameter with a given confidence. For example, one may be interested to obtain a bound such that he /she is 95% confident that the average life of the electric bulbs of a company is no less than one years. In such cases, we construct one-sided confidence intervals.

Let $X_1, X_2, \dots, X_n$ be a random sample of size n taken from the population having probability density(mass) function $f(x, \theta)$ and also let T_1 be a statistic such that

$$P[\theta \geq T_1] = 1 - \alpha$$

as shown in Fig. 9.2, then statistic T_1 is called lower confidence bound for parameter θ with confidence coefficient $(1-\alpha)$ and $[T_1, \infty)$ is called a lower one-sided $(1-\alpha)$ 100% confidence interval for parameter θ .

Similarly, let T_2 be a statistic such that

$$P[\theta \leq T_2] = 1 - \alpha$$

as shown in Fig. 9.3, then statistic T_2 is called upper confidence bound for parameter θ with confidence coefficient $(1-\alpha)$ and $(-\infty, T_2]$ is called an upper one-sided $(1-\alpha)$ 100% confidence interval for parameter θ .

Note 3: Generally, one-sided confidence intervals are rarely used so we focus on two-sided confidence interval in this course. **Generally, confidence interval means two-sided confidence interval unless or otherwise it is stated as one-sided.**

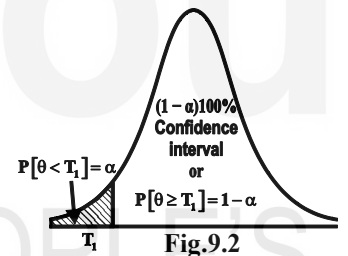


Fig.9.2

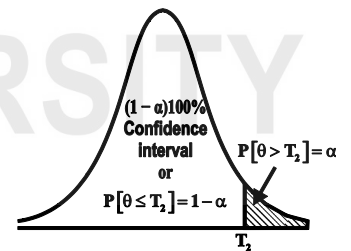


Fig. 9.3

Check Your Progress 2

- 1) What is the density function of the binomial population? What is the maximum likelihood estimator for p in a binomial population?

.....
.....
.....

- 2) List any five properties of maximum likelihood estimators.

.....
.....
.....

- 3) Find the length of the following confidence intervals:

- (i) $P[-1.65 \leq \mu \leq 3.0] = 0.95$ (ii) $P[-1.68 \leq \mu \leq 2.70] = 0.95$
(iii) $P[-1.70 \leq \mu \leq 2.54] = 0.95$ (iv) $P[-1.96 \leq \mu \leq 1.96] = 0.95$

.....
.....
.....

- 4) Find the lower and upper confidence limits and also the confidence coefficient of the following confidence intervals:

- (i) $P[0 \leq \theta \leq 1.5] = 0.90$ (ii) $P[-1 \leq \theta \leq 2] = 0.95$
(iii) $P[-2 \leq \theta \leq 2] = 0.98$ (iv) $P[-2.5 \leq \mu \leq 2.5] = 0.99$

.....
.....
.....

9.7 CONFIDENCE INTERVAL FOR POPULATION MEAN

There are so many problems in real life where it becomes necessary to obtain the confidence interval of the population mean. For example, an investigator may interest to find the interval estimate of the average income of the people living in a particular geographical area, a product manager may want to find the interval estimate of the average life of electric bulbs manufactured by a company, a pathologist may want to obtain the interval estimate of the mean time required to complete a certain analysis, etc.

For describing confidence interval for population mean, let $X_1, X_2, \dots, X_n$ be a random sample of size n taken from a normal population having mean μ and variance σ^2 . We can determine confidence interval for population mean μ under the following two cases:

1. When population variance σ^2 is known

2. When population variance σ^2 is unknown.

These two cases for one population are discussed in Sections 9.7.1 and 9.7.2, and for two populations in Sections 9.7.3 and 9.7.4.

9.7.1 Confidence Interval for Population Mean when Population Variance is Known

Let $X_1, X_2, \dots, X_n$ be a random sample of size n taken from normal population $N(\mu, \sigma^2)$ when σ^2 is known, that is, σ^2 has a specified value, say, σ_0^2 .

To find out the confidence interval for population mean, first of all, we search the statistic for estimating μ whose distribution is completely known. Generally, we use the value of statistic (sample mean) $\bar{X}$ to estimate the population mean μ and also it is a sufficient statistic for parameter θ . Therefore, we use $\bar{X}$ to make the pivotal quantity.

We know that when the parent population is normal $N(\mu, \sigma^2)$, then sampling distribution of the sample mean $\bar{X}$ is normally distributed with mean μ and variance σ^2/n , that is, if

$$X_i \sim N(\mu, \sigma^2)$$

then

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

and the variate

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

follows the normal distribution with mean 0 and variance unity. Therefore, the probability density function of Z is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}; -\infty < z < \infty$$

This can be used to compute the $(1-\alpha)$ 100% confidence interval of the population mean as:

$$\left[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] \quad \dots (1)$$

and corresponding limits are given by

$$\bar{X} \mp z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \dots (2)$$

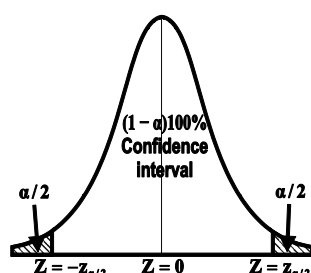


Fig. 9.4: Confidence Interval of Population Mean (Variance Known)

Note 4: In interval estimation generally, we have to find the 90%, 95%, 98% and 99% confidence intervals, therefore, the corresponding values of $z_{\alpha/2}$ are summaries in Table 9.2 and we will use these values directly when needed.

Table 9.2: Communally Used Values of Standard Normal Variate Z

$1 - \alpha$	0.90	0.95	0.98	0.99
α	0.10	0.05	0.02	0.01
$\alpha/2$	0.05	0.025	0.01	0.005
$z_{\alpha/2}$	1.645	1.96	2.33	2.58

For example, if we want to find the 99% confidence interval (two-sided) for μ then

$$1 - \alpha = 0.99 \Rightarrow \alpha = 0.01$$

For $\alpha = 0.01$, the value of $z_{\alpha/2} = z_{0.005}$ is 2.58 therefore, 99% confidence interval for μ is given by

$$\left[\bar{X} - 2.58 \frac{\sigma}{\sqrt{n}}, \bar{X} + 2.58 \frac{\sigma}{\sqrt{n}} \right]$$

The application of the above discussion can be seen in the following example.

Example 4: The mean life of the tyres manufactured by a company follows a normal distribution with standard deviation 3200 km. A sample of 250 tyres is taken and it is found that the average life of the tyres is 50000 km with a standard deviation of 3500 km. Establish the 99% confidence interval within which the mean life of tyres of the company is expected to lie.

Solution: Here, we are given that

$$n = 250, \sigma = 3200, \bar{X} = 50000, S = 3500$$

Since population standard deviation i.e. population variance σ^2 is known, therefore, we use $(1-\alpha)$ 100% confidence limits for population mean when population variance is known which are given by

$$\bar{X} \mp z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \dots(3)$$

where, $z_{\alpha/2}$ is the value of the variate Z having an area of $\alpha/2$ under the right tail of the probability curve of Z and for 99% confidence interval, we have $1 - \alpha = 0.99 \Rightarrow \alpha = 0.01$. For $\alpha = 0.01$ we have, $z_{\alpha/2} = z_{0.005} = 2.58$.

Therefore, the 99% confidence limits are

$$\bar{X} \mp 2.58 \frac{\sigma}{\sqrt{n}}$$

By putting the values of n , $\bar{X}$ and σ , the 99% confidence limits are

$$\begin{aligned} & 50000 \mp 2.58 \times \frac{3200}{\sqrt{250}} \\ & 50000 \mp 522.20 = 49477.80 \text{ and } 50522.20 \end{aligned}$$

Hence, 99% confidence interval within which the mean life of tyres of the company is expected to lie is

$$[49477.80, 50522.20]$$

9.7.2 Confidence Interval for Population Mean when Population Variance is Unknown

In the cases, described in the previous sub-section we assume that variance σ^2 of the normal population is known but in general, it is not known and in such a

situation the only alternative left is to estimate the unknown σ^2 . The value of sample variance (S^2) is used to estimate the σ^2 where,

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$$

In this situation, we know that the variate

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

follows t-distribution with $(n-1)$ df. Therefore, the probability density function of statistic t is given by

$$f(t) = \frac{1}{B\left(\frac{n-1}{2}, \frac{1}{2}\right) \sqrt{n-1}} \left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}}; \quad -\infty < t < \infty$$

This can be used to find $(1-\alpha)$ 100% confidence interval for population mean of the normal population when variance σ^2 is unknown as:

$$\left[\bar{X} - t_{(n-1), \alpha/2} \frac{S}{\sqrt{n}}, \quad \bar{X} + t_{(n-1), \alpha/2} \frac{S}{\sqrt{n}} \right] \quad \dots (4)$$

and corresponding limits are given by

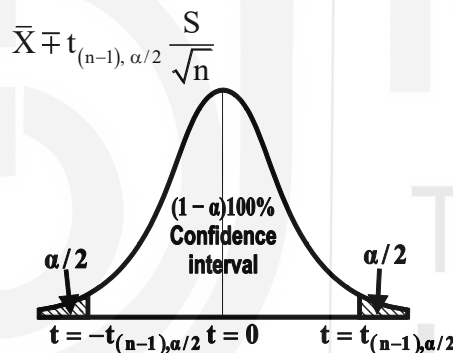


Fig. 9.5: Confidence Interval using t-Distribution (Variance Unknown)

Note 5: For different confidence intervals and different degrees of freedom, the values of $t_{(n-1), \alpha/2}$ are different. Therefore, for given values of α and n we read the tabulated value of t-statistic from the table of t-distribution (t-table) given in references.

For example, if we want to find the 95% confidence interval for μ when $n = 8$, then we have

$$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$$

From the t-table, for $\alpha = 0.05$ and $v = n - 1 = 7$, we have the value of t as:

$$t_{(n-1), \alpha/2} = t_{(7), 0.025} = 2.365.$$

You may observe in the t-table of the References that when the sample size is greater than 30 ($n > 30$) then all values of variate t are not given in this table so for convenience when n is sufficiently large (> 30) then we know that almost all the distributions are very closely approximated by a normal distribution. Thus, in this case t-distribution is also approximated a normal distribution. So the variate

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim N(0, 1) \quad \dots(5)$$

also follows the normal distribution with mean 0 and variance unity. Therefore, when population variance is unknown and the sample size is large then the $(1-\alpha)$ 100% confidence interval for population mean may be obtained by using the same procedure as we have followed in the case when σ^2 is known by taking S^2 in place of σ^2 which is given as

$$\left[\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}} \right] \quad \dots (6)$$

and corresponding limits are given by

$$\bar{X} \mp z_{\alpha/2} \frac{S}{\sqrt{n}} \quad \dots (7)$$

The following example will explain the application of t-distribution.

Example 5: It is known that the average weight of students of a Study Centre of IGNOU follows a normal distribution. To estimate the average weight, a sample of 10 students is taken from this Study Centre and measured their weights (in kg) which are given below:

48, 50, 62, 75, 80, 60, 70, 56, 52, 77

Compute the 95% confidence interval for the average weight of students of the Study Centre of IGNOU.

Solution: Since population variance is unknown, therefore, $(1-\alpha)$ 100% confidence limits for the average weight of students of Study Centre are given by

$$\bar{X} \mp t_{(n-1), \alpha/2} \frac{S}{\sqrt{n}}$$

$$\text{where, } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

Calculation for $\bar{X}$ and S :

S. No.	Weight (X)	$(X - \bar{X})$	$(X - \bar{X})^2$
1	48	-15	225
2	50	-13	169
3	62	-1	1
4	75	12	144
5	80	17	289
6	60	-3	9
7	70	7	49
8	56	-7	49
9	52	-11	121
10	77	14	196
Sum	$\sum X = 630$		$\sum (X - \bar{X})^2 = 1252$

From the above calculation, we have

$$\bar{X} = \frac{1}{n} \sum X = \frac{1}{10} \times 630 = 63$$

$$S^2 = \frac{1}{n-1} \sum (X - \bar{X})^2 = \frac{1}{9} \times 1252 = 139.11$$

$$\Rightarrow S = \sqrt{139.11} = 11.79$$

For 95% confidence interval, we have $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$. Also from t-table, we have, $t_{(n-1), \alpha/2} = t_{(9), 0.025} = 2.262$.

Thus, the 95% confidence limits are

$$\begin{aligned} \bar{X} \mp t_{(n-1), 0.025} \frac{S}{\sqrt{n}} &= 63 \mp 2.262 \times \frac{11.79}{\sqrt{10}} \\ &= 63 \mp 8.43 = 54.57 \text{ and } 71.43 \end{aligned}$$

Hence, the required 95% confidence interval for the average weight of students of the Study Centre of IGNOU is given by

$$[54.57, 71.43]$$

Example 6: The mean life of 100 electric bulbs produced by a company is 2550 hours with a standard deviation 54 hours. Find 95% confidence limits for population mean life of the electric bulbs produced by the company.

Solution: Here, we are given that

$$n = 100, \bar{X} = 2550, S = 54$$

Since population variance is unknown and the sample size is large (> 30), therefore, we can use $(1 - \alpha)$ 100 % confidence limits for population mean which are given by

$$\bar{X} \mp z_{\alpha/2} \frac{S}{\sqrt{n}}$$

where $z_{\alpha/2}$ is the value of the variate Z having an area of $\alpha/2$ under the right tail of the probability curve of Z . For 95% confidence limits, we have $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$ and for $\alpha = 0.05$, we have $z_{\alpha/2} = z_{0.025} = 1.96$.

Thus, the 95% confidence limits for mean life of electric bulbs are

$$\begin{aligned} \bar{X} \mp 1.96 \frac{S}{\sqrt{n}} &= 2550 \mp 1.96 \frac{54}{\sqrt{100}} \\ &= 2550 \mp 1.96 \times 5.4 \\ &= 2550 \mp 10.58 = 2539.42 \text{ and } 2560.58 \end{aligned}$$

9.7.3 Confidence Interval for Difference of Two Population Means when Samples are Independent and Population Variances are Known

Let $X_1, X_2, \dots, X_{n_1}$ be a random sample of size n_1 taken from normal population-I $N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \dots, Y_{n_2}$ be another independent random sample of size n_2 taken from normal population-II $N(\mu_2, \sigma_2^2)$. Let $\bar{X}$ and $\bar{Y}$ be the sample means

of random samples $X_1, X_2, \dots, X_{n_1}$ and $Y_1, Y_2, \dots, Y_{n_2}$ respectively then we know that sample means $\bar{X}$ and $\bar{Y}$ are normally distributed, that is,

$$\bar{X} \sim N(\mu_1, \sigma_1^2/n_1) \text{ and } \bar{Y} \sim N(\mu_2, \sigma_2^2/n_2)$$

Also, by the property of normal distribution, we have

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

And the variate

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

is normally distributed with mean 0 and variance unity. Since the distribution of Z is independent of the parameters so it can be taken as a pivotal quantity.

The required $(1 - \alpha)$ 100% confidence interval for $(\mu_1 - \mu_2)$ is

$$\left[(\bar{X} - \bar{Y}) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{X} - \bar{Y}) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right] \quad \dots (8)$$

and corresponding limits are given by

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \dots (9)$$

Note 6: When $\bar{Y}$ is greater than $\bar{X}$ then we take $(\bar{Y} - \bar{X})$ in place of $(\bar{X} - \bar{Y})$.

The following example will make you user friendly with the way how the above concepts can be used in a practical situation:

Example 7: A company manufacturing two types of bulbs and product manager tested a random sample of 50 bulbs of type I and 60 bulbs of type II and found the following information:

	Mean life (hrs.)	Population Standard Deviation (hrs.)
Type I	1300	41
Type II	1280	46

Obtain the 95% confidence interval for the difference of average life of two types of bulbs assuming that the distributions of average lives of two types of bulbs follow the normal distribution.

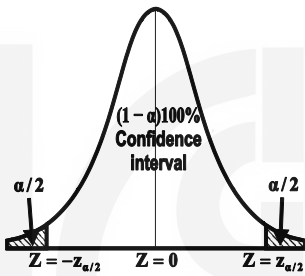
Solution: Here, we are given that

$$n_1 = 50, \quad \bar{X} = 1300, \quad \sigma_1 = 41$$

$$n_2 = 60, \quad \bar{Y} = 1280, \quad \sigma_2 = 46$$

Since population standard deviations, i.e. population variances of both the populations, are known so we use $(1 - \alpha)$ 100% confidence limits for the

Fig. 9.6



difference of population mean when population variances are known which are given by

$$(\bar{X} - \bar{Y}) \mp z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

For 95% confidence interval, $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$ and $\alpha/2 = 0.025$. Also $z_{\alpha/2} = 1.96$.

Thus, the 95% confidence limits for the difference of average lives of two types of bulbs are

$$\begin{aligned} (\bar{X} - \bar{Y}) \mp z_{0.025} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} &= (1300 - 1280) \mp 1.96 \sqrt{\frac{(41)^2}{50} + \frac{(46)^2}{60}} \\ &= 20 \mp 1.96 \sqrt{33.62 + 35.27} \\ &= 20 \mp 1.96 \times 8.3 \\ &= 20 \mp 16.27 \\ &= 3.73 \text{ and } 36.27 \end{aligned}$$

Hence, the required 95% confidence interval is

$$[3.73, 36.27]$$

9.7.4 Confidence Interval for Difference of Two Population Means when Population Variances are Unknown

Let $X_1, X_2, \dots, X_{n_1}$ be a random sample of size n_1 taken from normal population-I with mean μ_1 and unknown variance σ_1^2 and also let $Y_1, Y_2, \dots, Y_{n_2}$ be another independent random sample of size n_2 taken from a second normal population with mean μ_1 and unknown variance σ_2^2 . Then, keeping the scope of the course in mind, let us discuss the following two cases:

Case I: When $\sigma_1^2 = \sigma_2^2 = \sigma^2$

In this case, σ^2 is estimated by pooled sample variance S_p^2 where,

$$S_p^2 = \frac{1}{n_1 + n_2 - 2} [(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2]$$

$$\text{where, } S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2 \text{ and } S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

This can also be written as

$$S_p^2 = \frac{1}{n_1 + n_2 - 2} \left[\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 \right]$$

In this situation, the variate

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

follows t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom. Since the distribution of variate t is independent of the parameters, therefore, it can be taken as a pivotal quantity.

Hence, the required $(1-\alpha)$ 100% confidence interval is

$$\left[(\bar{X} - \bar{Y}) - t_{(n_1+n_2-2), \alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, (\bar{X} - \bar{Y}) + t_{(n_1+n_2-2), \alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right] \quad \dots (10)$$

and the corresponding limits are given by

$$(\bar{X} - \bar{Y}) \pm t_{(n_1+n_2-2), \alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \dots (11)$$

Note 7: When $\bar{Y}$ is greater than $\bar{X}$ then we take $(\bar{Y} - \bar{X})$ in place of $(\bar{X} - \bar{Y})$.

Case II: When $\sigma_1^2 \neq \sigma_2^2$

In this case, σ_1^2 & σ_2^2 are estimated by the values of their estimators S_1^2 & S_2^2 respectively, where

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2 \quad \& \quad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

In this situation, the pivotal quantity is difficult to obtain. But we know by the central limit theorem that for large sample sizes n_1 and n_2 (> 30) the variate

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0, 1)$$

follows normal distribution with mean 0 and variance unity, so $(1-\alpha)$ 100% confidence interval for difference of population means may be obtained by the procedure as same as we have followed in the case when σ_1^2 & σ_2^2 are known by taking S_1^2 & S_2^2 in place of σ_1^2 & σ_2^2 respectively. Then $(1-\alpha)$ 100% confidence interval for $(\mu_1 - \mu_2)$ is given by

$$\left[(\bar{X} - \bar{Y}) - z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, (\bar{X} - \bar{Y}) + z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right] \quad \dots (12)$$

and corresponding limits are given by

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad \dots (13)$$

Note 8: When $\bar{Y}$ is greater than $\bar{X}$ then we take $(\bar{Y} - \bar{X})$ in place of $(\bar{X} - \bar{Y})$.

The following examples will make you user friendly with the way how above concepts can be used in a practical situation:

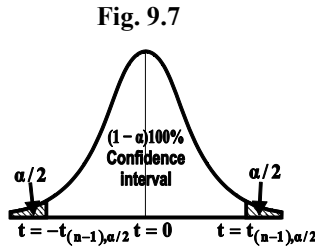


Fig. 9.7

Example 8: A random sample of 120 colonies was taken from city A and the average population per colony was found to be 540 with a standard deviation 18. Another sample of 150 colonies taken from another city B gave an average population 460 per colony with a standard deviation 15. Find 95% confidence interval for the difference of two population averages by assuming that both average populations follow the normal distributions.

Solution: Here, we are given that

$$n_1 = 120, \quad \bar{X} = 540, \quad S_1 = 18$$

$$n_2 = 150, \quad \bar{Y} = 460, \quad S_2 = 15$$

Since population variances are unknown, therefore, we use $(1-\alpha)$ 100% confidence limits for the difference of population mean when population variances are unknown which are given by

$$(\bar{X} - \bar{Y}) \mp z_{0.025} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

For 95% confidence interval, $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$ and $\alpha/2 = 0.025$. Also $z_{\alpha/2} = 1.96$.

So, 95% confidence limits for the difference of two population averages are given by

$$\begin{aligned} (540 - 460) \mp 1.96 \sqrt{\frac{(18)^2}{120} + \frac{(15)^2}{150}} \\ = 80 \mp 1.96 \sqrt{\frac{324}{120} + \frac{225}{150}} \\ = 80 \mp 1.96 \sqrt{2.70 + 1.50} \\ = 80 \mp 1.96 \times 2.05 \\ = 80 \mp 4.02 \\ = 75.98 \text{ and } 84.02 \end{aligned}$$

Hence, the required 95% confidence interval for $(\mu_1 - \mu_2)$ is

$$[75.98, 84.02]$$

9.8 CONFIDENCE INTERVAL FOR POPULATION PROPORTION

In last section, we have discussed the confidence interval for population mean. But in many real-world situations, in business and other areas, the data are collected in the form of counts or the collected data classified into two categories or groups according to an attribute or characteristic under study. Generally, such types of data are considered in terms of the proportion of elements/individuals/units/items that possess or not possess a given characteristic or attribute. For example, the proportion of female in the

population, proportion of diabetes patients in a hospital, proportion of Science books in a library, proportion of defective articles in a lot, etc.

In such situations, we deal population proportion instead of population mean and one may want to obtain the confidence interval for population proportion. For example, a sociologist may want to know the confidence interval for proportion of female in the population of a state. A doctor may want to know the confidence interval for proportion of diabetes patients in a hospital, a product manager may want to know the confidence interval for proportion of defective articles in a lot, etc.

Generally, population proportion is estimated by sample proportion.

Let $X_1, X_2, \dots, X_n$ be a random sample of size n taken from a population with population proportion P . Also, let X denotes the number of observations or elements that possess a certain attribute (successes) out of n observations of the sample then sample proportion p can be defined as

$$p = \frac{X}{n} \leq 1$$

The mean and variance of the sampling distribution of sample proportion are

$$E(p) = P \text{ and } \text{Var}(p) = \frac{PQ}{n}$$

where, $Q = 1 - P$.

But sample proportion is generally considered for a large sample so if the sample size is sufficiently large, such that $np > 5$ and $nq > 5$ then by central limit theorem, the sampling distribution of sample proportion p is approximately normally distributed with mean P and variance PQ/n .

Hence, $(1-\alpha)$ 100% confidence interval for population proportion is given by

$$\left[p - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}, p + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \right] \dots (14)$$

Therefore, corresponding confidence limits are

$$p \mp z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \dots (15)$$

This also defines the *Confidence Interval for the Bernoulli distribution*.

The following example will explain the application of the above discussion.

Example 9: A sample of 200 voters is chosen at random from all voters in a given city. 60% of them were in favour of a particular candidate. If a large number of voters cast their votes then find 99% and 95% confidence intervals for the proportion of voters in favour of a particular candidate.

Solution: Here, we are given

$$n = 200, p = 0.60$$

First, we check the condition of normality as

$$\therefore np = 200 \times 0.60 = 120 > 5 \text{ and } nq = 200 \times (1 - 0.60) = 200 \times 0.40 = 80 > 5$$

so $(1-\alpha)$ 100% confidence limits for the proportion are given by

$$p \mp z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

For 99% confidence interval, we have $1 - \alpha = 0.99 \Rightarrow \alpha = 0.01$. For $\alpha = 0.01$, we have $z_{0.005} = 2.58$ and for $\alpha = 0.05$, $z_{0.025} = 1.96$.

Therefore, 99% confidence limits of voters in favour of a particular candidate are

$$\begin{aligned} p \pm z_{0.005} \sqrt{\frac{p(1-p)}{n}} &= 0.60 \pm 2.58 \times \sqrt{\frac{0.60 \times 0.40}{200}} \\ &= 0.60 \pm 2.58 \times 0.03 \\ &= 0.60 \pm 0.08 = 0.52 \text{ and } 0.68 \end{aligned}$$

Hence, the required 99% confidence interval for the proportion of voters in favour of a particular candidate is given by

$$[0.52, 0.68]$$

Similarly, 95% confidence limits are given by

$$\begin{aligned} p \pm z_{0.025} \sqrt{\frac{p(1-p)}{n}} &= 0.60 \pm 1.96 \times 0.03 \\ &= 0.60 \pm 0.06 = 0.54 \text{ and } 0.66 \end{aligned}$$

Hence, 95% confidence interval for the proportion of voters in favour of a particular candidate is given by

$$[0.54, 0.66]$$

Check Your Progress 3

1. Certain refined oil is packed in tins holding 15 kg each. The filling machine maintains this but has a standard deviation 0.30 kg. A sample of 200 tins is taken from the production line. If the sample mean is 15.25 kg then find the 95% confidence interval for the average weight of oil tins.
.....
.....
2. Sample mean of weights (in kg) of 150 students of IGNOU is found to be 65 kg with standard deviation 12 kg. Find the 95% confidence limits in which the average weight of all students of IGNOU expected to lie.
.....
.....
3. A sample of the height of 2500 Nepalis has a mean of 68.50 inches and a standard deviation of 2.52 inches, while the sample of the height of 1600 Indians has a mean 70.25 inches and a standard deviation 2.58 inches. Find a 90% interval estimate for the difference of mean heights of both the countries by assuming that the height of the persons of both the countries is normally distributed.
.....
.....
.....
.....

4. A random sample of 400 apples was taken from a large consignment and 80 were found to be bad. Obtain the 99% confidence limits for the proportion of bad apples in the consignment.

.....

.....

.....

.....

9.9 DETERMINATION OF SAMPLE SIZE

To estimate some population parameter we have to draw a random sample. A natural question that may arise in your mind: “how large should my sample be?” And this is a very important question that is commonly asked. From a statistical point of view, the best answer to this question is “take as large a sample as you can afford. That is, if possible ‘sample’ the entire population that means study all units of the population under study because by taking all units of the population we will have all the information about the parameter and we will know the exact value of the population parameter which is better than any estimate of that parameter. Generally, this is impractical to take the entire population to be sampled due to economic constraints, time constraints and other limitations. So the answer, “take as large a sample as you can afford is best if we ignore all costs because, as you have studied in this unit, the larger the sample size, the smaller the standard error of the statistic, which means less uncertainty.

When the resources in terms of money and time are limited then the question is “how to find the minimum sample size which will satisfy some precision requirements”. In such cases, we should require first the answers to the following three questions about his / her requirement about the survey:

1. How close do you want your sample estimate to be to the unknown parameter? That means what should be the allowable difference between the sample estimate and the true value of the population parameter. This difference is known as sampling error or margin of error and represented by E .
2. The next question is, what do you want the confidence level to be so that the difference between the estimate and the parameter is less than or equal to E ? That is, 90%, 95%, 99%, etc.
3. The last question is what is the population variance or population proportion as may be the case?

When we have the answers of these three questions then we will get an answer of the minimum required sample size.

In this section, we will describe the procedure of determining of sample size for estimating population mean and population proportion.

Determination of Minimum Sample Size for Estimating Population Mean

For determining the minimum sample size for estimating the population mean we use a confidence interval. Since the confidence interval for the population mean depends upon the nature of the population and population variance (σ^2) is known or unknown, the following cases may arise:

Case I: Population is normal and population variance (σ^2) is known

If σ^2 is known and form of the population is normal then we know that $(1 - \alpha)100\%$ confidence interval for population mean is given by

$$P\left[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

$$\text{Also } P\left[-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha \quad \dots (16)$$

Since the normal distribution is symmetric, we can concentrate on the right-hand equality, that is:

$$\bar{X} - \mu \leq z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \dots (17)$$

This inequality implies that the largest value that the difference $\bar{X} - \mu$ can assume is $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$.

Also the difference between the estimator (sample mean $\bar{X}$) and the population parameter (population mean μ) is called the sampling error, so

$$E = \bar{X} - \mu = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \dots (18)$$

Solving this equation for n, we have

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2} \quad \dots (19)$$

When population size N is finite and sampling is to be done without replacement then finite population correction $\sqrt{\frac{N-n}{N-1}}$ is required so equation (20) becomes

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

which gives

$$n = \frac{N z_{\alpha/2}^2 \sigma^2}{E^2 (N-1) + z_{\alpha/2}^2 \sigma^2} \quad \dots (20)$$

If the finite population correction is ignored then equation (20) is reduced to equation (19).

Case II: Population is non-normal and population variance (σ^2) is known

If the population is not assumed to be normal and the population variance σ^2 is known then by central limit theorem we know that sampling distribution of mean is approximately normally distributed as sample size increases. So the above method can be used determining minimum sample size. Once the required sample size is obtained, we can check to see it that sample size greater than 30 and if it does, we may be confident that our method of solution was appropriate.

Case III: Population is normal or non-normal and population variance (σ^2) is unknown

In this case, we use value of sample variance (S^2) to estimate the population variance but S^2 is also calculated from a sample and we have not yet taken a sample. So in this case, determination of sample size is not directly obtained.

The most frequent methods for estimating σ^2 are as follows:

1. A pilot or preliminary sample may be drawn from the population under study and the variance computed from this sample may be used as an estimate of σ^2 .
2. The variance of previous or similar studies may be used to estimate σ^2 .

3. If the population is assumed to be normal then we may use the fact that the range is approximately equal to six times of standard deviation i.e. $\sigma = R/6$. This approximate require only knowledge of the largest and smallest value of the variable under study because range may be defined as

$$R = \text{largest value} - \text{smallest value}$$

Determination of Minimum Sample Size for Estimating Population Proportion

The method of determination of the minimum sample size for estimating population proportion is similar to that described in estimating the population mean. So the formula for minimum sample size is given by

$$n = \frac{z_{\alpha/2}^2 P(1-P)}{E^2} \quad \dots (21)$$

where P is the population proportion and $E = p - P$ is the sampling error or margin of error.

When the population is finite of size N and sampling is to be done without replacement then finite population correction $\sqrt{\frac{N-n}{N-1}}$ is required so

$$n = \frac{N z_{\alpha/2}^2 P(1-P)}{E^2 (N-1) + z_{\alpha/2}^2 P(1-P)} \quad \dots (22)$$

Here, the sample size is dependent upon the population proportion and it is generally unknown so the most frequent methods for estimating P are as follows:

1. A pilot or preliminary sample may be drawn from the population under study and the sample proportion computed from this sample may be used as an estimate of P.
2. The proportion of previous or similar studies may be used to estimate P.
3. Absence of any information, we may use $P = 0.5$.

Now, it is time to do some examples based on the determination of sample size.

Example 10: A hospital administrator wishes to estimate the mean weight of babies born in her hospital. How large a sample of birth records should be taken if she wants a 99 per cent confidence that the estimate is within the range of 0.4 pounds? Assume that a reasonable estimate of σ is 0.5 pounds.

Solution: Here, we are given that

$$E = \text{margin of error} = 0.4, \text{ confidence level} = 0.99 \text{ and } \sigma = 0.5$$

Also for 99% confidence, $1 - \alpha = 0.99 \Rightarrow \alpha = 0.01$ and $\alpha/2 = 0.005$, we have

$$z_{\alpha/2} = z_{0.005} = 2.58.$$

The hospital administrator wishes to obtain the minimum sample size for estimating the mean weight of babies born in her hospital so we have

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2} = \frac{(2.58)^2 (0.5)^2}{(0.4)^2} = 10.40 \sim 11$$

Hence, the hospital administrator should obtain a random sample of at least 11 babies.

Example 11: The manufacturers of a car want to estimate the proportion of people who are interested in a certain model. The company wants to know the population proportion, P , to within 0.05 with 95% confidence. Current company's records indicate that the proportion P is around 0.20. What is the minimum required sample size for this survey?

Solution: Here, we are given that

$$E = \text{margin of error} = 0.05, \text{ confidence level} = 0.95 \text{ and } P = 0.20$$

Also for 95% confidence, $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$ and $\alpha/2 = 0.025$, we have:

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The manufacturers of a car interested to obtain the minimum sample size for estimating the population proportion so the required formula is given below

$$n = \frac{z_{\alpha/2}^2 P(1-P)}{E^2} = \frac{(1.96)^2 (0.20)(0.80)}{(0.05)^2} = 245.86 \sim 246$$

Hence, the company should require a random sample of at least 246 people.

Check Your Progress 4

- 1) A survey is planned to determine the average annual family medical expenses of employees of a large company. The management of the company wishes to be 95% confident that the sample average is correct to within ± 100 Rs of the true average family expenses. A pilot study indicates that the standard deviation can be estimated as 400 Rs. How large a sample size is necessary?

.....

- 2) The manager of a bank in a small city would like to determine the proportion of the depositors per week. The manager wants to be 95% confident of being correct to within ± 0.10 of the true proportion of depositors per week. A guess is that the parentage of such depositors is about 8%, what sample size is needed?

.....

We now end this unit by giving a summary of what we have covered in it.

9.10 SUMMARY

In this unit, we have covered the following points:

1. The parameter space and the basic characteristics of an estimator.
2. Unbiasedness, consistency, efficiency and sufficiency of an estimator
3. The point estimation.
4. The method of maximum likelihood and its properties.
5. The interval estimation.
6. The method of obtaining confidence interval for population mean of a normal population when variance is known and unknown

7. The method of construction of confidence interval for difference of means of two normal populations when variances are known and unknown.
8. The concept of the shortest interval.
9. Determination of sample size.

9.11 SOLUTIONS / ANSWERS

Check Your Progress 1

- 1) Refer to Section 9.3.
- 2)
 - (i) Here, investigator estimates the average income Rs. 1.5 lakh i.e. he/she estimate average income with a single value, therefore, the investigator used the point estimation technique.
 - (ii) The product manager estimates the average life of electric bulbs with the help of two values 800 hours and 1000 hours, therefore, the product manager used the interval estimation technique.
 - (iii) A pathologist estimates the mean time required completing a certain analysis with the help of two values 30 minutes and 45 minutes, therefore, he/she used the interval estimation technique.

- 3) Refer to Section 9.4.

Check Your Progress 2

- 1) Let $X_1, X_2, \dots, X_n$ be a random sample of size n taken from $B(n, p)$ whose pmf is given by

$$P[X = x] = {}^nC_x p^x q^{n-x}; \quad x = 0, 1, \dots, n \text{ \& } q = 1 - p$$

ML estimator for parameter p is $\bar{X}/n$.

- 2) Refer to Section 9.5.
- 3) We know that the length of confidence interval $[T_1, T_2]$ is given by

$$L = T_2 - T_1$$

Therefore, in our case we have

- (i) $L = 3.0 - (-1.65) = 4.65$
 - (ii) $L = 2.70 - (-1.68) = 4.38$
 - (iii) $L = 2.54 - (-1.70) = 4.24$
 - (iv) $L = 1.96 - (-1.96) = 3.92$
- 4) We know that if we have confidence interval for parameter θ is

$$P[T_1 \leq \theta \leq T_2] = 1 - \alpha$$

then

$$\text{Lower confidence limit (LCL)} = T_1$$

$$\text{Upper confidence limit (UCL)} = T_2$$

$$\text{Confidence coefficient (CC)} = 1 - \alpha$$

Therefore, in our cases, we have

- (i) LCL = 0, UCL = 1.5 & CC = 0.90
- (ii) LCL = -1, UCL = 2 & CC = 0.95
- (iii) LCL = -2, UCL = 2 & CC = 0.98
- (iv) LCL = -2.5, UCL = 2.5 & CC = 0.99

Check Your Progress 3

- 1) Here, we are given that

$$n = 200, \sigma = 0.30, \bar{X} = 15.25$$

Since population standard deviation or variance is known so we use. (1- α) 100% confidence limits for population mean when population variance is known which are given by

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Where $z_{\alpha/2}$ is the value of the variate Z having an area of $\alpha/2$ under the right tail of the probability curve of Z. For 95% confidence interval, we have $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$ and for $\alpha = 0.05$, we have, $z_{\alpha/2} = z_{0.025} = 1.96$.

Thus, the 95% confidence limits for the average weight of oil tins are

$$\begin{aligned} \bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}} &= 15.25 \pm 1.96 \times \frac{0.30}{\sqrt{200}} \\ &= 15.25 \pm 0.04 = 15.21 \text{ and } 15.29 \end{aligned}$$

- 2) Here, we are given that

$$n = 150, \bar{X} = 65, S = 12$$

Since population variance is unknown and the sample size is large sample,

so we use (1- α) 100% confidence limits for population mean when population variance is unknown which are given by

$$\bar{X} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}$$

For 95% confidence interval, we have $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$ and for $\alpha = 0.05$, we have, $z_{\alpha/2} = z_{0.025} = 1.96$.

Thus, the 95% confidence limits in which the average weight of all students of a Study centre of IGNOU expected to lie are given by

$$\begin{aligned} \bar{X} \pm 1.96 \frac{S}{\sqrt{n}} &= 65 \pm 1.96 \times \frac{12}{\sqrt{150}} \\ &= 65 \pm 1.92 = 63.08 \text{ and } 66.92 \end{aligned}$$

Hence, required confidence limits are

$$63.08 \text{ and } 66.92.$$

$$n = 400, X = 80$$

$$p = \frac{X}{n} = \frac{80}{400} = \frac{1}{5} = 0.20$$

$\therefore np = 400 \times 0.20 = 80 > 5$ and $nq = 400 \times (1 - 0.20) = 400 \times 0.80 = 320 > 5$ so $(1-\alpha)100\%$ confidence limits for the proportion are given by

$$p \mp z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

For $\alpha = 0.01$, we have, $z_{\alpha/2} = z_{0.005} = 2.58$.

Therefore, 99% confidence limits are

$$\begin{aligned} & 0.20 \mp 2.58 \sqrt{\frac{0.20 \times 0.80}{400}} \\ &= 0.2 \mp 2.58 \times 0.02 \\ &= 0.2 \mp 0.05 = 0.15 \text{ and } 0.25 \end{aligned}$$

For 99% confidence interval
 $1 - \alpha = 0.99 \Rightarrow \alpha = 0.01$
and $\alpha/2 = 0.005$.

3) Here, we are given that

$$n_1 = 2500, \quad \bar{X} = 68.50, \quad S_1 = 2.52$$

$$n_2 = 1600, \quad \bar{Y} = 70.25, \quad S_2 = 2.58$$

Since population variances are unknown, therefore, we use $(1-\alpha) 100\%$ confidence limits for the difference of two population means when population variances are unknown which are given by

$$(\bar{X} - \bar{Y}) \mp z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

For 90% confidence interval, $1 - \alpha = 0.90 \Rightarrow \alpha = 0.10$ and $\alpha/2 = 0.05$. Also $z_{\alpha/2} = 1.645$.

Since in this case, $\bar{Y} > \bar{X}$ therefore, we take $(\bar{Y} - \bar{X})$ in place of $(\bar{X} - \bar{Y})$.

Thus, 90% confidence limits are given by

$$\begin{aligned} & (\bar{Y} - \bar{X}) \mp z_{0.05} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \\ &= (70.25 - 68.50) \mp 1.645 \times \sqrt{\frac{(2.52)^2}{2500} + \frac{(2.58)^2}{1600}} \\ &= 1.75 \mp 1.645 \times \sqrt{\frac{6.35}{2500} + \frac{6.66}{1600}} \\ &= 1.75 \mp 1.645 \times \sqrt{0.0025 + 0.0042} \\ &= 1.75 \mp 1.645 \times 0.08 \\ &= 1.62 \text{ and } 1.88 \end{aligned}$$

Hence, the required 90% confidence interval is

$$[1.62, 1.88]$$

4) We have

$$n = 400, X = 80$$

$$p = \frac{X}{n} = \frac{80}{400} = \frac{1}{5} = 0.20$$

$\therefore np = 400 \times 0.20 = 80 > 5$ and $nq = 400 \times (1 - 0.20) = 400 \times 0.80 = 320 > 5$ so $(1-\alpha)100\%$ confidence limits for the proportion are given by

$$p \mp z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

For $\alpha = 0.01$, we have, $z_{\alpha/2} = z_{0.005} = 2.58$.

Therefore, 99% confidence limits are

For 99% confidence interval

$1 - \alpha = 0.99 \Rightarrow \alpha = 0.01$
and $\alpha/2 = 0.005$.

$$\begin{aligned} &0.20 \mp 2.58 \sqrt{\frac{0.20 \times 0.80}{400}} \\ &= 0.2 \mp 2.58 \times 0.02 \\ &= 0.2 \mp 0.05 = 0.15 \text{ and } 0.25 \end{aligned}$$

Check Your Progress 4

1) Here, we are given that

$E = \text{margin of error} = 100$, confidence level = 0.95 and $\sigma = 400$

Also for 95% confidence, $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$ and $\alpha/2 = 0.025$, we have $z_{\alpha/2} = z_{0.025} = 1.96$.

The management of the company wishes to obtain the minimum sample size for estimating the average annual family medical expenses of employees of the company so the required formula is given below

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2} = \frac{(1.96)^2 (400)^2}{(100)^2} = 61.47 \sim 62$$

Hence, the management of the company should require a random sample of at least 62 employees.

2) Here, we are given that

$E = \text{margin of error} = 0.1$, confidence level = 0.95 and $P = 0.08$

Also for 95% confidence, $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$ and $\alpha/2 = 0.025$, we have $z_{\alpha/2} = z_{0.025} = 1.96$.

The manager wants to obtain the minimum sample size for determining the proportion so the required formula is given below

$$n = \frac{z_{\alpha/2}^2 P(1-P)}{E^2} = \frac{(1.96)^2 (0.08)(0.92)}{(0.1)^2} = 28.27 \sim 29$$

Hence, the manager should require a random sample of at least 29 depositors.

9.12 REFERENCES

IGNOU Course Material

1. Post-Graduate Diploma in Applied Statistics (PGDAST), MST 004: Statistical Inference, Block 2: Estimation UNIT 5: Introduction to Estimation (for Sections 9.2 and 9.3 and related portions in Sections 9.0 , 9.1, 9.10, 9.11 and Check Your Progress questions)
2. Post-Graduate Diploma in Applied Statistics (PGDAST), MST 004: Statistical Inference, Block 2: Estimation UNIT 6: Point Estimation (for Sections 9.4 and 9.5 and related portions in Sections 9.0 , 9.1, 9.10, 9.11 and Check Your Progress questions)
3. Post-Graduate Diploma in Applied Statistics (PGDAST), MST 004: Statistical Inference, Block 2: Estimation UNIT 7: Interval Estimation for One Population (for Sections 9.6 to 9.8 and related portions in Sections 9.0 , 9.1, 9.10, 9.11 and Check Your Progress questions)
4. Post-Graduate Diploma in Applied Statistics (PGDAST), MST 004: Statistical Inference, Block 2: Estimation UNIT 8: Interval Estimation for Two Population (for Section 9.7 and related portions in Sections 9.0 , 9.1, 9.10, 9.11 and Check Your Progress questions)

Reference Book

5. “Introduction to Probability and Statistics for Engineers and Scientists”, 3rd Edition, Sheldon M. Ross, Elsevier Academic Press, 2004