

**Centre for Distance & Online Education  
(CDOE)**

**Bachelor of Commerce**

**BCOM 402**

**BUSINESS STATISTICS-II**



**Guru Jambheshwar University of Science &  
Technology, HISAR-125001**

**CONTENTS**

<b>Lesson No.</b>	<b>Name of Topic</b>	<b>Page No.</b>
1	Probability Theory	3
2	Probability Distributions-I	36
3	Probability Distributions-II	61
4	Sampling and Sampling Methods	84
5	Sampling Distributions	116
6	Testing of Hypotheses	169
7	Non-Parametric Tests	213
8	Index Number	237
9	Analysis of Time Series	280



Subject: Business Statistics-II	
Course Code: BCOM 402	Author: Anil Kumar
Lesson: 01	Vetter: Dr. Karam Pal
<b>PROBABILITY THEORY</b>	

## **STRUCTURE**

- 1.0 Learning Objectives
- 1.1 Introduction
  - 1.1.1 Some Basic Concepts
  - 1.1.2 Approaches to Probability Theory
  - 1.1.3 Probability Rules
  - 1.1.4 Bayes' Theorem
- 1.2 Some Counting Concepts
- 1.3 Check your Progress
- 1.4 Summary
- 1.5 Keywords
- 1.6 Self-Assessment Test
- 1.7 Answers to check your progress
- 1.8 References/Suggested Readings

## **1.0 LEARNING OBJECTIVES**

After going through this lesson, students will be able to:

- Understand the concept of probability
- Appreciate the relevance of probability theory in decision-making under conditions of uncertainty
- Understand and use the different approaches to probability as well as different probability rules for calculating probabilities in different situations.

## **1.1 INTRODUCTION**

Life is full of uncertainties. 'Probably', 'likely', 'possibly', 'chance' *etc.* is some of the most commonly used terms in our day-to-day conversation. All these terms more or less convey the same sense - "*the*



*situation under consideration is uncertain and commenting on the future with certainty is impossible*". Decision-making in such areas is facilitated through formal and precise expressions for the uncertainties involved. For example, product demand is uncertain but study of demand spelled out in a form amenable for analysis may go a long to help analyze, and facilitate decisions on sales planning and inventory management. Intuitively, we see that if there is a high chance of a high demand in the coming year, we may decide to stock more. We may also take some decisions regarding the price increase, reducing sales expenses *etc.* to manage the demand. However, in order to make such decisions, we need to quantify the chances of different quantities of demand in the coming year. Probability theory provides us with the ways and means to quantify the uncertainties involved in such situations.

A probability is a quantitative measure of uncertainty - a number that conveys the strength of our belief in the occurrence of an uncertain event.

Since uncertainty is an integral part of human life, people have always been interested - consciously or unconsciously - in evaluating probabilities.

Having its origin associated with gamblers, the theory of probability today is an indispensable tool in the analysis of situations involving uncertainty. It forms the basis for inferential statistics as well as for other fields that require quantitative assessments of chance occurrences, such as quality control, management decision analysis, and almost all areas in physics, biology, engineering and economics or social life.

### 1.1.1 Some Basic Concepts

Probability, in common parlance, refers to the chance of occurrence of an event or happening. In order that we are able to compute it, a proper understanding of certain basic concepts in probability theory is required. These concepts are an *experiment*, a *sample space*, and an *event*.

## EXPERIMENT

An experiment is a process that leads to one of several possible outcomes. An outcome of an experiment is some observation or measurement.

The term experiment is used in probability theory in a much broader sense than in physics or chemistry. Any action, whether it is the drawing a card out of a deck of 52 cards, or reading the temperature, or measurement of a product's dimension to ascertain quality, or the launching of a new product in the market, constitute an experiment in the probability theory terminology.



The experiments in probability theory have three things in common:

- there are two or more outcomes of each experiment
- it is possible to specify the outcomes in advance
- there is uncertainty about the outcomes

For example, the product we are measuring may turn out to be undersize or right size or oversize, and we are not certain which way it will be when we measure it. Similarly, launching a new product involves uncertain outcome of meeting with a success or failure in the market.

A single outcome of an experiment is called a **basic outcome** or an **elementary event**. Any particular card drawn from a deck is a basic outcome.

## SAMPLE SPACE

The sample space is the universal set  $S$  pertinent to a given experiment. It is the set of all possible outcomes of an experiment.

So each outcome is visualized as a sample point in the sample space. The sample spaces for the above experiments are:

Experiment	Sample Space
Drawing a Card	{all 52 cards in the deck}
Reading the Temperature	{all numbers in the range of temperatures}
Measurement of a Product's Dimension	{undersize, outsize, right size}
Launching of a New Product	{success, failure}

## EVENT

An event, in probability theory, constitutes one or more possible outcomes of an experiment. It is a subset of a sample space. It is a set of basic outcomes. We say that the event occurs if the experiment gives rise to a basic outcome belonging to the event.

For the experiment of drawing a card, we may obtain different events A, B, and C like:

A: The event that card drawn is king of club

B: The event that card drawn is red

C: The event that card drawn is ace



In the first case, out of the 52 sample points that constitute the sample space, only one sample point or outcome defines the event, whereas the number of outcomes used in the second and third case is 13 and 4 respectively.

### 1.1.2 Approaches to Probability Theory

Three different approaches to the definition and interpretation of probability have evolved, mainly to cater to the three different types of situations under which probability measures are normally required.

We will study these approaches with the help of examples of distinct types of experiments.

Consider the following situations marked by three distinct types of experiments. The events that we are interested in, within these experiments, are also given.

#### Situation I:

Experiment: Drawing a Card Out of a Deck of 52 Cards

Event A: On any draw, a king is there

#### Situation II

Experiment: Administering a Taste Test for a New Soup

Event B: A consumer likes the taste

#### Situation III

Experiment: Commissioning a Solar Power Plant

Event C: The plant turns out to be a successful venture

### Situation I : THE CLASSICAL APPROACH

The first situation is characterized by the fact that for a given experiment we have a sample space with equally likely basic outcomes. When a card is drawn out of a well-shuffled deck, every one of the cards (the basic outcomes) is as likely to occur as any other. This type of situations, marked by the presence of "*equally likely*" outcomes, gave rise to the **Classical Approach** to the probability theory. In the Classical Approach, probability of an event is defined as the *relative size* of the event with respect to the size of the sample space. Since there are 4 kings and there are 52 cards, the size of A is 4 and the size of the sample space is 52. Therefore, the probability of A is equal to  $4/52$ .

The rule we use in computing probabilities, assuming equal likelihood of all basic outcomes, is as follows:

Probability of the event A:



$$P(A) = \frac{n(A)}{N(S)} \quad \dots\dots\dots(6-1)$$

where  $n(A)$  = the number of outcomes favorable to the event A  
 $n(S)$  = total number of outcomes

### **Situation II : THE RELATIVE FREQUENCY APPROACH**

If we try to apply the classical definition of probability in the second experiment, we find that we cannot say that consumers will equally like the taste of the soup. Moreover, we do not know as to how many persons have been tested. This implies that we should have the past data on people who were administered the soup and the number that liked the taste. In the absence of past data, we have to undertake an experiment, where we administer the taste test on a group of people to check its effect.

The **Relative Frequency Approach** is used to compute probability in such cases. As per this approach, the probability of occurrence of an event is given by the observed relative frequency of an event in a very large number of trials. In other words, the probability of occurrence of an event is the ratio of the number of times the event occurs to the total number of trials. The probability of the event B:

$$P(B) = \frac{n}{N} \quad \dots\dots\dots(6-2)$$

Where  $n$  = the number of times the event occurs  
 $N$  = total number of trials

It is appreciated in this approach that, in order to take such a measure, we should have the soup tested for a large number of people. In other words, the total number of trials in the experiment should be very large.

### **Situation III : THE SUBJECTIVE APPROACH**

The third situation seems apparently similar to the second one. We may be tempted here to apply the Relative Frequency Approach. We may calculate the probability of the event that the venture is a success as the ratio of number of successful ventures to the total number of such ventures undertaken *i.e.* the relative frequency of successes will be a measure of the probability.

However, the calculation here presupposes that either

- (a) it is possible to do an experiment with such ventures, or
- (b) that past data on such ventures will be available



In practice, a solar power plant being a relatively new development involving the latest technology, past experiences are not available. Experimentation is also ruled out because of high cost and time involved, unlike the taste testing situation. In such cases, the only way out is the **Subjective Approach** to probability. In this approach, we try to assess the probability from our own experiences. We may bring in any information to assess this. In the situation cited, we may, perhaps, look into the performance of the commissioning authority in other new and related technologies.

Therefore the Subjective Approach involves personal judgment, information, intuition, and other subjective evaluation criteria. A physician assessing the probability of a patient's recovery and an expert assessing the probability of success of a merger offer are both making a personal judgment based upon what they know and feel about the situation. The area of subjective probability - which is relatively new, having been first developed in the 1930s - is somewhat controversial. One person's subjective probability may very well be different from another person's subjective probability of the same event. We may note here that since the assessment is a purely subjective one, it will vary from person to person and, therefore, subjective probability is also called **Personal Probability**.

### **Three Approaches – A Comparative View**

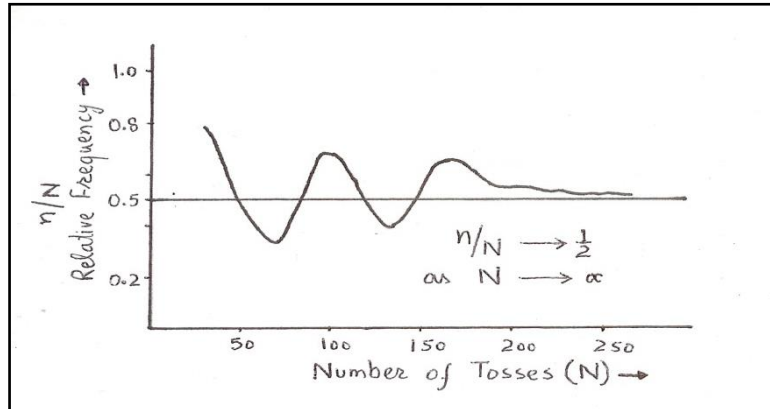
As already noted, the different approaches have evolved to cater to different kinds of situations. So these approaches are not contradictory to one another. In fact, these complement each other in the sense that where one fails, the other becomes applicable. These are identical inasmuch as probability is defined as a ratio or a weight assigned to the occurrence of an event. However, in contrast to the Subjective measure of the third approach, the first two approaches - Classical and Relative Frequency - provide an objective measure of probability in the sense that no personal judgment is involved.

We can bring out the commonality between the Classical Approach and the Relative Frequency Approach with the help of an example. Let us assume that we are interested in finding out the chances of getting a head in the toss of a coin. By now, you would have come up with the answer by the Classical Approach, using the argument, that there are two outcomes, heads and tails, which are equally likely. Hence, given that a head can occur only once, the probability is  $\frac{1}{2}$  : Consider the following alternative line of argument, where the probability can be estimated using the Relative Frequency





Approach. If we toss the coin for a sufficiently large number of times and note down the number of times the head occurs, the proportion of times that a head occurs will give us the required probability.



**Figure 1-1**  $P(H) = n/N \rightarrow 1/2$  as  $N \rightarrow \infty$

Thus, given our definition of the approaches, we find both the arguments to be valid. This brings out, in a way, the commonality between the Relative Frequency and the Classical Approach. The difference, however, is that the probability computed by using the Relative Frequency Approach will be tending to be  $1/2$  with a large number of trials; moreover an experiment is necessary in this case. In comparison, in the Classical Approach, we know apriori that the chances are  $1/2$ , based on our assumption of "equally likely" outcomes.

### Example 1-1

A fair coin is tossed twice. Find the probabilities of the following events:

- (a) A, getting two heads
- (b) B, getting one head and one tail
- (c) C, getting at least one head or one tail
- (d) D, getting four heads

**Solution:** Being a Two-Trial Coin Tossing Experiment, it gives rise to the following  $O^n = 2^n = 4$ , possible equally likely outcomes:

HH    HT    TH    TT

Thus, for the sample space  $N(S) = 4$

We can use the Classical Approach to find out the required probabilities.



- (a) For the event A, the number of favourable cases are:

$$n(A) = 1 \quad \{ HH \}$$

So the required probability

$$P(A) = \frac{n(A)}{N(S)} = \frac{1}{4}$$

- (b) For the event B, the number of favourable cases are:

$$n(B) = 2 \quad \{ HT, TH \}$$

So the required probability

$$P(B) = \frac{n(B)}{N(S)} = \frac{2}{4} = \frac{1}{2}$$

- (c) For the event C, the number of favourable cases are:

$$n(C) = 4 \quad \{ HH, HT, TH, TT \}$$

So the required probability

$$P(C) = \frac{n(A)}{N(S)} = \frac{4}{4} = 1$$

- (d) For the event D, the number of favourable cases are:

$$n(D) = 0$$

So the required probability

$$P(D) = \frac{n(D)}{N(S)} = \frac{0}{4} = 0$$

It may be noted that the occurrence of C is certainty, whereas D is an impossible event.

### **Example 1-2**

A newspaper boy wants to find out the chances that on any day he will be able to sell more than 90 copies of *The Times of India*. From his dairy where he recorded the daily sales of the last year, he finds out that out of 365 days, on 75 days he had sold 80 copies, on 144 days he had sold 85 copies, on 62 days he had sold 95 copies and on 84 days he had sold 100 copies of *The Times of India*. Find out the required probability for the newspaper boy.

**Solution:** Taking the Relative Frequency Approach, we find:



Sales(Event)	No. of Days (Frequency)	Relative Frequency
80	75	75/365
85	144	144/365
95	62	62/365
100	84	84/365

Thus, the number of days when his sales were more than 90 = (62 + 84) days = 146 days

So the required probability

$$P(\text{Sales} > 90) = \frac{n}{N} = \frac{146}{365} = 0.4$$

### Probability Axioms

All the three approaches to probability theory share the same basic axioms. These axioms are fundamental to probability theory and provide us with unified approach to probability.

The axioms are:

- (a) The probability of an event A, written as  $P(A)$ , must be a number between zero and one, both values inclusive. Thus

$$0 \leq P(A) \leq 1 \quad \dots\dots\dots(1-3)$$

- (b) The probability of occurrence of one or the other of all possible events is equal to one. As  $S$  denotes the sample space or the set of all possible events, we write

$$P(S) = 1. \quad \dots\dots\dots(1-4)$$

Thus in tossing a coin once;  $P(\text{a head or a tail}) = 1$ .

- (c) If two events are such that occurrence of one implies that the other cannot occur, then the probability that either one or the other will occur is equal to the sum of their individual probabilities. Thus, in a coin-tossing situation, the occurrence of a head rules out the possibility of occurrence of tail. These events are called **mutually exclusive events**. In such cases then, if A and B are the two events respectively, then

$$P(A \text{ or } B) = P(A) + P(B)$$

$$\text{i.e. } P(\text{Head or Tail}) = P(\text{Head}) + P(\text{Tail})$$



It follows from the last two axioms that if two mutually exclusive events form the sample space of the experiment, then

$$P(A \text{ or } B) = P(A) + P(B) = 1; \text{ thus } P(\text{Head}) + P(\text{Tail}) = 1$$

If two or more events together define the total sample space, the events are said to be collectively exhaustive.

Given the above axioms, we may now define probability as a function, which assigns probability value  $P$  to each sample point of an experiment abiding by the above axioms. Thus, the axioms themselves define probability.

### ***Interpretation of a Probability***

From our discussion so far, we can give a general definition of probability:

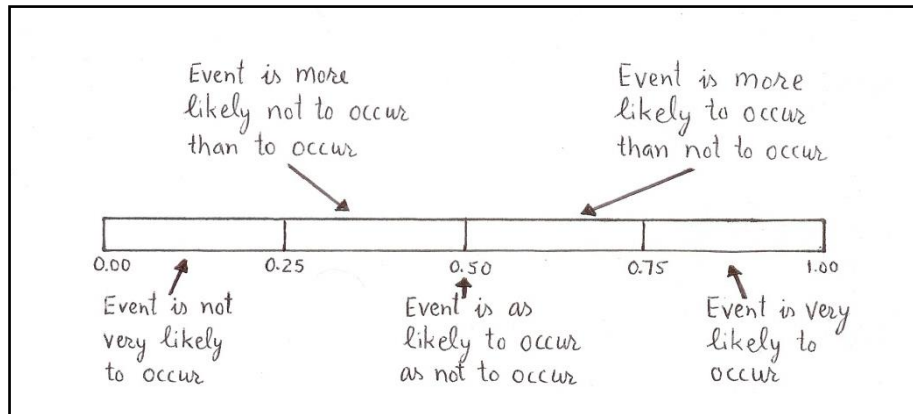
***Probability is a measure of uncertainty. The probability of event A is a quantitative measure of the likelihood of the event's occurring.***

We have also seen that 0 and 1, both values inclusive, sets the range of values that the probability measure may take. In other words  $0 \leq P(A) \leq 1$

When an event cannot occur (impossible event), its probability is zero. The probability of the empty set is zero:  $P(\Phi) = 0$ . In a deck where half the cards are red and half are black, the probability of drawing a green card is zero because the set corresponding to that event is the empty set: there are no green cards.

Events that are certain to occur have probability 1.00. The probability of the entire sample space  $S$  is equal to 1.00:  $P(S) = 1.00$ . If we draw a card out of a deck, 1 of the 52 cards in the deck will certainly be drawn, and so the probability of the sample space, the set of all 52 cards, is equal to 1.00.

Within the range of values 0 to 1, the greater the probability, the more confidence we have in the occurrence of the event in question. A probability of 0.95 implies a very high confidence in the occurrence of the event. A probability of 0.80 implies a high confidence. When the probability is 0.5, the event is as likely to occur as it is not to occur. When the probability is 0.2, the event is not very likely to occur. When we assign a probability of 0.05, we believe the event is unlikely to occur, and so on. Figure 1-2 is an informal aid in interpreting probability.



**Figure 1-2 Interpretation of a Probability**

Note that probability is a measure that goes from 0 to 1. In everyday conversation we often describe probability in less formal terms. For example, people sometimes talk about **odds**. If the odds are 1 to 1, the probability is  $\frac{1}{1+1}$  i.e.  $\frac{1}{2}$ ; if the odds are 1 to 2, the probability is  $\frac{1}{1+2}$  i.e.  $\frac{1}{3}$ ; and so on. Also, people sometimes say, "The probability is 80 percent." Mathematically, this probability is 0.80.

### 1.1.3 Probability Rules

We have seen how to compute probabilities in certain situations. The nature of the events was relatively simple, so that direct application of the definition of probability could be used for computation. Quite often, we are interested in the probability of occurrence of more complex events. Consider for example, that you want to find the probability that a king or a club will occur in a draw from a deck of 52 cards. Similarly, on examining couples with two children, if one child is known as a boy, you may be interested in the probability of the event of both the children being boys. These two situations, we find, are not as simple as those discussed in the earlier section. As a sequel to the theoretical development in the field of probability, certain results are available which help us in computing probabilities in such situations. Now we will explore these results through examples.

### THE UNION RULE

A very important rule in probability theory, the **Rule of Unions** (also called **Addition Theorem**) allows us to write the probability of the union of two events in terms of the probabilities of the two events and the probability of their intersection.

Consider two events A and B defined over the sample space S, as shown in Figure 1-3.

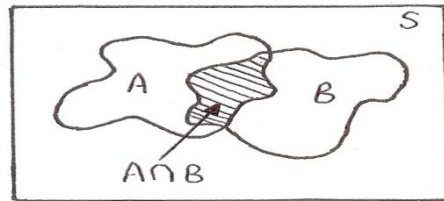


Figure 1-3 Two Overlapping Events A and B

We may define

$$\begin{aligned}
 P(A \cup B) &= \frac{n(A \cup B)}{N(S)} \\
 &= \frac{n(A) + n(B) - n(A \cap B)}{N(S)} \\
 &= \frac{n(A)}{N(S)} + \frac{n(B)}{N(S)} - \frac{n(A \cap B)}{N(S)} \\
 &= P(A) + P(B) - P(A \cap B)
 \end{aligned}$$

Thus, the rule of unions is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad \dots\dots\dots(1-5)$$

The probability of the intersection of two events  $P(A \cap B)$  is called their **joint probability**. The meaning of this rule is very simple and intuitive: When we add the probabilities of A and B, we are measuring, or counting, the probability of their intersection *twice*—*once* when measuring the relative size of A within the sample space and *once* when doing this with B. Since the relative size, or probability, of the intersection of the two sets is counted twice, we subtract it once so that we are left with the true probability of the union of the two events.

The rule of unions is especially useful when we do not have the sample space for the union of events but do have the separate probabilities.

### Example 1-3

A card is drawn from a well-shuffled pack of playing cards. Find the probability that the card drawn is either a club or a king.

Solution: Let A be the event that a club is drawn and B the event that a king is drawn. Then,



$$\begin{aligned}P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\&= 13/52 + 4/52 - 1/52 \\&= 16/52 \\&= 4/13\end{aligned}$$

**Example 1-4**

Suppose your chance of being offered a certain job is 0.45, your probability of getting another job is 0.55, and your probability of being offered both jobs is 0.30. What is the probability that you will be offered at least one of the two jobs?

**Solution:** Let A be the event that the first job is offered and B the event that the second job is offered. Then,

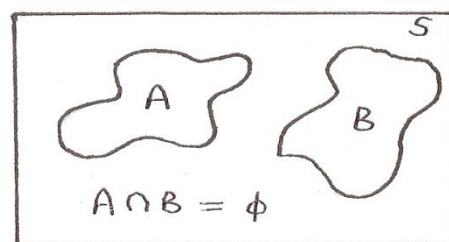
$$P(A) = 0.45 \quad P(B) = 0.55 \quad \text{and } P(A \cap B) = 0.30$$

So, the required probability is given as:

$$\begin{aligned}P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\&= 0.45 + 0.55 - 0.30 \\&= 0.70\end{aligned}$$

**Mutually Exclusive Events**

When the sets corresponding to two events are disjoint (*i.e.*, have no intersection), the two events are called **mutually exclusive** (see Figure 6-4).



**Figure 1-4 Two Mutually Exclusive Events A and B**

For mutually exclusive events, the probability of the intersection of the events is zero. This is so because the intersection of the events is the empty set, and we know that the probability of the empty set is zero.



For mutually exclusive events A and B:

$$P(A \cap B) = 0 \quad \dots\dots\dots(1.6)$$

This fact gives us a special rule for unions of mutually exclusive events. Since the probability of the intersection of the two events is zero, there is no need to subtract  $P(A \cap B)$  when the probability of the union of the two events is computed. Therefore, For mutually exclusive events A and B:

$$P(A \cup B) = P(A) + P(B) \quad \dots\dots\dots(1.7)$$

This is not really a new rule since we can always use the rule of unions for the union of two events: If the events happen to be mutually exclusive, we subtract zero as the probability of the intersection.

### **Example 1-5**

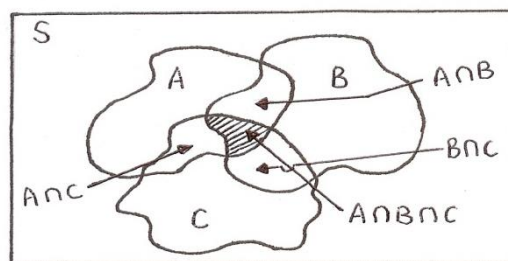
A card is drawn from a well-shuffled pack of playing cards. Find the probability that the card drawn is either a king or a queen.

**Solution:** Let A be the event that a king is drawn and B the event that a queen is drawn. Since A and B are two mutually exclusive events, we have,

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) \\ &= 4/52 + 4/52 = 8/52 = 2/13 \end{aligned}$$

We can extend the Rule of Unions to three (or more) events. Let A, B, and C be the three events defined over the sample space S, as shown in Figure 6-5 Then, the Rule of Unions is

$$\begin{aligned} P(A \cup B \cup C) &= \\ P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C) &\dots\dots\dots(1.8) \end{aligned}$$



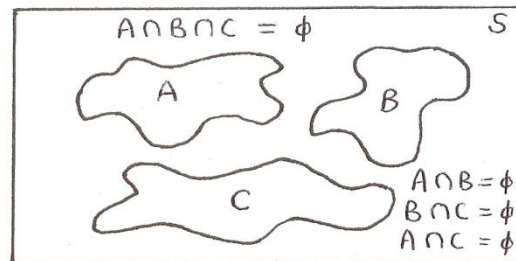
**Figure 1-5 Three Overlapping Events A, B and C**





When the three events are mutually exclusive (see Figure 8-6), the Rule of Unions is

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) \quad \dots\dots\dots(1.9)$$



**Figure 1-6 Three Mutually Exclusive Events A, B and C**

### **Example 1-6**

A card is drawn from a well-shuffled pack of playing cards. Find the probability that the card drawn is

- (a) either a heart or an honour or king
- (b) either an ace or a king or a queen

**Solution:** (a) Let A be the event that a heart is drawn, B the event that an honour is drawn and C the event that a king is drawn. So we have

$$n(A) = 13 \quad n(B) = 20 \quad n(C) = 4$$

$$n(A \cap B) = 5 \quad n(B \cap C) = 4 \quad n(A \cap C) = 1$$

and  $n(A \cap B \cap C) = 1$

The required probability (using Eq. (6.8) is

$$P(A \cup B \cup C) = 13/52 + 20/52 + 4/52 - 5/52 - 4/52 - 1/52 + 1/52$$

$$= 28/52$$

$$= 7/13$$

(b) Let A be the event that an ace is drawn, B the event that a king is drawn and C the event that a queen is drawn. So we have

$$n(A) = 4 \quad n(B) = 4 \quad n(C) = 4$$

Since A, B and C are mutually exclusive events, the required probability (using Eq. (6.9) is

$$P(A \cup B \cup C) = 4/52 + 4/52 + 4/52$$

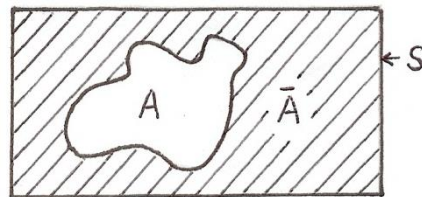
$$= 12/52$$



$$= 3/13$$

### THE COMPLEMENT RULE

The **Rule of Complements** defines the probability of the complement of an event in terms of the probability of the original event. Consider event  $A$  defined over the sample space  $S$ . The complement of set  $A$ , denoted by  $\bar{A}$ , is a subset, which contains all outcomes, which do not belong to  $A$  (see Figure 1-7).



**Figure 1-7 Complement of an Event**

In other words	$A + \bar{A} = S$	
so	$P(A + \bar{A}) = P(S)$	
or	$P(A) + P(\bar{A}) = 1$	
or	$P(\bar{A}) = 1 - P(A)$	.....(1.10)

Eq. (8.10) is our Rule of Complements. As a simple example, if the probability of rain tomorrow is 0.3, then the probability of no rain tomorrow must be  $1 - 0.3 = 0.7$ . If the probability of drawing a king is  $4/52$ , then the probability of the drawn card's not being a king is  $1 - 4/52 = 48/52$ .

### Example 1-7

Find the probability of the event of getting a total of less than 12 in the experiment of throwing a die twice.

**Solution:** Let  $A$  be the event of getting a total 12.

Then we have,

$$A = \{6,6\} \quad \text{and} \quad P(A) = 1/36$$

The event of getting a total of less than 12 is the complement of  $A$ , so the required probability is

$$P(\bar{A}) = 1 - P(A)$$



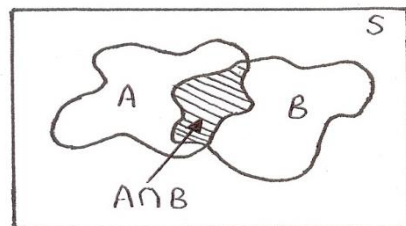
$$P(\bar{A}) = 1 - 1/36$$

$$P(\bar{A}) = 35/36$$

### THE CONDITIONAL PROBABILITY RULE

As a measure of uncertainty, probability depends on information. We often face situations where the probability of an event A is influenced by the information that another event B has occurred. Thus, the probability we would give the event "Xerox stock price will go up tomorrow" depends on what we know about the company and its performance; the probability is *conditional* upon our information set. If we know much about the company, we may assign a different probability to the event than if we know little about the company. We may define the probability of event A *conditional upon* the occurrence of event B. In this example, event A may be the event that the stock will go up tomorrow, and event B may be a favorable quarterly report.

Consider two events A and B defined over the sample space S, as shown in Figure 1-8



**Figure 1-8 Conditional Probability of Event A**

Thus, the probability of event A given the occurrence of event B is

$$P(A/B) = \frac{n(A \cap B)}{n(B)}$$

$$P(A/B) = \frac{n(A \cap B)/N}{n(B)/N}$$

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \quad \dots\dots\dots(1.11)$$

The vertical line in  $P(A/B)$  is read *given*, or *conditional upon*.



Therefore, the probability of event A given the occurrence of event B is defined as the probability of the intersection of A and B, divided by the probability of event B.

### Example 1-8

For an experiment of throwing a die twice, find the probability:

- (a) of the event of getting a total of 9, given that the die has shown up points between 4 and 6 (both inclusive)
- (b) of the event of getting points between 4 and 6 (both inclusive), given that a total of 9 has already been obtained

**Solution:** Let getting a total 9 be the event A and the die showing points between 4 and 6 (both inclusive) be the event B

Thus,  $N(S) = 36$  and  $A = \{(3,6) (4,5) (5,4) (6,3)\}$

$B = \{(4,4) (4,5) (4,6) (5,4) (5,5) (5,6) (6,4) (6,5) (6,6)\}$

and  $P(A \cap B) = \{(4,5) (5,4)\}$

So  $n(A) = 4$      $n(B) = 9$      $n(A \cap B) = 2$

So the required probabilities are

$$(a) \quad P(A / B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A / B) = \frac{2 / 36}{9 / 36}$$

$$P(A / B) = \frac{2}{9}$$

$$(b) \quad P(B / A) = \frac{P(B \cap A)}{P(A)}$$

$$P(B / A) = \frac{2 / 36}{4 / 36}$$

$$P(B / A) = \frac{1}{2}$$



### THE PRODUCT RULE

The **Product Rule** (also called **Multiplication Theorem**) allows us to write the probability of the simultaneous occurrence of two (or more) events.

In the conditional probability rules

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B/A) = \frac{P(B \cap A)}{P(A)}$$

$A \cap B$  or  $B \cap A$  is the event A and B occur simultaneously. So rearranging the conditional probability rules, we have our **Product Rule**

$$P(A \cap B) = P(A/B).P(B) \text{ and } P(A \cap B) = P(B/A).P(A) \quad \dots\dots(1.12)$$

The Product Rule states that the probability that both A and B will occur simultaneously is equal to the probability that B (or A) will occur multiplied by the conditional probability that A (or B) will occur, when it is known that B (or A) is certain to occur or has already occurred.

#### Example 1-9

A box contains 10 balls out of which 2 are green, 5 are red and 3 are black. If two balls are drawn at random, one after the other without replacement, from the box. Find the probabilities that:

- (a) both the balls are of green color
- (b) both the balls are of black color
- (c) both the balls are of red color
- (d) the first ball is red and the second one is black
- (e) the first ball is green and the second one is red

**Solution:** (a)  $P(G_1 \cap G_2) = P(G_2 / G_1).P(G_1)$

$$= \frac{1}{9} \times \frac{2}{10}$$

$$= \frac{1}{45}$$

(b)  $P(B_1 \cap B_2) = P(B_2 / B_1).P(B_1)$

$$= \frac{2}{9} \times \frac{3}{10}$$

$$= \frac{1}{15}$$



$$(c) \quad P(R_1 \cap R_2) = P(R_2 / R_1) \cdot P(R_1)$$

$$= \frac{4}{9} \times \frac{5}{10}$$

$$= \frac{2}{9}$$

$$(d) \quad P(R_1 \cap B_2) = P(B_2 / R_1) \cdot P(R_1)$$

$$= \frac{3}{9} \times \frac{5}{10}$$

$$= \frac{1}{6}$$

$$(e) \quad P(G_1 \cap R_2) = P(R_2 / G_1) \cdot P(G_1)$$

$$= \frac{5}{9} \times \frac{2}{10}$$

$$= \frac{1}{9}$$

### **Example 1-10**

A consulting firm is bidding for two jobs, one with each of two large multinational corporations. The company executives estimate that the probability of obtaining the consulting job with firm A, event A, is 0.45. The executives also feel that if the company should get the job with firm A, then there is a 0.90 probability that firm B will also give the company the consulting job. What are the company's chances of getting both jobs?

**Solution:** We are given  $P(A) = 0.45$ . We also know that  $P(B / A) = 0.90$ , and we are looking for  $P(A \cap B)$ , which is the probability that both A and B will occur.

$$\text{So} \quad P(A \cap B) = P(B / A) \cdot P(A)$$

$$P(A \cap B) = 0.90 \times 0.45$$

$$= 0.405$$

### **Independent Events**

Two events are said to be *independent* of each other if the occurrence or non-occurrence of one event in any trial does not affect the occurrence of the other event in any trial. Events A and B are *independent* of each other if and only if the following three conditions hold:

**Conditions for the independence of two events A and B:**

$$P(A/B) = P(A) \quad \dots\dots\dots(1.13a)$$

$$P(B/A) = P(B) \quad \dots\dots\dots(1.13b)$$

$$\text{and } P(A \cap B) = P(A).P(B) \quad \dots\dots\dots(1.14)$$

The first two equations have a clear, intuitive appeal. The top equation says that when A and B are independent of each other, then the probability of A stays the same even when we know that B has occurred - it is a simple way of saying that knowledge of B tells us nothing about A when the two events are independent. Similarly, when A and B are independent, then knowledge that A has occurred gives us absolutely no information about B and its likelihood of occurring.

The third equation, however, is the most useful in applications. It tells us that when A and B are independent (and only when they are independent), we can obtain the probability of the joint occurrence of A and B (*i.e.* the probability of their intersection) simply by multiplying the two separate probabilities. This rule is thus called the ***Product Rule for Independent Events***.

As an example of independent events, consider the following: Suppose I roll a single die. What is the probability that the number 5 will turn up? The answer is 1/6. Now suppose that I told you that I just tossed a coin and it turned up heads. What is now the probability that the die will show the number 5? The answer is unchanged, 1/6, because events of the die and the coin are independent of each other. We see that  $P(5/H) = P(5)$ , which is the first rule above.

The rules for union and intersection of two independent events can be extended to sequences of more than two events.

**Intersection Rule**

The probability of the intersection of several independent events  $A_1, A_2, \dots$  is just the product of separate probabilities *i.e.*

$$P(A_1 \cap A_2 \cap A_3) = P(A_1).P(A_2).P(A_3)\dots\dots\dots(1.15)$$

**Union Rule**

The probability of the union of several independent events  $A_1, A_2, \dots$  is given by the following equation

$$P(A_1 \cup A_2 \cup A_3 \cup \dots\dots\dots) = 1 - P(\overline{A_1}).P(\overline{A_2}).P(\overline{A_3})\dots\dots\dots(1.16)$$



The union of several events is the event that at least one of the events happens.

### Example 1-11

A problem in mathematics is given to five students A, B, C, D and E. Their chances of solving it are  $1/2$ ,  $1/3$ ,  $1/3$ ,  $1/4$  and  $1/5$  respectively. Find the probability that the problem will

- (a) not be solved
- (b) be solved

**Solution:** (a) The problem will not be solved when none of the students solve it. So the required probability is:

$$\begin{aligned} P(\text{problem will not be solved}) &= P(\bar{A}).P(\bar{B}).P(\bar{C}).P(\bar{D}).P(\bar{E}) \\ &= (1 - 1/2).(1 - 1/3).(1 - 1/3).(1 - 1/4).(1 - 1/5) \\ &= 2/15 \end{aligned}$$

(b) The problem will be solved when at least one of the students solve it. So the required probability is:

$$\begin{aligned} P(A \cup B \cup C \cup D \cup E) &= 1 - P(\bar{A}).P(\bar{B}).P(\bar{C}).P(\bar{D}).P(\bar{E}) \\ &= 1 - 2/15 \\ &= 13/15 \end{aligned}$$

#### 1.1.4 Bayes' Theorem

As we have already noted in the introduction, the basic objective behind calculating probabilities is to help us in making decisions by quantifying the uncertainties involved in the situations. Quite often, whether it is in our personal life or our work life, decision-making is an ongoing process. Consider for example, a seller of winter garments, who is interested in the demand of the product. In deciding on the amount he should stock for this winter, he has computed the probability of selling different quantities and has noted that the chance of selling a large quantity is very high. Accordingly, he has taken the decision to stock a large quantity of the product. Suppose, when finally the winter comes and the season ends, he discovers that he is left with a large quantity of stock. Assuming that he is in this business, he feels that the earlier probability calculation should be updated given the new experience to help him decide on the stock for the next winter.

Similar to the situation of the seller of winter garment, situations exist where we are interested in an event on an ongoing basis. Every time some new information is available, we do revise our odds mentally. This revision of probability with added information is formalised in probability theory with





the help of famous Bayes' Theorem. The theorem, discovered in 1761 by the English clergyman Thomas Bayes, has had a profound impact on the development of statistics and is responsible for the emergence of a new philosophy of science. Bayes himself is said to have been unsure of his extraordinary result, which was presented to the Royal Society by a friend in 1763 - after Bayes' death. We will first understand *The Law of Total Probability*, which is helpful for derivation of Bayes' Theorem.

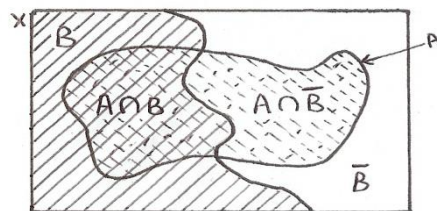
### The Law of Total Probability

Consider two events A and B. Whatever may be the relation between the two events, we can *always* say that the probability of A is equal to the probability of the intersection of A and B, plus the probability of the intersection of A and the complement of B (event  $\bar{B}$ ).

$$P(A) = P(A \cap B) + P(A \cap \bar{B})$$

or  $P(A) = P(A/B).P(B) + P(A/\bar{B}).P(\bar{B}) \dots\dots\dots(1.17)$

The sets B and  $\bar{B}$  form a **partition** of the sample space. A partition of a space is the division of the sample space into a set of events that are mutually exclusive (disjoint sets) and cover the whole space. Whatever event B may be, either B or  $\bar{B}$  must occur, but not both. Figure 1-9 demonstrates this situation and the law of total probability.



**Figure 1-9 Total Probability of Event A**

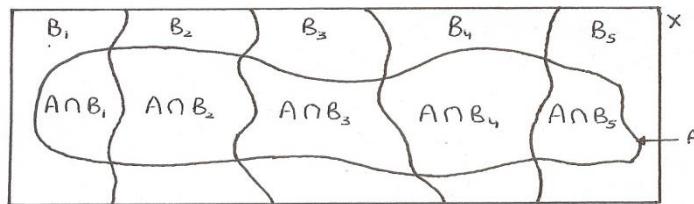
The law of total probability may be extended to more complex situations, where the sample space X is partitioned into more than two events. Say, we have partition of the space into a collection of  $n$  sets  $B_1, B_2, \dots, B_n$ . The law of total probability in this situation is:

$$P(A) = \sum_{i=1}^n P(A \cap B_i)$$



or 
$$P(A) = \sum_{i=1}^n P(A / B_i) \cdot P(B_i) \quad \dots\dots\dots (1.18)$$

Figure 1-10 shows the partition of a sample space into five events  $B_1, B_2, B_3, B_4$  and  $B_5$ ; and shows their intersections with set  $A$ .



**Figure 1-10 Total Probability of Event A**

We can demonstrate the rule with a more specific example. Let us define  $A$  as the event that an honour card is drawn out of a deck of 52 cards (the honour cards are the aces, kings, queens, jacks and 10). Letting  $H, C, D$ , and  $S$  denote the events that the card drawn is a heart, club, diamond, or spade, respectively, we find that the probability of an honour card is:

Heart	Diamond	Spade	Club
1 2 3	1 2 3	1 2 3	1 2 3
4 5 6	4 5 6	4 5 6	4 5 6
7 8 9	7 8 9	7 8 9	7 8 9
10 J Q	10 J Q	10 J Q	10 J Q
K A	K A	K A	K A

$A \cap H$        $A \cap D$        $A \cap S$        $A \cap C$

**Figure 1-11 Total Probability of Event A: An Honour Card**

$$\begin{aligned}
 P(A) &= P(A \cap H) + P(A \cap C) + P(A \cap D) + P(A \cap S) \\
 &= 5/52 + 5/52 + 5/52 + 5/52 \\
 &= 20/52 \\
 &= 5/13
 \end{aligned}$$

which is what we know the probability of an honour card to be just by counting 20 honour cards out of a total of 52 cards in the deck. The situation is shown in Figure 6-11.

As can be seen from the figure, the event  $A$  is the set addition of the intersections of  $A$  with each of the four sets  $H, D, C$ , and  $S$ .

**Example 1-12**

A market analyst believes that the stock market has a 0.70 probability of going up in the next year if the economy should do well, and a 0.20 probability of going up if the economy should not do well during the year. The analyst believes that there is a 0.80 probability that the economy will do well in the coming year. What is the probability that stock market will go up next year?

**Solution:** Let U be the event that the stock market will go and W is the event that the economy will do well in the coming year.

Then

$$\begin{aligned}
 P(U) &= P(U / W).P(W) + P(U / \bar{W}).P(\bar{W}) \\
 &= (0.70)(0.80) + (0.20)(0.20) \\
 &= 0.56 + 0.04 \\
 &= 0.60
 \end{aligned}$$

**BAYES' THEOREM**

We will now develop the Bayes' theorem. Bayes' theorem is easily derived from the law of total probability and the definition of conditional probability.

By definition of conditional probability, we have

$$P(B / A) = \frac{P(B \cap A)}{P(A)} \quad \dots\dots\dots(1.19)$$

By product rule, we have

$$P(B \cap A) = P(A \cap B) = P(A / B).P(B) \quad \dots\dots\dots(1.20)$$

Substituting Eq.(6.19) in Eq.(6.20), we have

$$P(B / A) = \frac{P(A / B).P(B)}{P(A)} \quad \dots\dots\dots(1.21)$$

By the law of total probability, we have

$$P(A) = P(A / B).P(B) + P(A / \bar{B}).P(\bar{B})$$

Substituting this expression for P(A) in the denominator of Eq.(1.21), we have the Bayes' theorem

$$P(B / A) = \frac{P(A / B).P(B)}{P(A / B).P(B) + P(A / \bar{B}).P(\bar{B})} \quad \dots\dots\dots(1.22)$$



Thus the theorem allows us to reverse the conditionality of events: we can obtain the probability of B given A from the probability of A given B (and other information). As we see from the theorem, the probability of B given A is obtained from the probabilities of B and  $\bar{B}$  and from the conditional probabilities of A given B and A given  $\bar{B}$ .

The probabilities  $P(B)$  and  $P(\bar{B})$  are called **prior probabilities** of the events B and  $\bar{B}$ ; the probability  $P(B/A)$  is called the **posterior probability** of B. It is possible to write Bayes' theorem in terms of  $\bar{B}$  and A, thus giving the posterior probability of  $\bar{B}$ ,  $P(\bar{B}/A)$ . Bayes' theorem may be viewed as a means of transforming our prior probability of an event B into a posterior probability of the event B - posterior to the known occurrence of event A.

The Bayes' theorem can be extended to a partition of more than two sets. This is done by using the law of total probability involving a partition in sets  $B_1, B_2, \dots, B_n$ . The resulting form of Bayes' theorem is:

$$P(B_i / A) = \frac{P(A / B_i) \cdot P(B_i)}{\sum_{i=1}^n P(A / B_i) \cdot P(B_i)} \quad \dots\dots(1.23)$$

The theorem gives the probability of one of the sets in the partition  $B_i$ , given the occurrence of event A.

### **Example 1-13**

An Economist believes that during periods of high economic growth, the Indian Rupee appreciates with probability 0.70; in periods of moderate economic growth, it appreciates with probability 0.40; and during periods of low economic growth, the Rupee appreciates with probability 0.20. During any period of time the probability of high economic growth is 0.30; the probability of moderate economic growth is 0.50 and the probability of low economic growth is 0.20. Suppose the Rupee value has been appreciating during the present period. What is the probability that we are experiencing the period of (a) high, (b) moderate, and (c) low, economic growth?

**Solution:** Our partition consists of three events: high economic growth (event H), moderate economic growth (event M) and low economic growth (event L). The prior probabilities of these events are:

$$P(H) = 0.30 \quad P(M) = 0.50 \quad P(L) = 0.20$$

Let A be the event that the rupee appreciates. We have the conditional probabilities

$$P(A / H) = 0.70 \quad P(A / M) = 0.40 \quad P(A / L) = 0.20$$



By using the Bayes' theorem we can find out the required probabilities

$P(H/A)$ ,  $P(M/A)$  and  $P(L/A)$

(a)  $P(H/A)$

$$\begin{aligned} P(H/A) &= \frac{P(A/H).P(H)}{P(A/H).P(H) + P(A/M).P(M) + P(A/L).P(L)} \\ &= \frac{(0.70)(0.30)}{(0.70)(0.30) + (0.40)(0.50) + (0.20)(0.20)} \\ &= 0.467 \end{aligned}$$

(b)  $P(M/A)$

$$\begin{aligned} P(M/A) &= \frac{P(A/M).P(M)}{P(A/H).P(H) + P(A/M).P(M) + P(A/L).P(L)} \\ &= \frac{(0.40)(0.50)}{(0.70)(0.30) + (0.40)(0.50) + (0.20)(0.20)} \\ &= 0.444 \end{aligned}$$

(c)  $P(L/A)$

$$\begin{aligned} P(L/A) &= \frac{P(A/L).P(L)}{P(A/H).P(H) + P(A/M).P(M) + P(A/L).P(L)} \\ &= \frac{(0.20)(0.20)}{(0.70)(0.30) + (0.40)(0.50) + (0.20)(0.20)} \\ &= 0.089 \end{aligned}$$

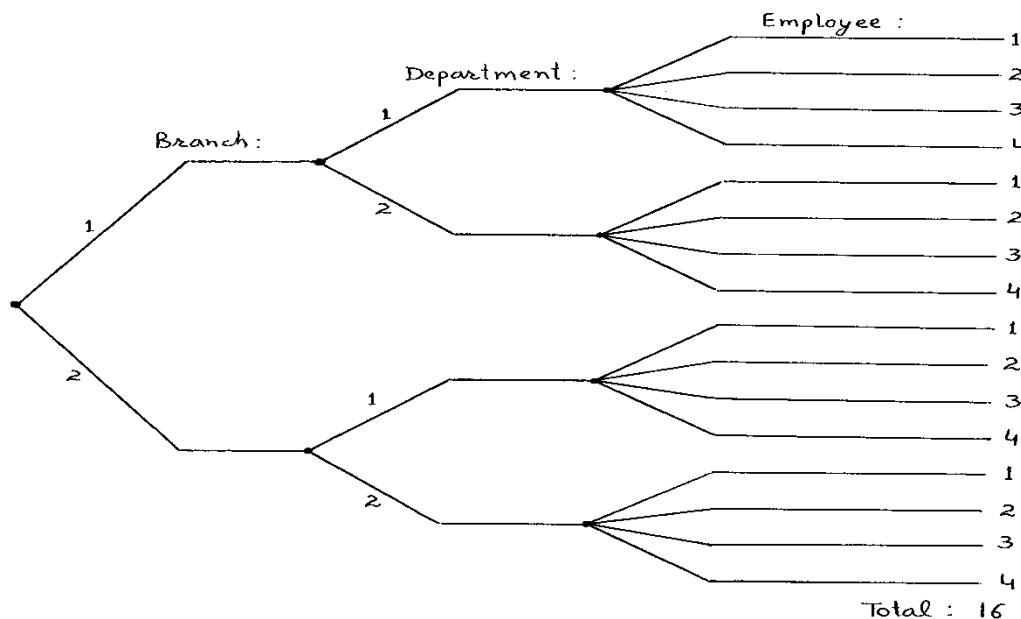
## 1.2 SOME COUNTING CONCEPTS

If there are  $n$  events and event  $i$  can occur in  $N_i$  possible ways, then the number of ways in which the sequence of  $n$  events may occur is

$$N_1. N_2. N_3 \dots \dots \dots N_n \quad \dots \dots \dots (1.24)$$

Suppose that a bank has two branches, each branch has two departments, and each department has four employees. Then there are  $(2)(2)(4)$  choices of employees, and the probability that a particular one will be randomly selected is  $1/(2)(2)(4) = 1/16$ .

We may view the choice as done sequentially: First a branch is randomly chosen, then a department within the branch, and then the employee within the department. This is demonstrated in the tree diagram in Figure 1-12.



For any positive integer  $n$ , we define  **$n$  factorial** as

$$n(n-1)(n-2) \dots 1 \quad \dots \dots \dots (1.25)$$

We denote  $n$  factorial by  $n!$ . The number  $n!$  is the number of ways in which  $n$  objects can be ordered. By definition,  $0! = 1$ .

For example,  $5!$  is the number of possible arrangements of five objects. We have  $5! = (5)(4)(3)(2)(1) = 120$ . Suppose that five applications arrive at a center on the same day, all written at different times. What is the probability that they will be read in the order in which they were written? Since there are 120 ways to order five applications, the probability of a particular order (the order in which the applications were written) is  $1/120$ .

**Permutations** are the possible ordered selections of  $r$  objects out of a total of  $n$  objects. The number of permutations of  $n$  objects taken  $r$  at a time is denoted by  ${}^n P_r$

$${}^n P_r = \frac{n!}{(n-r)!} \quad \dots \dots \dots (1.26)$$

Suppose that 4 people are to be randomly chosen out of 10 people who agreed to be interviewed in a market survey. The four people are to be assigned to four interviewers. How many possibilities are there? The first interviewer has 10 choices, the second 9 choices, the third 8, and the fourth 7. Thus,



there are  $(10)(9)(8)(7) = 5,040$  selections. We can see that this is equal to  $n(n-1)(n-2) \dots (n-r+1)$ , which is equal to  ${}^nP_r = \frac{n!}{(n-r)!}$ .

If choices are made randomly, the probability of any predetermined assignment of 4 people out of a group of 10 is  $1/5,040$ .

If choices are made randomly, the probability of any predetermined assignment of 4 people out of a group of 10 is  $1/5,040$ .

**Combinations** are the possible selections of  $r$  items from a group of  $n$  items regardless of the order of selection. The number of combinations is denoted by  ${}^nC_r$  and is read  $n$  choose  $r$ . We define the number of combinations of  $r$  out of  $n$  elements as

$${}^nC_r = \frac{n!}{r!(n-r)!} \quad \dots\dots\dots(1.27)$$

Suppose that 3 out of the 10 members of the board of directors of a large corporation are to be randomly selected to serve on a particular task committee. How many possible selections are there? Using Eq.

(6.27), we find that the number of combinations is  ${}^nC_r = \frac{n!}{r!(n-r)!} = 10!/(3!7!) = 120$ .

If the committee is chosen in a truly random fashion, what is the probability that the three-committee members chosen will be the three senior board members? This is 1 combination out of a total of 120, so the answer is  $1/120 = 0.00833$ .

### 1.3 CHECK YOUR PROGRESS

1. Subjective Approach involves personal judgment, information, intuition, and other ..... criteria.
2. Rule of Unions allows us to write the probability of the union of two events in terms of the probabilities of the two events and the probability of their ..... .
3. The Rule of Complements defines the probability of the ..... of an event in terms of the probability of the original event.
4. The Product Rule allows us to write the probability of the ..... occurrence of two (or more) events.
5. Bayes' theorem, discovered in 1761 by the English clergyman ..... .

### 1.4 SUMMARY

In life, we may also take some decisions regarding the price increase, reducing sales expenses *etc.* to manage the demand. However, in order to make such decisions, we need to quantify the chances of



different quantities of demand in the coming year. Probability theory provides us with the ways and means to quantify the uncertainties involved in such situations. Three different approaches to the definition and interpretation of probability have evolved, mainly to cater to the three different types of situations under which probability measures are normally required. There are different rules to find the probability like union rule, complement rule and product rule. Every time some new information is available, we do revise our odds mentally. This revision of probability with added information is formalized in probability theory with the help of famous Bayes' Theorem.

### 1.5 KEYWORDS

**Experiment:** It is a process that leads to one of several possible outcomes. An outcome of an experiment is some observation or measurement.

**Sample space:** It is the universal set  $S$  pertinent to a given experiment. It is the set of all possible outcomes of an experiment.

**Event:** It is a subset of a sample space. It is a set of basic outcomes. It can say that the event occurs if the experiment gives rise to a basic outcome belonging to the event.

**Classical Approach:** In this, probability of an event is defined as the *relative size* of the event with respect to the size of the sample space.

**Relative Frequency Approach:** It is used to compute probability in such cases. As per this approach, the probability of occurrence of an event is given by the observed relative frequency of an event in a very large number of trials.

**Permutations:** It is the possible ordered selections of  $r$  objects out of a total of  $n$  objects.

**Combinations:** It is the possible selections of  $r$  items from a group of  $n$  items regardless of the order of selection.

### 1.6 SELF-ASSESSMENT TEST

1. Explain what you understand by the term 'probability'. How the concept of probability is relevant to decision making under uncertainty?
2. What are different approaches to the definition of probability? Are these approaches contradictory to one another? Which of these approaches you will apply for calculating the probability that:
  - (a) A leap year selected at random, will contain 53 Monday.
  - (b) An item, selected at random from a production process, is defective.





- (c) Mr. Bhupinder S. Hooda will win the assembly election from Kiloi.
3. With the help of an example explain the meaning of the following:
- (a) Random experiment, and sample space
  - (b) An event as a subset of sample space
  - (c) Equally likely events
  - (d) Mutually exclusive events.
  - (e) Exhaustive events
  - (f) Elementary and compound events.
4. A proofreader is interested in finding the probability that the number of mistakes in a page will be less than 10. From his past experience he finds that out of 3600 pages he has proofed, 200 pages contained no errors, 1200 pages contained 5 errors, and 2200 pages contained 11 or more errors. Can you help him in finding the required probability?
5. State and develop the Addition Theorem of probability for:
- (a) mutually exclusive events
  - (b) overlapping events
  - (c) complementary events
6. Explain the concept of conditional probability with the help of a suitable example.
7. State and develop the Multiplication Theorem of probability for:
- (a) dependent events
  - (b) independent event
8. State the Bayes' Theorem of probability. Using an appropriate example, develop the Bayesian probability rule and generalize it.
9. What do you understand by permutations and combinations?
- (a) In how many ways we can select three players out of 12 players of the Indian Cricket team, for playing in the World XI team?
  - (b) In how many ways can a sub-committee of 2 out of 6 members of the executive committee of the employees' association be constituted?
10. What is the probability that a non leap year, selected at random, will contain
- (a) 52 Sundays? (b) 53 Sundays? (c) 54 Sundays?



11. A card is drawn at random from well shuffled deck of 52 cards, find the probability that
- (a) the card is either a club or diamond
  - (b) the card is not a king
  - (c) the card is either a face card or a club card.
12. From a well-shuffled deck of 52 cards, two cards are drawn at random.
- (a) If the cards are drawn simultaneously, find the probability that these consists of (i) both clubs, (ii) a king and a queen, (iii) a face card and a 8.
  - (b) If the cards are drawn one after the other with replacement. Find the probability that these consists of (i) both clubs, (ii) a king and a queen, (iii) a face card and a 8.
13. A problem in mathematics is given to four students A, B,C, and D their chances of solving it are  $\frac{1}{2}$  ,  $\frac{1}{3}$ ,  $\frac{1}{4}$  and  $\frac{1}{5}$  respectively. Find the probability that the problem will
- (a) be solved
  - (b) not be solved
14. The odds that A speaks the truth are 3:2 and the odds that B does so are 7:3. In what percentage of cases are they likely to
- (a) contradict each other on an identical point?
  - (b) agree each other on an identical point?
15. Among the sales staff engaged by a company 60% are males. In terms of their professional qualifications, 70% of males and 50% of females have a degree in marketing. Find the probability that a sales person selected at random will be
- (a) a female with degree in marketing
  - (b) a male without degree in marketing
16. A and B play for a prize of Rs. 10,000. A is to throw a die first and is to win if he throws 1: If A fails, B it to throw and is to win if he throws 2 or 1. If B fails, A is to throw again and to win if he throws 3, 2 or 1: and so on. Find their respective expectations.
17. A factory has three units A, B, and C. Unit A produces 50% of its products, and units B and C each produces 25% of the products. The percentage of defective items produced by A, B, and C units are 3%, 2% and 1%, respectively. If an item is selected at random from the total production of the factory is found defective, what is the probability that it is produced by:
- (a) Unit A                      (b) Unit B                      (c) Unit C



## **1.7 ANSWERS TO CHECK YOUR PROGRESS**

1. Subjective evaluation
2. Intersection
3. Complement
4. Simultaneous
5. Thomas Bayes

## **1.8 REFERENCES/SUGGESTED READINGS**

1. Statistics (Theory & Practice) by Dr. B.N. Gupta. Sahitya Bhawan Publishers and Distributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing Company Ltd., New Delhi.
3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill Publishing Company Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., New Delhi.
5. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.
6. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.
7. Statistics for Management by Richard I. Levin and David S. Rubin. Prentice Hall of India Pvt. Ltd., New Delhi.
8. Statistics for Business and Economics by Kohlar Heinz. Harper Collins., New York.



Subject: Business Statistics-II	
Course Code: BCOM 402	Author: Anil Kumar
Lesson: 02	Vetter: Dr. Karam Pal
<b>PROBABILITY DISTRIBUTIONS-I</b>	

## STRUCTURE

### 2.0 Learning Objectives

#### 2.1 Introduction

##### 2.1.1 Discrete Probability Distribution

##### 2.1.2 Bernoulli Random Variable

#### 2.2 The Binomial Distribution

##### 2.2.1 Conditions for a Binomial Random Variable

##### 2.2.2 Binomial Probability Function

##### 2.2.3 Characteristics of a Binomial Distribution

##### 2.2.4 Importance of the Binomial Distribution

##### 2.2.5 Fitting a Binomial Distribution

#### 2.3 The Poisson Distribution

##### 2.3.1 Characteristics of Poisson Distribution

##### 2.3.2 Role of the Poisson Distribution

##### 2.3.3 Fitting a Poisson Distribution

#### 2.4 Check your Progress

#### 2.5 Summary

#### 2.6 Keywords

#### 2.7 Self-Assessment Test

#### 2.8 Answers to check your progress

#### 2.9 References/Suggested Readings

## 2.0 LEARNING OBJECTIVES

After going through this lesson, students will be able to:



- Understand the concept of random variable and discrete probability distributions.
- Appreciate the usefulness of probability distributions in decision-making
- Identify situations where Binomial and Poisson probability distributions can be applied.

## 2.1 INTRODUCTION

In many situations, our interest does not lie in the outcomes of an experiment as such; we may find it more useful to describe a particular property or attribute of the outcomes of an experiment in numerical terms. For example, out of three births; our interest may be in the matter of the probabilities of the number of boys. Consider the sample space of 8 equally likely sample points.

GGG	GGB	GBG	BGG
GBB	BGB	BBG	BBB

Now look at the variable *“the number of boys out of three births”*. This number varies among sample points in the sample space and can take values 0,1,2,3, and it is random –given to chance.

*“A random variable is an uncertain quantity whose value depends on chance.”*

A random variable may be...

- **Discrete** if it takes only a countable number of values. For example, number of dots on two dice, number of heads in three coin tossing, number of defective items, number of boys in three births and so on.
- **Continuous** if can take on any value in an interval of numbers (*i.e.* its possible values are unaccountably infinite). For example, measured data on heights, weights, temperature, and time and so on.

A random variable has a probability law - a rule that assigns probabilities to different values of the random variable. This probability law - the probability assignment is called the **probability distribution** of the random variable. We usually denote the random variable by  $X$ . In this lesson, we will discuss discrete probability distributions. Continuous probability distributions will be discussed in the next lesson.



### 2.1.1 Discrete Probability Distribution

The random variable  $X$  denoting “*the number of boys out of three births*”, we introduced in the introduction of the lesson, is a discrete random variable; so it will have a discrete probability distribution. It is easy to visualize that the random variable  $X$  is a function of sample space. We can see the correspondence of sample points with the values of the random variable as follows:

BBB ( $X=0$ )	GGB GBG BGG ( $X=1$ )
GBB BGB BBG ( $X=2$ )	BBB ( $X=3$ )

The correspondence between sample points and the value of the random variable allows us to determine the probability distribution of  $X$  as follows:

$P(X=0) = 1/8$  since one out of 8 equally likely points leads to  $X = 0$

$P(X=1) = 3/8$  since three out of 8 equally likely points leads to  $X = 1$

$P(X=2) = 3/8$  since three out of 8 equally likely points leads to  $X = 2$

$P(X=3) = 1/8$  since one out of 8 equally likely points leads to  $X = 3$

The above probability statement constitute the probability distribution of the random variable  $X = \text{number of boys in three births}$ . We may appreciate how this probability law is obtained simply by associating values of  $X$  with sets in the sample space. (For example, the set GGB, GBG, BGG leads to  $X = 1$ ). We may write down the probability distribution of  $X$  in table format (see Table 2-1) or we may plot it graphically by means of probability Histogram (see Figure 2-1a) or a Line chart (see Figure 2-1b).



Table 2-1

## Probability Distribution of the Number of Boys out of Three Births

No. of Boys $X$	Probability $P(X)$
0	$1/8$
1	$3/8$
2	$3/8$
3	$1/8$

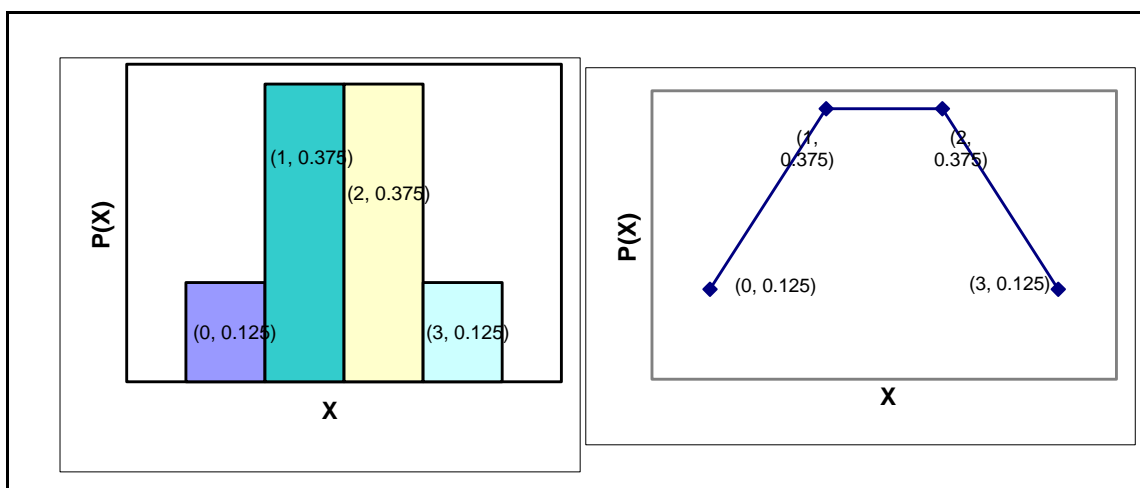


Figure 2-1 Probability Distribution of the Number of Boys out of Three Births

The probability distribution of a discrete random variable  $X$  must satisfy the following two conditions:

1.  $P(X = x) \geq 0$  for all values  $x$

2.  $\sum_{all\ x} P(X = x) = 1$

These conditions must hold because the  $P(X = x)$  values are probabilities. First condition specifies that all probabilities must be greater than or equal to zero, as we know from Lesson 6.



For the second condition, we note that for each value  $x$ ,  $P(x) = P(X = x)$  is the probability of the event that the random variable equals  $x$ . Since by definition all  $x$  means all the values the random variable  $X$  may take, and since  $X$  may take on only one value at a time, the occurrences of these values are mutually exclusive events, and one of them must take place. Therefore, the sum of all the probabilities  $P(X = x)$  must be 1.00.

### ***Cumulative Distribution Function***

The probability distribution of a discrete random variable lists the probabilities of occurrence of different values of the random variable. We may be interested in *cumulative* probabilities of the random variable. That is, we may be interested in the probability that the value of the random variable is *at most* some value  $x$ . This is the sum of all the probabilities of the values  $i$  of  $X$  that are less than or equal to  $x$ .

The *cumulative distribution function* (also called *cumulative probability function*)  $F(X = x)$  of a discrete random variable  $X$  is

$$F(X = x) = P(X \leq x) = \sum_{\text{all } i \leq x} P(i)$$

For example, to find the probability of at most two boys out of three births, we have

$$\begin{aligned} F(X = 2) &= P(X \leq 2) = \sum_{\text{all } i \leq 2} P(i) \\ &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= 1/3 + 3/8 + 3/8 \\ &= 7/8 \end{aligned}$$

### ***Expected Value and Variance of a Discrete Random Variable***

The expected value of a discrete random variable  $X$  is equal to the sum of all values of the random variable, each value multiplied (weighted) by its probability.

$$\mu = E(X) = \sum_{\text{all } x} x \cdot P(x)$$

The variance of a discrete random variable is given by

$$\sigma^2 = V(X) = E[(X - \mu)^2] = \sum_{\text{all } x} (x - \mu)^2 \cdot P(x)$$

In the same way we can calculate the other summary measures *viz.* skewness, kurtosis and moments.



***Probability Distributions are Theoretical Distributions***

Consider a random variable  $X$  that measures the “number of heads” in a three-trial coin tossing experiment. The probability distribution of  $X$  will be

$X$	:	0	1	2	3
$P(X)$	:	1/8	3/8	3/8	1/8

Now imagine this experiment is repeated 200 times, we may expect ‘no head’ and ‘three heads’ will each occur 25 times; ‘one head’ and ‘two heads’ each will occur 75 times. Since these results are what we expect on the basis of theory, the resultant distribution is called a ***theoretical or expected distribution***.

However, when the experiment is actually performed 200 times, the results, which we may actually obtain, will normally differ from the theoretically expected results. It is quite possible that in actual experiment ‘no head’ and ‘three heads’ may occur 20 and 28 times respectively and ‘one head’ and ‘two heads’ may occur 66 and 86 times respectively. The distribution so obtained through actual experiment is called the ***empirical or observed distribution***.

In practice, however, assessing the probability of every possible value of a random variable through actual experiment can be difficult, even impossible, especially when the probabilities are very small. But we may be able to find out what type of random variable the one at hand is by examining the causes that make it random. Knowing the type, we can often approximate the random variable to a standard one for which convenient formulae are available.

The proper identification of experiments with certain known processes in Probability theory can help us in writing down the probability distribution function. Two such processes are the ***Bernoulli Process*** and the ***Poisson Process***. The standard discrete probability distributions that are consequent to these processes are the ***Binomial*** and the ***Poisson*** distribution. We will now look into the conditions that characterize these processes, and examine the standard distributions associated with the processes. This will enable us to identify situations for which these distributions apply.

Let us first study the ***Bernoulli random variable***, named so in honor of the mathematician Jakob Bernoulli (1654-1705). It is the building block for other random variables and the resulting distributions we will study in this lesson.



### 2.1.2 Bernoulli Random Variable

Suppose an operator uses a lathe to produce pins, and the lathe is not perfect in the sense that it does not always produce a good pin. Rather, it has a probability  $p$  of producing a good pin and  $(1 - p)$  of producing a defective one. Let us denote a good pin as “*success*” and a defective pin as “*failure*”.

Just after the operator produces one pin, it is inspected; let  $X$  denote the “*number of good pins produced*” i. e. “*the number of successes*”.

Now analyzing the trial- “*inspecting a pin*” and our random variable  $X$ -“*number of successes*”, we note two important points:

- The trial- “*inspecting a pin*” has only two possible outcomes, which are mutually exclusive. Such a trial, whose outcome can only be either a success or a failure, is a ***Bernoulli trial***. In other words, the sample space of a Bernoulli trial is  $S = \{\text{success, failure}\}$ .
- The random variable,  $X$ , that measures number of successes in one Bernoulli trial, is a ***Bernoulli random variable***. Clearly,  $X$  is 1 if the pin is good and 0 if it is defective.

It is easy to derive the probability distribution of Bernoulli random variable

$$\begin{array}{lcl} X & : & 0 \quad 1 \\ P(X) & : & p \quad 1-p \end{array}$$

If  $X$  is a Bernoulli random variable, we may write

$$X \sim \text{BER}(p)$$

Where  $\sim$  is read as “is distributed as” and BER stands for Bernoulli.

A Bernoulli random variable is too simple to be of immediate practical use. But it forms the building block of the ***Binomial random variable***, which is quite useful in practice. The binomial random variable in turn is the basis for many other useful cases, such as ***Poisson random variable***.

## 2.2 THE BINOMIAL DISTRIBUTION

In the real world we often make several trials, not just one, to achieve one or more successes. Let us consider such cases of several trials.

Consider  $n$  number of *identically and independently distributed* Bernoulli random variables  $X_1, X_2, \dots, X_n$ . Here, *identically* means that they all have the same  $p$ , and *independently* means that the value of one  $X$  does not in any way affect the value of another. For example, the value of  $X_2$  does not



affect the value of  $X_3$  or  $X_8$  and so on. Such a sequence of identically and independently distributed Bernoulli variables is called a ***Bernoulli Process***.

Suppose an operator produces  $n$  pins, one by one, on a lathe that has probability  $p$  of making a good pin at each trial, the sequence of numbers (1 or 0) denoting the good and defective pins produced in each of the  $n$  trials is a Bernoulli process. For example, in the sequence of nine trials denoted by 001011001, the third, fifth, sixth and ninth are good pins, or successes. The rest are failures.

In practice, we are usually interested in the total number of good pins rather than the sequence of 1's and 0's. In the example above, four out of nine are good. In the general case, let  $X$  denote the total number of good pins produced in  $n$  trials. We then have  $X = X_1 + X_2 + \dots + X_n$  where all  $X_i \sim \text{BER}(p)$  and are independent.

The random variable that counts the number of successes in many independent, identical Bernoulli trials is called a Binomial Random Variable.

### 2.2.1 Conditions for a Binomial Random Variable

We may appreciate that the condition to be satisfied for a binomial random variable is that ***the experiment should be a Bernoulli Process***.

Any uncertain situation or experiment that is marked by the following three properties is known as a Bernoulli Process:

- There are only two mutually exclusive and collectively exhaustive outcomes in the experiment *i.e.*  
 $S = \{\text{success, failure}\}$
- In repeated trials of the experiment, the probabilities of occurrence of these events remain constant
- The outcomes of the trials are independent of one another

The probability distribution of Binomial Random Variable is called the ***Binomial Distribution***

### 2.2.2 BINOMIAL PROBABILITY FUNCTION

Now we will develop the distribution of our Binomial random variable. To describe the distribution of Binomial random variable we need two parameters,  $n$  and  $p$  we write  $X \sim B(n, p)$  to indicate that  $X$  is Binomially distributed with  $n$  number of independent trials and  $p$  probability of success in each trial. The letter  $B$  stands for binomial.



Let us analyze the probability that the number of successes  $X$  in the  $n$  trials is exactly  $x$  (obviously number of failures are  $n-x$ ) i.e.  $X = x$  and  $x = 0, 1, 2, \dots, n$ ; as  $n$  trials are made, at the best all  $n$  can be successes.

Now we know that there are  ${}^nC_x$  ways of getting  $x$  successes out of  $n$  trials. We also observe that each of these  ${}^nC_x$  possibilities has  $p^x(1-p)^{n-x}$  probability of occurrence corresponding to  $x$  successes and  $(n-x)$  failures. Therefore,

$$P(X = x) = {}^nC_x p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, 2, \dots, n$$

This equation is the Binomial probability formula. If we denote the probability of failure as  $q$  then the Binomial probability formula is

$$P(X = x) = {}^nC_x p^x q^{n-x} \quad \text{for } x = 0, 1, 2, \dots, n$$

We may write down the Binomial probability distribution in table format (see Table 2-2)

**Table 2-2**

<b>Binomial Distribution of <math>X</math></b>	
$X = x$	$P(X = x)$
0	${}^nC_0 p^0 q^n$
1	${}^nC_1 p^1 q^{n-1}$
...	...
$x$	${}^nC_x p^x q^{n-x}$
...	...
...	...
$n$	${}^nC_n p^n q^0$

Each of the term for  $x = 0, 1, 2, \dots, n$  correspond to the Binomial expansion of  $(p + q)^n$

### 2.2.3 Characteristics of a Binomial Distribution

#### 1. Expected Value or Mean



The expected value or the mean, denoted by  $\mu$ , of a Binomial distribution is computed as  $E(X) = \mu =$

$$\sum_{x=0}^n x.P(x). \text{ An evaluation of } \mu \text{ will show that } \mu = np.$$

## 2. Variance

The variance, denoted by  $\sigma^2$ , of a Binomial distribution is computed as

$$\begin{aligned} V(X) &= \sigma^2 = E[(X - \mu)^2] \\ &= \sum_{x=0}^n (x - \mu)^2 . P(x) \end{aligned}$$

An evaluation of  $\sigma^2$  will show that  $\sigma^2 = npq$

## 3. Moments about the Origin

The  $r^{th}$  moment about the origin denoted by  $m_r^0$ , of a Binomial distribution is computed as:  $m_r^0 =$

$$\sum_{x=0}^n x^r . P(x). \text{ For example:}$$

$$(a) \text{ First moment about the origin will be } m_1^0 = \sum_{x=0}^n x.P(x) = np = \mu.$$

$$(b) \text{ Second moment about the origin will be } m_2^0 = \sum_{x=0}^n x^2 . P(x) = n(n-1)p^2 + np.$$

## 4. Moments about the Mean

The  $r^{th}$  moment about the mean denoted by  $m_r^\mu$ , of a binomial distribution is computed as:  $m_r^\mu =$

$$\sum_{x=0}^n (x - \mu)^r . P(x). \text{ For example,}$$

$$(a) \text{ First moment about the mean will be } m_1^\mu = \sum_{x=0}^n (x - \mu)^1 . P(x) = 0$$

$$(b) \text{ Second moment about the mean will be } m_2^\mu = \sum_{x=0}^n (x - \mu)^2 . P(x) = npq = \sigma^2$$

$$(c) \text{ Third moment about the mean will be } m_3^\mu = \sum_{x=0}^n (x - \mu)^3 . P(x) = npq(q-p)$$

$$(d) \text{ Fourth moment about the mean will be } m_4^\mu = \sum_{x=0}^n (x - \mu)^4 . P(x) = 3(npq)^2 + npq(1-6pq)$$



## 5. Skewness

To bring out the skewness of a Binomial distribution we can calculate, moment coefficient of skewness,

$\gamma_1$

$$\gamma_1 = \sqrt{\beta_1} = \sqrt{\frac{(m_3^\mu)^2}{(m_2^\mu)^3}} = \frac{m_3^\mu}{(\sqrt{m_2^\mu})^3} = \frac{npq(q-p)}{(\sqrt{npq})^3} = \frac{q-p}{\sqrt{npq}}$$

Evaluating  $\gamma_1 = \frac{q-p}{\sqrt{npq}}$  we note:

- the Binomial distribution is skewed to the right *i.e.* has positive skewness when  $\gamma_1 > 0$ , which is so when  $p < q$
- the Binomial distribution is skewed to the left *i.e.* has negative skewness when  $\gamma_1 < 0$ , which is so when  $p > q$
- the Binomial distribution is symmetrical *i.e.* has no skewness when  $\gamma_1 = 0$ , which is so when  $p = q$ . Thus,  $n$  being the same, the degree of skewness in a Binomial distribution tends to vanish as  $p$  approaches  $\frac{1}{2}$  *i.e.* as  $p \rightarrow \frac{1}{2}$
- for a given value of  $p$ , as  $n$  increases the Binomial distribution moves to the right, flattens and spreads out As  $n \rightarrow \infty$ ,  $\gamma_1 \rightarrow 0$ , the distribution tends to be symmetrical.

## 6. Kurtosis

A measure of kurtosis of the Binomial distribution is given by the moment coefficient of kurtosis  $\gamma_2$

$$\gamma_2 = \beta_2 - 3 = \frac{m_4^\mu}{(m_2^\mu)^2} - 3 = \frac{3n^2 p^2 q^2 + npq(1-6pq)}{n^2 p^2 q^2} - 3 = \frac{1-6pq}{npq}$$

Evaluating  $\gamma_2 = \frac{1-6pq}{npq}$  we note

- the Binomial distribution is leptokurtic when  $\gamma_2 > 0$ , which is so when  $6pq < 1$ .
- the Binomial distribution is platykurtic when  $\gamma_2 < 0$ , which is so when  $6pq > 1$ .
- the Binomial distribution is mesokurtic when  $\gamma_2 = 0$ , which is so when  $6pq = 1$ .

## 7. Normal approximation of the Binomial distribution



If  $n$  is large and if neither of  $p$  or  $q$  is too close to zero, the Binomial distribution can be closely approximated by a Normal distribution with standardized variable  $Z = \frac{X - np}{\sqrt{npq}}$ .

### 8. Poisson approximation of the Binomial distribution

Binomial distribution can reasonably be approximated by the Poisson distribution when  $n$  is infinitely large and  $p$  is infinitely small i. e. when  $n \rightarrow \infty$  and  $p \rightarrow 0$ .

#### 2.2.4 Importance of Binomial Distribution

*The binomial probability distribution is a discrete probability distribution that is useful in describing an enormous variety of real life events. For example, a quality control inspector wants to know the probability of defective light bulbs in a random sample of 10 bulbs if 10% of the bulbs are defective. He can quickly obtain the answer from tables of the binomial distribution. The binomial distribution can be used when:*

- *The outcome or results of each trial in the process are featured as one of two types of possible outcomes. In other words, they are attributes.*
- *The possibility of outcome of any trial does not change and is independent of the results of previous trials.*

#### 2.2.5 Fitting a Binomial Distribution

*When a binomial distribution is to be fitted to observe data, the following procedure is adopted:*

1. *Determine the value of  $p$  and  $q$ . if one of these values is known the other can be found out by the simple relationship  $p = (1 - q)$ , and  $q = (1 - p)$ . When  $p$  and  $q$  are equal, the distribution is symmetrical. For  $p$  and  $q$  may be interchanged without alternating the value of any terms and consequently terms equidistant from the two ends of the series are equal. If  $p$  and  $q$  are unequal, the distribution is skew. If  $p$  is less than  $\frac{1}{2}$ , the distribution is positively skewed and when  $p$  is more than  $\frac{1}{2}$  the distribution is negatively skewed.*
2. *Expand the binomial  $(q + p)^n$ . The power  $n$  is equal to one less than the number of terms in the expected binomial. Thus when two coins are tossed ( $n = 2$ ) there will be three terms in the binomial. Similarly, when four coins are tossed ( $n = 4$ ) there will be five terms and so on.*



3. Multiply each term of the expanded binomial by  $N$  (the total frequency) in order to obtain the expected frequency in each category.

The following examples illustrate the procedure:

### **Example 2-1**

Assuming the probability of male birth as  $\frac{1}{2}$ , find the probability distribution of number of boys out of 5 births.

- (a) Find the probability that a family of 5 children have
- at least one boy
  - at most 3 boys
- (b) Out of 760 families with 5 children each find the expected number of families with (i) and (ii) above

**Solution:** Let the random variable  $X$  measures the number of boys out of 5 births. Clearly  $X$  is a binomial random variable. So we apply the Binomial probability function to calculate the required probabilities.

$$X \sim B(5, \frac{1}{2})$$

$$P(X = x) = {}^nC_x p^x q^{n-x} \text{ for } x = 0, 1, 2, 3, 4, 5$$

The probability distribution of  $X$  is given below

$X = x$	:	0	1	2	3	4	5
$P(X = x)$	:	1/32	5/32	10/32	10/32	5/32	1/32

- (a) The required probabilities are

$$(i) \quad P(X \geq 1) = 1 - P(X = 0)$$

$$= 1 - 1/32$$

$$= 31/32$$

$$(ii) \quad P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

$$= 1/32 + 5/32 + 10/32 + 10/32$$

$$= 26/32$$

- (b) Out of 760 families with 5 children, the expected number of families with

$$(i) \quad \text{at least one boy} = 760 * P(X \geq 1)$$

$$= 760 * 31/32$$





$$= 730$$

$$(ii) \quad \text{at most 3 boys} = 760 * P(X \leq 3)$$

$$= 760 * 26/32$$

$$= 720$$

**Example 2.2** Eight coins are tossed at a time 256 times. Number of heads observed at each throw is recorded and the result are given below. Find the expected frequencies. What are the theoretical values of mean and standard deviation? Calculate also the mean and standard deviation of the observed frequencies.

Number of heads at a throw	Frequency	Number of heads at a throw	Frequency
0	2	5	56
1	6	6	32
2	30	7	10
3	52	8	1
4	67		

**Solution:** The chance of getting a head in a single throw of one coin is  $\frac{1}{2}$ , hence  $p = \frac{1}{2}$ ,  $q = \frac{1}{2}$ ,  $n = 8$ ,  $N = 256$ .

By expanding  $256(\frac{1}{2} + \frac{1}{2})^8$  we shall get the expected frequencies of 1, 2, ....., 8 heads (successes).

Number of Heads (X)	Frequency = $N \times {}^nC_r q^{n-r} p^r$
0	$256 (\frac{1}{2})^8 = 1$
1	$256 \times {}^8C_1 (\frac{1}{2})^7 (\frac{1}{2})^1 = 8$
2	$256 \times {}^8C_2 (\frac{1}{2})^6 (\frac{1}{2})^2 = 28$
3	$256 \times {}^8C_3 (\frac{1}{2})^5 (\frac{1}{2})^3 = 56$
4	$256 \times {}^8C_4 (\frac{1}{2})^4 (\frac{1}{2})^4 = 70$
5	$256 \times {}^8C_5 (\frac{1}{2})^3 (\frac{1}{2})^5 = 56$
6	$256 \times {}^8C_6 (\frac{1}{2})^2 (\frac{1}{2})^6 = 28$
7	$256 \times {}^8C_7 (\frac{1}{2})^1 (\frac{1}{2})^7 = 8$
8	$256 \times {}^8C_8 (\frac{1}{2})^0 (\frac{1}{2})^8 = 1$
$n = 8$	Total (N) = 256



The mean of the above distribution is  $np = 8 \times \frac{1}{2} = 4$ . The standard deviation is:

$$\text{Standard Deviation} = \sqrt{npq} = \sqrt{\frac{1}{2} \times \frac{1}{2} \times 8} = \sqrt{2} = 1.414$$

These are the mean and standard deviation of the expected frequency distribution. The mean and the standard deviation of the observed frequency distribution shall be:

X	F	d	fd	fd <sup>2</sup>
0	2	-4	-8	32
1	6	-3	-18	54
2	30	-2	-60	120
3	52	-1	-52	52
4	67	0	0	0
5	56	+1	+56	56
6	32	+2	+64	128
7	10	+3	+30	90
8	1	+4	+4	16
	N = 256		$\sum fd = 16$	$\sum fd^2 = 548$

$$\bar{X} = A + \frac{\sum fd}{N} = 4 + \frac{16}{256} = 4.0625$$

$$\sigma = \sqrt{\frac{fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = \sqrt{\frac{548}{256} - \left(\frac{16}{256}\right)^2} = \sqrt{2.141 - 0.04} = \sqrt{2.137} = 1.462$$

## 2.3 THE POISSON DISTRIBUTION

Poisson Distribution was developed by a French Mathematician Simeon D Poisson (1781-1840). If a random variable  $X$  is said to follow a **Poisson Distribution**, then its probability distribution is given by

$$P(X = x) = \frac{e^{-\mu} \mu^x}{x!} \quad x = 0, 1, 2, \dots$$

Where,  $x$  is the number of successes

$\mu$  is the mean of the Poisson distribution and

$e = 2.71828$  (the base of natural logarithms)



The random variable  $X$  counts the number of successes in **Poisson Process**. A Poisson process corresponds to a Bernoulli process under the following conditions:

- the number of trials  $n$ , is infinitely large i.e.  $n \rightarrow \infty$
- the constant probability of success  $p$ , for each trial is infinitely small i.e.  $p \rightarrow 0$  (obviously  $q \rightarrow 1$ )
- $np = \mu$  is finite

We can develop the Poisson probability rule from the Binomial probability rule under the above conditions.

Let us consider a Bernoulli process with  $n$  trials and probability of success in any trial

$p = \frac{\mu}{n}$ , where  $\mu \geq 0$ . Then, we know that the probability of  $x$  successes in  $n$  trials is given by

$$\begin{aligned}
 P(X = x) &= {}^nC_x \left(\frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{n-x} = \frac{n!}{x!(n-x)!} \left(\frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{n-x} \\
 &= \frac{n[n-1][n-2]\dots\dots\dots[n-(x-1)]}{x!} \left(\frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{n-x} \\
 &= \frac{\mu^x}{x!} \left[\frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \dots\dots\dots \frac{n-(x-1)}{n}\right] \left(1 - \frac{\mu}{n}\right)^{n-x} \\
 &= \frac{\mu^x}{x!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots\dots\dots \left(1 - \frac{x-1}{n}\right) \left(1 - \frac{\mu}{n}\right)^n \left(1 - \frac{\mu}{n}\right)^{-x}
 \end{aligned}$$

Now if  $n \rightarrow \infty$ , then the terms,  $\left(1 - \frac{1}{n}\right); \left(1 - \frac{2}{n}\right); \dots\dots\dots; \left(1 - \frac{x-1}{n}\right)$  and  $\left(1 - \frac{\mu}{n}\right)^{-x}$  will all be tending to 1

and  $\left(1 - \frac{\mu}{n}\right)^n \rightarrow e^{-\mu}$  if  $n \rightarrow \infty$  Thus we have  $P(X = x) = \frac{e^{-\mu} \mu^x}{x!}$  **where,  $x = 0, 1, 2, \dots\dots\dots$**

This equation is the probability distribution function of Poisson distribution.



Thus, we have seen that to describe the distribution of Poisson random variable we need only one parameter  $\mu$ , we write If  $X \sim POI(\mu)$ , Then  $P(X = x) = \frac{e^{-\mu} \mu^x}{x!}$   $x = 0, 1, 2, \dots$  We may write down the Poisson probability distribution in table format (see Table 2-3)

Table 2-3

Poisson Distribution of  $X$ 

$X = x$	$P(X = x)$
0	$e^{-\mu}$
1	$\mu e^{-\mu}$ or $\mu P(X = 0)$
2	$\frac{\mu^2}{2!} e^{-\mu}$ or $\frac{\mu}{2} P(X = 1)$
...	...
...	...
$X$	$\frac{\mu^x}{x!} e^{-\mu}$ or $\frac{\mu}{x} P(X = x-1)$
...	...
...	...

Poisson distribution may be expected in situations where the chance of occurrence of any event is small, and we are interested in the occurrence of the event and not in its non-occurrence. For example, number of road accidents, number of defective items, number of deaths in flood or because of snakebite or because of a rare disease *etc.* In these situations, we know about the occurrence of an event although its probability is very small, but we do not know how many times it does not occur. For instance, we can say that two road accidents took place today, but it is almost impossible to say as to how many times,



accident fails to take place. The reason is that the number of trials is very large here and the nature of event is of rare type. The Poisson random variable  $X$ , counts the number of times a rare event occurs during a fixed interval of time or space.

### 2.3.1 Characteristics of a Poisson Distribution

#### 1. Expected Value or Mean

The expected value or the mean, denoted by  $\mu$ , of a Poisson distribution is computed as  $E(X) = \mu =$

$\sum_{all\ x} x.P(x)$  An evaluation of *mean* will show that it is always  $\mu$  itself.

#### 2. Variance

The variance, denoted by  $\sigma^2$ , of a Poisson distribution is computed as  $V(X) = \sigma^2 = E[(X - \mu)^2] =$

$\sum_{all\ x} (x - \mu)^2 .P(x)$  An evaluation of  $\sigma^2$  will show that  $\sigma^2 = \mu$

#### 3. Moments about the Origin

The  $r^{th}$  moments about the origin denoted by  $m_r^0$ , of a Poisson distribution is computed as:

$$m_r^0 = \sum_{all\ x} x^r .P(x)$$

For example, (a) First moment about the origin will be

$$\begin{aligned} m_1^0 &= \sum_{all\ x} x.P(x) \\ &= \mu \end{aligned}$$

(b) Second moment about the origin will be

$$\begin{aligned} m_2^0 &= \sum_{all\ x} x^2 .P(x) \\ &= \mu + \mu^2 \end{aligned}$$

#### 4. Moments about the Mean

The  $r^{th}$  moments about the mean denoted by  $m_r^\mu$ , of a Poisson distribution is computed as:

$$m_r^\mu = \sum_{all\ x} (x - \mu)^r .P(x)$$



For example, (a) First moment about the mean will be

$$\begin{aligned} m_1^\mu &= \sum_{\text{all } x} (x - \mu)^1 \cdot P(x) \\ &= 0 \end{aligned}$$

(b) Second moment about the mean will be

$$\begin{aligned} m_2^\mu &= \sum_{\text{all } x} (x - \mu)^2 \cdot P(x) \\ &= \sigma^2 \\ &= \mu \end{aligned}$$

(c) Third moment about the mean will be

$$\begin{aligned} m_3^\mu &= \sum_{\text{all } x} (x - \mu)^3 \cdot P(x) \\ &= \mu \end{aligned}$$

(d) Fourth moment about the mean will be

$$\begin{aligned} m_4^\mu &= \sum_{\text{all } x} (x - \mu)^4 \cdot P(x) \\ &= 3\mu^2 + \mu \end{aligned}$$

## 5. Skewness

To bring out the skewness we can calculate, moment coefficient of skewness,  $\gamma_1$

$$\gamma_1 = \sqrt{\beta_1} = \sqrt{\frac{(m_3^\mu)^2}{(m_2^\mu)^3}} = \frac{m_3^\mu}{(\sqrt{m_2^\mu})^3} = \frac{1}{\sqrt{\mu}}$$

Evaluating  $\gamma_1 = \frac{1}{\sqrt{\mu}}$  we note that Poisson distribution is always skewed to the right *i.e.* has positive

skewness which is so as it is a distribution of rare events.

The degree of skewness in a Poisson distribution decreases as the value of  $\mu$  increases.

## 6. Kurtosis

A measure of kurtosis of the Poisson distribution is given by the moment coefficient of kurtosis  $\gamma_2$



$$\gamma_2 = \beta_2 - 3 = \frac{m_4^\mu}{(m_2^\mu)^2} - 3 = \frac{1}{\sqrt{\mu}}$$

Evaluating  $\gamma_2 = \frac{1}{\sqrt{\mu}}$  we note that the Poisson distribution is leptokurtic.

### 7. Poisson approximation of the Binomial distribution

Poisson distribution can reasonably approximate Binomial distribution when  $n$  is infinitely large and  $p$  is infinitely small *i. e.* when

$$n \rightarrow \infty \text{ and } p \rightarrow 0$$

#### 2.3.2 Role of Poisson Distribution

The Poisson distribution is used in practice in a wide variety of problems where there are infrequently occurring events with respect to time, area, volume or similar units. Some practical situations in which Poisson distribution can be used are given below:

1. It is used in quantity control statistics to count the number of defects of an item.
2. It is used in biology and physics to count the number of bacteria and to count the number of particles emitted from a radioactive substances.
3. It is used in insurance problems to count the number of casualties.
4. It is used in call center or telephonic companies in waiting-time problems to count the number of incoming telephones calls or incoming customers.
5. It is used in count the number of traffic arrivals such as trucks at terminals, airplanes at airports, ships at docks and so forth.
6. It is used in determining the number of deaths in a district in a given period say a year, by a rare diseases.
7. It is used in count the number of typographical errors per page in typed material, number of deaths as a result of road accident etc.
8. It is used in dealing with the inspection of manufactured products with the probability that any one piece is defective is very small and the lots are very large.



*In general, the Poisson distribution explains the behaviour of those discrete variants where the probability of occurrence of the event is small and the total number of possible cases is sufficiently large.*

### **2.3.3 Fitting of a Poisson Distribution**

*The process of fitting a Poisson distribution is very simple. We have just to obtain the value of  $m$ , i.e., the average occurrence, and calculate the frequency of 0 successes. The other frequencies can be very easily calculated as follows:*

$$N(P_0) = Ne^{-m}$$

$$N(P_1) = N(P_0) \times m/1$$

$$N(P_2) = N(P_1) \times m/2$$

$$N(P_3) = N(P_2) \times m/3$$

$$N(P_4) = N(P_3) \times m/4$$

**Example 2.3** *At a parking place the average number of car-arrivals during a specified period of 15 minutes is 2. If the arrival process is well described by a Poisson process, find the probability that during a given period of 15 minutes*

- (a) no car will arrive
- (b) atleast two cars will arrive
- (c) atmost three cars will arrive
- (d) between 1 and 3 cars will arrive

**Solution:** Let  $X$  denote the number of cars arrivals during the specified period of 15 minutes. So

$$X \sim POI(\mu)$$

We apply the Poisson probability function  $P(X = x) = \frac{e^{-\mu} \mu^x}{x!}$   $x = 0, 1, 2, \dots$  to calculate the required probabilities.





$$\begin{aligned} \text{(a)} \quad P(\text{no car will arrive}) &= P(X = 0) = \frac{e^{-2} 2^0}{0!} \\ &= 0.1353 \end{aligned}$$

$$\begin{aligned} \text{(b)} \quad P(\text{atleast two cars will arrive}) &= P(X \geq 2) \\ &= 1 - [P(X = 0) + P(X = 1)] \\ &= 1 - \left[ \frac{e^{-2} 2^0}{0!} + \frac{e^{-2} 2^1}{1!} \right] \\ &= 1 - [0.1353 + 0.2707] \\ &= 1 - 0.4060 \\ &= 0.5740 \end{aligned}$$

$$\begin{aligned} \text{(c)} \quad P(\text{atmost three cars will arrive}) &= P(X \leq 3) \\ &= \sum_{x=0}^3 \frac{e^{-2} 2^x}{x!} \\ &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\ &= 0.8571 \end{aligned}$$

$$\begin{aligned} \text{(d)} \quad P(\text{between 1 and 3 cars will arrive}) &= P(1 \leq X \leq 3) \\ &= P(X \leq 3) - P(X = 0) \\ &= \sum_{x=0}^3 \frac{e^{-2} 2^x}{x!} - \frac{e^{-2} 2^0}{0!} \\ &= 0.8571 - 0.1353 \\ &= 0.7218 \end{aligned}$$

## 2.4 CHECK YOUR PROGRESS

1. The Binomial distribution is skewed to the right *i.e.* has ..... when  $\gamma_1 > 0$ , which is so when  $p < q$ .
2. The degree of skewness in a Poisson distribution ..... as the value of  $\mu$  increases.
3. Poisson distribution can reasonably approximate ..... when  $n$  is infinitely large and  $p$  is infinitely small.
4. If  $n$  is large and if neither of  $p$  or  $q$  is too close to zero, the Binomial distribution can be closely approximated by a ..... with standardized variable.



5. The random variable that counts the number of successes in many independent, ..... is called a Binomial Random Variable.

## 2.5 SUMMARY

A random variable has a probability law - a rule that assigns probabilities to different values of the random variable. This probability law - the probability assignment is called the probability distribution of the random variable. The probability distribution of a discrete random variable lists the probabilities of occurrence of different values of the random variable. The proper identification of experiments with certain known processes in Probability theory can help us in writing down the probability distribution function. Two such processes are the Bernoulli Process and the Poisson Process. The standard discrete probability distributions that are consequent to these processes are the Binomial and the Poisson distribution. There are different characteristics of both.

## 2.6 KEYWORDS

**Discrete:** It takes only a countable number of values.

**Continuous:** It take on any value in an interval of numbers (*i.e.* it's possible values are unaccountably infinite).

**Binomial Random Variable:** The random variable that counts the number of successes in many independent, identical Bernoulli trials is called a Binomial Random Variable.

**Poisson distribution:** It may be expected in situations where the chance of occurrence of any event is small, and we are interested in the occurrence of the event and not in its non-occurrence.

## 2.7 SELF-ASSESSMENT TEST

1. Explain what you understand by random experiment and a random variable. Briefly explain the following:
  - a. Discrete and continuous random variables
  - b. Discrete probability distribution.
2. "Binomial random variable measures the number of successes in a Bernoulli Process". Explain this statement. Also develop and generalize Binomial probability rule with the help of an example.
3. State the important properties of a Binomial distribution. Give examples of some of the important area where Binomial distribution is used.



4. Under what condition can the Poisson distribution approximate Binomial distribution? Develop the Poisson probability rule from the Binomial probability rule under these conditions.
5. List some of the important areas where Poisson distribution is used. Also state the important properties of a Poisson distribution.
6. On an average a machine produces 20 % defective item find the probability that a random sample of 4 items consists of
  - (a) none to four defective items
  - (b) atleast 3 defective items
  - (c) almost 2 defective items.

Out of 200 samples of 4 items, find the expected number of samples with (a), (b), and (c) above

7. A gardener knows from his personal experiences that 2% of seedlings fail to service on transplantation. Find the mean, standard deviation and moment coefficient of skewness of the distribution of rate of failure to service in a sample of 400 seedlings.
8. If the sum of mean and variance of a binomial distribution of 5 trials is  $7/5$ , find the binomial distribution.
9. The mean and variance of a binomial distribution are 2 and 1.5 respectively. Find the probability of
  - (a) 2 successes
  - (b) atleast 2 successes
  - (c) at most 2 successes.
10. 150 random samples of 4 units each are inspected for number of defective item. The results are:
 

Number of defective items	:	0	1	2	3	4
Number of Samples	:	28	62	46	10	4

Fit a binomial distribution to the observed data.

11. The probability that a particular injection will have reaction to an individual is 0.002. Find the probability that out of 1000 individuals (a) no, (b) 1, (c) at least 1, and (d) almost 2; individuals will have reaction from the injection.
12. In a razor blades manufacturing factory, there is small chance of  $1/500$  for any blade to be defective. The blades are supplied in packets of 10. Find the approximate number of packets containing (a) no, (b) 1, and (c) 2 defective blades in a consignment of 10,000 packets.
13. If  $P(x = 1) = P(x = 2)$ , for a distribution of Poisson random variable  $X$ . Find the mean of the distribution.
14. The distribution of typing mistakes committed by a typist is given below:

Number of mistakes (X)	:	0	1	2	3	4	5
------------------------	---	---	---	---	---	---	---



Number of pages ( $f$ ) : 142 156 67 27 5 1

Fit a Poisson distribution and find the expected frequencies.

## 2.8 ANSWERS TO CHECK YOUR PROGRESS

1. Positive skewness
2. Decreases
3. Binomial distribution
4. Normal distribution
5. Identical Bernoulli trials

## 2.9 REFERENCES/SUGGESTED READINGS

1. Statistics (Theory & Practice) *by* Dr. B.N. Gupta. Sahitya Bhawan Publishers and Distributors (P) Ltd., Agra.
2. Statistics for Management *by* G.C. Beri. Tata McGraw Hills Publishing Company Ltd., New Delhi.
3. Business Statistics *by* Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill Publishing Company Ltd., New Delhi.
4. Statistics for Business and Economics *by* R.P. Hooda. MacMillan India Ltd., New Delhi.
5. Business Statistics *by* S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.
6. Statistical Method *by* S.P. Gupta. Sultan Chand and Sons., New Delhi.
7. Statistics for Management *by* Richard I. Levin and David S. Rubin. Prentice Hall of India Pvt. Ltd., New Delhi.
8. Statistics for Business and Economics *by* Kohlar Heinz. Harper Collins., New York.



Subject: Business Statistics-II	
Course Code: BCOM 402	Author: Anil Kumar
Lesson: 03	Vetter: Dr. Karam Pal
<b>PROBABILITY DISTRIBUTIONS-II</b>	

## **STRUCTURE**

- 3.0 Learning Objectives
- 3.1 Introduction
  - 3.1.1 Continuous Probability Distribution
  - 3.1.2 The Normal Distribution
- 3.2 The Standard Normal Distribution
  - 3.2.1 The Standard Area Table
  - 3.2.2 Finding Probabilities of Standard Normal Distribution
  - 3.2.3 Finding the Value of Z given a Probability
- 3.3 The Transformation of Normal Random Variables
- 3.4 Check your Progress
- 3.5 Summary
- 3.6 Keywords
- 3.7 Self-Assessment Test
- 3.8 Answers to check your progress
- 3.9 References/Suggested Readings

## **3.0 LEARNING OBJECTIVES**

After going through this lesson, students will be able to:

- Understand the concept of continuous random variable and Normal distribution
- Appreciate the usefulness of normal distribution in decision-making
- Identify situations where normal probability distribution can be applied.



### 3.1 INTRODUCTION

We have learnt that a probability distribution is basically a convenient representation of the different values a random variable may take, together with their respective probabilities of occurrence. In the last lesson, we have examined situations involving discrete random variables and the resulting discrete probability distributions. Consider the following random variables that we have taken up in the last lesson:

1. Number of Successes ( $X_1$ ) in a Bernoulli's Process
2. Number of Successes ( $X_2$ ) in a Poisson Process

In the first case, Binomial random variable  $X_1$  could take only finite number of integer values;  $0, 1, 2, \dots, n$ ; whereas in the second case, Poisson random variable  $X_2$  could take an infinite number of integer value;  $0, 1, 2, 3, \dots$ . The random variables  $X_1$  and  $X_2$  are discrete, in the sense that they could be listed in a sequence, finite or infinite. In contrast to these, let us consider a situation, where the variable of interest may take any value within a given range. Suppose we are planning for measuring the variability of an automatic bottling process that fills  $\frac{1}{2}$ -liter ( $500 \text{ cm}^3$ ) bottles with cola. The variable, say  $X$ , indicating the deviation of the actual volume from the normal (average) volume can take any real value - positive or negative; integer or decimal. This type of random variable, which can take an infinite number of values in a given range, is called a **continuous random variable**, and the probability distribution of such a variable is called a **continuous probability distribution**. The concepts and assumption inherent in the treatment of such distributions are quite different from those used in the context of a discrete distribution. In the present lesson, after understanding the basic concepts of continuous distributions, we will discuss Normal distribution - an important continuous distribution that is applicable to many real-life processes.

#### 3.1.1 Continuous Probability Distribution

Consider our planning for measuring the variability of the automatic bottling process that fills  $\frac{1}{2}$ -liter ( $500 \text{ cm}^3$ ) bottles with cola. The random variable  $X$  indicates 'the deviation of the actual volume from the normal (average) volume.' Let us, for some time, measure our random variable  $X$  to the nearest one  $\text{cm}^3$ .

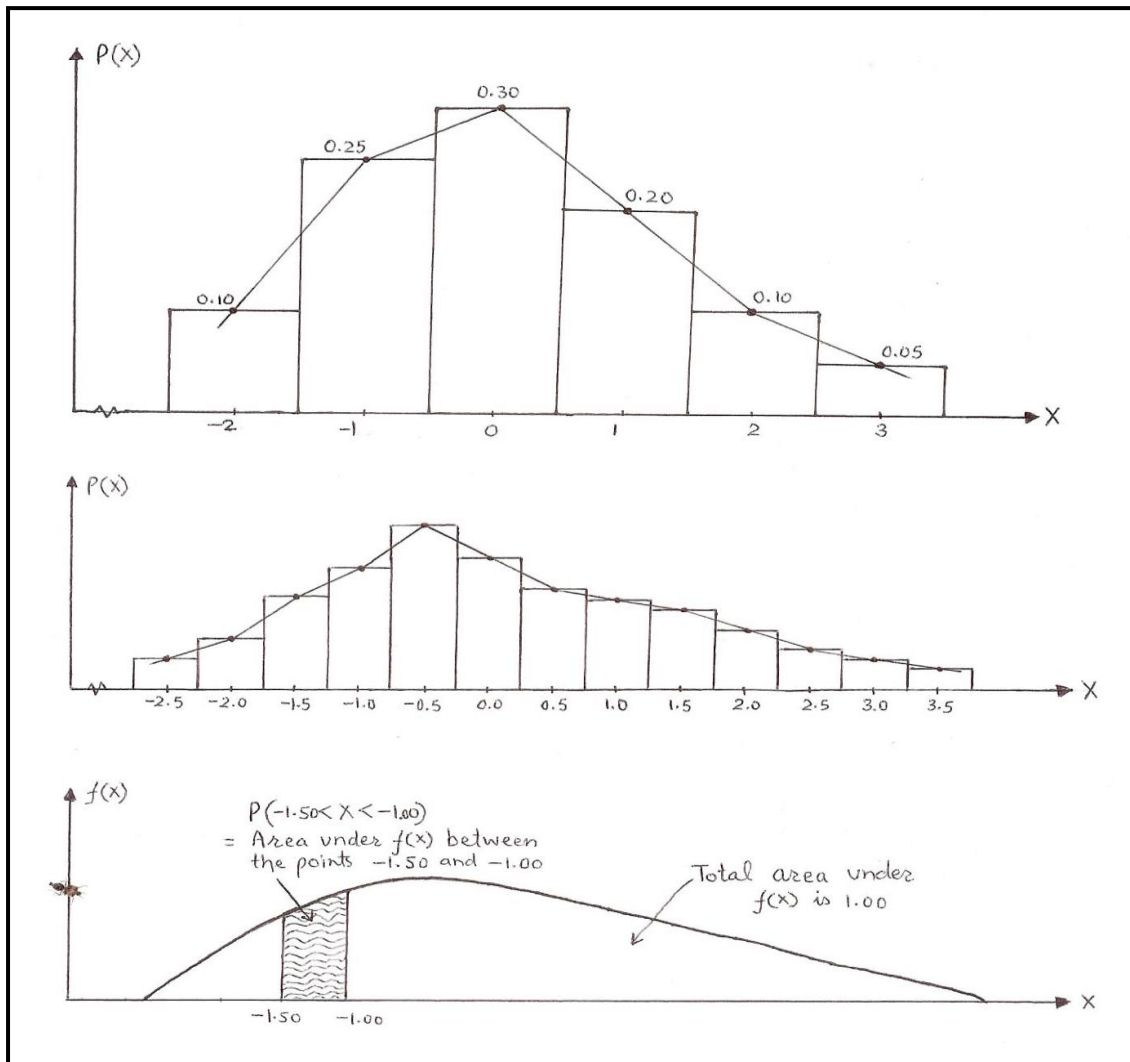


Figure 3-1 Histograms of the Distribution of X as Measurements is refined to Smaller and Smaller Intervals of Volume, and the Limiting Density Function  $f(x)$

Suppose Figure 8-1a represent the histogram of the probability distribution of X. The probability of each value of X is the area of the rectangle over the value. Since the rectangle will have the same base, the height of each rectangle is proportional to the probability. The probabilities also add to 1.00 as required for a probability distribution.

Volume is a continuous random variable; it can take on any value measured on an interval of numbers. Now let us imagine the process of refining the measurement scale of X to the nearest  $1/2 \text{ cm}^3$ , the nearest  $1/8 \text{ cm}^3$ ... and so on. Obviously, as the process of refining the measurement scale continues, the



number of rectangles in the histogram increases and the width of each rectangle decreases. The probability of each value is still measured by the area of the rectangle above it, and the total area of all rectangles remains 1.00. As we keep refining our measurement scale, the discrete distribution of  $X$  tends to a continuous probability distribution. The step like surface formed by the tops of the rectangles in the histogram tends to a smooth function. This function is denoted by  $f(x)$  and is called the **probability density function** of the continuous random variable  $X$ . The density function is the limit of the histograms as the number of rectangles approaches infinity and the width of each rectangle approaches zero. The density function of the limiting continuous variable  $X$  is shown in Figure 6-1 *i.e.* the values  $X$  can assume between the intervals  $-2.00$  to  $-3.00$  approaches infinity. The probability that  $X$  assumes a particular value (Say  $X = 1.5$ ) approaches zero. Probabilities are still measured as areas under the curve. The probability that deviation will be between  $-1.50$  and  $-1.00$  is the area under  $f(x)$  between the points  $x = -1.50$  and  $x = -1.00$ . Let us now make some formal definitions.

***A continuous random variable is a random variable that can take on any value in an interval of numbers.***

The probabilities associated with a continuous random variable  $X$  are determined by the **probability density function** of the random variable. The function, denoted by  $f(x)$ , has the following properties:

1.  $f(x) = 0$  for all  $x$
2. The probability that  $X$  will be between two numbers  $a$  and  $b$  is equal to the area under  $f(x)$  between  $a$  and  $b$ .

$$P(a < X < b) = \int_a^b f(x).dx$$

3. The total area under the entire curve of  $f(x)$  is equal to 1.00.

$$P(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f(x).dx = 1.00$$

When the sample space is continuous, the probability of any single given value is zero. For a continuous random variable, therefore, the probability of occurrence of any given value is zero. We see this from property 2, noting that the area under a curve between a point and itself is the area of a line, which is zero. ***For a continuous random variable, non-zero probabilities are associated only with intervals of numbers.***





We define the cumulative distribution function  $F(x)$  for a continuous random variable similarly to the way we defined it for a discrete random variable:  $F(x)$  is the probability that  $X$  is less than (or equal to)  $x$ .

Thus, the **cumulative distribution function** of a continuous random variable:

$F(x) = P(X \leq x)$  = area under  $f(x)$  between the *smallest* possible value of  $X$  (often  $-\infty$ ) and point  $x$

$$= \int_{-\infty}^x f(x).dx$$

The cumulative distribution function  $F(x)$  is a smooth, non-decreasing function that increases from 0 to 1.00.

The expected value of a continuous random variable  $X$ , denoted by  $E(X)$ , and its variance, denoted by  $V(X)$ , require the use of calculus for their computation. Thus

$$E(X) = \int_{-\infty}^{\infty} x.f(x).dx$$

$$V(X) = \int_{-\infty}^{\infty} [x - E(x)]^2 .f(x).dx$$

### 3.1.2 The Normal Distribution

*The Normal Distribution is the most versatile of all the continuous probability distributions. It is being widely used in all data-based research in the field of agriculture, trade, business and industry. It is found to be useful in characterizing uncertainties in many real-life processes, in statistical inferences, and in approximating other probability distributions.*

A large number of random variables occurring in practice can be approximated to the normal distribution.

***A random variable that is affected by many independent causes, and the effect of each cause is not overwhelmingly large compared to other effects, closely follow a normal distribution.***

The lengths of pins made by an automatic machine; the times taken by an assembly worker to complete the assigned task repeatedly; the weights of baseballs; the tensile strengths of a batch of bolts; and the volumes of cola in a particular brand of canned cola - are good examples of normally distributed random variables. All of these are affected by several independent causes where the effect of each cause



is small. This knowledge helps us in calculating the probabilities of different events in varied situations, which in turn is useful for decision-making.

In many real life situations, we face the problem of making statistical inferences about processes based on limited data. Limited data is basically a sample from the full body of data on the process. Irrespective of how the full body of data is distributed, it has been found that the Normal Distribution can be used to characterize the sampling distribution of many of the sample statistics. (we will see it in next few lessons). This helps considerably in Statistical Inferences.

Finally, the Normal Distribution can be used to approximate certain probability distributions. This helps considerably in simplifying the probability calculations.

### Probability Density Function

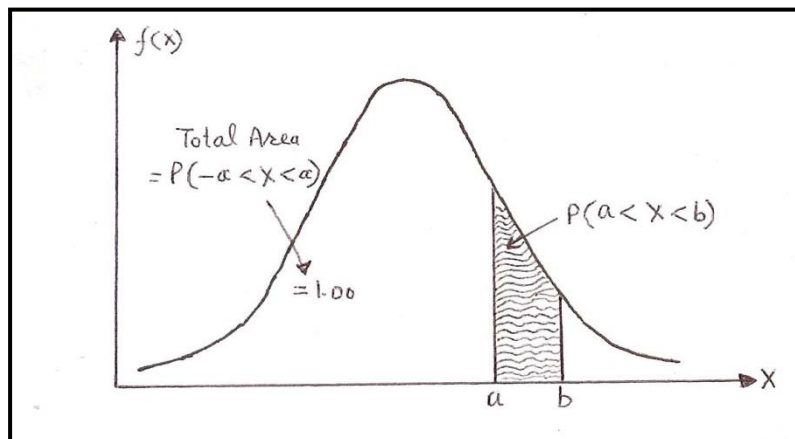
If  $X$  is normally distributed with mean  $\mu$  and variance  $\sigma^2$ , we write  $X \sim N(\mu, \sigma^2)$  and the probability density function  $f(x)$  is given by the formula:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\alpha < x < +\alpha$$

In the equation  $e$  is the base of natural logarithm, equal to 2.71828.... By substituting desired values for  $\mu$  and  $\sigma$ , we can get any desired density function. For example, a distribution with mean 80 and standard deviation 5 will have the density function.

$$f(x) = \frac{1}{\sqrt{2\pi}5} e^{-\frac{1}{2}\left(\frac{x-100}{5}\right)^2} \quad -\alpha < x < +\alpha$$

This function when plotted (see Figure 6-2) will give the famous bell-shaped mesokurtic normal curve.





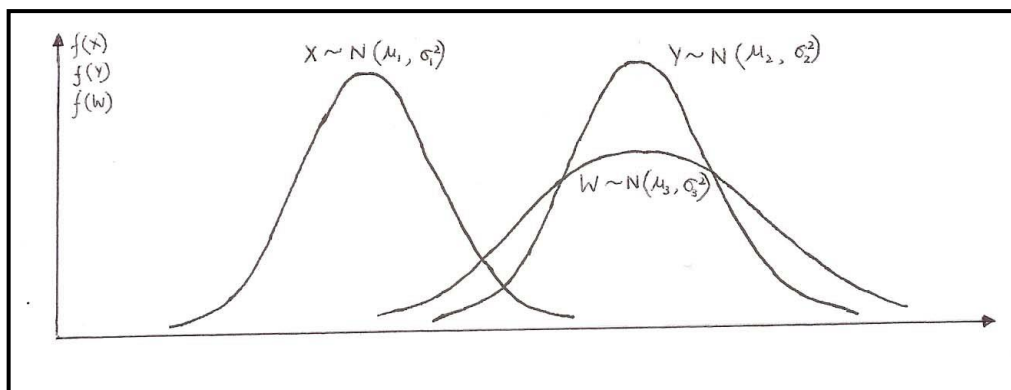
**Figure 3-2 A Normal Distribution with  $\mu = 80$  and  $\sigma = 5$**

Many mathematicians have worked on the mathematics behind the normal distribution and have made many independent discoveries. In the initial stages, the normal distribution was developed by Abraham De Moivre (1667-1754). His work was later taken up by Pierre S Laplace (1749-1827). But the discovery of equation for the normal density function is attributed to Carl Friedrich Gauss (1777-1855), who did much work with the formula. In science books, this distribution is often called the Gaussian distribution.

We will now examine the properties of the Normal distribution.

### ***Properties of Normal Distribution***

1. The normal curve is not a single curve representing only one continuous distribution. Obviously, it represents a family of normal curves; since for each different value of  $\mu$  and  $\sigma$ , there is a specific normal curve different in its positioning on the  $X$ -axis and the extent of spread around the mean. Figure 8-3 shows three different normal distributions – with different shapes and positions.



***Figure 3-3 Three Different Normal Distribution***

2. The normal curve is bell-shaped and perfectly symmetric about its mean. As a result 50% of the area lies to the right of mean and balance 50% to the left of mean. Perfect symmetry, obviously, implies that mean, median and mode coincide in case of a normal distribution. The normal curve gradually tapers off in height as it moves in either direction away from the mean, and gets closer to the  $X$ -axis.
3. The normal curve has a (relative) kurtosis of 0, which means it has average peakedness and is mesokurtic.



4. Theoretically, the normal curve never touches the horizontal axis and extends to infinity on both sides. That is the curve is asymptotic to  $X$ -axis.
5. If several independent random variables are normally distributed, then their sum will also be normally distributed. The mean of the sum will be the sum of all the individual means, and by virtue of the independence, the variance of the sum will be the sum of all the individual variances.

If  $X_1, X_2, \dots, X_n$  are independent normal variables, then their sum  $S$  will also be a normal variable with

$$E(S) = E(X_1) + E(X_2) + \dots + E(X_n)$$

$$\text{and } V(S) = V(X_1) + V(X_2) + \dots + V(X_n)$$

6. If a normal variable  $X$  undergoes a linear change in scale such as  $Y = aX + b$ , where  $a$  and  $b$  are constants and  $a \neq 0$ ; the resultant  $Y$  variable will also be normally distributed with mean  $= a E(X) + b$  and Variance  $= a^2 V(X)$

We can combine the above two properties.

If  $X_1, X_2, \dots, X_n$  are independent random variables that are normally distributed, then the random variable  $Q$  defined as

$Q = a_1X_1 + a_2X_2 + \dots + a_nX_n + b$  will also be normally distributed with

$$E(Q) = a_1E(X_1) + a_2E(X_2) + \dots + a_nE(X_n) + b$$

$$\text{and } V(Q) = a_1^2 V(X_1) + a_2^2 V(X_2) + \dots + a_n^2 V(X_n)$$

Let us see the application of this result with the help of an example.

### **Example 3-1**

A cost accountant needs to forecast the unit cost of a product for the next year. He notes that each unit of the product requires 8 labor hours and 5 kg of raw material. In addition, each unit of the product is assigned an overhead cost of Rs 200. He estimates that the cost of a labor hour next year will be normally distributed with an expected value of Rs 45 and a standard deviation of Rs 2; the cost of raw material will be normally distributed with an expected value of Rs 60 and a standard deviation of Rs 3. Find the distribution of the unit cost of the product. Find its expected value and variance.

**Solution:** Since the cost of labor  $L$  may not influence the cost of raw material  $M$ , we can assume that the two are independent. This makes the unit cost of the product  $Q$  a random variable. So if

$$L \sim N(45, 2^2) \quad \text{and} \quad M \sim N(60, 3^2)$$



Then,  $Q = 8L + 5M + 200$  will follow normal distribution with

$$\begin{aligned}\text{Mean} &= E(Q) = 8E(L) + 5E(M) + 200 \\ &= 8(45) + 5(60) + 200 \\ &= 950\end{aligned}$$

$$\begin{aligned}\text{Variance} &= V(Q) = 8^2V(L) + 5^2V(M) \\ &= 80(4) + 25(9) \\ &= 625\end{aligned}$$

So  $Q \sim N(950, 25^2)$

7. Same important area relationships under normal curve are

Area between  $\mu - 1\sigma$  and  $\mu + 1\sigma$  is about 0.6826

Area between  $\mu - 2\sigma$  and  $\mu + 2\sigma$  is about 0.9544

Area between  $\mu - 3\sigma$  and  $\mu + 3\sigma$  is about 0.9974

Area between  $\mu - 1.96\sigma$  and  $\mu + 1.96\sigma$  is 0.95

Area between  $\mu - 2.58\sigma$  and  $\mu + 2.58\sigma$  is 0.99

### Importance of Normal Distribution

The normal distribution has long occupied a central place in the theory of statistics. Its importance will be clear from the following points:

1. It has the remarkable property stated in the Central Limit Theorem. According to this theorem as the sample size  $n$  increases, the distribution of mean,  $\bar{X}$  of a random sample taken from any population approaches to a normal distribution (with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ ). Thus, if samples of large size,  $n$  are drawn from a population that is not normally distributed, nevertheless the successive sample means will form themselves a distribution that is approximately normal. The Central Limit Theorem also applies to the distribution of Median and standard deviation but not range.
2. As  $n$  becomes large the normal distribution serves as a good approximation of many discrete distributions such as Binomial or Poisson model whenever the exact discrete probability is difficult to obtain or impossible to calculate accurately.
3. In theoretical statistics many problems can be solved only under the assumptions of a normal population.



4. The normal distribution has numerous mathematical properties which make it popular and comparatively easy to manipulate.
5. The normal distribution is used extensively in statistical quality control in on industry in setting up of control limits.

### 3.2 THE STANDARD NORMAL DISTRIBUTION

*There are infinitely many possible normal random variables and the resulting normal curves for different values of  $\mu$  and  $\sigma^2$ . So the range probability  $P(a < X < b)$  will be different for different normal curves. We can make use of integral calculus to compute the required range probability*

$$P(a < X < b) = \int_a^b f(x).dx$$

It may be appreciated that we can simplify this process of computing range probabilities to a great extent by tabulating the range probabilities. Since it is not practicable and indeed impossible to have separate probability tables for each of the infinitely many possible normal curves, we select one normal curve to serve as a **standard**. Probabilities associated with the range of values of this standard normal random variable are tabulated. A special transformation then allows us to apply the tabulated probabilities to *any* normal random variable. The standard normal random variable is denoted by a special name,  $Z$  (rather than the general name  $X$  we use for other random variables).

***We define the standard normal random variable  $Z$  as the normal random variable with mean = 0 and standard deviation = 1. We say***

$$Z \sim N(0, 1^2)$$

#### 3.2.1 Standard Area Tables

The probabilities associated with standard normal distribution are tabulated in two ways – say Type I and Type II tables, as shown in Figure 6-4. Type I Tables give the area between  $\mu = 0$  and any other  $z$  value, as shown by vertical hatched area in Figure 8-4a. The hatched area shown in figure is  $P(0 < Z < z)$ .

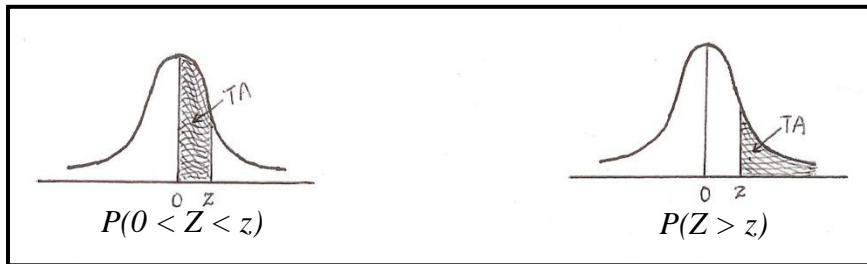


Figure 6-4 Standard Area Tables

Type II Tables give the area towards the tail-end of the standard normal curve beyond the ordinate at any particular  $z$  value. The hatched area shown in Figure 8-4b is  $P(Z > z)$ .

As the normal curve is perfectly symmetrical, the areas given by Type 1 Tables when subtracted from 0.5 will provide the same areas as given by Type II Tables and vice-versa.

$$i.e. \quad P(0 < Z < z) = 0.5 - P(Z > z).$$

### 3.2.2 Finding Probabilities of the Standard Normal Distribution

We will now illustrate the use of standard normal area tables for calculating the range probabilities. Probability of intervals is areas under the density curve  $f(z)$  over the intervals in question.

#### Example 3-2

Find the probability that the value of the standard normal random variable will be...

- |                        |                      |
|------------------------|----------------------|
| (a) between 0 and 1.74 | (b) less than -1.47  |
| (c) between 1.3 and 2  | (d) between -1 and 2 |

**Solution:** (a)  $P(Z \text{ is between } 0 \text{ and } 1.74)$

That is, we want  $P(0 < Z < 1.74)$ . In Figure 8-4a, substitute 1.74 for the point  $z$  on the graph. We are looking for the table area in the row labeled 1.7 and the column labeled 0.04. In the table, we find the probability 0.4591. Thus

$$P(0 < Z < 1.74) = 0.4591$$

(b)  $P(Z \text{ is less than } -1.47)$

That is, we want  $P(Z < -1.47)$ . By the symmetry of the normal curve, the area to the left of -1.47 is exactly equal to the area to the right of 1.47. We find

$$P(Z < -1.47) = P(Z > 1.47)$$



$$= 0.5000 - 0.4292$$

$$= 0.0808$$

(c)  $P(Z \text{ is between } 1.3 \text{ and } 2)$

That is, we want  $P(1.3 < Z < 2)$ . The required probability is the area under the curve between the two points 1.3 and 2. The table gives us the area under the curve between 0 and 1.3, and the area under the curve between 0 and 2. Areas are additive; therefore,

$$\begin{aligned} P(1.30 < Z < 2) &= \text{TA}(\text{for } 2.00) - \text{TA}(\text{for } 1.30) \\ &= P(0 < Z < 2) - P(0 < Z < 1.3) \\ &= 0.4772 - 0.4032 \\ &= 0.0740 \end{aligned}$$

(d)  $P(Z \text{ is between } -1 \text{ and } 2)$

That is, we want  $P(-1 < Z < 2)$ . The required probability is the area under the curve between the two points -1 and 2. The table gives us the area under the curve between 0 and 1, and the area under the curve between 0 and 2. Areas are additive; therefore,

$$\begin{aligned} P(-1 < Z < 2) &= P(-1 < Z < 0) + P(0 < Z < 2) \\ &= P(0 < Z < 1) + 0.4772 \\ &= 0.3413 + 0.4772 \\ &= 0.8185 \end{aligned}$$

In cases, where we need probabilities based on values with greater than second-decimal accuracy, we may use a linear interpolation between two probabilities obtained from the table.

### **Example 3-3**

Find  $P(0 \leq Z \leq 1.645)$

**Solution:**  $P(0 \leq Z \leq 1.645)$  is found as the midpoint between the two probabilities  $P(0 \leq Z \leq 1.64)$  and  $P(0 \leq Z \leq 1.65)$ . So

$$P(0 \leq Z \leq 1.645) = \frac{1}{2}[P(0 \leq Z \leq 1.64) + P(0 \leq Z \leq 1.65)]$$





$$\begin{aligned} &= \frac{1}{2}[0.4495 + 0.4505] \\ &= 0.45 \end{aligned}$$

### 3.2.3 Finding Values of Z Given a Probability

In many situations, instead of finding the probability that a standard normal random variable will be within a given interval; we may be interested in the reverse: finding an interval with a given probability. Consider the following examples.

#### Example 3-4

Find a value  $z$  of the standard normal random variable such that the probability that the random variable will have a value between 0 and  $z$  is 0.40.

**Solution:** We look inside the table for the value closest to 0.40. The closest value we find to 0.40 is the table area 0.3997. This value corresponds to 1.28 (row 1.2 and column .08).

So for  $P(0 < Z < z) = 0.40$ ;  $z = 1.28$

#### Example 3-5

Find the value of the standard normal random variable that cuts off an area of 0.90 to its left.

**Solution:** Since the area to the left of the given point  $z$  is greater than 0.50,  $z$  must be on the right side of 0. Furthermore, the area to the left of 0 all the way to  $-\infty$  is equal to 0.50. Therefore,  $TA = 0.90 - 0.50 = 0.40$ . We need to find the point  $z$  such that  $TA = 0.40$ .

We find that for  $TA = 0.40$ ;  $z = 1.28$ .

Thus  $z = 1.28$  cuts off an area of 0.90 to the left of standard normal curve.

#### Example 3-6

Find a 0.99 probability interval, symmetric about 0, for the standard normal random variable.

**Solution:** The required area between the two  $z$  values that are equidistant from 0 on either side is 0.99. Therefore, the area under the curve between 0 and the positive  $z$  value is  $TA = 0.99/2 = 0.495$ . We now look in our normal probability table for the area closest to 0.495. The area 0.495 lies exactly between the two areas 0.4949 and 0.4951, corresponding to  $z = 2.57$  and  $z = 2.58$ . Therefore, a simple linear interpolation between the two values gives us  $z = 2.575$ . The answer, therefore, is  $z = \pm 2.575$ .

So for  $P(-z < Z < z) = 0.99$ ;  $z = 2.575$



### 3.3 THE TRANSFORMATION OF NORMAL RANDOM VARIABLES

The importance of the standard normal distribution derives from the fact that any normal random variable may be transformed to the standard normal random variable. If we want to transform  $X$ , where  $X \sim N(\mu, \sigma^2)$ , into the standard normal random variable  $Z \sim N(0, 1^2)$ , we can do this as follows:

$$Z = \frac{X - \mu}{\sigma}$$

We move the distribution from its center of  $\mu$  to a center of 0. This is done by subtracting  $\mu$  from all the values of  $X$ . Thus, we shift the distribution  $\mu$  units back so that its new center is 0. To make the standard deviation of the distribution equal to 1, we divide the random variable by its standard deviation  $\sigma$ . The area under the curve adjusts so that the total remains the same. All probabilities (areas under the curve) adjust accordingly. Thus, the transformation from  $X$  to  $Z$  is achieved by first subtracting  $\mu$  from  $X$  and then dividing the result by  $\sigma$ .

#### Example 3-7

If  $X \sim N(50, 8^2)$ , find the probability that the value of the random variable  $X$  will be greater than 60

**Solution:**

$$\begin{aligned} P(X > 60) &= P\left(\frac{X - \mu}{\sigma} > \frac{60 - \mu}{10}\right) \\ &= P\left(Z > \frac{60 - 50}{10}\right) \\ &= P(Z > 1) \\ &= P(Z > 0) - P(0 < Z < 1) \\ &= 0.5000 - 0.3413 \\ &= 0.1587 \end{aligned}$$

#### Example 3-8

The weekly wage of 2000 workmen is normally distribution with mean wage of Rs 70 and wage standard deviation of Rs 5. Estimate the number of workers whose weekly wages are

- |     |                         |     |                         |     |
|-----|-------------------------|-----|-------------------------|-----|
| (a) | between Rs 70 and Rs 71 | (b) | between Rs 69 and Rs 73 | (c) |
|     | more than Rs 72         | (d) | less than Rs 65         |     |

**Solution:** Let  $X$  be the weekly wage in Rs, then



$$X \sim N(70, 5^2)$$

(a) The required probability to be calculated is  $P(70 < X < 71)$

$$\begin{aligned} \text{So } P(70 < X < 71) &= P\left(\frac{70 - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{71 - \mu}{\sigma}\right) \\ &= P\left(\frac{70 - 70}{5} < Z < \frac{71 - 70}{5}\right) \\ &= P(0 < Z < 0.2) \\ &= 0.0793 \end{aligned}$$

So the number of workers whose weekly wages are between Rs 70 and Rs 71  
 $= 2000 \times 0.0793$   
 $= 159$

(b) The required probability to be calculated is  $P(69 < X < 73)$

$$\begin{aligned} \text{So } P(69 < X < 73) &= P\left(\frac{69 - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{73 - \mu}{\sigma}\right) \\ &= P\left(\frac{69 - 70}{5} < Z < \frac{73 - 70}{5}\right) \\ &= P(-0.2 < Z < 0.6) \\ &= P(-0.2 < Z < 0) + P(0 < Z < 0.6) \\ &= P(0 < Z < 0.2) + P(0 < Z < 0.6) \\ &= 0.0793 + 0.2257 \\ &= 0.3050 \end{aligned}$$

So the number of workers whose weekly wages are between Rs 69 and Rs 73  
 $= 2000 \times 0.3050$   
 $= 68$

(c) The required probability to be calculated is  $P(X > 72)$

$$\begin{aligned} \text{So } P(X > 72) &= P\left(\frac{X - \mu}{\sigma} > \frac{72 - \mu}{\sigma}\right) \\ &= P\left(Z > \frac{72 - 70}{5}\right) \\ &= P(Z > 0.4) \\ &= 0.5 - P(0 < Z < 0.4) \end{aligned}$$



$$= 0.5 - 0.1554$$

$$= 0.3446$$

So the number of workers whose weekly wages are more than Rs 72

$$= 2000 \times 0.3446$$

$$= 689$$

(d) The required probability to be calculated is  $P(X < 65)$

$$\begin{aligned} \text{So } P(X < 65) &= P\left(\frac{X - \mu}{\sigma} < \frac{65 - \mu}{\sigma}\right) \\ &= P\left(Z < \frac{65 - 70}{5}\right) \\ &= P(Z < -1.0) \\ &= P(Z > 1.0) \\ &= P(Z > 0) - P(0 < Z < 1.0) \\ &= 0.5 - 0.3413 \\ &= 0.1567 \end{aligned}$$

So the number of workers whose weekly wages are less than Rs 65

$$= 2000 \times 0.1567$$

$$= 313$$

### The Inverse Transformation

The transformation  $Z = \frac{X - \mu}{\sigma}$  takes us from a random variable  $X$  with mean  $\mu$ , and standard deviation  $\sigma$  to the standard normal random variable. We also have an opposite, or inverse, transformation, which takes us from the standard normal random variable  $Z$  to the random variable  $X$  with mean  $\mu$  and standard deviation  $\sigma$ . The inverse transformation is given as

$$X = \mu + Z\sigma$$

We use the inverse transformation when we want to get from a given probability, the value or values of a normal random variable  $X$ .

### Example 3-9

The amount of fuel consumed by the engines of a jetliner on a flight between two cities is a normally distributed random variable  $X$  with mean  $\mu = 5.7$  tons and standard derivation  $\sigma = 0.5$  tons. Carrying



too much fuel is inefficient as it slows the plans. If, however, too little fuel is loaded on the plane, an emergency landing may be necessary. What should be the amount of fuel to load so that there is 0.99 probability that the plane will arrive at its destination without emergency landing?

**Solution:** Given that  $X \sim N(5.7, 0.5^2)$ ,

We have to find the value  $x$  such that

$$P(X < x) = 0.99$$

$$\text{or } P\left(\frac{X - \mu}{\sigma} < z\right) = 0.99$$

$$\begin{aligned} \text{or } P(Z < z) &= 0.99 \\ &= 0.5 + 0.49 \\ &= 0.5 + P(0 < Z < z) \end{aligned}$$

From the table, value of  $z$  is 2.33

$$\begin{aligned} \text{So } x &= \mu + z\sigma \\ x &= 5.7 + 2.33 \times 0.5 \\ x &= 6.865 \end{aligned}$$

Therefore, the plane should be loaded with 6.865 tons of fuel to give 0.99 probability that the fuel will last throughout the flight.

### Example 3-10

Monthly sale of beer at a bar is believed to be approximately normally distributed with mean 2450 units and standard 400 units. To determine the level of orders and stock, the management wants to find two values symmetrically on either side of mean, such that the probability that sales of beer during the month will be between the two values is

$$(a) \quad 0.95 \qquad (b) \quad 0.99$$

Find the required values.

**Solution:** Let  $X$  be the monthly sale of beer, then

$$X \sim N(2450, 400^2),$$

(a) We have to find the values  $x_1$  and  $x_2$  such that

$$P(x_1 < X < x_2) = 0.95$$



$$\text{or} \quad P\left(\frac{x_1 - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{x_2 - \mu}{\sigma}\right) = 0.95$$

$$\text{or} \quad P(z_1 < Z < z_2) = 0.95$$

$$\text{We know} \quad P(-1.96 < Z < 1.96) = 0.95$$

$$\text{So} \quad z_1 = -1.96 \quad \text{and} \quad z_2 = 1.96$$

Using the inverse transformation,

$$x_1 = \mu + z_1\sigma \quad \text{and} \quad x_2 = \mu + z_2\sigma$$

$$x_1 = 2450 + (-1.96)400 \quad x_2 = 2450 + (1.96)400$$

$$x_1 = 1666 \quad x_2 = 3234$$

Therefore, the management may be 95% sure that sales in any given month will be between 1666 and 3234 units.

(b) We have to find the values  $x_1$  and  $x_2$  such that

$$P(x_1 < X < x_2) = 0.99$$

$$\text{or} \quad P\left(\frac{x_1 - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{x_2 - \mu}{\sigma}\right) = 0.99$$

$$\text{or} \quad P(z_1 < Z < z_2) = 0.99$$

$$\text{We know} \quad P(-2.58 < Z < 2.58) = 0.99$$

$$\text{So} \quad z_1 = -2.58 \quad \text{and} \quad z_2 = 2.58$$

Using the inverse transformation,

$$x_1 = \mu + z_1\sigma \quad \text{and} \quad x_2 = \mu + z_2\sigma$$

$$x_1 = 2450 + (-2.58)400 \quad x_2 = 2450 + (2.58)400$$

$$x_1 = 1418 \quad x_2 = 3482$$

Therefore, the management may be 99% sure that sales in any given month will be between 1418 and 3482 units.

We can summarize the procedure of obtaining values of a normal random variable, given a probability, as:

- draw a picture of the normal distribution in question and the standard normal distribution
- in the picture, shade in the area corresponding to the probability



- use the table to find the  $z$  value (or values) that gives the required probability
- use the transformation from  $Z$  to  $X$  to get the appropriate value (or values) of the original normal random variable

### 3.4 CHECK YOUR PROGRESS

1. The probabilities associated with a continuous random variable  $X$  are determined by the ..... of the random variable.
2. For a continuous random variable, non-zero probabilities are associated only with ..... of numbers.
3. In science books, normal distribution is often called the .....
4. The normal curve has a (relative) kurtosis of 0, which means it has ..... and is mesokurtic.
5. We define the standard normal random variable  $Z$  as the ..... random variable with mean = 0 and standard deviation = 1.

### 3.5 SUMMARY

A continuous random variable is a random variable that can take on any value in an interval of numbers. The Normal Distribution is the most versatile of all the continuous probability distributions. It is being widely used in all data-based research in the field of agriculture, trade, business and industry. It is found to be useful in characterizing uncertainties in many real-life processes, in statistical inferences, and in approximating other probability distributions. A large number of random variables occurring in practice can be approximated to the normal distribution. A random variable that is affected by many independent causes, and the effect of each cause is not overwhelmingly large compared to other effects, closely follow a normal distribution. It may be appreciated that we can simplify this process of computing range probabilities to a great extent by tabulating the range probabilities. Since it is not practicable and indeed impossible to have separate probability tables for each of the infinitely many possible normal curves, we select one normal curve to serve as a standard. Probabilities associated with the range of values of this standard normal random variable are tabulated. The importance of the standard normal distribution derives from the fact that any normal random variable may be transformed to the standard normal random variable.



### 3.6 KEYWORDS

**Continuous random variable:** The variable, say  $X$ , indicating the deviation of the actual volume from the normal (average) volume can take any real value - positive or negative; integer or decimal. This type of random variable, which can take an infinite number of values in a given range, is called a continuous random variable.

**Continuous probability distribution:** The probability distribution of continuous random variable is called a continuous probability distribution.

**Standard Normal Distribution:** It defines the standard normal random variable  $Z$  as the normal random variable with mean = 0 and standard deviation = 1.

### 3.7 SELF-ASSESSMENT TEST

1. Define continuous probability distribution. State the properties of the probability density function of a continuous random variable.
2. (a) Define normal random variable. State the probability density function of a normal random variable.  
(b) List down important properties of a normal curve.
3. Discuss the role of normal distribution in statistical theory.
4. What do you mean by standard normal variable? Bring out the need for having a standard normal curve.
5. Find the probability that a standard normal variable will have a value  
(a) less than  $-8$                       (b) between  $-0.01$  and  $0.05$
6. A sensitive measuring device is calibrated so that errors in the measurements it provides are normally distributed with mean 0 and variance 1.00. Find the probability that a given error will be between  $-2$  and  $2$ .
7. The deviation of a magnetic needle from the magnetic pole in a certain area in northern Canada is a normally distributed random variable with mean 0 and standard deviation 1.00. What is the probability that the absolute value of the deviation from the north pole at a given moment will be more than 2.4?
8. Find two values of the standard normal random variable,  $z$  and  $-z$ , such that  
(a) the two corresponding "tail areas" of the distribution add to 0.01.





- (b) each tail have an area of 0.05
9. Let  $X$  be a normally distributed random variable with mean  $\mu = 16$  and standard deviation  $\sigma = 3$ . Find
- (a)  $P(8 < X < 18)$  (b)  $P(16 < X < 18)$  (c)  $P(X > 14)$
10. For a normally distributed random variable with mean -44 and standard deviation 16, find the probability that the value of the random variable will be
- (a) above 0 (b) -8 (c) below 0
11. A normal random variable has mean 0 and standard deviation 4. Find the probability that the random variable will be...
- (a) above 2.5 (b) between 2 and 3 (c) below 1
12. The time it takes an international telephone operator to place an overseas phone call is normally distributed with mean 45 seconds and standard deviation 8 seconds.
- (a) What is the probability that my call will go through in less than 1 minute?
- (b) What is the probability that my call will get through in less than 40 seconds?
- (c) What is the probability that I will have to wait more than 70 seconds for my call to go through?
13. The number of votes cast in favor of a controversial proposition is believed to be approximately normally distributed with mean 8,000 and standard deviation 1,000. The proposition needs at least 9,322 votes in order to pass. What is the probability that the proposition will pass? (Assume numbers are on a continuous scale.)
14. A manufacturing company regularly consumes a special type of glue purchased from a foreign supplier. From past experience, the materials manager notes that the company's demand for glue during the uncertain lead-time is normally distributed with a mean of 187.6 gallons and a standard deviation of 12.4 gallons. The company follows a policy of placing the order when the glue stock falls to a predetermined value, called "re-order point". If the demand during lead-time exceeds the reorder level, the glue would go 'stock-out' and production process would have to stop.
- (a) If the re-order point is kept at 187.6 gallons, what is the probability that a stock-out condition would occur?
- (b) If the reorder point is kept at 200 gallons, what is the probability that a stock-out condition would occur?



- (c) If the company wants to be 95% confident that the stock-out condition will not occur, what should be the reorder point? The reorder point minus the mean demand during lead-time is known as the "safety stock." What is the safety stock in this case?
- (d) If the company wants to be 99% confident that the stock-out condition will not occur, what should be the reorder point? What is the safety stock in this case?
15. If  $X$  is a normally distributed random variable with mean 125 and standard deviation 44, find a value  $x$  such that the probability that  $X$  will be less than  $x$  is 0.66.
16. For a normal random variable with mean 8.5 and standard deviation 0.4, find a point of the distribution such that there is a 0.95 probability that the value of the random variable will be above it.
17. For a normal random variable with mean 29,500 and standard deviation 48, find a point of the distribution such that the probability that the random variable will exceed this value is
- (a) 0.03                      (b) 0.25
18. Find two values of the normal random variable with mean 80 and standard deviation 5 lying symmetrically on either side of the mean and covering an area of 0.98 between them.
19. For  $X \sim N(32, 7^2)$ , find two values  $x_1$  and  $x_2$ , symmetrically lying on each side of the mean, with
- (a)  $P(x_1 < X < x_2) = 0.99$                       (b)  $P(x_1 < X < x_2) = 0.95$
20. The results of a given selection test exercise are summarized as
- (i) cleared with distinction = 8%
- (ii) cleared without distinction = 60%
- (iii) those who failed = 30%.
- A candidate gets failed if he/she obtains less than 40% marks, while one must obtain at least 75% marks to pass with distinction. Determine the mean and standard deviation of the distribution of marks, assuming the same to be normal.
21. The demand for gasoline at a service station is normally distributed with mean 27,009 gallons per day and standard deviation 4,530. Find two values that will give a symmetric 0.95 probability interval for the amount of gasoline demanded daily.
22. The percentage of protein in a certain brand of dog food is a normally distributed random variable with mean 11.2 % and standard deviation 0.6 %. The manufacturer would like to state on the



package that the product has a protein content of at least  $x_1$  % and no more than  $x_2$  %. He wants the statement to be true for 99% of the packages sold. Determine the values  $x_1$  and  $x_2$ .

### **3.8 ANSWERS TO CHECK YOUR PROGRESS**

1. Probability density function
2. Intervals
3. Gaussian distribution
4. Average peakedness
5. Normal

### **3.9 REFERENCES/SUGGESTED READINGS**

1. Statistics (Theory & Practice) by Dr. B.N. Gupta. Sahitya Bhawan Publishers and Distributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing Company Ltd., New Delhi.
3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill Publishing Company Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., New Delhi.
5. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.
6. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.
7. Statistics for Management by Richard I. Levin and David S. Rubin. Prentice Hall of India Pvt. Ltd., New Delhi.
8. Statistics for Business and Economics by Kohlar Heinz. Harper Collins., New York.



<b>Subject: Business Statistics-II</b>	
<b>Course Code: BCOM 402</b>	<b>Author: Dr. Pardeep Gupta</b>
<b>Lesson: 04</b>	<b>Vetter: Dr. B.S. Bodla</b>
<b>SAMPLING AND SAMPLING METHODS</b>	

## STRUCTURE

### 4.0 Learning Objectives

#### 4.1 Introduction

##### 4.1.1 Census Vs.Sampling method

##### 4.1.2 Definitions

#### 4.2 Probability samples vs. non-probability samples

##### 4.2.1 Probability sampling methods

##### 4.2.2 Non-probability sampling methods

##### 4.2.3 Determination of Sample size

#### 4.3 Sampling and Non-sampling errors

#### 4.4 Check your Progress

#### 4.5 Summary

#### 4.6 Keywords

#### 4.7 Self-Assessment Test

#### 4.8 Answers to check your progress

#### 4.9 References/Suggested readings

## 4.0 LEARNING OBJECTIVES

After going through this lesson, the students will be able to:

- Understand various terms associated with sampling
- Understand various methods of probability and non-probability Sampling
- Understanding of how to determine sample size.



## **4.1 INTRODUCTION**

Sampling is the procedure or process of selecting a sample from a population. A sampling can also be defined as the process of drawing a sample from a population and of compiling a suitable statistic from such a sample in order to estimate the parameter of the parent population and to test the significance of the statistic computed from such sample. When secondary data are not available for the problem under study, a decision may be taken to collect primary data by using any of the methods discussed in this lesson. The required information may be obtained by following either the census method or the sample method.

### **4.1.1 Census Vs. Sampling Method**

Sample is a part of the population from which it is selected. The process of selecting a sample is known as sampling. Thus, the sampling theory is a study of relationship that exists between the population and the samples drawn from the population. The complete enumeration, popularly known as census, may not be feasible either due to non-availability of time or because of high cost involved. Therefore, it becomes essential to draw inferences for the population on the basis of sample information. Thus, sampling helps us to get as much information as possible of the whole universe. The sampling also helps us in determining the reliability of the estimates. This can be done by drawing samples from the same parent population and comparing the results obtained from different samples.

In a survey of the entire population, data is collected from every elementary unit of the population. Suppose, one is studying the wage structure of the coal mining industry in the country, then one approach is to collect the data on wages of every worker in the coal industry. From this data, one can calculate the various characteristics of the population, such as average wage, the range and the variance, etc. This is referred as census survey. The advantages of the census approach are

- (i) every unit of the population is considered and the respective data on the various characteristics are compiled,
- (ii) the analysis made on the basis of census data is very accurate and reliable, and
- (iii) in one time studies of special importance, only census method is adopted in order to get accurate and reliable data. The data collected by this method becomes a data base for all future studies. This is one of the reasons why population data are collected once in a decade by the census method.



Although there are many advantages with the census method, the cost, effort and the time required to conduct census survey is very large, unless the population is very small, and in many cases it is so prohibitive that one rarely uses this method in surveys.

Sampling involves an examination of a small portion of the elementary units in a population. Although, a census operation gives a more reliable data, sampling method is more desired when

- (i) the population is very large, i.e., infinite and it would be impossible to conduct census surveys;
- (ii) when quick results are required it would be appropriate to conduct sample surveys rather than census surveys;
- (iii) in studies involving destruction of the elementary units under study, it would only be appropriate to go for sample testing. Items such as light bulbs and ammunition often must be destroyed as a part of testing process;
- (iv) cost of conducting surveys would be very prohibitive in census method, and therefore, it is advisable to carry out a sample survey, and lastly; and
- (v) some times accuracy may be lost because of the large size of the population. Sampling involves a small portion of the population and therefore, would involve very few people for conducting surveys and for data collection and compilation. This would not be so in the census method and the chances of committing errors would increase.

As the sampling involves less time and money, it would be possible to give attention to different characteristics of the elementary units. A sample using same money and time can produce a detailed study of lesser number of units. The process of sampling involves selecting a sample, collecting all relevant information, and finally drawing conclusions about the population from which the sample has been drawn.

#### **4.1.2 Definitions**

The surveys are concerned with the attributes of certain entities, such as business enterprises, human beings, etc. The attributes that are the object of the study are known as characteristics and the units possessing them are called the *elementary units*.

The aggregate of elementary units to which the conclusions of the study apply is termed as *population/universe*, and the units that form the basis of the sampling process are called sampling units. The sampling unit may be an elementary unit.



The sample is defined as an aggregate of sampling units actually chosen in obtaining a representative subset from which inferences about the population are drawn. *The frame*— a list or directory, defines all the sampling units in the universe to be covered. This frame is either constructed for the purpose of a particular survey or may consist of previously available description of the population; the latter is the commonly used method. For example, telephone directory can be used as a frame for conducting opinion surveys in a city or locality.

In order that, sampling results reflect the characteristics of the population, it is necessary that the sample selected for study should be

- (i) Truly representative, i.e., the selected sample truly represent the universe so that the results can be generalised;
- (ii) Adequate, i.e., the size of the sample or the sample size should be adequate enough to represent the various characteristics of the universe;
- (iii) Independent, i.e. the elementary units selected should be independent of one another and all units of the population should have the same chance of being selected in the sample; and lastly
- (iv) Homogeneous, i.e., there should not be any basic difference between the characteristics of the units in the sample and that of the population. This means that if two or more samples are drawn from the same population, the results should be more or less identical.

## 4.2 PROBABILITY SAMPLES VS. NON-PROBABILITY SAMPLES

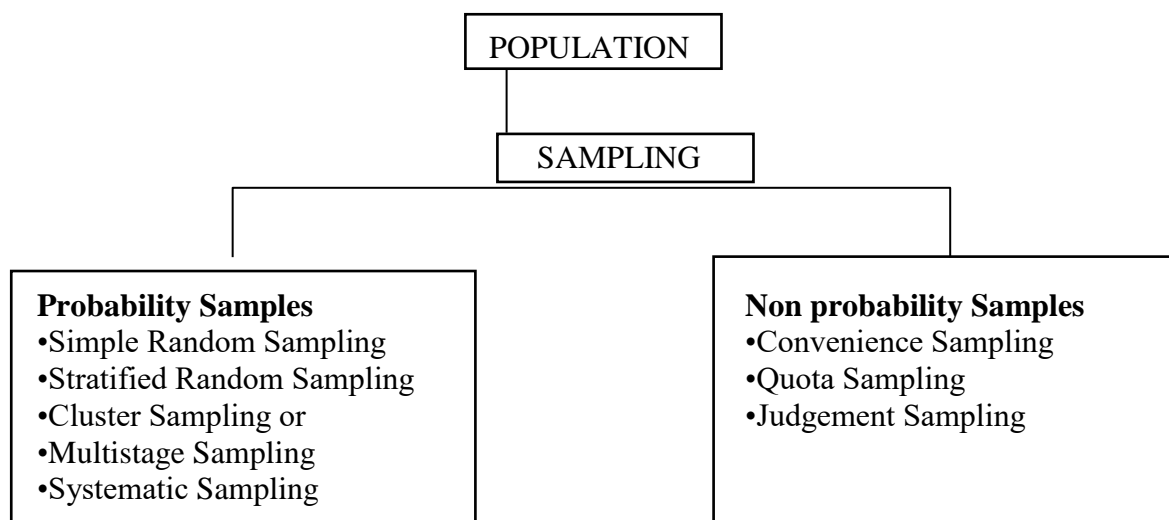
A probability sample is one for which the inclusion or exclusion of any individual element of the population depends upon the application of probability methods and not on a personal judgement. It is so designed and drawn that the probability of inclusion of an element is known. The essential feature of drawing such a sample is the randomness. As against the probability sample, we have a variety of other samples, termed as judgement samples, purposive samples, quota samples, etc. These samples have one common distinguishing feature: personal judgement rather than the random procedure to determine the composition of what is to be taken as a representative sample. The judgement affects the choice of the individual elements. All such samples are non-random, and no objective measure of precision may be attached to the results arrived at.

In a probability sampling, it is possible to estimate the error in the estimates and they can be minimized also. It is also possible to evaluate the relative efficiency of the various probability sampling designs.



Probability sampling does not depend upon the detailed information about population for its effectiveness. However, probability sampling requires a high level of skill and experience for its use. It also requires sufficient time and money to execute.

Non-probability sampling is a procedure of selecting a sample without the use of probability or randomisation. It is based on convenience, judgement, etc. The major difference between the two approaches is that it is possible to estimate the sampling variability in the case of probability sampling while it is not possible to estimate the same in the non-probability sampling. The classification of various probability and non-probability methods are shown in Fig. 4.1.



**Fig. 4.1 Classification of sampling schemes**

#### 4.2.1 Probability Sampling Methods

The various probability sampling methods are described as under:

##### (a) Simple Random Sampling Method

In simple random sampling, drawing of elements from the population is random and the choice of an element is made in such a way that every element has the same probability of being chosen. When the sample is so selected, every possible set of elements has the same chance of being drawn. With  $N$ , population size, fairly large, the number of such possible sets of size  $n$  is of course very large. This number is given by  ${}^N C_n$ . Of course, it is unnecessary in a specific case to compute the number of





possible sets of stated size that might be drawn from a given population, but the process of sample selection should be such that the probability of selection is the same for every such set.

The objective is to achieve randomness in drawing the individual elements of a sample for ensuring that all possible samples have the same chance of being selected. If we are to draw from a population containing  $N$  elementary units, the elementary unit also being a sampling unit, it is necessary that each of the  $N$  units should be individually numbered or otherwise distinctively designed. One of the approaches for drawing random sample of size  $n$  from a population of  $N$  units is to draw  $n$  cards from  $N$  cards which are numbered from 1 to  $N$  and mixed thoroughly. The sample size  $n$ , thus drawn, would constitute a simple random sample (SRS). Another popular method of selecting a random sample is by lottery method. In this method all the elements are named or numbered on a small slip of paper of identical shape and size. These slips are folded identically and mixed up well in a container. Number of slips of desired sample size is selected blindly from this container. Thus, the selection of elementary units depends purely on chance and no personal bias exists. We shall illustrate this method of selection of a sample with the following example: Suppose the warden of a student's hostel with 200 occupants wants to constitute a welfare committee with the members randomly selected. The lottery method of selecting these five members from a group of 200 would be first to prepare 200 slips of identical shape and size and write the name of each student on a slip. Fold these 200 slips identically and mix them well in a container. Then select five folded slips, from the container at random. The five students so selected would constitute a welfare committee of the hostel.

There are, however, some difficulties in these procedures. For, if  $N$  is large, the task becomes physically difficult. So it is desirable to use better methods for ensuring randomness. One such method is the use of random number tables.

### ***Use of random number tables***

If the  $N$  elements of a total population are numbered serially from 1 to  $N$ , a random sample may be most readily and reliably drawn by using a table of random numbers. Such tables enable us to select  $n$  numbers at random from the full list of serial numbers from 1 to  $N$ . In a random number table, digits in each column are in random order and so are the digits in each row. As the arrangement is random in all directions, it makes no difference where we begin in our selection of random numbers from such a table. However, the column arrangement is generally found more convenient for references.



Several random number tables are available for use. These numbers have been adequately tested for randomness. Among them, the most popular ones are:

- (i) Tippett's (1927) 10,400 sets of four-digit random numbers;
- (ii) Fisher and Yates (1938) table of random numbers with 1,500 sets of ten-digit random numbers; and
- (iii) Rand Corporation (1955) table of random numbers of 2,00,000 sets of five-digit random numbers.

Tippett's table of random numbers is most popularly used in practice. Given below are the first forty sets from Tippett's table as an illustration of the general appearance of random numbers:

2952	6641	3992	9792	7969	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2693
2670	7483	3408	2762	3563	1089	6913	7691
0560	5246	1112	6107	6008	8125	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

Tippett's numbers have been subjected to numerous tests and used in many investigations and their randomness has been well established for all practical purposes. An example to illustrate how Tippett's table of random numbers may be used is given below.

Suppose ten numbers from out of 0 and 80 are required. We start anywhere in the table and write down the numbers in pairs. The table can be read horizontally, vertically, diagonally or in any methodical way. Starting with the first and reading horizontally first we obtain 29, 52, 66, 41, 39, 92, 97, 92, 79, 69, 59, 11, 31, 70, 56, 24, 41, 67 and so on. Ignoring the numbers greater than 80, we obtain for one purpose ten random numbers, namely 29, 52, 66, 41, 39, 79, 69, 59, 11 and 31.

The sampling procedure described above is quite satisfactory for a small population. With a large population, the process of identification of numbers to each elementary sampling unit becomes very prohibitive with respect to both time and money. Moreover, the population is often geographically spread out or composed of clearly identified strata possessing unique characteristics. Whenever any of the above situations arise, alternative sampling schemes that are sophisticated combinations of simple random sampling provide significantly better results for the same expenditure and time. As a result, the simple random sampling method is not very frequently used in practice. However, the simple random sampling scheme is the basis of any other probabilistic sampling schemes.



### ***(b) Stratified Random Sampling Method***

In simple random sampling, the population to be sampled is treated as homogeneous and the individual elements are drawn at random from the whole universe. However, it is often possible and desirable to classify the population into distinctive classes or strata and then obtain a sample by drawing at random the specified number of sampling units from each of the classes thus constructed. This may be desirable because of our interest in the distinct classes of the universe as a whole.

In stratified random sampling, the population is sub-divided into strata before the sample is drawn. Strata are so designed that they should not overlap. A sample of specified size is drawn at random from the sampling units that make up each stratum. If a given stratum is of our interest, the corresponding sub-sample provides the basis for estimates concerning the attributes of the population stratum, or sub-universe from which it is drawn. The total of sub-samples constitutes the aggregate sample on which estimates of attributes of the entire population are based.

Stratified samples may be either proportional or non-proportional. In a proportional stratified sampling, the number of elements to be drawn from each stratum is proportional to the size of that stratum compared with the population. For example, if a sample size of 500 elementary units have to be drawn from a population with 10,000 units divided in four strata in the following way:

	<b>Population size</b>	<b>Sample size</b>
Stratum I =	2000	$500 \times 0.2 = 100$
Stratum II =	3000	$500 \times 0.3 = 150$
Stratum III =	4000	$500 \times 0.4 = 200$
Stratum IV =	1000	$500 \times 0.1 = 50$
	Total <span style="border: 1px solid black; padding: 2px;">10000</span>	<span style="border: 1px solid black; padding: 2px;">500</span>

Thus, the elements to be drawn from each stratum would be 100, 150, 200 and 50 respectively. Proportional stratification yields a sample that represents the population with respect to the proportion in each stratum in the population. Proportional stratified sampling yields satisfactory results if the dispersion in the various strata is of proportionately the same magnitude. If there is a significant difference in dispersion from stratum to stratum, sample estimates will be much more efficient if non-proportional stratified random sampling is used. Here, equal numbers of elements are selected from



each stratum regardless of how the stratum is represented in the population. Thus, in the earlier example, an equal number, i.e., 125, of elementary units will be drawn to constitute the sample.

A sample drawn by stratified random sampling scheme ensures a representative sample as the population is first divided into various strata and then a sample is drawn from each stratum. Stratified random sampling also ensures greater accuracy and it is maximum if each stratum is formed in such a way that it consists of uniform or homogeneous items. Compared with a simple random sample, a stratified sample can be more concentrated geographically, i.e., the elementary units from different strata may be selected in such a way that all of them are located in one geographical area. This would also reduce both time and cost involved in data collection. However, care should be exercised in dividing the population into various strata. Each stratum must contain, as far as possible, homogeneous units, as otherwise the reliability of the results would be lost.

In conclusion, stratification is an effective sampling device to the extent that it creates classes that are more homogeneous than the total. When this can be done, the classes are distinguished that differ among themselves in respect of a stated characteristic. Stratification may be futile if classes do not differ among themselves. Thus, there should be homogeneity within classes and heterogeneity between classes.

### ***(c) Cluster Sampling or Multistage Sampling***

Under this method, the random selection is made of primary, intermediate and final (or the ultimate) units from a given population or stratum. There are several stages in which the sampling process is carried out. At first, the first stage units are sampled by some suitable method, such as simple random sampling. Then, a sample of second stage unit is selected from each of the selected first stage units, again by some suitable method which may be same as or different from the method employed for the first stage units. Further stages may be added as required. The procedure may be illustrated as follows:

Suppose we want to take a sample of 5,000 households from the State of Haryana. At the first stage, the state may be divided into a number of districts and a few districts are selected at random. At the second stage, each district may be sub-divided into a number of villages and a sample of villages may be taken at random. At the third stage, a number of households may be selected from each of the villages selected at second stage. To take another example supposes in a particular survey, we wish to take a



sample of 10,000 students from a University. We may take colleges at the first stage, then draw departments at the second stage, and choose students as the third and last stage.

*Merits:* Multi-stage sampling introduces flexibility in the sampling method which is lacking in the other methods. It enables existing divisions and sub-divisions of the population to be used as units at various stages, and permits the field work to be concentrated and yet large area to be covered.

Another advantage of this method is that sub-division into second stage units need be carried out for only those first stage units which are included in the sample. It is, therefore, particularly valuable in surveys of under-developed areas where no frame is generally sufficiently detailed and accurate for subdivision of the material into reasonably small sampling units.

*Limitations:* However, a multi-stage sample is in general less accurate than a sample containing the same number of final stage units which have been selected by some suitable single stage process.

#### ***(d) Systematic Sampling***

Another sampling form, simple in design and execution, may be employed when the members of population to be sampled are arranged in order, the order corresponding to consecutive numbers. The arrangements of names in a telephone directory or income-tax returns in the income tax department are the illustrations of such orderings. A sample of suitable size is obtained by taking every unit say, seventh unit of the population, one of the first seven units in this ordered arrangement is chosen at random and the sample is completely by selecting every seventh unit from the rest of the list. If the first unit selected is the fifth, the researcher will include in his sample 12<sup>th</sup>, 19<sup>th</sup>, 26<sup>th</sup>, 33<sup>rd</sup>, etc. We can generalize the approach as follows: if the requirements of the survey call for the inclusion of one unit out of every  $m$  units in the population, a unit is chosen at random from the first  $m$  units, thereafter, every  $m$ th unit in the population when arranged in order, is included in the sample. This mode of selection is called systematic sampling,  $m$  is generally referred to as the sampling ratio, i.e., the ratio of the population size to the sample size. Symbolically  $m = \frac{N}{n}$ .

where  $N$  is the population size and  $n$  is the sample size. While calculating the value of  $m$ , we may get a fractional value. In such cases, it is rounded off to the nearest digit.



### *Which sampling scheme to select*

In sampling, one scheme is said to be more efficient than another when the sample estimates developed by the scheme tend to cluster more closely around the population parameter being estimated. An estimator of the population parameter should possess the following characteristics:

- (i) It should be unbiased: An estimator is unbiased when the expected (average) value of the sample statistic is equal to the population parameter being estimated.
- (ii) It should be efficient: Efficiency is with respect to sample size and it means that the sample estimates should be clustered as closely possible to the population parameter being estimated for a given sample size. For example, when the population is normally distributed, both the sample mean and the median are unbiased estimators of the population mean. However, for any given sample size, the sample means cluster more closely around the population mean than do the sample medians. Thus, both mean and the median are the unbiased estimators of the population mean. However, the sample mean is the unbiased efficient estimator of the population mean. In stratified random sampling, where stratification is meaningful, a stratified random sample will be more efficient than a simple random sample of the same size. A sampling design is considered efficient with respect to cost if the sample estimates cluster more closely around the population parameter being estimated than they would for any alternative sampling scheme involving equivalent rupee expenditure.
- (iii) It should be consistent: An estimator is considered to be consistent if the sample estimates cluster more and more closely around the population parameter being estimated as the sample size increases.

### *E Multy Stage Sampling Method*

It is a complex probability sampling technique that involves selecting a sample in multiple stages, often using different sampling methods at each stage. It is particularly useful for large, geographically dispersed populations where a straightforward sampling method (like simple random sampling) would be impractical or too costly.

#### **Key Features of Multistage Sampling:**

- **Multiple Stages:** As the name suggests, the sampling process is done in stages. At each stage, a sample is selected from a larger group (which might also be subdivided further in later stages).



- **Combination of Sampling Techniques:** Different sampling methods (e.g., simple random sampling, systematic sampling, stratified sampling) can be used at different stages of the sampling process.
- **Flexibility:** This method is highly adaptable and can be adjusted to the specific needs of the study and the structure of the population.

### Steps in Multistage Sampling:

#### 1. First Stage (Primary Sampling Unit Selection):

- The population is divided into large groups or clusters (e.g., districts, regions, or countries).
- These groups are then sampled randomly or systematically.
- For example, if you are studying a national population, you might divide the country into regions or states.

#### 2. Second Stage (Secondary Sampling Unit Selection):

- Once clusters are selected, these clusters are further divided into smaller subgroups or units (e.g., cities within a state, villages within a district).
- A random sample of these smaller units is selected from each cluster. You may apply a different sampling method here, such as simple random sampling or systematic sampling.

#### 3. Third Stage (Tertiary Sampling Unit Selection):

- This stage continues the process by subdividing the smaller groups even further, if needed (e.g., households within a city or individual people within a household).
- A sample is selected from these smaller units, either randomly or using another sampling technique.

#### 4. Additional Stages:

- More stages can be added, depending on the complexity and scale of the survey. For example, if a study requires very specific subsets of the population, additional layers can be included (e.g., sampling individuals from households or schools).

#### 5. Data Collection:

- After selecting the final sample at the last stage, data is collected from these selected units.

### Example of Multistage Sampling:

Imagine you are conducting a national survey on education in a country. The steps might look like this:



1. **First Stage:** Divide the country into regions (e.g., North, South, East, West). Select a few regions at random.
2. **Second Stage:** Within the selected regions, divide them into states or provinces. Randomly select a few states from each region.
3. **Third Stage:** In the selected states, divide them into cities or districts. Select a few cities or districts randomly.
4. **Fourth Stage:** In the selected cities, divide them into schools. Randomly select a few schools from each city.
5. **Fifth Stage:** Within the selected schools, select individual students or classrooms to survey.

At each stage, the sampling can be done using methods such as simple random sampling, stratified sampling, or systematic sampling.

**Advantages of Multistage Sampling:**

1. **Cost-effective:** It reduces the costs of sampling by first sampling large groups and then narrowing down the sample size progressively.
2. **Practical for Large Populations:** Multistage sampling is useful when the population is geographically spread out or very large, making it difficult to sample all individuals directly.
3. **Flexibility:** The method can be tailored to the specific needs of the study. Researchers can use different sampling techniques at each stage to improve efficiency and accuracy.
4. **Generalizability:** By sampling from different levels (e.g., regions, cities, households), multistage sampling ensures a broad representation of the population, leading to more reliable and generalizable results.

**Disadvantages of Multistage Sampling:**

1. **Complexity:** The process of dividing the population into multiple stages can be time-consuming and complicated. Each additional stage introduces more complexity to the design and analysis.
2. **Sampling Errors:** Since the sampling is done in stages, there is a chance of introducing sampling errors at each stage. The final sample may not always be as representative as desired.
3. **Increased Variability:** If one stage of sampling is not representative of the population, it can lead to biased results. For example, if certain regions are over- or under-represented, this could skew the findings.

Multistage sampling is commonly used in large-scale surveys and research projects, such as:





- **National or International Surveys:** Studies involving large countries or regions, like census data collection, public health surveys, and education research.
- **Market Research:** When businesses want to collect data from customers in different regions or from various demographics.
- **Social Science Research:** Large population studies on topics like poverty, employment, or housing.

Multistage sampling is a flexible, efficient method that can handle large, diverse populations. By breaking down the sampling process into several stages, it reduces costs and logistical challenges, though it can be more complex to implement and analyze. This method is particularly valuable when conducting studies on a national or global scale, where sampling at each level ensures a representative sample for accurate data collection and analysis.

#### 4.2.2 NON-PROBABILITY SAMPLING METHODS

There are six methods of sampling in this category. These are explained as follows:

##### 1. Convenience Sampling

In this scheme, a sample is obtained by selecting 'convenient' population elements. For example, a sample selected from the readily available sources or lists such as telephone directory or a register of the small scale industrial units, etc. will give us a convenient sample. In these cases, even if a random approach is used for identifying the units, the scheme will not be considered as simple random sampling. For example, if one studies the wage structure in a close by textile industry by interviewing a few selected workers, then the scheme adopted here is convenient sampling. The results obtained by convenience sampling method can hardly be said to be representative of the population parameters. Therefore, the results obtained are generally biased and unsatisfactory. However, convenient sampling approach is generally used for making pilot studies, particularly for testing a questionnaire and to obtain preliminary information about the population.

##### 2. Quota Sampling

In this method of sampling, the basic parameters which describe the population are identified first. Then the sample is selected which conform to these parameters. Thus, in a quota sample, quotas are fixed according to these parameters, and each field investigator is assigned with quotas of the number of units to be interviewed. Within the preassigned quotas, the selection of the sample elements depends on the



personal judgement. For example, if one is studying the consumer preferences for ice creams among children and college going students and supposes it is fixed to interview 250 individuals from each category. If the city has five colleges, one decides to fix up a quota of 50 students to be interviewed from each college. It entirely depends upon the interviewer who will constitute this sub-sample of 50 students in a college— they may be the first 50 students who visit the ice cream parlour or may be the 50 students who visit the parlour between 4 p.m. and 6 p.m., etc.

Quota sampling method has the advantage that the sample will conform to the selected parameters of the population. The cost and time involved in getting information from the sample will be relatively less for a quota sample but there are many weaknesses too. Some of these are:

- (i) It is difficult to validate the information gathered on the elementary units,
- (ii) It may be difficult to specify the characteristics of the population and therefore it may be hard to identify it,
- (iii) Even when the sample does conform to the characteristics used in the quotas, the sample may be distorted on other factors of importance in the study. For example, interviewing first 50 students or the last 50 students visiting the ice cream parlour can make a lot of difference particularly about their purchasing capacity, tastes, etc. This may completely distort the results.

Quota sampling method is generally used in public opinion studies, election forecast polls, as there is not sufficient time to adopt a probability sampling scheme.

### **3. Judgement Sampling**

Judgement sampling method can also be called as sampling by opinion. In this method, someone who is well acquainted with the population decides which members (elementary units) in his or her judgement would constitute a proper cross-section representing the parameters of relevance to the study. This method of sampling is generally used in studies involving performance of personnel. For example, if one is studying the performance of sales staff in a marketing organisation, the people here are classified into top grade, medium grade and low grade performers. Having specified qualities that are important in the study, the expert (possibly here the Vice-President-sales) indicates the people who, in his or her knowledge, would be representative of each of the three categories mentioned earlier. This, of course, is not a scientific method, but in the absence of better evidence, such a judgement method may have to be used.



#### **4. Snowball Sampling**

A method used when participants are hard to reach or part of a hidden population. Existing participants refer new participants, creating a "snowball" effect. The researcher starts with a small group of initial participants, who then refer others. This process continues as the sample "snowballs." Research on drug users might begin with one or two participants who then refer other drug users they know. This method is useful for studying populations that are difficult to access. The disadvantage of this method is that sample may become homogenous and biased, as participants are interconnected. This method is common in qualitative research, but it can be biased due to the reliance on initial participants' networks.

#### **5. Self-Selection Sampling**

According to this non probability sampling method, individuals volunteer to participate in the study, often by responding to a public advertisement or call for participation. Participants choose to join the study, usually by signing up or responding to an invitation. An online survey where people sign up to participate. This method is very easy to implement, especially for online studies. In this method, the sample may be biased since only those with a specific interest or motivation to participate will join. In self-selection sampling, participants choose themselves to be part of the study, often through advertisements or calls for volunteers. This method can lead to a biased sample because people who volunteer may differ systematically from those who do not.

#### **6. Expert Sampling**

A specific type of purposive sampling where experts in a particular field or subject matter are chosen to provide insights or information for the study. Researchers intentionally select individuals who have specialized knowledge or experience relevant to the research topic. A study on climate change might rely on interviews with climate scientists and experts in environmental policy. This method provides access to valuable expert knowledge. Also this method is subject to researcher bias in selecting experts, and may not reflect the broader population.

#### **Summary of Non-Probability Sampling Methods:**

- **Convenience Sampling:** Participants are chosen for ease of access.
- **Judgmental (Purposive) Sampling:** Participants are selected based on the researcher's judgment.



- **Quota Sampling:** Ensures specific subgroups are represented, but selection within groups is non-random.
- **Snowball Sampling:** Participants refer others, useful for hard-to-reach populations.
- **Self-Selection Sampling:** Participants volunteer to participate.
- **Expert Sampling:** Focuses on selecting experts in a field.

Each of these methods has its strengths and weaknesses, and the choice of method depends on the research goals, the population being studied, and practical considerations such as time, budget, and access to participants.

#### 4.2.3 Determination of sample size

We prefer samples to complete enumeration because of convenience and reduced cost of data collection. However, in sampling, there is a likelihood of missing some useful information about the population. For a high level of precision, we need to take a larger sample. How large should be the sample and what should be the level of precision? In specifying a sample size, care should be taken such that (i) neither so few are selected so as to render the risk of sampling error intolerably large, nor (ii) too many units are included, which would raise the cost of the study to make it inefficient. It is, therefore, necessary to make a trade-off between (i) increasing sample size, which would reduce the sampling error but increase the cost, and (ii) decreasing the sample size, which might increase the sampling error while decreasing the cost. Therefore, one has to make a compromise between obtaining data with greater precision and with that of lower cost of data collection. Several factors need to be considered before determining the sample size. The first and the foremost is the size of the error that would be tolerable for the purposes of decision-making. The second consideration would be the degree of confidence with the results of the study, i.e., if one wants to be 100 per cent confident of the results, the entire population must be studied. However, this is generally too impractical and costly. Therefore, one must accept something less than 100 per cent confidence. In practice, the confidence limits most often used are 99 per cent, 95 per cent and 90 per cent. Most commonly used confidence limit is 95 per cent. This means that there is a 5 per cent risk that the true population statistic is outside the range of possible error specified by the confidence interval. This 5 per cent risk appears to be acceptable in most of the decisions. Thus, for 95 per cent level of confidence,  $Z$  value is 1.96. The  $Z$  value can be obtained from normal probability distribution for a specified level of confidence. For determining the sample size, we make use of the following relationship:



$$\sigma_{\bar{x}} = \text{standard error of the estimate} = \frac{s}{\sqrt{n}}$$

$\sigma_{\bar{x}}$  can be calculated if we know the upper and lower confidence limits. Let these limits be Y, then  $Z \sigma_{\bar{x}} = Y$ .

Where Z is the value of the normal variate for a given confidence level. The procedure has been explained using the illustration given below:

**Illustration 7.1.** A state cooperative department is performing a survey to determine the annual salary earned by managers numbering 3000 in the cooperative sector within the state. How large a sample size it should take in order to estimate the mean annual earnings within plus and minus 1,000 and at 95 per cent confidence level? The standard deviation of annual earnings of the entire population is known to be Rs. 3,000.

**Solution.** As the desired upper and lower limit is Rs. 1,000, i.e., we want to estimate the annual earnings within plus and minus Rs. 1,000.

$$\square \quad z \sigma_{\bar{x}} = 1,000$$

As the level of confidence is 95 per cent, the Z value is 1.96

$$\square \quad 1.96 \sigma_{\bar{x}} = 1,000$$

$$\sigma_{\bar{x}} = \frac{1,000}{1.96} = 510.20$$

The standard error  $\sigma_{\bar{x}}$  is given by  $\square/\sqrt{n}$  where  $\square$  is the population standard deviation

$$\square \quad \frac{s}{\sqrt{n}} = 510.20$$

$$\text{i.e.,} \quad \frac{3000}{\sqrt{n}} = 510.20$$

$$\text{i.e.,} \quad \sqrt{n} = \frac{3000}{510.2} = 5.88$$

This gives  $n = 34.57$

Therefore, the desired sample size is about 35.

**Key Factors to Consider When Determining Sample Size:**

- 1. Population Size (N):** The total number of individuals in the population from which the sample is drawn. For small populations, you may need to adjust the sample size using finite population correction formulas. If you are conducting research on a small group of employees in a company (population size = 500), the sample size will differ from a larger population (e.g., the entire country).
- 2. Confidence Level (Z):** The confidence level represents the probability that the sample accurately reflects the population. Common confidence levels are 90%, 95%, and 99%. Z-scores for typical confidence levels: 90%  $\rightarrow Z = 1.645$ ; 95%  $\rightarrow Z = 1.96$  and 99%  $\rightarrow Z = 2.576$ .  
If you choose a 95% confidence level, you are 95% confident that your sample results reflect the true population values.
- 3. Margin of Error (E):** The margin of error (or precision) indicates how much the sample results can differ from the actual population value. It is usually expressed as a percentage, such as  $\pm 5\%$ . If the margin of error is 5%, the sample estimate could be 5% higher or lower than the true population value.
- 4. Population Proportion (p):** If you are estimating a proportion (e.g., the proportion of people who support a policy), you need to have an estimate of the population proportion (p). If unknown, a conservative estimate of 50% ( $p = 0.5$ ) is typically used, as it maximizes the required sample size. If you are studying voter preferences and know that 60% of the population supports a policy, you would use  $p = 0.6$ .
- 5. Standard Deviation ( $\sigma$ ):** In studies where you are estimating a mean rather than a proportion (e.g., average income), the standard deviation ( $\sigma$ ) of the population is required. If the standard deviation is unknown, you can use an estimate from previous studies or conduct a pilot study. If you're studying the average height of people in a city, you would need to know the population's standard deviation for height.

**Sample Size for Stratified Sampling**

Once the strata have been established, we are interested in the size of the stratified random sample. The size will depend upon whether the proportional or disproportional (optimal) sample is being taken.



A proportional stratified sample is one in which the sample units in a given stratum are allocated in proportion to the relative size of the stratum. The following formula is used for calculation of the proportional sample for each stratum

$$n_i = \frac{N_i}{N} \times n$$

Where  $n_i$  = number of sample units from stratum  $i$ ,  $N$  = the total number of units in the population,  $N_i$  = the total number of units in the stratum  $i$ ,  $n$  = sample size desired.

The standard error of mean is

$$\sigma_{\bar{x}} = \sqrt{\sum_{i=1}^k w_i^2 s_i^2 / n_i}$$

where  $w_i$  = the weight of stratum  $i = N_i/N$ ,  $n_i$  = the standard deviation of the  $i$ th stratum,  $k$  = the total number of strata. In case of disproportionate stratified sampling, the proportion of units in the sample stratum is not equal to the proportion of the population. The formula for sample allocation in this case is

$$n_i = \frac{w_i s_i n}{\sum_{i=1}^k w_i s_i}$$

Thus, the disproportionate stratified sample is more desirable if standard deviation ( $n_i$ ) of each stratum is known. The standard error of the mean of a disproportionate stratified sample is

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum (w_i s_i)^2}{\frac{1}{n_i}}}$$

It may be observed that the standard error for stratified sample is smaller than for *simple* random sample, i.e., much smaller samples may be utilized when the population has been stratified.

**Illustration 4.2.** In a market area, shops are divided into two categories, viz., those that have daily turnover of more than Rs. 2000 and those that have daily turnover of less than Rs. 2000 for the study of estimating the total sales in the area. The total number of shops in the first stratum are 420 and in the second stratum 180. A sample of 50 was selected, the standard deviation has been found to be 70 for first stratum and 95 for second stratum. What size of stratified random sample should be taken under proportional and disproportionate stratified sampling?



**Solution.** Under the proportional stratified sampling, the sample size is given by

$$n_i = \frac{N_i}{N} \times n$$

$$\text{and, therefore } n_1 = \frac{420}{600} \times 50 = 35$$

$$\text{and } n_2 = \frac{180}{600} \times 50 = 15$$

$$\begin{aligned} \text{The standard error } (\sigma_{\bar{x}}) &= \sqrt{\sum w_i^2 \frac{s_i^2}{n_i}} \\ &= \sqrt{(0.7)^2 \times \frac{(70)^2}{35} + \frac{(0.3)^2 \times (95)^2}{15}} \\ &= \sqrt{122.75} = 11.079 \end{aligned}$$

For disproportionate sampling, the sample size is given by:

$$n_i = \frac{w_i s_i n}{\sum w_i s_i}$$

$$n_1 = \frac{0.7 \times 70 \times 50}{0.7 \times 70 + 0.3 \times 95} = \frac{2450}{77.5} = 32.0$$

$$\text{and } n_2 = \frac{0.3 \times 95 \times 50}{0.7 \times 70 + 0.3 \times 95} = \frac{1425}{77.5} = 18.0$$

The standard error is given by

$$\begin{aligned} \sigma_{\bar{x}} &= \sqrt{\frac{\sum (w_i s_i)^2}{\sum n_i}} = \sqrt{\frac{(0.7 \times 70 + 0.3 \times 95)^2}{50}} \\ &= \sqrt{120.125} = 10.96 \end{aligned}$$

### Cost as a Factor in the Determination of the Sample Size

Another consideration in determining the sample size is the cost. Management may reduce the level of confidence in an attempt to reduce the cost of sampling. An illustration will clarify how cost of sampling can be reduced by reducing the sample size.

**Illustration 4.3.** In a market area there are 600 shops. A researcher wishes to estimate number of customers visiting these shops per day. The researcher wants to estimate the sampling error in the





number of customers visiting is no larger than  $\pm 10$  with probability of 0.95. The previous studies indicated that the standard deviation is 85 customers. If the cost per interview is Rs. 20 (this includes field work, supervision of interviewers, coding, editing and tabulation of results and report writing, etc.), calculate the total cost involved. Researcher is willing to sacrifice some accuracy in order to reduce cost. If he settles for an estimate with 0.90 probability, how much reduction in cost can be achieved?

**Solution.** For 95 per cent confidence levels,

$$Z \sigma_{\bar{x}} = Y$$

$$\text{i.e., } 1.96 \sigma_{\bar{x}} = 10.0$$

$$\sigma_{\bar{x}} = \frac{10}{1.96}$$

Now,  $\sigma_{\bar{x}}$  is given by  $s/\sqrt{n}$  and therefore, the sample size will be determined by the equation

$$\frac{s}{\sqrt{n}} = \frac{10}{1.96}$$

Since  $s = 85$ , we have

$$\frac{85}{\sqrt{n}} = \frac{10}{1.96}$$

$$n = 277.6$$

Thus, if the sample is taken as 278, the total cost involved will be  $278 \times 20 = \text{Rs. } 5560$ . As this cost is considered to be on the higher side by the researcher and in order to reduce the cost, the researcher has now settled to 90 per cent confidence level. At 90 per cent confidence level, the sample size can be calculated as follows:

$$Z \sigma_{\bar{x}} = 10$$

$$1.65 \sigma_{\bar{x}} = 10$$

$$1.65 \sigma_{\bar{x}} = 10$$

$$\text{or } \sigma_{\bar{x}} = \frac{10}{1.65}$$

$$\frac{s}{\sqrt{n}} = \frac{10}{1.65}$$



$$\begin{aligned} \text{i.e., } \frac{85}{\sqrt{n}} &= \frac{10}{1.65} \\ n &= 196.7 \end{aligned}$$

The cost of survey for this sample size will be  $197 \times 20 = \text{Rs. } 3940$ . Thus, we have observed that by reducing the confidence level from 95 per cent to 90 per cent, the researcher would reduce the cost from Rs. 5560 to Rs. 3940. The researcher may not like to reduce the confidence level further and so further cost reduction may not be desirable.

### 4.3 SAMPLING AND NON-SAMPLING ERRORS

The choice of a sample though may be made with utmost care, involves certain errors which may be classified into two types: (a) Sampling errors, and (b) Non-Sampling errors. These errors may occur in the collection, processing and analysis of data.

#### (a) Sampling Errors

A sampling error refers to the difference between a sample statistic (such as a sample mean or sample proportion) and the corresponding population parameter (such as the population mean or population proportion) that the sample is intended to represent. This error arises because a sample is only a subset of the entire population, and it is unlikely to perfectly reflect the characteristics of the population.

#### Key points about sampling error:

- 1. Natural variation:** Sampling error is a natural result of the fact that different samples taken from the same population may give slightly different results. It does not indicate a mistake or problem with the sampling process, but rather the inherent variability that occurs when studying a sample instead of the entire population.
- 2. Size of the sample:** The larger the sample size, the smaller the sampling error tends to be. This is because larger samples more closely approximate the population characteristics, reducing the discrepancy between the sample and population.
- 3. Standard error:** The standard deviation of the sampling error is known as the standard error. It can be used to quantify the variability in the sample statistics due to sampling error.
- 4. Bias:** Sampling error is different from sampling bias, which occurs when a sample is not representative of the population in a systematic way, leading to inaccurate conclusions.

Imagine a study to estimate the average height of adult men in a country. If you randomly sample 100 people, the average height you calculate will be close to, but not exactly the same as, the true population



average. The difference between your sample average and the true population average is the sampling error.

The formula to calculate the sampling error (for means) is:

$$\text{Standard Error (SE)} = \frac{\sigma}{\sqrt{n}}$$

Where:

- SE = Standard Error
- $\sigma$  = Standard deviation of the population
- n = Sample size

In conclusion, sampling error reflects the natural variation that occurs when making inferences about a population based on a sample. The error decreases as the sample size increases.

### Casuses of Sampling Error:

Sampling error arises due to various factors inherent in the process of selecting a sample from a population. These errors occur because a sample, by definition, is just a subset of the entire population, and it is unlikely to be a perfect representation of the population. Below are the primary causes of sampling error:

**1. Randomness of Sampling:** The most fundamental cause of sampling error is the random selection of individuals from the population. Even when the sampling process is unbiased and conducted properly, random variations mean that no two samples will be exactly the same. Different samples may yield slightly different results (e.g., a sample mean or proportion), which leads to sampling error. In a survey about public opinion, different random samples of the population may show slightly different estimates of the proportion of people who support a certain policy.

**2. Sample Size:** The size of the sample is a significant factor in determining the magnitude of sampling error. Smaller samples are more likely to exhibit greater variability, resulting in larger sampling errors. With smaller samples, there is a greater chance of the sample being unrepresentative of the population, as it may not capture the full diversity of the population. In a small sample of 10 people, the proportion of people who prefer a certain brand may be significantly different from the actual proportion in the population. In a larger sample of 1,000 people, the proportion is likely to be closer to the true value.

**3. Population Variability:** The variability or heterogeneity in the population can influence the sampling error. A population with a high degree of variability in key characteristics (such as income,



age, or opinion) will lead to larger potential sampling errors. A more diverse population requires larger samples to estimate characteristics accurately because the increased diversity introduces more potential for variation. If you're estimating the average income of a country and the population has a wide range of incomes, the sampling error will tend to be larger unless the sample size is sufficiently large.

**4. Sampling Method:** Improper or biased sampling methods can contribute to increased sampling error. If the method of selecting the sample is flawed (e.g., using convenience sampling or not selecting a truly random sample), the sample may not represent the population well. Even if the sample size is large, using a biased sampling method can lead to sampling error that does not reflect the true population parameters. If a survey about health habits is conducted by interviewing people at a gym, the sample is likely to be biased toward people who are more health-conscious, leading to sampling error in terms of the broader population's health habits.

**5. Non-Response Bias:** Non-responses occur when some individuals selected for the sample do not participate. This can be due to various factors, such as lack of interest, inability to respond, or unwillingness to participate. Non-respondents may differ from respondents in key characteristics, leading to a biased sample and thus, an increased sampling error. If a survey on political opinions has a low response rate, and the people who respond are more likely to have strong opinions, the sampling error will be larger because the sample does not accurately represent the overall population.

**6. Sampling Frame Issues (Frame Error):** The sampling frame is the list or set from which the sample is drawn. If the frame does not accurately represent the entire population (e.g., if some groups are excluded or overrepresented), it can cause sampling error. An incomplete or flawed sampling frame leads to systematic errors in the sample selection, meaning that some parts of the population may be overrepresented or underrepresented, which increases the error in the estimate. If a survey on household incomes uses a telephone directory as the sampling frame, it might exclude households without phones, leading to an unrepresentative sample and increased sampling error.

**7. Clustered Sampling:** When using cluster sampling (where the population is divided into clusters, and a sample of clusters is selected), the sampling error can arise if the clusters themselves are not homogenous and are too similar to each other. This leads to less diversity in the sample compared to the population, which increases sampling error. In contrast, using simple random sampling might capture more diverse elements of the population. If you're sampling schools in a city to estimate educational



outcomes, and you select only urban schools while ignoring rural ones, your sample may not reflect the diversity of educational experiences in the broader population, introducing sampling error.

**8. Over-Sampling or Under-Sampling of Subgroups:** Over-sampling or under-sampling particular subgroups within the population can contribute to sampling error. This happens when certain groups are intentionally or unintentionally given more or less representation in the sample than they have in the population. When subgroups are over-sampled, they may dominate the sample and cause an inflated estimate for the characteristic in question. Conversely, under-sampling can lead to underrepresentation of certain characteristics. If a survey about consumer preferences for a product is conducted by surveying a disproportionate number of young people and neglecting older individuals, the results might be biased, leading to a higher sampling error for the overall population.

**9. Measurement Error:** Although not strictly a sampling error, measurement errors (such as inaccurate data collection methods) can amplify the impact of sampling error. Poorly designed surveys, ambiguous questions, or faulty data collection tools can all contribute to errors in the sample that affect the overall estimate. Measurement errors can skew the sample data, which in turn increases the discrepancy between the sample statistic and the true population parameter. If a survey question is poorly worded, respondents might misinterpret it, leading to data that doesn't truly reflect their opinions, thus increasing sampling error.

Sampling error occurs due to factors such as random variability, the sample size, population variability, and issues with the sampling process. These errors can be minimized by using proper random sampling techniques, increasing sample size, and ensuring the sampling frame is representative. However, some causes, like random fluctuations, are inevitable and will always produce some degree of error. Understanding these causes is essential for interpreting statistical results and improving sampling methods.

### **(b) Non-Sampling Errors**

Non-sampling errors refer to errors that occur in the process of collecting, processing, or analyzing data, which are not due to the sampling process itself. Unlike sampling errors, which arise because the sample may not perfectly represent the population, non-sampling errors are caused by issues that can occur even if the entire population were surveyed. These errors can distort the results and make them inaccurate or unreliable.



### Causes of Non-Sampling Errors

1. **Measurement Error:** Measurement errors occur when the data collected is inaccurate or imprecise. This can happen due to faulty instruments, unclear survey questions, or incorrect recording of responses. These errors can lead to biased or incorrect data, which can affect the conclusions drawn from the study. A respondent misinterpreting a question on a survey or the interviewer recording the wrong response due to misunderstanding.
2. **Response Error:** This type of error occurs when respondents provide inaccurate or misleading answers, either unintentionally or intentionally. Response errors can occur due to misunderstanding the question, poor recall, or deliberate misreporting. A respondent might give an incorrect answer on a questionnaire because they misunderstood the question, or they might provide socially desirable responses (e.g., claiming to exercise more than they actually do).
3. **Non-Response Error:** This error arises when individuals who are selected for the sample do not respond or are unavailable to provide data. Non-responses can lead to a biased sample if the non-respondents have different characteristics from those who respond. If a survey about health behaviors has a low response rate among young people, the results might over-represent older, health-conscious individuals, leading to biased conclusions.
4. **Processing Errors:** These errors occur during the data entry, coding, or analysis stages. They can arise from mistakes made by those handling the data. Processing errors can introduce inaccuracies into the dataset, making the results unreliable. A typographical error when entering survey responses into a database or incorrectly coding a variable can lead to incorrect analysis results.
5. **Coverage Error (Frame Error):** This type of error occurs when the sampling frame (the list or group from which the sample is drawn) is incomplete or inaccurate, causing certain parts of the population to be excluded or overrepresented. If the sampling frame is not representative of the population, it can lead to bias in the sample, even if random sampling is used. Using a telephone directory as a sampling frame may exclude people who don't have landlines, leading to a sample that isn't representative of the population.
6. **Interviewer Bias:** This occurs when the interviewer's personal beliefs, attitudes, or behaviors influence the responses of the participants. Interviewer bias can skew the data by leading the respondent toward a particular answer, intentionally or unintentionally. An interviewer might emphasize certain



questions more than others or inadvertently show approval or disapproval of a response, thus affecting the accuracy of the information provided by the respondent.

**7. Questionnaire Design Error:** This error arises when the survey or questionnaire is poorly designed, leading to ambiguous, confusing, or biased questions. Poorly designed questions can confuse respondents, leading to inaccurate or invalid responses, or they may introduce bias into the survey results. A question like "How satisfied are you with our excellent customer service?" leads respondents to think that the service was excellent even if it wasn't, thus introducing bias in the responses.

**8. Follow-Up Error:** This occurs when follow-up attempts to contact non-respondents or clarify answers are either insufficient or incorrectly conducted. If follow-up surveys or contacts are poorly managed, this may lead to missing data or inaccurate responses, contributing to non-sampling error. If only a small subset of non-respondents is followed up on, and they are not representative of the entire non-responding group, it can skew the results.

**9. Social Desirability Bias:** Respondents may provide answers that they believe are socially acceptable or desirable, rather than what they truly think or do. This bias can distort the data, especially in surveys about sensitive topics such as health behaviors, income, or drug use. In a survey on alcohol consumption, respondents may underreport their actual alcohol intake to conform to social norms, leading to a misrepresentation of the data.

**10. Time-Related Errors:** Errors that arise from changes over time, especially in longitudinal studies, where the circumstances or behavior of participants may change after the data is collected. The data may become outdated or no longer applicable to the current population or conditions. A study on consumer behavior conducted in one year may no longer accurately reflect the behaviors of consumers in a different year due to changes in preferences, market conditions, or external events (like a recession).

Non-sampling errors can have a significant impact on the accuracy and reliability of survey or study results. Unlike sampling errors, which are due to the variability in choosing a sample from the population, non-sampling errors are caused by issues during data collection, processing, and analysis. These errors include problems like measurement errors, response biases, non-responses, and faulty data handling. To minimize non-sampling errors, it's important to design clear and unbiased surveys, use proper sampling methods, ensure high response rates, and carefully process the data.

The following table differentiate between sampling and non sampling error:

**TABLE 4.0 DIFFERENCE BETWEEN SAMPLING AND NON SAMPLING ERROR**

Feature	Sampling Error	Non Sampling Error
<b>Definition</b>	Error due to random variability in sampling.	Error from factors other than sampling.
<b>Cause</b>	Random selection of the sample	Errors in data collection, processing, or analysis.
<b>Effect</b>	Difference between sample statistic and population parameter.	Inaccuracy or bias in the data collected.
<b>Occurrence</b>	Happens during sampling.	Happens during any stage (design, collection, processing).
<b>Examples</b>	Standard error, margin of error.	Measurement errors, non-responses, processing errors.
<b>Reduction</b>	Increased sample size.	Improving survey design, data entry, and follow-up.
<b>Quantifiability</b>	Can be measured (e.g., confidence intervals).	Not easily quantifiable.

In essence, sampling errors are inherent and inevitable in any sample-based research but can be managed statistically, whereas non-sampling errors arise from human error, survey design flaws, or biases and can be more difficult to identify and control.

#### 4.4 CHECK YOUR PROGRESS

1. In simple random sampling, drawing of elements from the population is ..... and the choice of an element is made in such a way that every element has the same probability of being chosen.
2. In stratified random sampling, the population is sub-divided into ..... before the sample is drawn.
3. In convenience sampling, a sample is obtained by selecting ..... population elements.
4. In a quota sample, quotas are fixed according to these parameters, and each field investigator is assigned with quotas of the ..... to be interviewed.
5. One has to make a compromise between obtaining data with greater ..... and with that of lower cost of data collection.

#### 4.5 SUMMARY

The process of selecting a sample is known as sampling. Thus, the sampling theory is a study of relationship that exists between the population and the samples drawn from the population. The





complete enumeration, popularly known as census, may not be feasible either due to non-availability of time or because of high cost involved. A probability sample is one for which the inclusion or exclusion of any individual element of the population depends upon the application of probability methods and not on a personal judgement. It is so designed and drawn that the probability of inclusion of an element is known. The essential feature of drawing such a sample is the randomness. As against the probability sample, we have a variety of other samples, termed as judgement samples, purposive samples, quota samples, etc. These samples have one common distinguishing feature: personal judgement rather than the random procedure to determine the composition of what is to be taken as a representative sample. The judgement affects the choice of the individual elements. All such samples are non-random, and no objective measure of precision may be attached to the results arrived at.

Non-probability sampling is a procedure of selecting a sample without the use of probability or randomisation. It is based on convenience, judgement, etc. Several factors need to be considered before determining the sample size. The first and the foremost is the size of the error that would be tolerable for the purposes of decision-making. The second consideration would be the degree of confidence with the results of the study

#### **4.6 KEYWORDS**

**Elementary Units:** The attributes that are the object of the study are known as characteristics and the units possessing them are called the elementary units.

**Population:** The aggregate of elementary units to which the conclusions of the study apply is termed as population/universe.

**Sampling Unit:** The units that form the basis of the sampling process are called sampling units. The sampling unit may be an elementary unit.

**Sample:** The sample is defined as an aggregate of sampling units actually chosen in obtaining a representative subset from which inferences about the population are drawn.

**Frame:** A list or directory, defines all the sampling units in the universe to be covered.



**Cluster sampling or multistage sampling:** Under this method, the random selection is made of primary, intermediate and final (or the ultimate) units from a given population or stratum. There are several stages in which the sampling process is carried out.

**Judgement sampling method:** In this method, someone who is well acquainted with the population decides which members (elementary units) in his or her judgement would constitute a proper cross-section representing the parameters of relevance to the study.

#### 4.7 SELF-ASSESSMENT TEST

1. Describe the various methods of drawing a sample. Which one would you suggest and why?
2. Describe the importance of sampling. Critically examine the merits of probability sampling and non-probability sampling methods.
3. Specify and explain the factors that make sampling preferable to a complete census in a statistical investigation.
4. How would you determine the sample size for stratified sampling? Explain with the help of a suitable example.
5. To determine the effectiveness of the advertising campaign of a new VCR, management would like to know what percentage of the household are aware of the new brand. The advertising agency thinks that this figure is as high as 70 per cent. The management would like a 95% confidence interval and a margin of error no greater than plus or minus 2 per cent. (a) What sample size should be used for this study? (b) Suppose that management wanted to be 99 per cent confident but could tolerate an error of plus or minus 3 per cent. How would the sample size change?

#### 4.8 ANSWERS TO CHECK YOUR PROGRESS

1. Random
2. Strata
3. Convenient
4. Number of units
5. Precision

#### 4.9 REFERENCES/SUGGESTED READINGS

1. Statistical Methods by S.P. Gupta. Sultan Chand and Sons, New Delhi.



2. Statistics for MBA by T.R. Jain and Dr. S.C. Aggarwal. VK (India) Enterprises, New Delhi. First edition.
3. Business Statistics by Shenoy and Shenoy.
4. Levin, Richard I. and David S. Rubin: Statistics for Management, Prentice Hall, New Delhi.
5. Lawrance B. Moore: Statistics for Business & Economics, Harper Collins, NY.
6. Watsman Terry J. and Keith Parramor: Quantitative Methods in Finance International, Thompson Business Press, London.



Subject: Business Statistics-II	
Course code: BCOM 402	Author: Anil Kumar
Lesson: 05	Vetter: Dr. B. S. Bodla
<b>SAMPLING DISTRIBUTIONS</b>	

## STRUCTURE

### 5.0 Learning Objectives

### 5.1 Introduction

#### 5.1.1 Sampling Distribution of the Mean

##### 5.1.1.1 Sampling from Infinite Populations

##### 5.1.1.2 Sampling with Replacement from Finite Populations

##### 5.1.1.3 Sampling without Replacement from Finite Populations

#### 5.1.2 Central Limit Theorem

#### 5.1.3 Sampling Distribution of the Proportion

#### 5.1.4 Sampling Distribution of the Difference of Sample Means

#### 5.1.5 Sampling Distribution of the Difference of Sample Proportions

#### 5.1.6 Small Sampling Distributions

#### 5.1.7 Sampling Distribution of the Variance

##### 5.1.7.1 The Sample Variance

##### 5.1.7.2 The Chi-Square Distribution

##### 5.1.7.3 The $\chi^2$ Distribution in terms of Sample Variance $S^2$

### 5.2 F Distribution and ANOVA

### 5.3 t-Distributions and Z-Distributions

### 5.4 Check your Progress

### 5.5 Summary

### 5.6 Keywords

### 5.7 Self-Assessment Questions

### 5.8 Answers to check your progress

### 5.9 References/Suggested Readings



## 5.0 LEARNING OBJECTIVES

After going through this lesson, students will be able to:

- Understand the concept of sampling distributions.
- Meaning and the need of studying sampling distribution of a sample statistic.

## 5.1 INTRODUCTION

Having discussed the various methods available for picking up a sample from a population, we would naturally be interested in drawing statistical inferences - making generalizations about the population on the basis of a sample drawn from it. The generalizations to be made about the population are usually either by way of

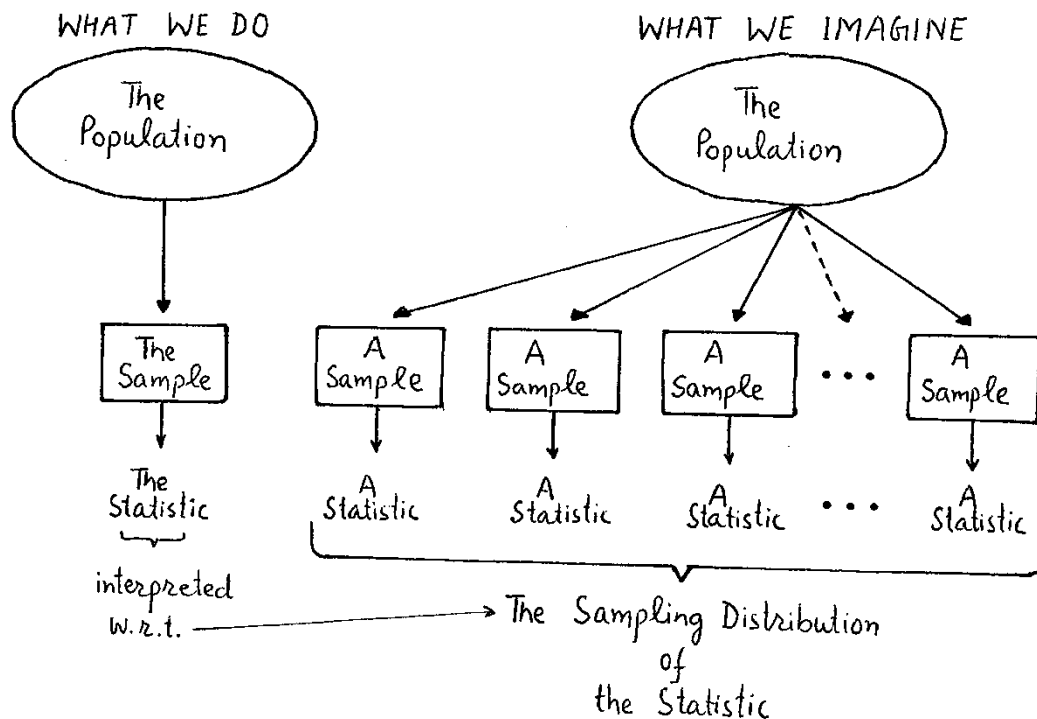
- Estimating the unknown population parameters, or
- Testing appropriate hypotheses stated in relation to population parameters in the light of sample data

These generalizations, together with the measurement of their reliability, are made in terms of the relationship between the values of any *sample statistic* and those of the corresponding *population parameters*. Population parameter is any number computed (or estimated) for the entire population viz. population mean, population median, population proportion, population variance and so on. Population parameter is unknown but fixed, whose value is to be estimated from the sample statistic that is known but random. Sample Statistic is any numbers computed from our sample data viz. sample mean, sample median, sample proportion, sample variance and so on.

It may be appreciated that no single value of the sample statistic is likely to be equal to the corresponding population parameter. This owes to the fact that the sample statistic being random, assumes different values in different samples of the same size drawn from the same population.

Referring to our earlier discussion on the concept of a random variable in the lessons on Probability Distributions, it is not difficult to see that *any sample statistics is a random variable* and, therefore, has a probability distribution better known as the Sampling Distribution of the statistic.

*The sampling distribution of a statistic is the probability distribution of all possible values the statistic may take when computed from random samples of the same size drawn from a specified population.*



**Figure 5-1 Sampling Distribution of a Statistic**

In reality, of course we do not have all possible samples and all possible values of the statistic. We have only one sample and one value of the statistic. This value is interpreted with respect to all other outcomes that might have happened, as represented by the sampling distribution of the statistic. In this lesson, we will refer to the sampling distributions of only the commonly used sample statistics like sample mean, sample proportion, sample variance *etc.*, which have a role in making inferences about the population.

### Why We Study Sampling Distributions?

Sample statistics form the basis of all inferences drawn about populations. Thus, sampling distributions are of great value in inferential statistics. The sampling distribution of a sample statistic possess well-defined properties which help lay down rules for making generalizations about a population on the basis of a single sample drawn from it. The variations in the value of sample statistic not only determine the shape of its sampling distribution, but also account for the element of error in statistical inference. If we know the probability distribution of the sample statistic, then we can calculate risks (error due to chance) involved in making generalizations about the population. With the help of the properties of



sampling distribution of a sample statistic, we can calculate the probability that the sample statistic assumes a particular value (if it is a discrete random variable) or has a value in a given interval. This ability to calculate the probability that the sample statistic lies in a particular interval is the most important factor in all statistical inferences. We will demonstrate this by an example.

Suppose we know that 40% of the population of all users of hair oil prefers our brand to the next competing brand. A "new improved" version of our brand has been developed and given to a random sample of 100 users for use. If 55 of these prefer our "new improved" version to the next competing brand, what should we conclude? For an answer, we would like to know the probability that the sample proportion in a sample of size 100 is as large as 55% or higher when the true population proportion is only 40%, *i.e.* assuming that the new version is no better than the old. If this probability is quite large, say 0.5, we might conclude that the high sample proportion *viz.* 55% is perhaps because of sampling errors and the new version is not really superior to the old. On the other hand, if this probability works out to a very small figure, say 0.001, then rather than concluding that we have observed a rare event we might conclude that the true population proportion is higher than 40%, *i.e.* the new version is actually superior to the old one as perceived by members of the population. To calculate this probability, we need to know the probability distribution of sample proportion *i.e.* the sampling distribution of the proportion.

Studying sampling distributions is crucial for understanding how sample statistics behave and how they relate to the population parameters. Sampling distributions form the foundation of statistical inference, enabling us to make reliable conclusions about a population based on sample data. Below are key reasons why we study sampling distributions:

**1. Foundation for Statistical Inference:** Sampling distributions provide a way to understand the behavior of sample statistics, such as the sample mean, variance, or proportion, across multiple samples drawn from the same population. With knowledge of the sampling distribution, we can make inferences about a population parameter (like the population mean or proportion) even when we have only a sample. If we know the sampling distribution of the sample mean, we can estimate the population mean and calculate confidence intervals and perform hypothesis tests.

**2. Understanding the Variability of Sample Statistics:** Every sample statistic (such as the mean or proportion) varies from sample to sample due to random chance. Sampling distributions help us understand this variability. By studying how the sample statistics behave over many samples, we can



assess how much variability to expect in our estimates, helping us determine if an observed statistic is likely due to random chance or represents a real effect in the population. Understanding the variability of the sample mean (its standard error) helps to estimate how much our sample mean is likely to differ from the true population mean.

**3. Estimation of Population Parameters:** Sampling distributions are used to estimate population parameters (e.g., the population mean, proportion, or standard deviation). By drawing multiple random samples and calculating the sample statistics (like the sample mean), we can estimate the population parameters more accurately. Sampling distributions tell us how good these estimates are likely to be. If you repeatedly draw samples and calculate sample means, you can use the sampling distribution of the mean to estimate the population mean and the likely error in the estimate.

**4. Standard Error and Confidence Intervals:** Sampling distributions provide the framework for calculating the standard error (the standard deviation of the sample statistic), which is key to constructing confidence intervals. The standard error measures the variability of the sample statistic, and understanding it allows us to calculate the range within which we expect the true population parameter to lie with a certain level of confidence. A confidence interval for the population mean is based on the sampling distribution of the sample mean and its standard error.

**5. Hypothesis Testing:** Sampling distributions are essential for hypothesis testing, as they allow us to determine the likelihood of obtaining a sample statistic under the null hypothesis. By comparing the observed sample statistic to the expected value under the null hypothesis (using the sampling distribution), we can calculate p-values and make decisions about whether to reject or fail to reject the null hypothesis. In testing whether a new drug is effective, the sampling distribution of the sample mean helps to determine whether the observed effect is statistically significant.

**6. Understanding the Central Limit Theorem (CLT):** The Central Limit Theorem states that the distribution of the sample mean (for sufficiently large sample sizes) will approximate a normal distribution, regardless of the shape of the population distribution. The CLT is a fundamental result that justifies the use of normal distribution approximations in inferential statistics, even when the underlying population is not normally distributed. Even if the population distribution is skewed, the distribution of the sample mean will become approximately normal as the sample size increases, enabling the use of z-scores and confidence intervals.





**7. Better Decision-Making:** By understanding sampling distributions, we can make more informed and reliable decisions based on sample data. Sampling distributions help assess how likely observed sample results are, giving decision-makers confidence in their estimates and predictions. In business, understanding the sampling distribution of customer satisfaction scores helps to make more confident decisions about product quality or service improvement.

**8. Improved Accuracy of Predictions:** Understanding how sample statistics behave allows for better predictions of future observations or sample statistics. By understanding the sampling distribution, we can determine the accuracy of predictions made from sample data, such as the likely range for future sample means or other statistics. In polling, knowing the sampling distribution of the sample mean helps predict the outcome of elections with a known margin of error.

**9. Understanding the Role of Sample Size:** The sampling distribution shows how the sample size affects the variability of sample statistics. Larger sample sizes reduce the variability of sample statistics and provide more reliable estimates of the population parameter. Studying sampling distributions helps us understand this relationship. A larger sample size leads to a smaller standard error, which in turn leads to more precise estimates of the population mean.

Studying sampling distributions is critical because they form the basis of statistical inference, helping us estimate population parameters, calculate confidence intervals, conduct hypothesis tests, and make accurate predictions. They allow us to understand the variability of sample statistics and provide a structured way to assess the reliability of sample-based estimates. Furthermore, concepts like the Central Limit Theorem and standard error are grounded in sampling distributions, enabling more accurate and confident decision-making across various fields, such as science, business, and social research.

### 5.1.1 Sampling Distribution of the Mean

Suppose we have a simple random sample of size  $n$ , picked up from a population of size  $N$ . We take measurements on each sample member in the characteristic of our interest and denote the observation as  $x_1, x_2, \dots, x_n$  respectively. The sample mean for this sample is defined as:

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

If we pick up another sample of size  $n$  from the same population, we might end up with a totally different set of sample values and so a different sample mean. Therefore, there are many (perhaps



infinite) possible values of the sample mean and the particular value that we obtain, if we pick up only one sample, is determined only by chance. In other words, the sample mean is a random variable. The possible values of this random variable depends on the possible values of the elements in the random sample from which sample mean is to be computed. The random sample, in turn, depends on the distribution of the population from which it is drawn. As a random variable,  $\bar{X}$  has a *probability distribution*. This probability distribution is the sampling distribution of  $\bar{X}$ .

*The sampling distribution of  $\bar{X}$  is the probability distribution of all possible values the random variable  $\bar{X}$  may take when a sample of size  $n$  is taken from a specified population.*

To observe the distribution of  $\bar{X}$  empirically, we have to take many samples of size  $n$  and determine the value of  $\bar{X}$  for each sample. Then, looking at the various observed values of  $\bar{X}$ , it might be possible to get an idea of the nature of the distribution. We will derive the distribution of  $\bar{X}$  in three cases:

- (a) Sampling from infinite populations
- (b) Sampling with replacement from finite populations
- (c) Sampling without replacement from finite populations

#### 5.1.1.1 Sampling from Infinite Populations

Let us assume we have a population, with mean  $\mu$  and variance  $\sigma^2$ , which is infinitely large. If we take a sample of size  $n$  with individual values  $x_1, x_2, \dots, x_n$ , then

$$\text{Sample Mean } (\bar{X}) = \frac{x_1 + x_2 + \dots + x_n}{n}$$

here  $x_1$  representing the first observed values in the sample, is a random variable since it may take any of the population values. Similarly  $x_2$ , representing the second observed value in sample is also a random variable since it may take any of the population values. In other words, we can say that  $x_i$ , representing the  $i^{\text{th}}$  observed value in the sample is a random variable.

Now when the population is infinitely large, whatever is the value of  $x_1$ , the distribution of  $x_2$  is not affected by it. This is true for any other pair of random variables as well. In other words;  $x_1, x_2, \dots, x_n$  are independent random variables and all are picked up from the same population.

$$\text{So } E(x_i) = \mu \quad \text{and} \quad \text{Var}(x_i) = \sigma^2 \quad \text{for } i = 1, 2, 3, \dots, n$$



Finally, we have

$$\begin{aligned}
 \mu_{\bar{x}} = E(\bar{X}) &= E\left(\frac{x_1 + x_2 + \dots + x_n}{n}\right) \\
 &= E\left(\frac{x_1}{n}\right) + E\left(\frac{x_2}{n}\right) + \dots + E\left(\frac{x_n}{n}\right) \quad [\text{as } E(A + B) = E(A) + E(B)] \\
 &= \frac{1}{n} E(x_1) + \frac{1}{n} E(x_2) + \dots + \frac{1}{n} E(x_n) \quad [\text{as } E(nA) = n E(A)] \\
 &= \frac{1}{n} \mu + \frac{1}{n} \mu + \dots + \frac{1}{n} \mu = \mu \quad \text{and} \\
 \sigma_{\bar{x}}^2 &= \text{Var}(\bar{X}) = \text{Var}\left(\frac{x_1 + x_2 + \dots + x_n}{n}\right) \\
 &= \text{Var}\left(\frac{x_1}{n}\right) + \text{Var}\left(\frac{x_2}{n}\right) + \dots + \text{Var}\left(\frac{x_n}{n}\right) \\
 &\quad [\text{as } \text{Var}(A + B) = \text{Var}(A) + \text{Var}(B)] \\
 &= \frac{1}{n^2} \text{Var}(x_1) + \frac{1}{n^2} \text{Var}(x_2) + \dots + \frac{1}{n^2} \text{Var}(x_n) \quad [\text{as } \text{Var}(nA) = n^2 \text{Var}(A)] \\
 &= \frac{1}{n^2} \sigma^2 + \frac{1}{n^2} \sigma^2 + \dots + \frac{1}{n^2} \sigma^2 \\
 &= \frac{\sigma^2}{n}
 \end{aligned}$$

$$\text{So, } \sigma_{\bar{x}} = SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

### 5.2.1.2 Sampling with Replacement from Finite Populations

The above results have been obtained under the assumption that the random variables  $x_1, x_2, \dots, x_n$  are independent. This assumption is valid when the population is infinitely large. It is also valid when the sampling is done with replacement, so that the population is back to the same form before the next sample member is picked up. Hence, if the sampling is done with replacement, we would again have:

$$\mu_{\bar{x}} = E(\bar{X}) = \mu \quad \text{and} \quad \sigma_{\bar{x}}^2 = \text{Var}(\bar{X}) = \frac{\sigma^2}{n} \quad \text{or} \quad \sigma_{\bar{x}} = SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$



### 5.1.1.3 Sampling without Replacement from Finite Populations

When sampling without replacement from a finite population, the probability distribution of the second random variable depends on what has been the outcome of the first pick and so on. In other words, the  $n$  random variables representing the  $n$  sample members do not remain independent, the expression for the variance of  $\bar{X}$  changes. The results in this case will be:

$$\mu_{\bar{x}} = E(\bar{X}) = \mu$$

$$\text{and } \sigma_{\bar{x}}^2 = \text{Var}(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} \quad \text{or} \quad \sigma_{\bar{x}} = S.D(\bar{X}) = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

By comparing these expressions with the ones derived above we find that the variance of  $\bar{X}$  is the same but further multiplied by a factor  $\frac{N-n}{N-1}$ . This factor is, therefore, known as the finite population multiplier or the correction factor. In practice, almost all the samples are picked up without replacement. Also, most populations are finite although they may be very large and so the variance of the mean should theoretically be found by using the expression given above. However, if the population size ( $N$ ) is large and consequently the sampling ratio ( $n/N$ ) small, then the finite population multiplier is close to 1 and is not used, thus treating large finite populations as if they were infinitely large. For example, if  $N = 100,000$  and  $n = 100$ , the finite population multiplier will be 0.9995, which is very close to 1 and the variance of the mean would, for all practical purposes, be the same whether the population is treated as finite or infinite. As a rule of that, the finite population multiplier may not be used if the sampling ratio ( $n/N$ ) is smaller than 0.05.

Above discussion on the sampling distribution of mean, presents two very important results, which we shall be using very often in statistical estimation and hypotheses testing. We have seen that the expected value of the sample mean is the same as the population mean. Similarly, that the variance of the sample mean is the variance of the population divided by the sample size (and multiplied by the correction factor when appropriate). The fact that the sampling distribution of  $\bar{X}$  has mean  $\mu$  is very important. It means that, *on the average*, the sample mean is equal to the population mean. The distribution of the statistic is *centered on* the parameter to be estimated, and this makes the statistic  $\bar{X}$  a good estimator of  $\mu$ . This fact will become clearer in the next lesson, where we will discuss estimators and their



properties. The fact that the standard deviation of  $\bar{X}$  is  $\sigma/\sqrt{n}$  means that as the sample size *increases*, the standard deviation of  $\bar{X}$  *decreases*, making  $\bar{X}$  more likely to be close to  $\mu$ . This is another desirable property of a good estimator, to be discussed in the next lesson.

If we take a large number of samples of size  $n$ , then the average value of the sample means tends to be close to the true population mean. On the other hand, if the sample size is increased then the variance of  $\bar{X}$  gets reduced and by selecting an appropriately large value of  $n$ , the variance of  $\bar{X}$  can be made as small as desired.

The standard deviation of  $\bar{X}$  is also called the *standard error of the mean*. It indicates the extent to which the observed value of sample mean can be away from the true value, due to sampling errors. For example, if the standard error of the mean is small, we may be reasonably confident that whatever sample mean value we have observed cannot be very far away from the true value.

Before discussing the shape of the sampling distribution of mean, let us verify the above results empirically, with the help of a simple example.

Consider a discrete uniform population consisting of the values 1, 2, and 3. If the random variable  $X$  represents these population values, its mean and variance is

$$\mu = \frac{\sum X_i}{N} = \frac{6}{3} = 2$$

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N} = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3} = \frac{2}{3}$$

### (a) Sampling with Replacement

If random samples of size  $n = 2$  are drawn with replacement from this population, we will have  $N^n = 3^2 = 9$  possible samples. These are shown in Box 5-1 along with the corresponding sample mean values, which vary from 1 to 3. The resulting distribution of  $\bar{X}$  is given below:

$\bar{X}$	:	1	1.5	2	2.5	3
$P(\bar{X})$	:	1/9	2/9	3/9	2/9	1/9



## Box 5-1

Sample No. 1 (1,1) $\bar{X} = 1$	Sample No. 2 (1,2) $\bar{X} = 1.5$	Sample No. 3 (1,3) $\bar{X} = 2$
Sample No. 4 (2,1) $\bar{X} = 1.5$	Sample No. 5 (2,2) $\bar{X} = 2$	Sample No. 6 (2,3) $\bar{X} = 2.5$
Sample No. 7 (3,1) $\bar{X} = 2$	Sample No. 8 (3,2) $\bar{X} = 2.5$	Sample No. 9 (3,3) $\bar{X} = 3$

Now we can find out the mean and variance of the sampling distribution, the necessary calculations are given in Table 8-1.

**Table 5-1 Calculations for  $\mu_{\bar{x}}$  and  $\sigma_{\bar{x}}^2$**

$\bar{X}$	$P(\bar{X})$	$X.P(\bar{X})$	$P(\bar{X})[\bar{X} - E(\bar{X})]^2$
1	1/9	1/9	1/9
1.5	2/9	3/9	2/36
2	3/9	6/9	0
2.5	2/9	5/9	2/36
3	1/9	3/9	1/9
	$\sum P(\bar{X}) = 1$	$\sum X.P(\bar{X}) = 2$	$\sum P(\bar{X})[\bar{X} - E(\bar{X})]^2 = 1/3$

So the mean of the sampling distribution,

$$\mu_{\bar{x}} = E(\bar{X}) = \sum X.P(\bar{X}) = 2 = \mu$$

and the variance of the sampling distribution,

$$\sigma_{\bar{x}}^2 = \text{Var}(\bar{X}) = \sum P(\bar{X})[\bar{X} - E(\bar{X})]^2 = 1/3 = \frac{2/3}{2} = \frac{\sigma^2}{n}$$

**(b) Sampling Without Replacement**

If random samples of size  $n = 2$  are drawn without replacement from this population, we will have  ${}^N P_n = {}^3 P_2 = 6$  possible samples. These are shown in Box 5-2 along with the corresponding sample mean values, which vary from 1.5 to 2.5.



Box 5-2

Sample No. 1 (1,2) $\bar{X} = 1.5$	Sample No. 2 (1,3) $\bar{X} = 2$	Sample No. 3 (2,1) $\bar{X} = 1.5$
Sample No. 4 (2,3) $\bar{X} = 2.5$	Sample No. 5 (3,1) $\bar{X} = 2$	Sample No. 6 (3,2) $\bar{X} = 2.5$

The resulting distribution of  $\bar{X}$  is given below:

$\bar{X}$	:	1.5	2	2.5
$P(\bar{X})$	:	2/6	2/6	2/6

Now we can find out the mean and variance of the sampling distribution, the necessary calculations are given in Table 5-2.

**Table 5-2 Calculations for  $\mu_{\bar{x}}$  and  $\sigma_{\bar{x}}^2$**

$\bar{X}$	$P(\bar{X})$	$X.P(\bar{X})$	$P(\bar{X}).[\bar{X} - E(\bar{X})]^2$
1.5	2/6	3/6	2/24
2	2/6	4/6	0
2.5	2/6	5/6	2/24
	$\sum P(\bar{X}) = 1$	$\sum X.P(\bar{X}) = 2$	$\sum P(\bar{X}).[\bar{X} - E(\bar{X})]^2 = 1/6$

So the mean of the sampling distribution,

$$\mu_{\bar{x}} = E(\bar{X}) = \sum X.P(\bar{X}) = 2 = \mu$$

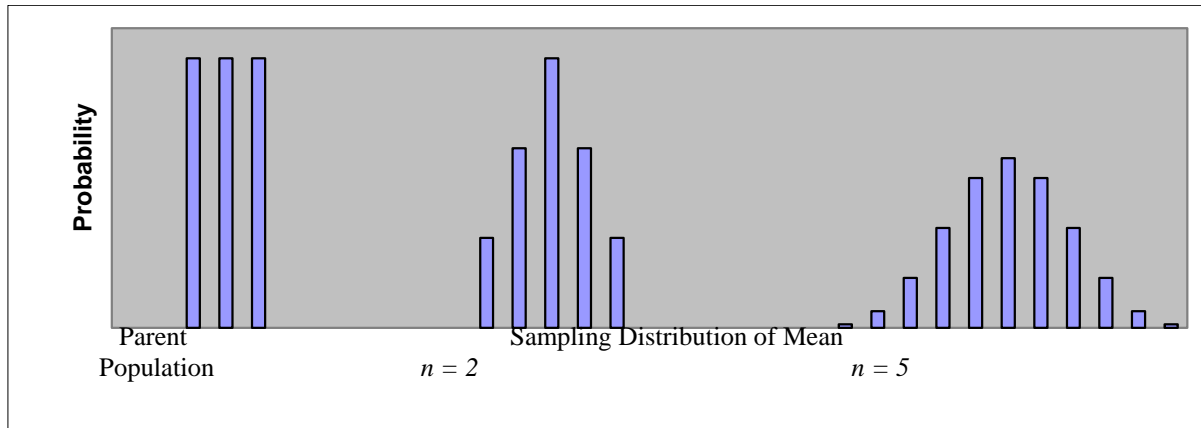
and the variance of the sampling distribution,

$$\sigma_{\bar{x}}^2 = Var(\bar{X}) = \sum P(\bar{X}).[\bar{X} - E(\bar{X})]^2 = 1/6 = \frac{2/3}{2} \cdot \frac{3-2}{3-1} = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

Now if we compare the shapes of the parent population and the resulting sampling distribution of mean, we find that although our parent population is uniformly distributed, the sampling distribution of mean is symmetrically distributed as shown in Figure 5-2.

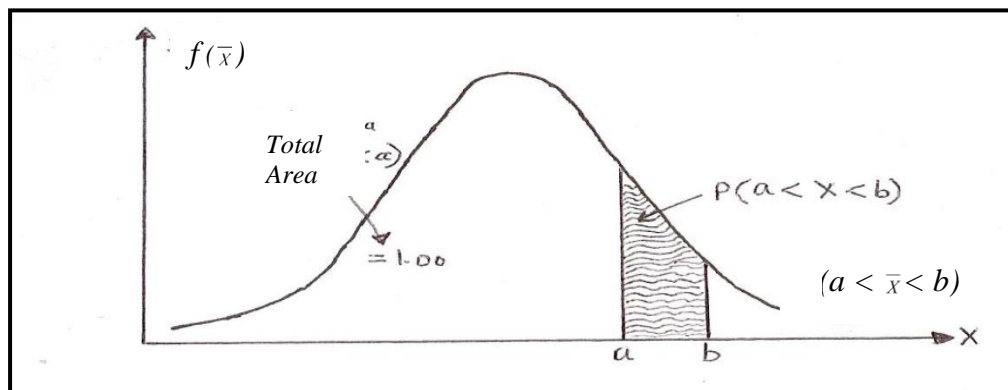
If we increase the sample size  $n$  we observe an interesting and important fact. As  $n$  increases.

- the possible values  $\bar{X}$  can assume increases, so the number of rectangles increases
- the probability that  $\bar{X}$  assumes a particular value decreases *i.e.* the width of rectangles decreases



**Figure 5-2 Parent Population and Sampling Distribution of Mean for  $n = 2$  and  $n = 5$**

In the limiting case when the sample size  $n$  increases infinitely, the particular values  $\bar{X}$  can assume approaches infinity and the probability that  $\bar{X}$  assumes a particular value approaches to zero. In other words, the limiting distribution of  $\bar{X}$  is normal distribution. Thus as  $n \rightarrow \infty$   $\bar{X} \sim N(\mu, \sqrt{\sigma^2/n})$



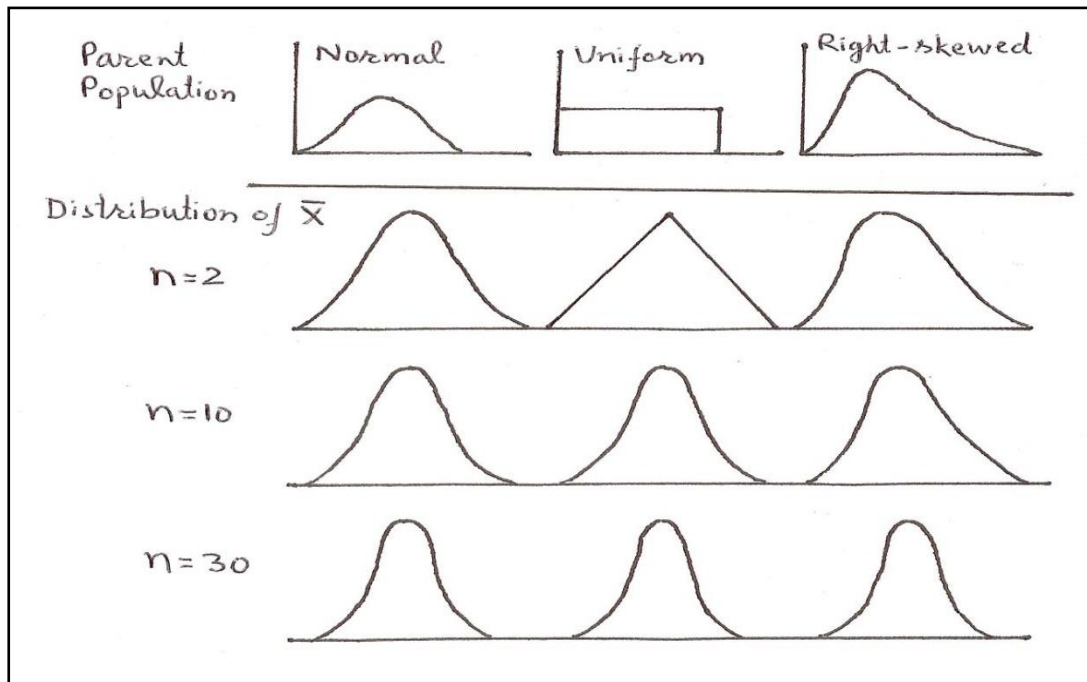
**Figure 5-3 Limiting Distribution of  $\bar{X}$**

### 5.1.2 THE CENTRAL LIMIT THEOREM

The result we just stated - *the limiting distribution of  $\bar{X}$  is the normal distribution* - is one of the most important results in statistics. It is popularly known as the *central limit theorem*. When sampling is done from a population with mean  $\mu$  and standard deviation  $\sigma$ , the sampling distribution of the sample mean  $\bar{X}$  tends to a normal distribution with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$  as the sample size  $n$  increases. For "Large Enough"  $n$ :  $\bar{X} \sim N(\mu, \sqrt{\sigma^2/n})$ . The central limit theorem is remarkable because it states that the distribution of the sample mean  $\bar{X}$  tends to a normal distribution *regardless of*



the distribution of the population from which the random sample is drawn. The theorem allows us to make probability statements about the possible range of values the sample mean may take. It allows us to compute probabilities of how far away  $\bar{X}$  may be from the population mean it estimates. We will extensively use the central limit theorem in the next two lessons about statistical estimation and testing of hypotheses.



**Figure 5-4 Sampling Distributions of  $\bar{X}$  for different Sample Sizes**

The central limit theorem says that, *in the limit*, as  $n$  goes to infinity ( $n \rightarrow \infty$ ), the distribution of  $\bar{X}$  becomes a normal distribution (regardless of the distribution of the population). The *rate at which* the distribution approaches a normal distribution does depend, however, on the shape of the distribution of the parent population:

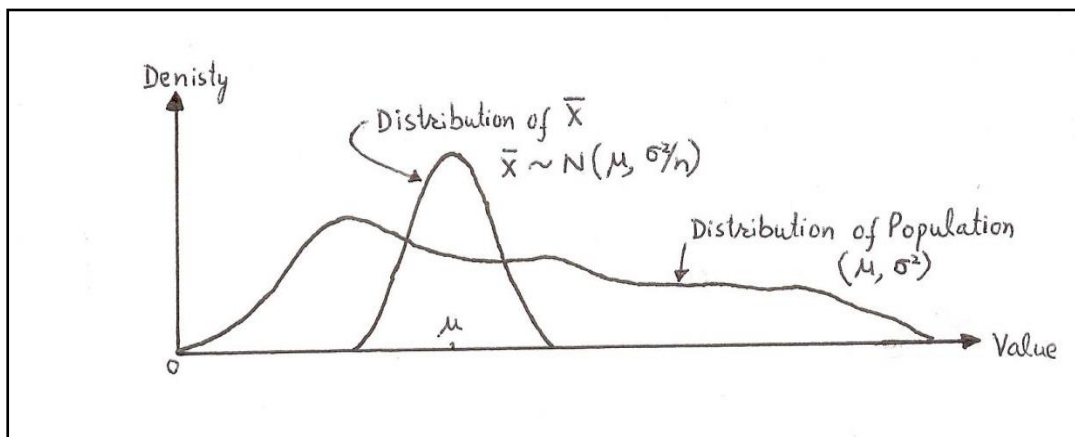
- if the population itself is normally distributed, the distribution of  $\bar{X}$  is normal for *any* sample size  $n$
- if the population distributions are very different from a normal distribution, a relatively large sample size is required to achieve a good normal approximation for the distribution of  $\bar{X}$

Figure 5-4 shows several parent population distributions and the resulting sampling distributions of  $\bar{X}$  for different sample sizes.

Since we often do not know the shape of the population distribution, it would be useful to have some general rule of thumb telling us when a sample is “Large Enough” that we may apply the central limit theorem:

*In general, a sample of 30 or more elements is considered “Large Enough” for the central limit theorem to be applicable.*

We emphasize that this is a *general*, and somewhat arbitrary, rule. A larger minimum sample size may be required for a good normal approximation when the population distribution is very different from a normal distribution. By the same reason, a smaller minimum sample size may suffice for a good normal approximation when the population distribution is close to a normal distribution.



**Figure 5-5 Population Distribution and the Sampling Distribution of  $\bar{X}$**

Figure 5-5 should help clarify the distinction between the population distribution and the sampling distribution of  $\bar{X}$ . The figure emphasizes the three aspects of the central limit theorem:

1. When the sample size is large enough, the sampling distribution of  $\bar{X}$  is normal
2. The expected value of  $\bar{X}$  is  $\mu$
3. The standard deviation of  $\bar{X}$  is  $\sigma/\sqrt{n}$

The last statement is the key to the important fact that as the sample size increases, the variation of  $\bar{X}$  about its mean  $\mu$  decreases. Stated another way, as we buy *more information* (take a larger sample), our *uncertainty* (measured by the standard deviation) about the parameter being estimated *decreases*.



### The History of the Central Limit Theorem

What we call the central limit theorem actually comprises several theorems developed over the years. The first such theorem was the discovery of the normal curve by Abraham De Moivre in 1733, when he discovered the normal distribution as the *limit of* the binomial distribution. The fact that the normal distribution appears as a limit of the binomial distribution as  $n$  increases is a form of the central limit theorem. Around the turn of the twentieth century, Liapunov gave a more general form of the central limit theorem, and in 1922 Lindeberg gave the final form we use in applied statistics. In 1935, W Feller gave the proof of the necessary condition of the theorem.

Let us now look at an example of the use of the central limit theorem.

#### Example 5-1

ABC Tool Company makes *Laser XR*; a special engine used in speedboats. The company's engineers believe that the engine delivers an average power of 220 horsepower, and that the standard deviation of power delivered is 15 horsepower. A potential buyer intends to sample 100 engines (each engine to be run a single time). What is the probability that the sample mean  $\bar{X}$  will be less than 217 horsepower?

**Solution:** Given that:

Population mean	$\mu = 220$ horsepower
Population standard deviation	$\sigma = 15$ horsepower
Sample size	$n = 100$

Here our random variable  $\bar{X}$  is normal (or at least approximately so, by the central limit theorem as our sample size is large).

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

$$\text{or } \bar{X} \sim N(220, \sqrt{15^2/100})$$

So we can use the standard normal variable  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  to find the required probability,

$$P(\bar{X} < 217) = P(Z < \frac{217 - 220}{15/\sqrt{100}}) = P(Z < -2) = 0.0228$$

So there is a small probability that the potential buyer's tests will result in a sample mean less than 217 horsepower.



### 5.1.3 SAMPLING DISTRIBUTION OF THE PROPORTION

Let us assume we have a binomial population, with a proportion  $p$  of the population possesses a particular attribute that is of interest to us. This also implies that a proportion  $q (=1-p)$  of the population does not possess the attribute of interest. If we pick up a sample of size  $n$  with replacement and found  $x$  successes in the sample, the sample proportion of success ( $\bar{p}$ ) is given by  $\bar{p} = \frac{x}{n}$  in which  $x$  is a binomial random variable, the possible value of this random variable depends on the composition of the random sample from which  $\bar{p}$  is computed. The probability of  $x$  successes in the sample of size  $n$  is given by a binomial probability distribution, viz.

$$P(x) = {}^nC_x p^x q^{n-x}$$

Since  $\bar{p} = \frac{x}{n}$  and  $n$  is fixed (determined before the sampling) the distribution of the number of successes ( $x$ ) leads to the distribution of  $\bar{p}$ .

*The sampling distribution of  $\bar{p}$  is the probability distribution of all possible values the random variable  $\bar{p}$  may take when a sample of size  $n$  is taken from a specified population.*

The expected value and the variance of  $x$  i.e. number of successes in a sample of size  $n$  is known to be:

$$E(x) = np; \quad \text{Var}(x) = npq$$

Finally we have mean and variance of the sampling distribution of  $\bar{p}$

$$\mu_{\bar{p}} = E(\bar{p}) = E\left(\frac{x}{n}\right) = \frac{1}{n} E(x) = \frac{1}{n} \cdot np = p \quad \text{and}$$

$$\sigma_{\bar{p}}^2 = \text{Var}(\bar{p}) = \text{Var}\left(\frac{x}{n}\right) = \frac{1}{n^2} \cdot \text{Var}(x) = \frac{1}{n^2} \cdot npq = \frac{pq}{n} \quad \sigma_{\bar{p}} = SD(\bar{p}) = \sqrt{\frac{pq}{n}}$$

When sampling is without replacement, we can use the finite population correction factor, so sampling distribution of  $\bar{p}$  has its

Mean  $\mu_{\bar{p}} = p$

Variance  $\sigma_{\bar{p}}^2 = \frac{pq}{n} \cdot \left(\frac{N-n}{N-1}\right)$



Standard deviation  $\sigma_{\bar{p}} = \sqrt{\frac{pq}{n} \cdot \frac{N-n}{N-1}}$

As the sample size  $n$  increases, the central limit theorem applies here as well. The *rate at which* the distribution approaches a normal distribution does depend, however, on the shape of the distribution of the parent population.

- if the population is symmetrically distributed, the distribution of  $\bar{p}$  approaches the normal distribution relatively fast
- if the population distributions are very different from a symmetrical distribution, a relatively large sample size is required to achieve a good normal approximation for the distribution of  $\bar{p}$

In order to use the normal approximation for the sampling distribution of  $\bar{p}$ , the sample size needs to be large. A commonly used rule of thumb says that the normal approximation to the distribution of  $\bar{p}$  may be used only if *both  $np$  and  $nq$  are greater than 5*. We now state the central limit theorem when sampling for the population proportion  $\bar{p}$ .

*When sampling is done from a population with proportion  $p$ , the sampling distribution of the sample proportion  $\bar{p}$  approaches to a normal distribution with proportion  $p$  and standard deviation  $\sqrt{pq/n}$  as the sample size  $n$  increases.*

For "Large Enough"  $n$ :  $\bar{p} \sim N(p, \sqrt{pq/n})$

The estimated standard deviation of  $\bar{p}$  is also called its *standard error*. We demonstrate the use of the theorem in Example 10-2

### **Example 5-2**

A manufacturer of screws has noticed that on an average 0.02 proportion of screws produced are defective. A random sample of 400 screws is examined for the proportion of defective screws. Find the probability that the proportion of the defective screws ( $\bar{p}$ ) in the sample is between 0.01 and 0.03?

#### **Solution:**

Given that:

Population proportion  $p = 0.02$



So

$$q = 0.08 (= 1-0.02)$$

Sample size

$$n = 400$$

Since the population is infinite and also the sample size is large, the central limit theorem applies. So

$$\bar{p} \sim N(p, \sqrt{pq/n})$$

$$\bar{p} \sim N(0.02, \sqrt{(0.02)(0.08)/400})$$

We can find the required probability using standard normal variable  $Z = \left( \frac{\bar{p} - p}{\sqrt{pq/n}} \right)$

$$\begin{aligned} P(0.01 < \bar{p} < 0.03) &= P\left( \frac{0.01 - 0.02}{\sqrt{\frac{(0.02)(0.08)}{400}}} < Z < \frac{0.03 - 0.02}{\sqrt{\frac{(0.02)(0.08)}{400}}} \right) \\ &= P\left( \frac{-0.01}{0.007} < Z < \frac{0.01}{0.007} \right) \\ &= P(-1.43 < Z < 1.43) \\ &= 2 P(0 < Z < 1.43) \\ &= 0.8472 \end{aligned}$$

So there is a very high probability that the sample will result in a proportion between 0.01 and 0.03.

#### 5.1.4 Sampling Distribution of the Difference of Sample Means

In order to bring out the sampling distribution of the difference of sample means, let us assume we have two populations labeled as 1 and 2. So that

$\mu_1$  and  $\mu_2$  denote the two population means.

$\sigma_1$  and  $\sigma_2$  denote the two population standard deviations

$n_1$  and  $n_2$  denote the two sample sizes

$\bar{X}_1$  and  $\bar{X}_2$  denote the two sample means

Let us consider independent random sampling from the populations so that the sample sizes need not be same for both populations.

Since  $\bar{X}_1$  and  $\bar{X}_2$  are random variables so is their difference  $\bar{X}_1 - \bar{X}_2$ . As a random variable,  $\bar{X}_1 - \bar{X}_2$  has a *probability distribution*. This probability distribution is the sampling distribution of  $\bar{X}_1 - \bar{X}_2$ .



The sampling distribution of  $\bar{X}_1 - \bar{X}_2$  is the probability distribution of all possible values the random variable  $\bar{X}_1 - \bar{X}_2$  may take when independent samples of size  $n_1$  and  $n_2$  are taken from two specified populations.

### Mean and Variance of $\bar{X}_1 - \bar{X}_2$

$$\mu_{\bar{X}_1 - \bar{X}_2} = E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2 \quad \text{and}$$

$$\begin{aligned} \sigma_{\bar{X}_1 - \bar{X}_2}^2 &= \text{Var}(\bar{X}_1 - \bar{X}_2) = \text{Var}(\bar{X}_1) + \text{Var}(\bar{X}_2) \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}; \text{ when sampling is with replacement} \\ &= \frac{\sigma_1^2}{n_1} \cdot \left( \frac{N_1 - n_1}{N_1 - 1} \right) + \frac{\sigma_2^2}{n_2} \cdot \left( \frac{N_2 - n_2}{N_2 - 1} \right); \text{ when sampling is without replacement} \end{aligned}$$

As the sample sizes  $n_1$  and  $n_2$  increases, the central limit theorem applies here as well. So we state the central limit theorem when sampling for the difference of population means  $\bar{X}_1 - \bar{X}_2$

When sampling is done from two populations with means  $\mu_1$  and  $\mu_2$  and standard deviations  $\sigma_1$  and  $\sigma_2$  respectively, the sampling distribution of the difference of sample means  $\bar{X}_1 - \bar{X}_2$  approaches to a normal distribution with mean  $\mu_1 - \mu_2$  and standard deviation  $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$  as the sample sizes  $n_1$  and  $n_2$  increases.

$$\text{For "Large Enough" } n_1 \text{ and } n_2: \bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

The estimated standard deviation of  $\bar{X}_1 - \bar{X}_2$  is also called its *standard error*. We demonstrate the use of the theorem in Example 10-3.

### **Example 5-3**

The makers of Duracell batteries claims that the size AA battery lasts on an average of 45 minutes longer than Duracell's main competitor, the Energizer. Two independent random samples of 100 batteries of each kind are selected. Assuming  $\sigma_1 = 84$  minutes and  $\sigma_2 = 67$  minutes, find the probability



that the difference in the average lives of Duracell and Energizer batteries based on samples does not exceed 54 minutes.

**Solution:** Given that:

$$\mu_1 - \mu_2 = 45$$

$$\sigma_1 = 84 \text{ and } \sigma_2 = 67$$

$$n_1 = 100 \text{ and } n_2 = 100$$

Let  $\bar{X}_1$  and  $\bar{X}_2$  denote the two sample average lives of Duracell and Energizer batteries respectively.

Since the population is infinite and also the sample sizes are large, the central limit theorem applies.

$$i.e \quad \bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

$$\bar{X}_1 - \bar{X}_2 \sim N(45, \sqrt{\frac{84^2}{100} + \frac{67^2}{100}})$$

So we can find the required probability using standard normal variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$\begin{aligned} \text{So } P(\bar{X}_1 - \bar{X}_2 < 54) &= P(Z < \frac{54 - 45}{\sqrt{\frac{84^2}{100} + \frac{67^2}{100}}}) \\ &= P(Z < 0.84) = 1 - 0.20045 = 0.79955 \end{aligned}$$

So there is a very high probability that the difference in the average lives of Duracell and Energizer batteries based on samples does not exceed 54 minutes.

### 5.1.5 Sampling Distribution of the Difference of Sample Proportions

Let us assume we have two binomial populations labeled as 1 and 2. So that

$p_1$  and  $p_2$  denote the two population proportions

$n_1$  and  $n_2$  denote the two sample sizes

$\bar{p}_1$  and  $\bar{p}_2$  denote the two sample proportions





Let us consider independent random sampling from the populations so that the sample sizes need not be same for both populations.

Since  $\bar{p}_1$  and  $\bar{p}_2$  are random variables so is their difference  $\bar{p}_1 - \bar{p}_2$ . As a random variable,  $\bar{p}_1 - \bar{p}_2$  has a *probability distribution*. This probability distribution is the sampling distribution of  $\bar{p}_1 - \bar{p}_2$ .

*The sampling distribution of  $\bar{p}_1 - \bar{p}_2$  is the probability distribution of all possible values the random variable  $\bar{p}_1 - \bar{p}_2$  may take when independent samples of size  $n_1$  and  $n_2$  are taken from two specified binomial populations.*

### Mean and Variance of $\bar{p}_1 - \bar{p}_2$

$$\mu_{\bar{p}_1 - \bar{p}_2} = E(\bar{p}_1 - \bar{p}_2) = E(\bar{p}_1) - E(\bar{p}_2) = p_1 - p_2$$

$$\sigma_{\bar{p}_1 - \bar{p}_2}^2 = \text{Var}(\bar{p}_1 - \bar{p}_2) = \text{Var}(\bar{p}_1) + \text{Var}(\bar{p}_2)$$

$$= \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}; \text{ when sampling is with replacement}$$

$$= \frac{p_1 q_1}{n_1} \cdot \left( \frac{N_1 - n_1}{N_1 - 1} \right) + \frac{p_2 q_2}{n_2} \cdot \left( \frac{N_2 - n_2}{N_2 - 1} \right); \text{ when sampling is without replacement}$$

As the sample sizes  $n_1$  and  $n_2$  increases, the central limit theorem applies here as well. So we state the central limit theorem when sampling for the difference of population proportions  $\bar{p}_1 - \bar{p}_2$

*When sampling is done from two populations with proportions  $p_1$  and  $p_2$  respectively, the sampling distribution of the difference of sample proportions  $\bar{p}_1 - \bar{p}_2$  approaches to a normal distribution with*

*mean  $p_1 - p_2$  and standard deviation  $\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$  as the sample sizes  $n_1$  and  $n_2$  increases.*

$$\text{For "Large Enough" } n_1 \text{ and } n_2: \quad \bar{p}_1 - \bar{p}_2 \sim N(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}})$$

The estimated standard deviation of  $\bar{p}_1 - \bar{p}_2$  is also called its *standard error*. We demonstrate the use of the theorem in Example 10-4.

### Example 5-4



It has been experienced that proportions of defaulters (in tax payments) belonging to business class and professional class are 0.20 and 0.15 respectively. The results of a sample survey are:

	Business class	Professional class
Sample size:	$n_1 = 400$	$n_2 = 420$
Proportion of defaulters:	$\bar{p}_1 = 0.21$	$\bar{p}_2 = 0.14$

Find the probability of drawing two samples with a difference in the two sample proportions larger than what is observed.

**Solution:** Given that:

$$\begin{aligned}
 p_1 &= 0.20 & p_2 &= 0.15 \\
 q_1 &= 1 - 0.20 = 0.80 & q_2 &= 1 - 0.15 = 0.85 \\
 n_1 &= 400 & n_2 &= 420 \\
 \bar{p}_1 &= 0.21 & \bar{p}_2 &= 0.14
 \end{aligned}$$

Since the population is infinite and also the sample sizes are large, the central limit theorem applies. *i.e.*

$$\begin{aligned}
 \bar{p}_1 - \bar{p}_2 &\sim N(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}) \\
 \bar{p}_1 - \bar{p}_2 &\sim N(0.05, \sqrt{\frac{(0.20)(0.80)}{400} + \frac{(0.15)(0.85)}{420}})
 \end{aligned}$$

So we can find the required probability using standard normal variable  $Z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$

$$\begin{aligned}
 P(\bar{p}_1 - \bar{p}_2 > 0.07) &= P(Z > \frac{0.07 - 0.05}{\sqrt{\frac{(0.20)(0.80)}{400} + \frac{(0.15)(0.85)}{420}}}) \\
 &= P(Z > 0.76) \\
 &= 0.22363
 \end{aligned}$$

So there is a low probability of drawing two samples with a difference in the two sample proportions larger than what is observed.



### 5.1.6 SMALL SAMPLING DISTRIBUTIONS

Up to now we were discussing the large sampling distributions in the sense that the various sampling distributions can be well approximated by a normal distribution for “Large Enough” sample sizes. In other words, the  $Z$ -statistic is used in statistical inference when sample size is large. It may, however, be appreciated that the sample size may be prohibited from being large either due to physical limitations or due to practical difficulties of sampling costs being too high. Consequently, for our statistical inferences, we may often have to contend ourselves with a small sample size and limited information. The consequences of the sample being small;  $n < 30$ ; are that

- the central limit theorem ceases to operate, and
- the sample variance  $S^2$  fails to serve as an unbiased estimator of  $\sigma^2$

Thus, the basic difference which the sample size makes is that while the sampling distributions based on large samples are approximately normal and sample variance  $S^2$  is an unbiased estimator of  $\sigma^2$ , the same does not occur when the sample is small.

It may be appreciated that the small sampling distributions are also known as exact sampling distributions, as the statistical inferences based on them are not subject to approximation. However, the assumption of population being normal is the basic qualification underlying the application of small sampling distributions.

In the category of small sampling distributions, the Binomial and Poisson distributions were already discussed in lesson 9. Now we will discuss three more important small sampling distributions – the chi-square, the  $F$  and the student  $t$ -distribution. The purpose of discussing these distributions at this stage is limited only to understanding the variables, which define them and their essential properties. The application of these distributions will be highlighted in the next two lessons.

The small sampling distributions are defined in terms of the concept of degrees of freedom. We will discuss this before concept proceeding further.

#### Degrees of Freedom ( $df$ )

The concept of degrees of freedom ( $df$ ) is important for many statistical calculations and probability distributions. We may define  $df$  associated with a sample statistic as *the number of observations contained in a set of sample data which can be freely chosen*. It refers to *the number of independent variables which vary freely without being influenced by the restrictions imposed by the sample statistic(s) to be computed*.



Let  $x_1, x_2, \dots, x_n$  be  $n$  observations comprising a sample whose mean  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  is a value known to us.

Obviously, we are free to assign any value to  $n-1$  observation out of  $n$  observations. Once the value are freely assigned to  $n-1$  observations, freedom to do the same for the  $n^{\text{th}}$  observation is lost and its value is automatically determined as

$n^{\text{th}}$  observation  $= n\bar{x} - \text{sum of } n-1 \text{ observations} = n\bar{x} - \sum_{i=1}^{n-1} x_i$  As the value of  $n^{\text{th}}$  observation must satisfy

the restriction  $\sum_{i=1}^n x_i = n\bar{x}$  We say that one degree of freedom,  $df$  is lost and the sum  $n\bar{x}$  of  $n$  observations has  $n-1$   $df$  associated with it.

For example, if the sum of four observations is 10, we are free to assign any value to three observations only, say,  $x_1 = 2, x_2 = 1$  and  $x_3 = 4$ . Given these values, the value of fourth observation is automatically determined as

$$x_4 = \sum_{i=1}^4 x_i - (x_1 + x_2 + x_3)$$

$$x_4 = 10 - (2 + 1 + 4)$$

$$x_4 = 3$$

Sampling essentially consists of defining various sample statistics and to make use them in estimating the corresponding population parameters. In this respect, degrees of freedom may be defined as *the number of  $n$  independent observations contained in a sample less the number of parameters  $m$  to be estimated on the basis of that sample information, i.e.  $df = n - m$ .*

For example, when the population variance  $\sigma^2$  is not known, it is to be estimated by a particular value of its estimator  $S^2$ ; the sample variance. The number of observations in the sample being  $n$ ,  $df = n - m = n - 1$  because  $\sigma^2$  is the only parameter (i.e.  $m = 1$ ) to be estimated by the sample variance.

### 5.1.7 SAMPLING DISTRIBUTION OF THE VARIANCE

We will now discuss the sampling distribution of the variance. We will first introduce the concept of the sample variance as an unbiased estimator of population variance and then present the chi-square distribution, which helps us in working out probabilities for the sample variance.



### 5.1.7.1 The Sample Variance

By now it is implicitly clear that we use the sample mean to estimate the population mean and sample proportion to estimate the population proportion, when those parameters are unknown. Similarly, we use a sample statistic called the sample variance to estimate the population variance.

As will see in the next lesson on Statistical Estimation a sample statistic is an unbiased estimator of the population parameter when the expected value of sample statistic is equal to the corresponding population parameter it estimates.

Thus, if we use the sample variance  $S^2$  as an unbiased estimator of population variance  $\sigma^2$  Then,  $E(S^2) = \sigma^2$

However, it can be shown empirically that while calculating  $S^2$  if we divide the sum of square of deviations from mean (SSD) i.e.  $\sum_{i=1}^n (x - \bar{x})^2$  by  $n$ , it will not be an unbiased estimator of  $\sigma^2$  and

$$E\left(\frac{\sum_{i=1}^n (x - \bar{x})^2}{n}\right) = \frac{n-1}{n} \sigma^2 = \sigma^2 - \frac{\sigma^2}{n} \quad \text{i.e.} \quad \frac{\sum_{i=1}^n (x - \bar{x})^2}{n} \text{ will underestimate the population variance } \sigma^2 \text{ by}$$

the factor  $\frac{\sigma^2}{n}$ . To compensate for this downward bias we divide  $\sum_{i=1}^n (x - \bar{x})^2$  by  $n-1$ , so that

$S^2 = \frac{\sum_{i=1}^n (x - \bar{x})^2}{n-1}$  is an unbiased estimator of population variance  $\sigma^2$  and we have:

$$E\left(\frac{\sum_{i=1}^n (x - \bar{x})^2}{n-1}\right) = \sigma^2$$

In other words *to get the unbiased estimator of population variance  $\sigma^2$ , we divide the sum  $\sum_{i=1}^n (x - \bar{x})^2$  by the degree of freedom  $n-1$ .*

### 5.1.7.2 The Chi-Square Distribution



Let  $X$  be a random variable representing  $N$  values of a population, which is normally distributed with mean  $\mu$  and variance  $\sigma^2$ , i. e.

$$X = \{X_1, X_2, \dots, X_N\}$$

We may draw a random sample of size  $n$  comprising  $x_1, x_2, \dots, x_n$  values from this population. As brought out in section 10.2, each of the  $n$  sample values  $x_1, x_2, \dots, x_n$  can be treated as an independent normal random variable with mean  $\mu$  and variance  $\sigma^2$ . In other words,

$$x_i \sim N(\mu, \sigma^2) \quad \text{where } i = 1, 2, \dots, n$$

Thus each of these  $n$  normally distribution random variable may be standardized so that

$$Z_i = \frac{x_i - \mu}{\sigma} \sim N(0, 1) \quad \text{where } i = 1, 2, \dots, n$$

A sample statistic  $U$  may, then, be defined as

$$U = Z_1^2 + Z_2^2 + \dots + Z_n^2$$

$$U = \sum_{i=1}^n Z_i^2$$

$$U = \sum_{i=1}^n \left( \frac{x_i - \mu}{\sigma} \right)^2$$

Which will take different values in repeated random sampling. Obviously,  $U$  is a random variable. It is called chi-square variable, denoted by  $\chi^2$ . Thus *the chi-square random variable is the sum of several independent, squared standard normal random variables*. The chi-square distribution is the probability distribution of chi-square variable. So, *The chi-square distribution is the probability distribution of the sum of several independent, squared standard normal random variables*. The chi-square distribution is defined as

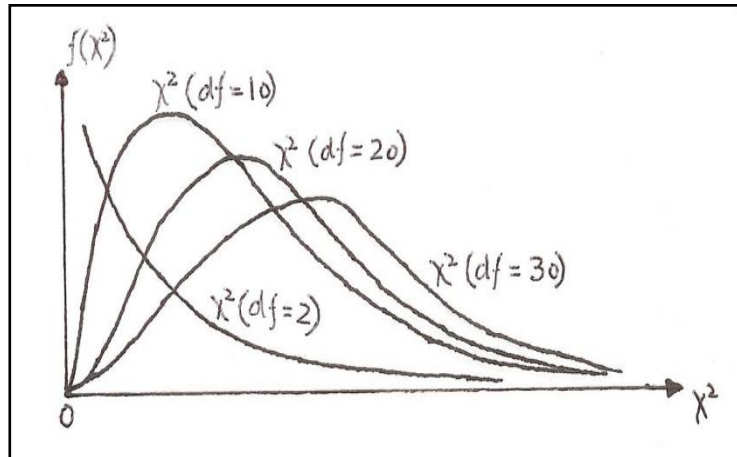
$$f(\chi^2) = C e^{-\frac{1}{2}\chi^2} (\chi^2)^{\frac{n}{2}-1} d\chi^2 \quad \text{for } \chi^2 \geq 0$$

where  $e$  is the base of natural logarithm,  $n$  denotes the sample size (or the number of independent normal random variables).  $C$  is a constant to be so determined that the total area under the  $\chi^2$  distribution is unity.  $\chi^2$  values are determined in terms of degrees of freedom,  $df = n$

### Properties of $\chi^2$ Distribution



1. A  $\chi^2$  distribution is completely defined by the number of degrees of freedom,  $df = n$ . So there are many  $\chi^2$  distributions each with its own  $df$ .
2.  $\chi^2$  is a sample statistic having no corresponding parameter, which makes  $\chi^2$  distribution a non-parametric distribution.
3. As a sum of squares the  $\chi^2$  random variable cannot be negative and is, therefore, bounded on the left by zero.



**Figure 5-6  $\chi^2$  Distribution with Different Numbers of  $df$**

4. The mean of a  $\chi^2$  distribution is equal to the degrees of freedom  $df$ . The variance of the distribution is equal to twice the number of degrees of freedom  $df$ .

$$E(\chi^2) = n; \text{ Var } (\chi^2) = 2n$$

5. Unless the  $df$  is large, a  $\chi^2$  distribution is skewed to the right. As  $df$  increases, the  $\chi^2$  distribution looks more and more like a normal. Thus for large  $df$

$$\chi^2 \sim N(n, \sqrt{2n}^2)$$

Figure 10-6 shows several  $\chi^2$  distributions with different numbers of  $df$ . In general, for  $n \geq 30$ , the probability of  $\chi^2$  taking a value greater than or less than a particular value can be approximated by using the normal area tables.

6. If  $\chi_1^2, \chi_2^2, \chi_3^2, \dots, \chi_k^2$  are  $k$  independent  $\chi^2$  random variables, with degrees of freedom  $n_1, n_2, n_3, \dots, n_k$ . Then their sum  $\chi_1^2 + \chi_2^2 + \chi_3^2 + \dots + \chi_k^2$  also possesses a  $\chi^2$  distribution with  $df = n_1 + n_2 + n_3 + \dots + n_k$ .

### 5.1.7.3 The $\chi^2$ Distribution in terms of Sample Variance $S^2$

We can write



$$\begin{aligned}
 \sum_{i=1}^n \left( \frac{x_i - \mu}{\sigma} \right)^2 &= \frac{1}{\sigma^2} \sum_{i=1}^n [(x_i - \bar{x}) + (\bar{x} - \mu)]^2 \\
 &= \frac{1}{\sigma^2} \sum_{i=1}^n [(x_i - \bar{x})^2 + (\bar{x} - \mu)^2 + 2(x_i - \bar{x})(\bar{x} - \mu)] \\
 &= \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{1}{\sigma^2} \sum_{i=1}^n (\bar{x} - \mu)^2 + \frac{2}{\sigma^2} (\bar{x} - \mu) \sum_{i=1}^n (x_i - \bar{x}) \\
 &= \frac{(n-1)S^2}{\sigma^2} + \left( \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \right)^2
 \end{aligned}$$

$$\left[ \text{since } \sum_{i=1}^n (x_i - \bar{x})^2 = (n-1)S^2 ; \sum_{i=1}^n (\bar{x} - \mu) = n(\bar{x} - \mu) \text{ and } \sum_{i=1}^n (x_i - \bar{x}) = 0 \right]$$

Now, we know that the LHS of the above equation is a random variable which has chi-square

distribution, with  $df = n$ . We also know that if  $\bar{x} \sim N(\mu, \sqrt{\sigma^2/n})$  Then  $\left( \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \right)^2$

will have a chi-square distribution with  $df = 1$ , Since the two terms on the RHS are independent,

$\frac{(n-1)S^2}{\sigma^2}$  will also has a chi-square distribution with  $df = n-1$ . One degree of freedom is lost because all

the deviations are measured from  $\bar{x}$  and not from  $\mu$ .

### Expected Value and Variance of $S^2$

In practice, therefore, we work with the distribution of  $\frac{(n-1)S^2}{\sigma^2}$  and not with the distribution of  $S^2$  directly.

Since  $\frac{(n-1)S^2}{\sigma^2}$  has a chi-square distribution with  $df = n-1$

$$\text{So } E\left[ \frac{(n-1)S^2}{\sigma^2} \right] = n-1$$

$$\frac{n-1}{\sigma^2} E(S^2) = n-1$$

$$E(S^2) = \sigma^2$$





$$\text{Also } \text{Var}\left[\frac{(n-1)S^2}{\sigma^2}\right] = 2(n-1)$$

Using the definition of variance, we get

$$E\left[\frac{(n-1)S^2}{\sigma^2} - E\left(\frac{(n-1)S^2}{\sigma^2}\right)\right]^2 = 2(n-1)$$

$$\text{or } E\left[\frac{(n-1)S^2}{\sigma^2} - (n-1)\right]^2 = 2(n-1)$$

$$\text{or } E\left[\frac{(n-1)^2 S^4}{\sigma^4} + (n-1)^2 - 2(n-1)\frac{(n-1)S^2}{\sigma^2}\right]^2 = 2(n-1)$$

$$\text{or } \frac{(n-1)^2}{\sigma^4} E[S^4 + \sigma^4 - 2S^2\sigma^2]^2 = 2(n-1)$$

$$\text{or } \frac{(n-1)^2}{\sigma^4} E(S^2 - \sigma^2)^2 = 2(n-1)$$

$$\text{or } E(S^2 - \sigma^2)^2 = \frac{2(n-1)}{(n-1)^2} \sigma^4$$

$$\text{So } \text{Var}(S^2) = \frac{2\sigma^4}{n-1}$$

It may be noted that the conditions necessary for the central limit theorem to be operative in the case of sample variance  $S^2$  are quite restrictive. For the sampling distribution of  $S^2$  to be approximately normal requires not only that the parent population is normal, but also that the sample is at least as large as 100.

### **Example 5-5**

In an automated process, a machine fills cans of coffee. The variance of the filling process is known to be 30. In order to keep the process in control, from time to time regular checks of the variance of the filling process are made. This is done by randomly sampling filled cans, measuring their amounts and computing the sample variance. A random sample of 101 cans is selected for the purpose. What is the probability that the sample variance is between 21.28 and 38.72?

**Solution:** We have Population variance  $\sigma^2 = 30$ ,  $n = 101$



We can find the required probability by using the chi-square distribution

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

$$\begin{aligned} \text{So, } P(21.28 < S^2 < 38.72) &= P\left(\frac{(101-1)21.28}{30} < \chi^2 < \frac{(101-1)38.72}{30}\right) \\ &= P(70.93 < \chi^2 < 109.06) \\ &= P(\chi^2 > 70.93) - P(\chi^2 > 109.06) \\ &\approx 0.990 - 0.025 = 0.965 \end{aligned}$$

Since our population is normal and also sample size is quite large, we can also estimate the required probability using normal distribution.

We have  $S^2 \sim \left(\sigma^2, \sqrt{\frac{2\sigma^4}{n-1}}\right)$

$$\begin{aligned} \text{So } P(21.28 < S^2 < 38.72) &= P\left(\frac{21.28 - \sigma^2}{\sqrt{\frac{2\sigma^4}{n-1}}} < Z < \frac{38.72 - \sigma^2}{\sqrt{\frac{2\sigma^4}{n-1}}}\right) \\ &= P\left(\frac{21.28 - 30}{\sqrt{\frac{2 \times 30 \times 30}{101-1}}} < Z < \frac{38.72 - 30}{\sqrt{\frac{2 \times 30 \times 30}{101-1}}}\right) \\ &= P\left(\frac{-8.72}{4.36} < Z < \frac{8.72}{4.36}\right) \\ &= P(-2 < Z < 2) \\ &= 2P(0 < Z < 2) \\ &= 2 \times 0.4772 = 0.9544 \end{aligned}$$

Which is approximately the same as calculated above using  $\chi^2$  distribution



## 5.2 The $f$ -Distribution and Analysis of Variance (ANOVA)

Let us assume two normal population with variances  $\sigma_1^2$  and  $\sigma_2^2$  repetitively. For a random sample of size  $n_1$  drawn from the first population, we have the chi-square variable.

$$\chi_1^2 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \text{ which process a } \chi^2 \text{ distribution with } v_1 = n_1 - 1 \text{ df}$$

Similarly, for a random sample of size  $n_2$  drawn from the second population, we have the chi-square variable

$$\chi_2^2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \text{ which process a } \chi^2 \text{ distribution with } v_2 = n_2 - 1 \text{ df}$$

A new sample statistic defined as  $F = \frac{\chi_1^2 / v_1}{\chi_2^2 / v_2}$

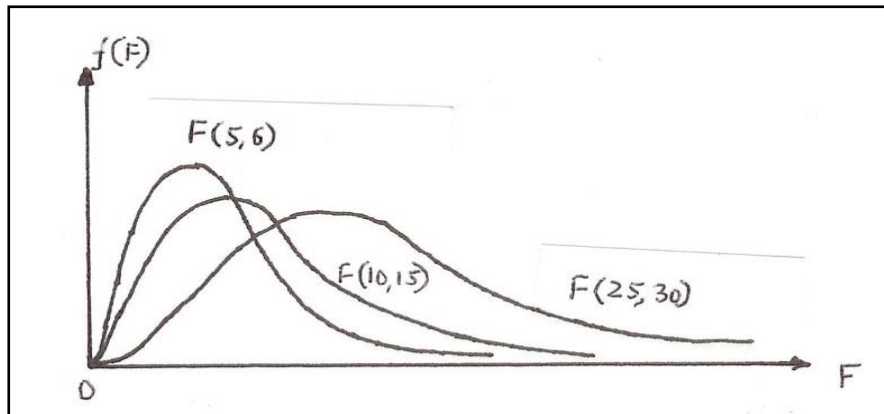
It is a random variable known as  **$F$  statistic**, named in honor of the English statistician Sir Ronald A Fisher.

Being a random variable it has a probability distribution, which is known as  **$F$  distribution**.

*The  $F$  distribution is the distribution of the ratio of two chi-square random variables that are independent of each other, each of which is divided by its own degrees of freedom.*

### Properties of $F$ - Distribution

1. The  $F$  distribution is defined by two kinds of degrees of freedom – the degrees of freedom of the numerator always listed as the first item in the parentheses and the degrees of freedom of the denominator always listed as the second item in the parentheses. So there are a large number of  $F$  distributions for each pair of  $v_1$  and  $v_2$ . Figure 10-7 shows several  $F$  distributions with different  $v_1$  and  $v_2$ .
2. As a ratio of two squared quantities, the  $F$  random variable cannot be negative and is, therefore, bounded on the left by zero.



**Figure 5-7 F- Distribution with different  $\nu_1$  and  $\nu_2$**

3. The  $F_{(\nu_1, \nu_2)}$  has no mean for  $\nu_2 \leq 2$  and no variance for  $\nu_2 \leq 4$ . However, for  $\nu_2 > 2$ , the mean and for  $\nu_2 > 4$ , the variance is given as

$$E(F_{(\nu_1, \nu_2)}) = \frac{\nu_2}{\nu_2 - 2} \quad \text{Var}(F_{(\nu_1, \nu_2)}) = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}$$

4. Unless the  $\nu_2$  is large, a  $F$  distribution is skewed to the right. As  $\nu_2$  increases, the  $F$  distribution looks more and more like a normal. In general, for  $\nu_2 \geq 30$ , the probability of  $F$  taking a value greater than or less than a particular value can be approximated by using the normal area tables.
5. The  $F$  distributions defined as  $F_{(\nu_1, \nu_2)}$  and as  $F_{(\nu_2, \nu_1)}$  are reciprocal of each other.

$$\text{i.e.} \quad F_{(\nu_1, \nu_2)} = \frac{1}{F_{(\nu_2, \nu_1)}}$$

### Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) is a statistical technique specially designed to test whether the means of more than two quantitative populations are equal. This technique is developed by R.A. Fisher in 1920's. It consists of classifying and cross-classifying statistical results and testing whether the means of a specified classification differ significantly. ANOVA is a method which separates the variations ascribable to one set of causes from the variations ascribable to other set of causes. In other words, analysis of variance is a method of splitting the total variations of a data into constituent parts in which measures different sources of variations. ANOVA enables us to analyze the total variations of data into components which may be attributed to various sources or causes of variation. The total variation is split up into the following two components:



- a) Variation within the subgroup of samples.
- b) Variations between the subgroups of the samples.

After obtaining the above two variations in viz., a) and b), these two variations are tested for their significance by F-test which is also known as Variance Ratio Test.

### Objects of Analysis of Variance

The first objects of analysis of variance is to obtain a measure of the total variation within the series and the second object is to find a measure of variation between or among the components. Then the test of significance of difference between the variations in two series or may be measured. In other words, with ANOVA technique, we can test the hypothesis that the means of all the components constituting a population are equal to the mean of the population or that the samples have come from the same population.

### Computation of Test Statistic

The actual analysis of variance is carried out on the basis of ratio between two variances. The variance ratio is obtained by the dividing the variance between the samples by the variance within the samples. This ratio forms the test statistic known as F-Statistic, i.e.,

$$F - \text{Statistic} = \frac{\text{Variance between the samples}}{\text{Variance within the samples}}$$

### Assumptions of Analysis of Variance (ANOVA)

The underlying assumptions for the study of analysis of variance are:

1. Each of the samples are a simple random sample.
2. Population from which the samples are selected is normally distributed. If however, the sample sizes are large enough, this assumption of normality is not required.
3. Each one of the samples are independent of other samples.
4. Each one of the populations has the same variance ( $\sigma^2_1 = \sigma^2_2 = \dots = \sigma^2_n$ ) and identical means ( $\mu_1 = \mu_2 = \mu_3 = \dots = \mu_n$ ).
5. The effect of various components are additive.

### Uses of ANOVA Table

The ANOVA table showing the source of variation, the sum of squares, degree of freedom, mean square (variance) and the formula for the F-ratio is known as ANOVA table. It is used to test whether the



means of a number of populations (more than two) are equal. We know that t-statistic is used for testing whether two population means are equal. Thus, the analysis of variance of test may be taken as an extension of t-test for the case of more than two population means.

### Classification of Analysis of Variance

The analysis of variance is mainly carried on under the following two classifications:

- a) One way Analysis of Variance or One way classification
- b) Two way Analysis of Variance or Two way classified Data or Manifold Classification

### One way Classification

In one way classification, the data are classified according to only one criterion. In this classification, the influence of only one attributes or factors considered. There are two methods of one way analysis of variance. They are:

- a) Direct Method
- b) Shortcut Method

### Direct Method

The following steps are required under the direct method of one way classification of analysis of variance (ANOVA):

#### 1. Set Null Hypothesis and Alternate Hypothesis:

**Null Hypothesis:** The means of the populations from which p samples are drawn equal to one another. The notation for Null hypothesis will be as:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

**Alternate Hypothesis:** At least two of the means of the populations are unequal or all the  $\mu_i$ 's are not equal. The notation for Alternate hypothesis will be as:

$$H_1: \mu_1 \neq \mu_2 \neq \dots \neq \mu_k$$

- #### 2. Calculate Variance Between the Samples:
- The variance between samples (groups) measures the difference between the sample means of each group and the overall mean weighted by the number of observations in each group. The variance between samples taken into account the random variations from observation to observation. It also measures difference from one group to another. The sum of squares between samples are denoted by SSC. For calculating variance between the samples we take the total of the square of the deviations of the means of various samples from the



grand average and divide this total by degree of freedom. Thus the steps in calculating variance between samples will be:

- Calculate the mean of each sample i.e.,  $\bar{X}_1, \bar{X}_2$ , etc.
- Calculate the grand average  $\bar{\bar{X}}$ , pronounced “X double bar”. Its value obtained as follows:

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \dots}{N_1 + N_2 + N_3 + \dots}$$

- Take the difference between the means of the various samples and the grand average
- Square these deviations and obtain the total which will give sum of squares between the samples
- Divide the total obtained in previous step by the degree of freedom. The degree of freedom will be one less than the number of samples, i.e., if there are 4 samples than the degree of freedom will be  $4 - 1 = 3$  or  $v = k - 1$ , where  $k$  = number of samples.

**3. Calculate Variance within the Samples:** The variance (or sum of squares) within samples measures those inter-sample differences due to chance only. It is denoted by SSE. The variance within samples (groups) measures variability around the mean of each group. For calculating variance within the samples, we take the total of the sum of squares of the deviation of various items from the mean values of the respective samples and divide this total by the degree of freedom. Thus steps involved in calculating variance within the samples will be:

- Calculate the mean of each sample i.e.,  $\bar{X}_1, \bar{X}_2, \bar{X}_3$ , etc.
- Take the deviations of various items in a sample from the mean values of the respective samples;
- Square these deviations and obtain the total which will give sum of squares within the samples;
- Divide the total obtained in previous step by the degree of freedom. The degree of freedom is obtained by deduction from the total number of items, the number of samples, i.e.,  $v = N - K$ , where  $K$  refers to the number of samples and  $N$  refers to total number of all the observations;

**4. Calculate the Ratio F as Follows:**

$$F - \text{Statistic} = \frac{\text{Variance between the samples}}{\text{Variance within the samples}}$$

Symbolically,

$$F = \frac{S_1^2}{S_2^2}$$



The F ratio measures the ratio of the variance between groups to the variance within groups. The variance between the sample means is the numerator and the variance within the sample means is the denominator. If there is no real difference from group to group, any sample difference will be explainable by random variation and the variance between groups should be close to the variance within groups. However, if there is a real difference between the groups, the variance between groups will be significantly larger than the variance within groups.

- 5. Compare the F value with Table value:** After calculation of F value, it is compared with the table value of F for the degrees of freedom at a certain significance (Generally at 5%) level. If calculated value of F is greater than the table value, the difference in sample means is significant. In other words, the samples do not come from the same population. If the calculated value of F is less than the table value, the difference is not significant and variation has arisen due to fluctuations of simple sampling.

The following ANOVA table summarize calculations for sums of squares, together with the r numbers of degrees of freedom and mean squares.

Source of variation	Sum of Squares	Degree of Freedom ( $\nu$ )	Mean Square	F
Between Samples	SSC	$\nu = c - 1$	$MSC = SSC/(c - 1)$	$F = MSC/MSE$
Within Samples	SSE	$\nu = n - c$	$MSE = SSE/(n - c)$	
Total	SST	$n - 1$		

SST = Total sum of squares of variations

SSC = Sum of squares between samples (columns)

SSE = Sum of squares within samples (rows)

MSC = Mean sum of squares between samples

MSE = Mean sum of squares within samples

The same procedure for analysis of variance is applicable for both the equal and unequal sample sizes.

The following example will illustrate the procedure:

#### **Example 5.6**





To assess the significance of possible variation in performance in a certain test between the convert schools of a city, a common test was given to a number of students taken at random from the senior fifth class of each of the four schools concerned. The results are given below. Make an analysis of variance of data.

Schools			
A	B	C	D
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15

**Solution:**

**Computation of Grand Mean**

Sample 1 $X_1$	Sample 2 $X_2$	Sample 3 $X_3$	Sample 4 $X_4$
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15
Total = 45	50	60	65
$\bar{X} = 9$	10	12	13

$$\text{Grand Mean or } \bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4}{N}$$

Where,  $\bar{X}_1, \bar{X}_2$ , etc., represents the mean of each sample and N the number of samples.

$$\text{Grand Mean or } \bar{\bar{X}} = \frac{9 + 10 + 12 + 13}{4} = \frac{44}{4} = 11$$

**Variation between samples**



To calculate variation between samples, calculate the square of the deviation of the various samples from the grand mean or average. The mean of sample 1 is 9 but the grand mean is 11. So, we will take the difference between 9 and 11 and square it. Similarly, other three samples means difference with grand mean calculated and squared. Thus we have the following table:

**Computation of variation between samples**

Sample 1 $(\bar{X}_1 - \bar{X})^2$	Sample 2 $(\bar{X}_2 - \bar{X})^2$	Sample 3 $(\bar{X}_3 - \bar{X})^2$	Sample 4 $(\bar{X}_4 - \bar{X})^2$
4	1	1	4
4	1	1	4
4	1	1	4
4	1	1	4
4	1	1	4
<b>20</b>	<b>5</b>	<b>5</b>	<b>20</b>

Sum of the squares between the samples =  $20 + 5 + 5 + 20 = 50$

Mean sum of squares between the samples is  $50/(4 - 1) = 16.7$  (because there are four samples and the degrees of freedom are  $4 - 1 = 3$ ).

### Variance within the Samples

Here, we find the sum of the squares is the deviations of various items in a sample from the mean values of respective samples. Thus, for the first sample, then mean is 9 and so we will take deviations from 10 and so on. The squared deviations are given in the following table:

**Computation of Variance within the sample**

Sample 1		Sample 2		Sample 3		Sample 4	
$X_1$	$(X_1 - \bar{X}_1)^2$	$X_2$	$(X_2 - \bar{X}_2)^2$	$X_3$	$(X_3 - \bar{X}_3)^2$	$X_4$	$(X_4 - \bar{X}_4)^2$
8	1	12	4	18	36	13	0
10	1	11	1	12	0	9	16
12	9	9	1	16	16	12	1
8	1	14	16	6	66	16	9
7	4	4	36	8	16	15	4
	$\sum(X_1 - \bar{X}_1)^2$ = 16		$\sum(X_2 - \bar{X}_2)^2$ = 58		$\sum(X_3 - \bar{X}_3)^2$ = 104		$\sum(X_4 - \bar{X}_4)^2$ = 30



Total sum of squares within the samples =  $16 + 58 + 104 + 30 = 208$

Mean sum of squares within the samples =  $208/20 - 4 = 208/16 = 13$

It is advisable to check up the calculations by finding out total variation. Total variation is calculated by taking the squares of the deviation of each item from the general or grand mean or average.

### Computation of Total Variation

Sample 1		Sample 2		Sample 3		Sample 4	
$X_1$	$(X_1 - \bar{X})^2$	$X_2$	$(X_2 - \bar{X})^2$	$X_3$	$(X_3 - \bar{X})^2$	$X_4$	$(X_4 - \bar{X})^2$
8	9	12	1	18	49	13	4
10	1	11	0	12	1	9	4
12	1	9	4	16	25	12	1
8	9	14	9	6	25	16	25
7	16	4	49	8	9	15	16
$\Sigma(X_1 - \bar{X})^2 = 36$		$\Sigma(X_2 - \bar{X})^2 = 63$		$\Sigma(X_3 - \bar{X})^2 = 109$		$\Sigma(X_4 - \bar{X})^2 = 50$	

Total sum of squares =  $36 + 36 + 109 + 50 = 258$

Degree of freedom =  $20 - 1 = 19$

When we add the sum of squares between samples and sum of squares within samples, we get the same total, i.e.,  $50 + 208 = 258$ . Hence, our calculations are right.

All the above results can be tabulated as follows:

### Summary of above Results

Source of variation	Sum of Squares	Degree of Freedom	Mean Square
Between Samples	50	3	16.7
Within Samples	208	16	13.0
Total	258	19	

$$F \text{ Value} = \frac{\text{Variance between the samples}}{\text{Variance within the samples}} = \frac{16.7}{13} = 1.285$$

The table value of F for  $v_1 = 3$  and  $v_2 = 16$  degree of freedom at 5% significance level = 3.24. The calculated value of F is less than the table value and hence the difference in the mean values of the sample is not significant, i.e., the samples could have come from the same universe.

### Short cut Method

The above method of calculating variance between the samples and variance within the samples are cumbersome or difficult. Generally, this method is not in practice because it is time consuming process. An easier method known as short cut method is usually followed which reduces considerably the computational work. The computations by the short cut method shall be as follows:



## Computation by Short cut Method

Sample 1		Sample 2		Sample 3		Sample 4	
$X_1$	$X_1^2$	$X_2$	$X_2^2$	$X_3$	$X_3^2$	$X_4$	$X_4^2$
8	64	12	144	18	324	13	169
10	100	11	121	12	144	9	81
12	144	9	81	16	256	12	144
8	64	14	196	6	36	16	256
7	49	4	16	8	64	15	225
$\sum X_1$ = 45	$\sum X_1^2$ = 421	$\sum X_2$ = 50	$\sum X_2^2$ = 558	$\sum X_3$ = 60	$\sum X_3^2$ = 824	$\sum X_4$ = 65	$\sum X_4^2$ = 875

The sum of all the items of various samples =  $\sum X_1 + \sum X_2 + \sum X_3 + \sum X_4$   
 $= 45 + 50 + 60 + 65 = 220$

Correction factor =  $T^2 / N = (220)^2 / 20 = 48400 / 20 = 2420$ .

Total sum of squares =  $\sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \sum X_4^2 - \text{Correction factor}$   
 $= 421 + 558 + 824 + 875 - 2420$   
 $= 2678 - 2420 = 258 \text{ (as above)}$

Sum of squares between the samples

$$\begin{aligned}
 &= \frac{(\sum X_1)^2}{N} + \frac{(\sum X_2)^2}{N} + \frac{(\sum X_3)^2}{N} + \frac{(\sum X_4)^2}{N} - \frac{T^2}{N} \\
 &= \frac{(45)^2}{5} + \frac{(50)^2}{5} + \frac{(60)^2}{5} + \frac{(65)^2}{5} - 2420 \\
 &= \frac{2025}{5} + \frac{2500}{5} + \frac{3600}{5} + \frac{4225}{5} - 2420 \\
 &= \frac{12350}{5} - 2420 = 2470 - 2420 = 50 \text{ (as before)}
 \end{aligned}$$

Sum of squares within the samples

$$\begin{aligned}
 &= \text{Total sum of squares} - \text{sum of squares between samples} \\
 &= 258 - 50 = 208 \text{ (as before)}
 \end{aligned}$$

**Coding of Data:** It refers to the addition, multiplication, subtraction or division of data by a constant. While making analysis of variance, it should be noted that the final quantity tested is a ratio. This means that the original measurement can be coded to simplify calculations without the need for any subsequent adjustments of the results. The previous example 10.6 would illustrate the point as follows:



Let us take 10 as common for the question take in example 10.6. the coded data are given below and the calculations are done therefrom:

### Coded Data

A (X <sub>1</sub> )	B (X <sub>2</sub> )	C (X <sub>3</sub> )	D (X <sub>4</sub> )
-2	+2	+8	+3
0	+1	+2	-1
+2	-1	+6	+2
-2	+4	-4	+6
-3	-6	-2	+5
$\sum X_1 = 45$	$\sum X_2 = 0$	$\sum X_3 = 10$	$\sum X_4 = 15$
$\bar{X} = 1$	0	2	3

$$\text{Grand Mean or } \bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4}{N} = \frac{-1 + 0 + 2 + 3}{4} = 1$$

### Sum of Squares between samples:

#### Computation of variation between samples

$(\bar{X}_1 - \bar{\bar{X}})^2$	$(\bar{X}_2 - \bar{\bar{X}})^2$	$(\bar{X}_3 - \bar{\bar{X}})^2$	$(\bar{X}_4 - \bar{\bar{X}})^2$
4	1	1	4
4	1	1	4
4	1	1	4
4	1	1	4
4	1	1	4
$\sum (\bar{X}_1 - \bar{\bar{X}})^2 = 20$	$\sum (\bar{X}_2 - \bar{\bar{X}})^2 = 5$	$\sum (\bar{X}_3 - \bar{\bar{X}})^2 = 5$	$\sum (\bar{X}_4 - \bar{\bar{X}})^2 = 20$

Sum of squares between samples = 20 + 5 + 5 + 20 = 50.

Mean squares between samples = 50 / (4 - 1) = 50/3 = 16.7 (as before)

### Sum of Square within samples

#### Computation of Variance within the sample

Sample 1		Sample 2		Sample 3		Sample 4	
X <sub>1</sub>	$(X_1 - \bar{X}_1)^2$	X <sub>2</sub>	$(X_2 - \bar{X}_2)^2$	X <sub>3</sub>	$(X_3 - \bar{X}_3)^2$	X <sub>4</sub>	$(X_4 - \bar{X}_4)^2$
-2	1	+2	4	+8	36	+3	0
0	1	+1	1	+2	0	-1	16
+2	9	-1	1	+6	16	+2	1
-2	1	+4	16	-4	66	+6	9
-3	4	-6	36	-2	16	+5	4
	$\sum (X_1 - \bar{X}_1)^2$		$\sum (X_2 - \bar{X}_2)^2$		$\sum (X_3 - \bar{X}_3)^2$		$\sum (X_4 - \bar{X}_4)^2$



	= 16		= 58		= 104		= 30
--	------	--	------	--	-------	--	------

Total sum of squares within the samples =  $16 + 58 + 104 + 30 = 208$

Mean squares within the samples =  $208/20-4 = 208/16 = 13$  (as before)

### Two Way Analysis of Variance (ANOVA)

In a manifold classification, we consider two or more characteristics or attributes. When it is believed that two independent factors might have an effect on the response variable of interest, it is possible to design the test so that an analysis of variance can be used to test for the effects of the two factors simultaneously. Such a test is called a two factor analysis of variance. With two factor analysis of variance, we can test two sets of hypothesis with the same data at the same time.

In a two way classification the data are classified according to two different criteria or factors. The procedure for analysis of variance is somewhat different than the one followed while dealing with problems of one way classification. In a two way classification the ANOVA table takes the following form:

Source of variation	Sum of Squares	Degree of Freedom	Mean Square	F Ratio
Between Samples	SSC	$c - 1$	$MSC = SSC/(c - 1)$	$MSC/MSE$
Between Rows	SSR	$r - 1$	$MSR = SSR/(r - 1)$	$MSR/MSE$
Residual or error	SSE	$(c - 1)(r - 1)$	$MSE = SSE/(r - 1)(c - 1)$	
Total	SST	$n - 1$		

SSC = Sum of squares between columns

SSR = Sum of squares between rows

SSE = Sum of squares due to error

SST = Total sum of squares

The sum of squares for the source 'Residual' is obtained by subtracting from the total sum of squares, the sum of squares between columns and rows, i.e.,  $SSE = SST - (SSC + SSR)$ .

The total number of degree of freedom =  $n - 1$  or  $cr - 1$ , where  $c$  refers to the number of columns and  $r$  refers to number of rows. So,

Number of degree of freedom between columns =  $(c - 1)$

Number of degree of freedom between rows =  $(r - 1)$



Number of degree of freedom for residual =  $(c - 1)(r - 1)$

The total sum of squares, sum of squares for between columns and sum of squares for between rows are obtained in the same way as before.

Residual or error sum of square = Total sum of squares – Sum of squares between columns – Sum of squares between rows.

The F value are calculated as follows:

$$F \text{ value} = \frac{MSC}{MSE}$$

In which  $v_1 = (c - 1)$  and  $v_2 = (c - 1)(r - 1)$

$$F \text{ value} = \frac{MSR}{MSE}$$

Where,  $v_1 = (r - 1)$  and  $v_2 = (c - 1)(r - 1)$

It should be carefully noted that  $v_1$  may not be same in both cases - in one case  $v_1 = (c - 1)$  and another case  $v_2 = (r - 1)$ .

The calculated values of F are compared with the table values. If calculated value of F is greater than the table value at pre-assigned level of significance, the null hypothesis is rejected, otherwise accepted. It would be clear from above that in problem involving two way classification, Residual is the measuring rod for testing significance. It represents the magnitude of variation due to forces called 'chance'. The following examples would illustrate the procedure:

### **Example 10.7**

Perform a two-way ANOVA on the data given below:

Plots of Land	Treatment			
	A	B	C	D
I	38	40	41	39
II	45	42	49	36
III	40	38	42	42

(Use coding method subtracting 40 from the given numbers)



**Solution:** Let us take the hypothesis that there is no significant difference in the treatment and plots of land. Applying analysis of variance technique.

On subtracting 40 from each value, we get

Plots of Land	Treatment				Total
	A	B	C	D	
I	-2	0	+1	-1	-2
II	+5	+2	+9	-4	+12
III	0	-2	+2	+2	+2
Total	+3	0	+12	-3	+12

$$\text{Correction factor} = \frac{T^2}{N} = \frac{(12)^2}{12} = \frac{144}{12} = 12$$

**Sum of squares between treatments:**

$$\begin{aligned}
 &= \frac{(3)^2}{3} + \frac{(0)^2}{3} + \frac{(12)^2}{3} + \frac{(-3)^2}{3} - \frac{T^2}{N} \\
 &= \frac{9}{3} + \frac{0}{3} + \frac{144}{3} + \frac{9}{3} - 12 \\
 &= 3 + 0 + 48 + 3 - 12 = 42
 \end{aligned}$$

Degree of freedom = 4 – 1 = 3

**Sum of squares between plot of land:**

$$\begin{aligned}
 &= \frac{(-2)^2}{3} + \frac{(12)^2}{3} + \frac{(2)^2}{3} - \frac{T^2}{N} \\
 &= \frac{4}{3} + \frac{144}{3} + \frac{4}{3} - 12 \\
 &= 1 + 36 + 1 - 12 = 26
 \end{aligned}$$

Degree of freedom = (3 – 1) = 2

**Total sum of squares**

$$\begin{aligned}
 &= (-2)^2 + (5)^2 + (0)^2 + (0)^2 + (2)^2 + (-2)^2 + (1)^2 + (9)^2 + (2)^2 + (-1)^2 + (-4)^2 + (2)^2 - \frac{T^2}{N} \\
 &= 4 + 25 + 0 + 0 + 4 + 4 + 1 + 81 + 4 + 1 + 16 + 4 - 12 = 132
 \end{aligned}$$

The following ANOVA table summarize the all calculation related to two way classification analysis of variance table:





Source of variation	Sum of Squares	Degree of Freedom	Mean Square	F Ratio
Between Samples	42	3	MSC = 14	MSC/MSE = 14/10.67 = 1.312
Between Rows	26	2	MSR = 13	MSR/MSE = 13/10.67 = 1.218
Residual or error	64	6	MSE = 10.67	
Total	132	11		

For (3, 6) d. f.  $F_{(0.05)} = 4.76$  and for (2, 6) d. f.  $F_{(0.05)} = 5.14$ . The calculated values of F are less than the table value at 5% level of significance. The hypothesis is accepted. Hence, there is no significant difference between treatments and plots of land.

### 5.3 THE $t$ -DISTRIBUTION AND Z-DISTRIBUTION

#### t-Distribution

Let us assume a normal population with mean  $\mu$  and variance  $\sigma^2$ . If  $x_i$  represent the  $n$  values of a sample drawn from this population. Then

$$Z_i = \frac{x_i - \mu}{\sigma} \sim N(0, 1) \text{ where } i = 1, 2, \dots, n$$

$$\text{And, } U = \sum_{i=1}^n \left( \frac{x_i - \mu}{\sigma} \right)^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} \sim \chi^2 (n-1 \text{ df}) \quad \text{where } i = 1, 2, \dots, n$$

A new sample statistic  $T$  may, then, be defined as

$$T = \frac{\frac{x_i - \mu}{\sigma}}{\sqrt{\frac{1}{n-1} \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2}}}$$

$$T = \frac{x_i - \mu}{\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}}$$

$$T = \frac{x_i - \mu}{S}$$



This statistic - *the ratio of the standard normal variable Z to the square root of the  $\chi^2$  variable divided by its degree of freedom* - is known as '**t**' statistic or **student 't'** statistic, named after the pen name of Sir W S Gosset, who discovered the distribution of the quantity.

The random variable  $\frac{x_i - \mu}{S}$  follows **t-distribution** with  $n-1$  degrees of freedom.

$$\frac{x_i - \mu}{S} \sim t(n-1 \text{ df}) \quad \text{where } i = 1, 2, \dots, n$$

### The t-distribution in terms of Sampling Distribution of Sample Mean

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

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

Putting  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  for  $\frac{x_i - \mu}{\sigma}$  in  $T = \frac{\frac{x_i - \mu}{\sigma}}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 / \sigma^2}}$ , we get

$$T = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 / \sigma^2}}$$

or

$$T = \frac{(\bar{X} - \mu) / \sigma}{\frac{1}{\sigma} \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n(n-1)}}}$$



or

$$T = \frac{\bar{X} - \mu}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}}$$

or

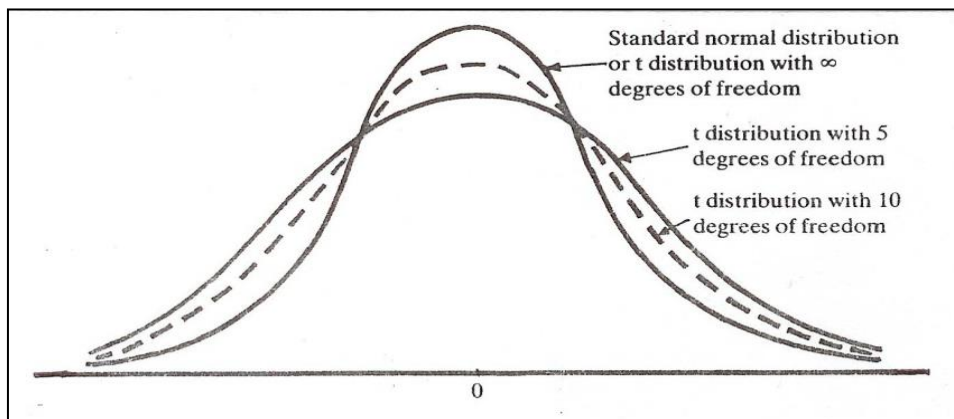
$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

When defined as above, T again follows ***t-distribution*** with  $n-1$  degrees of freedom.

$$\frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t(n-1 \text{ df}) \quad \text{where } i = 1, 2, \dots, n$$

### Properties of *t*- Distribution

1. The *t*-distribution like Z distribution, is unimodal, symmetric about mean 0, and the *t*- variable varies from  $-\infty$  and  $\infty$
2. The *t*-distribution is defined by the degrees of freedom  $\nu = n-1$ , the df associated with the distribution are the *df* associated with the sample standard deviation.



**Figure 5-8 *t*-Distribution with different *df***

3. The *t*-distribution has no mean for  $n = 2$  i.e. for  $\nu = 1$  and no variance for  $n \leq 3$  i.e. for  $\nu \leq 2$ .

However, for  $\nu > 1$ , the mean and for  $\nu > 2$ , the variance is given as  $E(T) = 0$ ;  $Var(T) = \frac{\nu}{\nu - 2}$ .



4. The variance  $\frac{v}{v-2}$  of the  $t$ -distribution must always be greater than 1, so it is more variable as against  $Z$  distribution which has variance 1. This follows from the fact that while  $Z$  values vary from sample to sample owing to the change in the  $\bar{X}$  alone, the variation in  $T$  values are due to changes in both  $\bar{X}$  and  $S$ .
5. The variance of  $t$ -distribution approaches 1 as the sample size  $n$  tends to increase. In general, for  $n \geq 30$ , the variance of  $t$ -distribution is approximately the same as that of  $Z$  distribution. In other words the  $t$ -distribution is approximately normal for  $n \geq 30$ .

### 5.4 CHECK YOUR PROGRESS

1. The variance of  $t$ -distribution approaches 1 as the sample size  $n$  tends to.....
2. As a ratio of two squared quantities, the  $F$  random variable cannot be negative and is, therefore,.....on the left by zero.
3. The  $F$  distributions defined as  $F_{(v_1, v_2)}$  and as  $F_{(v_2, v_1)}$  are.....of each other.
4. The  $t$ -distribution has no mean for  $n = 2$  i.e. for  $v = 1$  and no.....for  $n \leq 3$  i.e. for  $v \leq 2$ .
5. When sampling without replacement from a finite population, the probability distribution of the second random variable depends on what has been the..... of the first pick and so on.

### 5.5 SUMMARY

Population parameter is any number computed (or estimated) for the entire population *viz.* population mean, population median, population proportion, population variance and so on. Population parameter is unknown but fixed, whose value is to be estimated from the sample statistic that is known but random. Sample Statistic is any numbers computed from our sample data *viz.* sample mean, sample median, sample proportion, sample variance and so on. The sampling distribution of a statistic is the probability distribution of all possible values the statistic may take when computed from random samples of the same size drawn from a specified population. The sampling distributions of only the commonly used sample statistics like sample mean, sample proportion, sample variance *etc.*, which have a role in making inferences about the population. The  $F$  distribution is the distribution of the ratio of two chi-square random variables that are independent of each other, each of which is divided by its own degrees of freedom. The ratio of the standard normal variable  $Z$  to the square root of the  $\chi^2$



variable divided by its degree of freedom - is known as 't' statistic or student 't' statistic, named after the pen name of Sir W S Gosset, who discovered the distribution of the quantity.

## 5.6 KEYWORDS

**Sampling Distributions:** The sampling distribution of a statistic is the probability distribution of all possible values the statistic may take when computed from random samples of the same size drawn from a specified population.

**T statistics:** The ratio of the standard normal variable  $Z$  to the square root of the  $\chi^2$  variable divided by its degree of freedom - is known as 't' statistic or student 't'.

**F distribution:** It is the distribution of the ratio of two chi-square random variables that are independent of each other, each of which is divided by its own degrees of freedom.

**Central Limit Theorem:** When sampling is done from a population with mean  $\mu$  and standard deviation  $\sigma$ , the sampling distribution of the sample mean  $\bar{X}$  tends to a normal distribution with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$  as the sample size  $n$  increases.

**Sampling distribution of difference of sample mean:** The sampling distribution of  $\bar{X}_1 - \bar{X}_2$  is the probability distribution of all possible values the random variable  $\bar{X}_1 - \bar{X}_2$  may take when independent samples of size  $n_1$  and  $n_2$  are taken from two specified populations.

**Degree of freedom:** It refers to the number of independent variables which vary freely without being influenced by the restrictions imposed by the sample statistic(s) to be computed.

## 5.7 SELF-ASSESSMENT TEST

1. What is a sampling distribution, and what are the uses of sampling distributions?
2. How does the size of population and the kind of random sampling determine the shape of the sampling distributions?
3. (a) A sample of size  $n = 5$  is selected from a population. Under what conditions is the sampling distribution of  $\bar{X}$  normal?  
(b) Suppose the population mean is  $\mu = 105$  and the population standard deviation is 20. What are the expected value and the standard deviation of  $\bar{X}$ ?
4. What is the most significant aspect of the central limit theorem? Discuss the practical utility of central limit theorem in applied statistics.



5. Under what conditions is the central limit theorem most useful in sampling for making statistical inferences about the population mean?
6. If the population mean is 1,247, the population variance is 10,000, and the sample size is 100, what is the probability that  $\bar{X}$  will be less than 1,230?
7. When sampling is from a population with standard deviation  $\sigma = 55$ , using a sample of size  $n = 150$ , what is the probability that  $\bar{X}$  will be at least 8 units away from the population mean  $\mu$ ?
8. The Colosseum, once the most popular monument in Rome, dates from about AD 70. Since then, earthquakes have caused considerable damage to the huge structure, and engineers are currently trying to make sure the building will survive future shocks. The Colosseum can be divided into several thousand small sections. Suppose that the average section can withstand a quake measuring 3.4 on the Richter scale with a standard deviation of 1.5. A random sample of 100 sections is selected and tested for the maximum earthquake force they can withstand. What is the probability that the average section in the sample can withstand an earthquake measuring at least 3.6 on the Richter scale?
9. On June 10, 1997, the average price per share on the Big Board Composite Index in New York rose 15 cents. Assume the population standard deviation that day was 5 cents. If a random sample of 50 stocks is selected that day, what is the probability that the average price change in this sample was a rise between 14 and 16 cents?
10. An economist wishes to estimate the average family income in a certain population. The population standard deviation is known to be Rs 4,000, and the economist uses a random sample of size  $n = 225$ . What is the probability that the sample mean will fall within Rs 750 of the population mean?
11. When sampling is done from a population with population proportion  $p = 0.2$ , using a sample size  $n = 15$ , what is the sampling distribution of  $\bar{p}$ ? Is it reasonable to use a normal approximation for this sampling distribution? Explain.
12. When sampling is done for the proportion of defective items in a large shipment, where the population proportion is 0.18 and the sample size is 200, what is the probability that the sample proportion will be at least 0.20?
13. A study of the investment industry claims that 55% of all mutual funds outperformed the stock market as a whole last year. An analyst wants to test this claim and obtains a random sample of 280



mutual funds. The analyst finds that only 108 of the funds outperformed the market during the year. Determine the probability that another random sample would lead to a sample proportion as low as or lower than the one obtained by the analyst, assuming the proportion of all mutual funds that outperformed the market is indeed 0.55.

14. In recent years, convertible sport coupes have become very popular in Japan. Toyota is currently shipping Celicas to Los Angeles, where a customizer does a roof lift and ships them back to Japan. Suppose that 25% of all Japanese in a given income and lifestyle category are interested in buying Celica convertibles. A random sample of 100 Japanese consumers in the category of interest is to be selected. What is the probability that at least 20% of those in the sample will express an interest in a Celica convertible?
15. What are the limitations of small samples?
16. What do you understand by small sampling distributions? Why are the small sampling distributions called exact distributions?
17. What do you understand by the concept of degrees of freedom?
18. Define the  $\chi^2$  statistic. What are important properties of  $\chi^2$  distribution?
19. Define the  $F$  statistic. What are important properties of  $F$  distribution?
20. Define the  $t$  statistic. What are important properties of  $t$ -distribution? How does  $t$  statistic differ from  $Z$  statistic?

## 5.8 ANSWERS TO CHECK YOUR PROGRESS

1. Increase
2. Bounded
3. Reciprocal
4. Variance
5. Outcome

## 5.9 REFERENCES/SUGGESTED READINGS

1. Statistics (Theory & Practice) by Dr. B.N. Gupta. Sahitya Bhawan Publishers and Distributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing Company Ltd., New Delhi.



3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill Publishing Company Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., New Delhi.
5. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.
6. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.
7. Statistics for Management by Richard I. Levin and David S. Rubin. Prentice Hall of India Pvt. Ltd., New Delhi.
8. Statistics for Business and Economics by Kohlar Heinz. Harper Collins., New York.





Subject: Business Statistics-II	
Course code: BCOM 402	Author: Anil Kumar
Lesson: 06	Vetter: Dr. Karam Pal
<b>TESTING OF HYPOTHESES</b>	

## STRUCTURE

- 6.0 Learning Objectives
- 6.1 Introduction
  - 6.1.1 The Null and the Alternative Hypothesis
  - 6.1.2 Some Basic Concepts
  - 6.1.3 Critical Region in Terms of Test Statistic
  - 6.1.4 General Testing Procedure
  - 6.1.5 Tests of Hypotheses about Population Means
  - 6.1.6 Tests of Hypotheses about Population Proportions
  - 6.1.7 Tests of Hypotheses about Population Variances
  - 6.1.8 The Comparison of Two Populations
- 6.2 Solved Problems
- 6.3 Check your Progress
- 6.4 Summary
- 6.5 Keywords
- 6.6 Self-Assessment Test
- 6.7 Answers to check your progress
- 6.8 References/Suggested Readings

## 6.0 LEARNING OBJECTIVES

After going through this lesson, the students will be able to:

- Understand the concept of hypotheses testing
- Specify the most appropriate test of hypothesis in a given situation
- Apply the procedure and make inferences from the results.



## 6.1 INTRODUCTION

Closely related to Statistical Estimation discussed in the preceding lesson, Testing of Hypotheses is one of the most important aspects of the theory of decision-making. In the present lesson, we will study a class of problems where the decision made by a decision maker depends primarily on the strength of the evidence thrown up by a random sample drawn from a population. We can elaborate this by an example where the operations manager of a cola company has to decide whether the bottling operation is under statistical control or it has gone out of control (and needs some corrective action). Imagine that the company sells cola in bottles labeled *1-liter*, filled by an automatic bottling machine. The implied claim that on the average each bottle contains  $1,000 \text{ cm}^3$  of cola may or may not be true.

- If the claim is true, the process is said to be under statistical control. It is in the interest of the company to continue the bottling process
- If the claim is not true *i.e.* the average is either more than or less than  $1,000 \text{ cm}^3$ , the process is said to be gone out of control. It is in the interest of the company to halt the bottling process and set right the error

Therefore, to decide about the status of the bottling operation, the operations manager needs a tool, which allows him to test such a claim.

Testing of Hypotheses provides such a tool to the decision maker. If the operations manager were to use this tool, he would collect a sample of filled bottles from the on-going bottling process. The sample of bottles will be evaluated and based on the strength of the evidence produced by the sample; the operations manager will accept or reject the implied claim and accordingly make the decision. The implied claim ( $\mu = 1,000 \text{ cm}^3$ ) is a hypothesis that needs to be tested and the statistical procedure, which allows us to perform such a test, is called *Hypothesis Testing* or *Testing of Hypotheses*.

### What is a Hypothesis?

A thesis is some thing that has been proven to be true. A hypothesis is something that has not yet been proven to be true. It is some statement about a population parameter or about a population distribution. Our hypothesis for the example of bottling process could be: ***“The average amount of cola in the bottles is equal to  $1,000 \text{ cm}^3$ ”***

This statement is tentative as it implies some assumption, which may or may not be found valid on verification. Hypothesis testing is the process of determining whether or not a given hypothesis is true.



If the population is large, there is no way of analyzing the population or of testing the hypothesis directly. Instead, the hypothesis is tested on the basis of the outcome of a random sample.

### 6.1.1 THE NULL AND THE ALTERNATIVE HYPOTHESIS

As stated earlier, a hypothesis is a statement about a population parameter or about a population distribution. In any testing of hypotheses problem, we are faced with a pair of hypotheses such that one and only one of them is always true. One of this pair is called the null hypothesis and the other one the alternative hypothesis.

A null hypothesis is an assertion about the value of a population parameter. It is an assertion that we hold as true unless we have sufficient statistical evidence to conclude otherwise. For example, a null hypothesis might assert that the population mean is equal to 1,000. Unless we obtain sufficient evidence that it is not 1,000, we will accept it as 1,000. We write the null hypothesis compactly as:

$$H_0: \mu = 1,000$$

Where the symbol  $H_0$  denotes the null hypothesis.

The alternative hypothesis is the negation of the null hypothesis. For the null hypothesis  $H_0: \mu = 1,000$ , the alternative hypothesis is  $\mu \neq 1000$ . We will write it as

$$H_1: \mu \neq 1,000$$

We use the symbol  $H_1$  (or  $H_a$ ) to denote the alternative hypothesis.

The null and alternative hypotheses assert exactly opposite statements. Obviously, both  $H_0$  and  $H_1$  cannot be true and one of them will always be true. Thus, rejecting one is equivalent to accepting the other. At the end of our testing procedure, if we come to the conclusion that  $H_0$  should be rejected, this also amounts to saying that  $H_1$  should be accepted and vice versa. It is not difficult to identify the pair of hypotheses relevant in any decision situation. Can any one of the two be called the null hypothesis? The answer is a big NO — because the roles of  $H_0$  and  $H_1$  are not symmetrical.

The possible outcomes of a test can be summarized as:

**Either:**

**Accept  $H_0$**  -a weak conclusion without any evidence in as a reasonable possibility support of  $H_0$  **or:**

**Reject  $H_0$**  and -a strong conclusion with strong evidence Accept  $H_1$  against  $H_0$

To better understand the role of null and alternative hypotheses, we can compare the process of hypothesis testing with the process by which an accused person is judged to be innocent or guilty. The



person before the bar is assumed to be “*innocent until proven guilty*” So using the language of hypothesis testing, we have:

$H_0$ : *The person is innocent*

$H_1$ : *The person is guilty*

The outcomes of the trial process may result

- Accepting  $H_0$  of innocence: when there was not enough evidence to convict. However, it does not prove that the person is truly innocent
- Rejecting  $H_0$  and accepting  $H_1$  of guilt: when there is enough evidence to rule out innocence as a possibility and to strongly establish guilt

The jury acquitted Michael Jackson, on June 13, of all charges against him in the child molestation case. In other words, using the language of hypothesis testing the jury had to accept the null hypothesis  $H_0$ : *Michael Jackson is innocent* because the prosecution could not prove their case against  $H_0$  of innocence. In a trial case we do not have to rule out guilt in order to find someone innocent, but we do have to rule out innocence in order to find someone guilty. On the similar lines, we do not have to rule out  $H_1$  in order to accept  $H_0$ ; but we do have to rule out  $H_0$  in order to accept  $H_1$ . Thus, it is clear that the two hypotheses - null and alternative - are not interchangeable; each one plays a different, a special role. So it becomes more important to be clear about what the null and alternative hypotheses should be in a given situation, or else the test is meaningless.

One can conceptualize the whole procedure of testing of hypothesis as trying to answer one basic question: Is the sample evidence strong enough to enable us to reject  $H_0$ ? This means that  $H_0$  will be rejected only when there is strong sample evidence against it. However, if the sample evidence is not strong enough, we shall conclude that we cannot reject  $H_0$  and so we accept  $H_0$  by default. Thus,  $H_0$  is accepted even without any evidence in support of it whereas it can be rejected only when there is overwhelming evidence against it. In other words, the decision maker is somewhat biased towards the null hypothesis and he does not mind accepting the null hypothesis. However, he would reject the null hypothesis only when the sample evidence against the null hypothesis is too strong to be ignored.

The null hypothesis is called by this name because in many situations, acceptance of this hypothesis would lead to null action. Thus, one way to ensure what the null hypothesis should be is to note that...



...if the null hypothesis is true, then no corrective action would be necessary. If the alternative hypothesis is true, then some corrective action would be necessary. Recall our example of the cola-company in which an automatic bottling machine fills *1-liter* bottles with cola. Now consider three different situations:

**Situation I:** The operations manager wants to test the average amount filled, in order to know whether the process is under statistical control.

In this situation, the operations manager will have to take corrective action when the average is either more than or less than  $1,000 \text{ cm}^3$ . Only when the average equals  $1,000 \text{ cm}^3$ , no corrective action is necessary. So we have

$$H_0: \mu = 1,000 \text{ cm}^3$$

$$H_1: \mu \neq 1,000 \text{ cm}^3$$

**Situation II:** A consumer advocate suspects that the average amount of cola is less than  $1,000 \text{ cm}^3$  and wants to test it.

In this situation, if the average amount of cola is greater than or equal to  $1,000 \text{ cm}^3$ , no corrective action is needed, but if the average amount is less than  $1,000 \text{ cm}^3$ , the company has to halt the bottling process and set right the error. So, in this case, we have

$$H_0: \mu \geq 1,000 \text{ cm}^3$$

$$H_1: \mu < 1,000 \text{ cm}^3$$

**Situation III:** The owner of the company suspects that the machine is wasting cola by filling more than  $1,000 \text{ cm}^3$  on the average and wants to test it.

From the owner's point of view, no corrective action is necessary if the average is less than or equal to  $1,000 \text{ cm}^3$ . And, therefore, in this case we have

$$H_0: \mu \leq 1,000 \text{ cm}^3$$

$$H_1: \mu > 1,000 \text{ cm}^3$$

As the bottling example indicates, there are three possible cases for the null hypothesis, involving  $\geq$ ,  $\leq$  and  $=$  relationships. The exact null hypothesis should be finalized before any evidence is gathered, or the test will not be valid. Data snooping - formulating the null and alternative hypotheses at one's convenience after collecting and looking at the evidence - is unethical.



### 6.1.2 SOME BASIC CONCEPTS

We will now discuss some concepts, which are essential for setting up a procedure for testing of hypotheses.

#### TYPE I AND TYPE II ERRORS

After the null and alternative hypotheses are spelled out, the next step is to gather evidence from a random sample of the population. An important limitation of making inferences from the sample data is that we cannot be 100% confident about it. Since variations from one sample to another can never be eliminated until the sample is as large as the population itself, it is possible that the conclusion drawn is incorrect which leads to an error. As shown in Table 6-1 below, there can be two types of errors.

*Table 6-1 Type I and Type II Errors of Hypothesis Testing*

Decision based on Sample	States of Population	
	$H_0$ True	$H_0$ False
Accept $H_0$	Correct decision (No Error)	Wrong Decision (Type II Error)
Reject $H_0$	Wrong Decision (Type I Error)	Correct Decision (No Error)

#### Type I Error

In the context of statistical testing, the wrong decision of rejecting a true null hypothesis is known as Type I Error. If the operations manager rejects  $H_0$  and concludes that the process has gone out of control, when in reality it is under control, he would be making a type I error.

#### Type II Error

The wrong decision of accepting (not rejecting, to be more accurate) a false null hypothesis is known as Type II Error. If the operations manager does not reject  $H_0$  and concludes that the process is under control, when in reality it has gone out of control, he would be making a type II error.



Both the type I and type II errors are undesirable and should be reduced to the minimum. Let us analyse how we can minimize the chances of type I and type II errors. It may be easily realized that it is possible, even with imperfect sample evidence, to reduce the probability of type I error all the way down to zero. Just accept the null hypothesis; no matter what the evidence is. Since we will never reject any null hypothesis, we will never reject a true null hypothesis and thus we will never commit a type I error! However, it is obvious that this would be foolish. If we always accept a null hypothesis, then given a false null hypothesis, no matter how wrong it is, we are sure to accept it. In other words, our probability of committing a type II error will be 1. Similarly, we find it foolish to reduce the probability of type II error all the way down to zero by always rejecting a null hypothesis, for we would then reject every true null hypothesis, no matter how right it is. Our probability of type I error will be 1. Therefore, we cannot and should not try to completely avoid either type of error. We should plan, organize, and settle for some small, optimal probability of each type of error. Before we discuss this issue, we need to learn a few more concepts.

In hypothesis testing, Type I and Type II errors refer to mistakes that can occur when making decisions based on sample data. These errors are related to the outcome of a statistical test and how it compares the sample data with the null hypothesis.

### 1. Type I Error (False Positive):

- **Definition:** A Type I error occurs when the null hypothesis is **rejected** when it is actually **true**.
- **Explanation:** In simpler terms, it's a false alarm. You conclude that there is an effect or difference (reject the null hypothesis) when, in reality, there isn't one.
- **Example:** In a medical test, a Type I error would occur if you conclude that a patient has a disease (rejecting the null hypothesis that they don't have it) when, in fact, they are healthy.
- **Symbol:** Type I error is denoted by  $\alpha$  (alpha), also called the significance level.
- **Consequences:** A Type I error leads to false conclusions about the presence of an effect or relationship, which can lead to unnecessary actions or interventions.
- **Control:** The probability of committing a Type I error is controlled by setting the significance level ( $\alpha$ ), often at 0.05, meaning there is a 5% risk of incorrectly rejecting the null hypothesis.

### 2. Type II Error (False Negative):

- **Definition:** A Type II error occurs when the null hypothesis is **not rejected** when it is actually **false**.



- **Explanation:** In this case, you fail to detect a true effect or difference. You conclude that there is no effect or relationship (fail to reject the null hypothesis), when in fact, there is one.
- **Example:** In a medical test, a Type II error would occur if you conclude that a patient does not have a disease (failing to reject the null hypothesis) when, in fact, they do have it.
- **Symbol:** Type II error is denoted by  $\beta$  (beta).
- **Consequences:** A Type II error means missing an important effect or relationship, leading to missed opportunities for discovery or intervention.
- **Control:** The probability of committing a Type II error ( $\beta$ ) depends on several factors, such as sample size, effect size, and significance level. Power ( $1-\beta$ ) is the probability of correctly rejecting a false null hypothesis.

#### Key Differences Between Type I and Type II Errors:

Aspect	Type I Error (False Positive)	Type II Error (False Negative)
<b>What Happens</b>	Rejecting the null hypothesis when it is true.	Failing to reject the null hypothesis when it is false.
<b>Symbol</b>	$\alpha$ (alpha) - significance level	$\beta$ (beta) - probability of Type II error.
<b>Consequences</b>	False conclusion of an effect or difference.	Failing to detect a true effect or difference.
<b>Example</b>	Concluding that a drug works when it doesn't.	Concluding that a drug doesn't work when it actually does.
<b>Probability</b>	Controlled by the significance level ( $\alpha$ ).	Related to the test's power ( $1 - \beta$ ).
<b>Control Method</b>	Set a significance level (usually 0.05).	Increase sample size, effect size, or power.

#### 3. Trade-Off Between Type I and Type II Errors:

- **Inverse Relationship:** There is typically an inverse relationship between Type I and Type II errors. If you decrease the probability of a Type I error (by lowering  $\alpha$ ), the probability of a Type II error ( $\beta$ ) tends to increase, and vice versa. For example, if you make the significance level more stringent (e.g., reducing  $\alpha$  from 0.05 to 0.01), it becomes harder to reject the null hypothesis, potentially increasing the risk of a Type II error.





- **Balancing Errors:** In practice, researchers aim to balance Type I and Type II errors depending on the context and consequences of each error. The choice of significance level ( $\alpha$ ) and the desired power of the test help strike this balance.

#### 4. Power of a Test:

- **Power** ( $1-\beta$ ) is the probability of correctly rejecting a false null hypothesis. A higher power means there is a greater chance of detecting a true effect and thus a lower chance of making a Type II error.
- Power depends on factors such as:
  - **Sample size:** Larger samples typically reduce the probability of Type II errors and increase power.
  - **Effect size:** Larger effects are easier to detect, reducing the likelihood of Type II errors.
  - **Significance level ( $\alpha$ ):** Higher  $\alpha$  increases power but also increases the risk of a Type I error.

Understanding Type I and Type II errors is essential for designing effective hypothesis tests and interpreting their results. Minimizing both errors is key to ensuring that conclusions drawn from statistical tests are accurate and reliable, particularly in fields like medical research, social sciences, and economics where consequences of errors can be significant. Balancing the risks of these errors depends on the context, the cost of each type of error, and the power of the test.

#### TEST STATISTIC AND THE $p$ -VALUE

Consider the case of owner's suspicion related to our bottling process example. The null and alternative hypotheses in this case are:

$$H_0: \mu \leq 1,000$$

$$H_1: \mu > 1,000$$

Suppose the population variance is 25 and a random sample of size 100 yields a sample mean of 1,000.5. Because the sample mean is more than 1,000, the evidence goes against the null hypothesis ( $H_0$ ). Can we reject  $H_0$  based on this evidence?

- if we reject it, there is some chance that we might be committing a type I error, and
- if we accept it, there is some chance that we might be committing a type II error.

Then what can we do? We should ask a natural question at this situation- "*What is the probability that  $H_0$  can still be true despite the evidence?*" The question asks for the "credibility" of  $H_0$  in light of



unfavorable evidence. However, due to mathematical complexities, it is not possible to compute the probability that  $H_0$  is true. We, therefore, settle for a question that comes very close.

*“When the actual  $\mu = 1,000$ , and with sample size 100, what is the probability of getting a sample mean that is more than or equal to 1000.5?”*

The answer to this question is then taken, as the "credibility rating" of  $H_0$ . Analyzing the question carefully, we note an important aspect:

The condition assumed is  $\mu = 1,000$ ; although  $H_0$  states  $\mu \leq 1,000$ . The reason for assuming  $\mu = 1,000$  is that ***it gives the most benefit of doubt to  $H_0$*** . If we assume  $\mu = 999$ , for instance, the probability of the sample mean being more than or equal to 1,000.5 will only be smaller, and  $H_0$  will only have less credibility. Thus the assumption  $\mu = 1,000$  gives the maximum credibility to  $H_0$ .

Now using our knowledge of sampling distribution of sample mean, we can easily answer our question.

Since population variance is known and sample size is large enough, the Central Limit Theorem is applicable here *i. e.*

$$\bar{X} \sim N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$$

and the standard normal variable  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  is to be used to calculate the required probability

$$P(\bar{X} \geq 1,000.5)$$

$$\text{So } P(\bar{X} \geq 1,000.5) = P\left(Z \geq \frac{1,000.5 - 1,000}{5/\sqrt{100}}\right)$$

$$= P(Z \geq 1.00)$$

$$= 0.1587$$

$$\approx 0.16$$

So the answer to our question is 16%. That is, there is a 16% chance for a sample of size 100 to yield a sample mean more than or equal to 1000.5 when the actual  $\mu = 1,000$ . Statisticians call this 16% the ***p***-



**value.** In other words *p-value-the probability of observing a sample statistic as extreme as the one observed if the null hypothesis is true-*.

is a kind of "credibility rating" of  $H_0$  in light of the evidence. A  $p$ -value of zero means  $H_0$  is certainly false and a  $p$ -value of 1 means that  $H_0$  is certainly true. A  $p$ -value of 16% means that there is roughly 16% probability that  $H_0$  is true, despite the evidence. Conversely, we can be roughly 84% confident that  $H_0$  is false in light of the evidence. The implication is that if we reject  $H_0$ , then there is about an 84% chance that we are doing the right thing, and about a 16% chance that we are committing a type I error. The formal definition of the  $p$ -value follows:

*Given a null hypothesis and sample evidence with sample size  $n$ , the **p-value** is the probability of getting a sample evidence with the same  $n$  that is equally or more unfavorable to the null hypothesis while the null hypothesis is actually true. The  $p$ -value is calculated giving the null hypothesis the maximum benefit of doubt.*

The random variable, as  $Z$  in this case, used to calculate the  $p$ -value is called test statistic. The formal definition of the test statistic follows:

*A **test statistic** is a random variable calculated from the sample evidence, which follows a well-known distribution and thus can be used to calculate the  $p$ -value.*

Most of the time, the test statistic we use will be  $Z$ ,  $t$ ,  $\chi^2$ , or  $F$ . The distributions of these random variables are well known and we can calculate the  $p$ -value.

Up to this point it is very much clear that statistical hypothesis is always stated with reference to a population parameter (mean, proportion or variance). The appropriate random variable calculated from the sample evidence acts as a test statistic and provide the means to decide whether statistical hypothesis is to be rejected or accepted.

### **THE SIGNIFICANCE LEVEL- $\alpha$**

From our discussion on  $p$ -value, it becomes clear that the  $p$ -value of a test *i.e.* the credibility of the null hypothesis varies with actual observed value of the sample statistic. This fact necessitates having a *policy* for rejecting  $H_0$  based on  $p$ -value. The most common policy in statistical hypothesis testing is to establish a **significance level**, denoted by  $\alpha$ , and to reject  $H_0$  when the  $p$ -value falls below it. When this policy is followed, one can be sure that the maximum probability of type I error is  $\alpha$ .

***Policy: When the  $p$ -value is less than  $\alpha$ , reject  $H_0$***



In other words, we can say that the rejection region for  $H_0$  is the area under the curve where the  $p$ -value is less than  $\alpha$ . This region is also called critical region. The standard values for  $\alpha$  are 10%, 5%, and 1%. Suppose  $\alpha$  is set at 5%. In the preceding example, for a sample mean of 1,000.5 the  $p$ -value was 16%, and  $H_0$  will not be rejected. For a sample mean of 1001 the  $p$ -value will be 2.28%, which is below  $\alpha = 5\%$ . Hence  $H_0$  will be rejected. Let us analyze in some detail the implications of using a significance level  $\alpha$  for rejecting a null hypothesis.

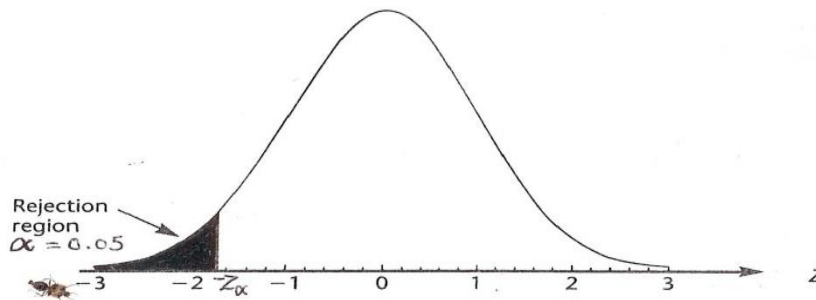
- The first thing to note is that *if we do not reject  $H_0$ , this does not prove that  $H_0$  is true*. For example, if  $\alpha = 5\%$  and the  $p$ -value = 6%, we will not reject  $H_0$ . But there is only about 6% chance that  $H_0$  is true, which is hardly proof that  $H_0$  is true. It may be possible that  $H_0$  is false and by not rejecting it, we are committing a type II error. For this reason, we should say "*We cannot reject  $H_0$  at an  $\alpha$  of 5%*" rather than "*We accept  $H_0$ .*"
- The second thing to note is that  $\alpha$  is the maximum probability of type I error we set for ourselves. Since  $\alpha$  is the maximum  $p$ -value at which we reject  $H_0$ , it is the maximum probability of committing a type I error. In other words, setting  $\alpha = 5\%$  means that we are willing to put up with up to 5% chance of committing a type I error.
- The third thing to note is that the selected value of  $\alpha$  indirectly determines the probability of type II error as well. In general, *other things remaining the same, increasing the value of  $\alpha$  will decrease the probability of type II error*. This should be intuitively obvious. For example, increasing  $\alpha$  from 5% to 10% means that in those instances with  $p$ -value in the range 5% to 10% the  $H_0$  that would not have been rejected before would now be rejected. Thus, some cases of false  $H_0$  that escaped rejection before may not escape now. As a result, the probability of type II error will decrease
- The fourth thing to note about  $\alpha$  is the meaning of  $(1 - \alpha)$ . If we set  $\alpha = 5\%$ , then  $(1 - \alpha) = 95\%$  is the minimum **confidence level** that we set in order to reject  $H_0$ . In other words, we want to be *at least 95% confident* that  $H_0$  is false before we reject it.

**One-Tailed and Two-Tailed Tests :** Consider the null and alternative hypotheses:

$$H_0: \mu \geq 1,000$$

$$H_1: \mu < 1,000$$

In this case, we will reject  $H_0$  only when  $X$  is significantly less than 1,000 or only when  $Z$  falls significantly below zero. Thus the rejection occurs only when  $Z$  takes a significantly low value in the *left tail* of its distribution. Such a case where rejection occurs in the *left tail* of the distribution of the test statistic is called a **left-tailed** test, as seen in Figure 6-1.



**Figure 6-1 A Left-tailed Test**

In the case of a left-tailed test, the  $p$ -value is the area to the left of the calculated value of the test statistic.

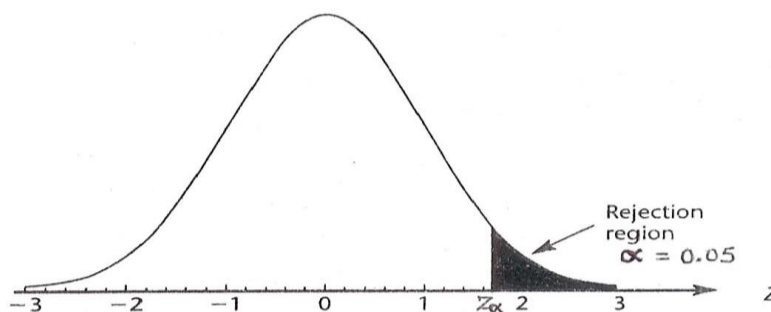
Now consider the case where the null and alternative hypotheses are:

$$H_0: \mu \leq 1,000$$

$$H_1: \mu > 1,000$$

In this case, we will reject  $H_0$  only when  $X$  is significantly more than 1,000 or only when  $Z$  is significantly greater than zero. Thus the rejection occurs only when  $Z$  takes a significantly high value in the *right tail* of its distribution.

Such a case where rejection occurs in the *right tail* of the distribution of the test statistic is called a **right-tailed** test, as seen in Figure 6-2.



**Figure 6-2 A Right-tailed Test**

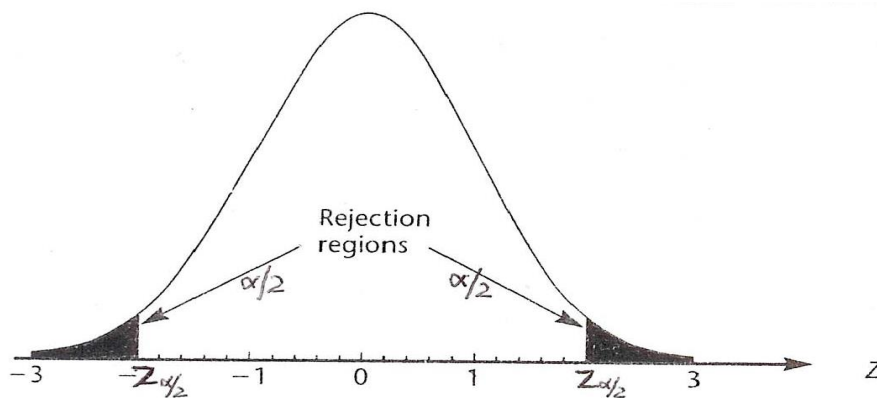


In the case of a right-tailed test, the  $p$ -value is the area to the right of the calculated value of the test statistic. In left-tailed and right-tailed tests, rejection occurs only on one tail. Hence each of them is called a **one-tailed test**. Finally, consider the case where the null and alternative hypotheses are:

$$H_0: \mu = 1,000$$

$$H_1: \mu \neq 1,000$$

In this case, we have to reject  $H_0$  in both cases, that is, whether  $X$  is significantly less than or greater than 1,000. Thus, rejection occurs when  $Z$  is significantly less than or greater than zero, which is to say that rejection occurs on both tails. Therefore, this case is called a **two-tailed test**. See Figure 6-3, where the shaded areas are the rejection regions.



**Figure 6-3 A Two-tailed Test**

In the case of a two-tailed test, the  $p$ -value is twice the tail area. If the calculated value of the test statistic falls on the left tail, then we take the area to the left of the calculated value and multiply it by 2. If the calculated value of the test statistic falls on the right tail, then we take the area to the right of the calculated value and multiply it by 2. For example, if the calculated  $Z = +1.75$ , the area to the right of it is 0.0401. Multiplying that by 2, we get the  $p$ -value as 0.0802.

### Selecting Optimal $\alpha$

All tests of hypotheses hinge upon this concept of the significance level and it is possible that a null hypothesis can be rejected at  $\alpha = 5\%$  whereas the same evidence is not strong enough to reject the null hypothesis at  $\alpha = 1\%$ . In other words, the inference drawn can be sensitive to the significance level used. We should note that selecting a value for  $\alpha$  is a question of compromise between type I and type II error probabilities. In practice, the significance level is supposed to be arrived at after considering the



cost consequences of type I error and type II error. However, most of the time the costs are difficult to estimate since they depend, among other things, on the unknown actual value of the parameter being tested. Thus, arriving at a "calculated" optimal value for  $\alpha$  is impractical. Instead, we follow an intuitive approach of assigning one of the three standard values, 1%, 5%, and 10%, to  $\alpha$ .

In the intuitive approach, we try to estimate the relative costs of the two types of errors. For example, suppose we are testing the average tensile strength of a large batch of bolts produced by a machine to see if it is above the minimum specified. Here type I error will result in rejecting a good batch of bolts and the cost of the error is roughly equal to the cost of the batch of bolts. Type II error will result in accepting a bad batch of bolts and its cost can be high or low depending on how the bolts are used.

If the bolts are used to hold together a structure, then the cost is high because defective bolts can result in the collapse of the structure, causing great damage. In this case, we should strive to reduce the probability of type II error more than that of type I error. *In such cases where type II error is more costly, we keep a large value for  $\alpha$ , namely, 10%.*

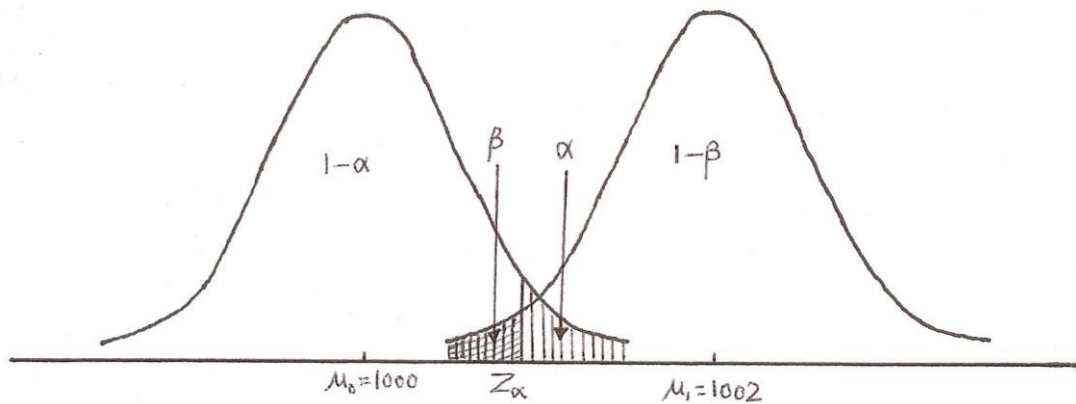
On the other hand, if the bolts are used to secure the lids on trash cans, then the cost of type II error is not high and we should strive to reduce the probability of type I error more than that of type II error. *In such cases where type I error is more costly, we keep a small value for  $\alpha$ , namely, 1%.* Then there are cases where we are not able to determine which type of error is more costly. *If the costs are roughly equal, or if we have not much knowledge about the relative costs of the two types of errors, then we keep  $\alpha = 5\%$ .*

### **$\beta$ and Power of the Test**

Denoted by  $\beta$ , Type II error is committed when a wrong decision is taken in accepting a false null hypothesis. It is the probability of accepting  $H_0$  when it should have rejected for being false. It should be noted that  $\beta$  depends on the actual value of the parameter being tested, the sample size, and  $\alpha$ . Let us see exactly how it depends. Consider the null and alternative hypotheses

$$H_0: \mu \leq 1,000$$

$$H_1: \mu > 1,000$$



**Figure 6-4 Type II Error:  $H_0: \mu \leq 1,000$  and actual  $\mu = 1,002$**

Suppose the actual value of  $\mu = \mu_1$  (say 1,002), such that  $\mu_1 > 1,000$ . Obviously,  $H_0$  is false. The cross-hatched area under the normal curve centered at  $\mu_1$  in Figure 12-4 is then the probability of accepting  $H_0$  when it is false. This area - in the acceptance region of the normal curve centered at  $\mu_0 = 1,000$ ; represents the probability that the observed sample mean  $\bar{X}$  falls in the acceptance region when  $\mu = \mu_1$  (1,002), that is when  $H_0$  is false.

Given the acceptance region  $(1 - \alpha)$  for the normal curve centered at  $\mu = \mu_0 = 1,000$ , a careful analysis of figure reveals the following.

- The value of  $\beta$  decreases as  $\mu_1$  move away from  $\mu_0$ , displaying the entire normal curve centered at  $\mu_1$  farther and farther away from the normal curve centered at  $\mu_0$ .
- The value of  $\beta$  tends to increase as  $\mu_1$  moves nearer to  $\mu_0$ . A limit is reached when  $\mu_1$  coincides with  $\mu_0$ , and the entire acceptance region  $(1 - \alpha)$  for  $\mu = \mu_0$  will represent the value of  $\beta$ . This is important conclusion in the sense that when  $H_0$  is true for  $\mu = \mu_0$ , the entire acceptance region is Type II error. Hence when  $H_0$  is true,  $\beta = 1 - \alpha$  and  $\alpha = 1 - \beta$ .
- The un-shaded area under the normal curve centered at  $\mu_1$ , which falls outside the acceptance region for  $\mu = \mu_0$ , represents the probability of rejecting  $H_0$  when it is false for  $\mu = \mu_1$ . This complement of  $\beta$ ;  $(1 - \beta)$  is known as the *power* of the test. *The power of a test is the probability that a false null hypothesis will be detected by the test.*



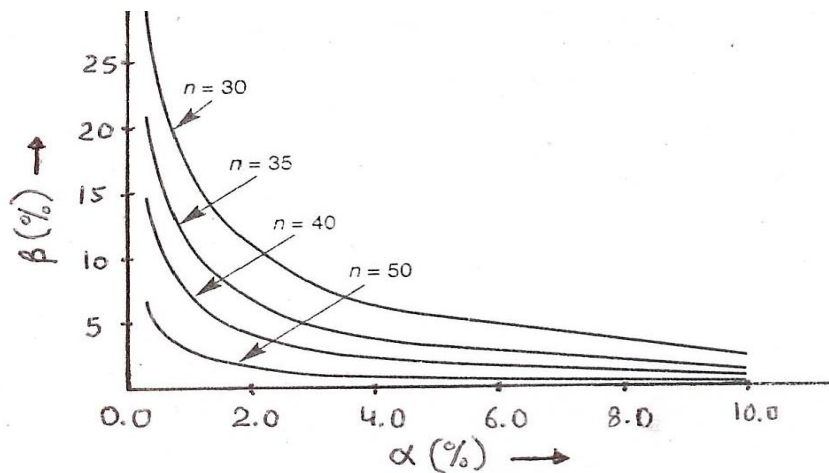


- A change in the level of significance  $\alpha$  means a change in the acceptance region  $(1 - \alpha)$ , which obviously implies a change in the cross hatched area *i.e.*  $\beta$ . In other words, the smaller the  $\alpha$ , the larger the  $\beta$  and vice-versa. Type I and type II errors are, therefore negatively related. Type I error and the power of the test  $(1 - \beta)$  are, however, positively related. Thus, the smaller the probability ( $\alpha$ ) of rejecting  $H_0$  when it is true, the smaller is the probability  $(1 - \beta)$  of rejecting  $H_0$  when it is false.

### Sample Size

In the discussion above we said that we can keep a  $\alpha$  low or a  $\beta$  low depending on which type of error is more costly. What if both types of error are costly and we want to have low  $\alpha$  as well as low  $\beta$ ? The only way to do this is to make our evidence more reliable, which can be done only by increasing the sample size. If the sample size increases, then the evidence becomes more reliable and the probability of any error will decrease.

Figure 6-5 shows the relationship between  $\alpha$  and  $\beta$  for various values of sample size  $n$ . As  $n$  increases, the curve shifts downwards reducing both  $\alpha$  and  $\beta$ . Thus, when the costs of both types of error are high, the best policy is to have a large sample and a low  $\alpha$ , such as 1%.



**Figure 6-5  $\beta$  versus  $\alpha$  for Various Values of  $n$**

After understanding the basic concepts of testing of hypotheses, we are now, able to develop tests concerning different population parameters. Under different conditions the test procedures have to be developed differently and different test statistics are used for testing. Before proceeding further let us define the critical region in terms of test statistic, which is often more helpful in many situations.



### 6.1.3 CRITICAL REGION IN TERMS OF TEST STATISTIC

We have seen that the most common policy in statistical hypothesis testing is to establish a **significance level- $\alpha$** . We decide to reject or not to reject the null hypothesis  $H_0$  by comparing the  $p$ -value with the significance level. We define the critical or rejection region as:

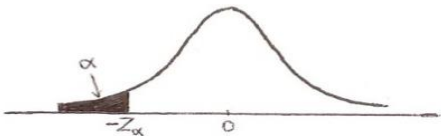
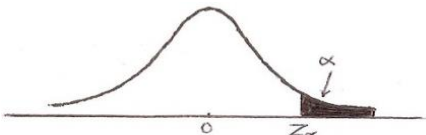
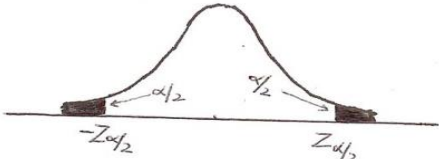
**Critical Region:  $p\text{-value} < \alpha$**

But in many situations we find it more useful to define the critical region in terms of test statistic. We, then, decide to reject or not to reject the null hypothesis  $H_0$  by comparing the observed value of the test statistic with the cut-off value or the critical value of the test statistic.

#### Z-test

When in the testing of hypotheses, we use the random variable  $Z$  for calculating the  $p$ -value and for defining the critical region of the test; we call the test as  $Z$ -test. The critical region in terms of  $Z$  are summarized in Table 6-2

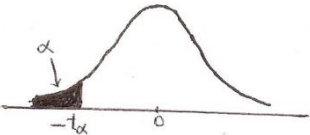
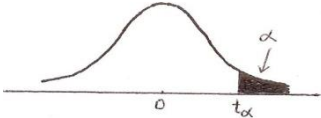
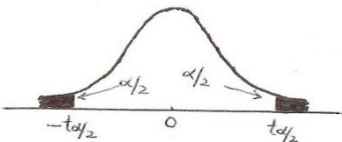
Table 6-2 Critical Region of Z-test

Test		Critical Region
Left-tailed		$Z < -Z_\alpha$
Right-tailed		$Z > Z_\alpha$
Two-tailed		$Z > Z_{\alpha/2}$ and $Z < -Z_{\alpha/2}$

#### t-test

When in the testing of hypotheses, we use the random variable  $t$  for calculating the  $p$ -value and for defining the critical region of the test; we call the test as  $t$ -test. The critical region in terms of  $t$  are summarized in Table 6-3

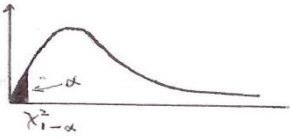
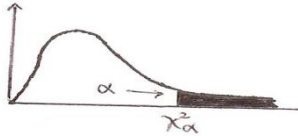
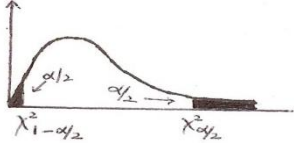
Table 6-3 Critical Region of  $t$ -test

Test		Critical Region
Left-tailed		$t < -t_\alpha$
Right-tailed		$t > t_\alpha$
Two-tailed		$t > t_{\alpha/2}$ and $t < -t_{\alpha/2}$

 $\chi^2$ -test

When in the testing of hypotheses, we use the random variable  $\chi^2$  for calculating the  $p$ -value and for defining the critical region of the test; we call the test as  $\chi^2$ -test. The critical region in terms of  $\chi^2$  are summarized in Table 12-4

Table 6-4 Critical Region of  $\chi^2$ -test

Test		Critical Region
Left-tailed		$\chi^2 < \chi^2_{1-\alpha}$
Right-tailed		$\chi^2 > \chi^2_\alpha$
Two-tailed		$\chi^2 > \chi^2_{\alpha/2}$ and $\chi^2 < \chi^2_{1-\alpha/2}$



### F-test

When in the testing of hypotheses, we use the random variable  $F$  for calculating the  $p$ -value and for defining the critical region of the test; we call the test as  $F$ -test. The critical region in terms of  $F$  are summarized in Table 6-5

**Table 6-5 Critical Region of  $F$ -test**

Test	Critical Region
Left-tailed 	$F < F_{1-\alpha}(n_1-1, n_2-1)$ <i>i.e.</i> $F < F_{\alpha}(n_2-1, n_1-1)$
Right-tailed 	$F > F_{\alpha}(n_1-1, n_2-1)$
Two-tailed 	$F > F_{\alpha/2}(n_1-1, n_2-1)$ and $F < F_{1-\alpha/2}(n_1-1, n_2-1)$ <i>i.e.</i> $F < F_{\alpha/2}(n_2-1, n_1-1)$

#### 6.1.4 GENERAL TESTING PROCEDURE

We have learnt a number of important concepts about hypothesis testing. We are now in a position to lay down a general testing procedure in a more systematic way. By now it should be clear that there are basically two phases in testing of hypothesis - in the first phase, we design the test and set up the conditions under which we shall reject the null hypothesis. In the second phase, we use the sample evidence and draw our conclusion as to whether the null hypothesis can be rejected. The detailed steps involved are as follows:

**Step 1:** State the Null and the Alternate Hypotheses. *i.e.*  $H_0$  and  $H_1$



**Step 2:** Specify a level of significance  $\alpha$

**Step 3:** Choose the test statistic and define the critical region in terms of the test statistic

**Step 4:** Make necessary computations

- calculate the observed value of the test statistic
- find the  $p$ - value of the test

**Step 5:** Decide to accept or reject the null hypothesis either

- by comparing the  $p$ - value with  $\alpha$  or
- by comparing the observed value of the test statistic with the cut-off value or the critical value of the test statistic.

### 6.1.5 TESTS OF HYPOTHESES ABOUT POPULATION MEANS

When the null hypothesis is about a population mean, the test statistic can be either  $Z$  or  $t$ . If we use  $\mu_0$  to denote the claimed population mean the null hypothesis can be any of the three usual forms:

$$\begin{array}{ll} H_0: & \mu = \mu_0 \quad \text{two-tailed test} \\ H_0: & \mu \geq \mu_0 \quad \text{left-tailed test} \\ H_0: & \mu \leq \mu_0 \quad \text{right-tailed test} \end{array}$$

#### Cases in Which the Test Statistic is $Z$

1. The population standard deviation,  $\sigma$ , is known and the population is normal.
2. The population standard deviation,  $\sigma$ , is known and the sample size,  $n$ , is at least 30 (The population need not be normal). The formula for calculating the test statistic  $Z$  in both these cases is

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

3. The population is normal and the population standard deviation,  $\sigma$ , is unknown, but the sample standard deviation,  $S$ , is known and the sample size,  $n$ , is large enough. The formula for calculating the test statistic  $Z$  in this case is

$$Z = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

#### Cases in Which the Test Statistic is $t$



1. The population is normal and the population standard deviation,  $\sigma$ , is unknown, but the sample standard deviation,  $S$ , is known and the sample size,  $n$ , is small.
2. The population is not normal and the population standard deviation,  $\sigma$ , is unknown, but the sample standard deviation,  $S$ , is known and the sample size,  $n$ , large enough.

The formula for calculating the test statistic  $t$  in both these cases is

$$t = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

The degrees of freedom for this  $t$  is  $(n-1)$

### 6.1.6 TESTS OF HYPOTHESES ABOUT POPULATION PROPORTIONS

When the null hypothesis is about a population proportion, the test statistic can be either the Binomial random variable or its Poisson or Normal approximation. If we use  $p_0$  to denote the claimed population proportion the null hypothesis can be any of the three usual forms:

$H_0: p = p_0$  two-tailed test

$H_0: p \geq p_0$  left-tailed test

$H_0: p \leq p_0$  right-tailed test

#### Cases in which the Test Statistic is Binomial Random Variable $X$

The Binomial distribution can be used whenever we are able to calculate the necessary binomial probabilities. When the Binomial distribution is used, the number of successes  $X$  serves as the test statistic. It is conveniently applicable to problems where sample size,  $n$ , is small and  $p_0$  is neither very close to 0 nor to 1.

#### Cases in which the Test Statistic is Poisson Random Variable $X$

The Poisson approximation of Binomial distribution is conveniently applicable to problems where sample size,  $n$ , is large and  $p_0$  is either very close to 0 or to 1. When the Poisson distribution is used, the number of successes  $X$  serves as the test statistic.

Note that the Binomial random variable or its Poisson approximation  $X$  follows a *discrete* distribution, and recall that the  $p$ -value is the probability of the test statistic being *equally or more unfavorable* to  $H_0$



than the value obtained from the evidence. For example, for a right-tailed test with  $H_0: p \leq 0.5$ , the  $p$ -value =  $P(X \geq \text{observed number of successes})$ .

### Cases in Which the Normal Approximation is to be used

The Normal approximation of Binomial distribution is conveniently applicable to problems where sample size,  $n$ , is large and  $p_0$  is neither very close to 0 nor to 1. When the normal distribution is used, the test statistic  $Z$  is calculated as:

$$Z = \frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

### 6.1.7 TESTS OF HYPOTHESES ABOUT POPULATION VARIANCES

When the null hypothesis is about a population variance, the test statistic is  $\chi^2$ . If we use  $\sigma_0$  to denote the claimed population proportion the null hypothesis can be any of the three usual forms:

$H_0:$	$\sigma = \sigma_0$	two-tailed test
$H_0:$	$\sigma \geq \sigma_0$	left-tailed test
$H_0:$	$\sigma \leq \sigma_0$	right-tailed test

The formula for calculating the test statistic  $\chi^2$  is:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

The degree of freedom for this  $\chi^2$  is  $(n - 1)$ .

### 6.1.8 THE COMPARISON OF TWO POPULATIONS

Almost daily we compare products, services, investment opportunities, management styles and so on. In all such situations we are interested in the comparisons of two populations with respect to some population parameter - the population mean, the population proportion, or the population variance. Now we will learn how to conduct such comparisons in an objective and meaningful way.

#### TESTING FOR DIFFERENCE BETWEEN MEANS

When we want to arrive at same conclusion about the difference between two population means, we draw one sample from each of the population. The samples drawn may be dependent on each other or these may be independent of each other.



### Dependent Samples- Paired Observations

In many situations, we can design our test in such a way that the samples drawn are dependent on each other and our observations come from two populations and are paired in some way. In general, when possible, it is often advisable to pair the observations, as this makes the experiment more precise. We can see the advantage of pairing observations with the helps of an example.

Consider a sales manager who wants to know if display at point of purchase helps in increasing the sales of his product. He may design the experiment in two ways:

**Design I:** He picks up a sample of, say 12, retail shops with no display at point of purchase. Similarly he picks up a sample of, say 10, retail shops with display at point of purchase. He will note his observations from both samples independently of each other.

**Design II:** He picks-up a random sample, of say 11, retail shops and note down the observations about weekly sale in each of these shops. Next he introduces display at point of purchase at each of these shops and again observes the weekly sales in them.

Obviously design II much better, as this tends to remove much of the extraneous variations in sales – the variation in the location of the shop, experimental conditions and other extraneous factors. Now after eliminating the effect of all other major factors, we can attribute the difference only to the ‘*treatment*’ we are studying -the display at point of purchase.

Let us label the two populations as 1 and 2. Under the situation of paired observations, it is easy to see that the variable in which we are interested is the differences between the two observations *i.e.*  $d = x_1 - x_2$ . In other words our two-population comparison test is reduced to a hypothesis test about one parameter - the difference between the means of two populations’ *i.e.*  $\mu_d = \mu_1 - \mu_2$

Thus the null hypothesis can be any of the three usual forms:

$$\begin{array}{llll}
 H_0: & \mu_1 - \mu_2 = \mu_{d0} & \text{or} & \mu_d = \mu_{d0} & \text{two-tailed test} \\
 H_0: & \mu_1 - \mu_2 \geq \mu_{d0} & & \mu_d \geq \mu_{d0} & \text{left-tailed test} \\
 H_0: & \mu_1 - \mu_2 \leq \mu_{d0} & & \mu_d \leq \mu_{d0} & \text{right-tailed test}
 \end{array}$$

The test statistic can be either  $t$  or  $Z$ .

### Cases in Which the Test Statistic is $t$





The population standard deviation of the difference,  $\sigma_d$ , is not known and the sample size,  $n$ , is small. The formula for calculating the test statistic  $t$  is

$$t = \frac{\bar{d} - \mu_{d_0}}{S_d / \sqrt{n}}$$

The degrees of freedom for this  $t$  is  $(n-1)$

### Cases in Which the Test Statistic is Z

The sample size,  $n$ , is large and/or we happen to know the population standard deviation of the difference,  $\sigma_d$ . The formula for calculating the test statistic  $t$  is

$$Z = \frac{\bar{d} - \mu_{d_0}}{S_d / \sqrt{n}}$$

or

$$Z = \frac{\bar{d} - \mu_{d_0}}{\sigma_d / \sqrt{n}}$$

### Independent Samples

When independent random sample are taken, the sample size need not be same for both populations. Let us label the two populations as 1 and 2. So that

$\mu_1$  and  $\mu_2$  denote the two population means.

$\sigma_1$  and  $\sigma_2$  denote the two population standard deviations

$n_1$  and  $n_2$  denote the two sample sizes

$\bar{X}_1$  and  $\bar{X}_2$  denote the two sample means

$S_1$  and  $S_2$  denote the two sample standard deviations

If we use  $(\mu_1 - \mu_2)_0$  to denote the claimed difference between the two population means, then the null hypothesis can be any of the three usual forms:

$$H_0: \mu_1 - \mu_2 = (\mu_1 - \mu_2)_0 \quad \text{two-tailed test}$$

$$H_0: \mu_1 - \mu_2 \geq (\mu_1 - \mu_2)_0 \quad \text{left-tailed test}$$

$$H_0: \mu_1 - \mu_2 \leq (\mu_1 - \mu_2)_0 \quad \text{right-tailed test}$$



The test statistic can be either  $Z$  or  $t$ .

### Cases in Which the Test Statistic is $Z$

1. The population standard deviations;  $\sigma_1$  and  $\sigma_2$ ; are known and both the populations are normal.
2. The population standard deviations;  $\sigma_1$  and  $\sigma_2$ ; are known and the sample sizes;  $n_1$  and  $n_2$ ; are both at least 30 (The population need not be normal).

The formula for calculating the test statistic  $Z$  in both these cases is

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

### Cases in Which the Test Statistic is $t$

The populations are normal; the population standard deviations;  $\sigma_1$  and  $\sigma_2$ ; are unknown, but the sample standard deviations;  $S_1$  and  $S_2$ ; are known. The formula for calculating the test statistic  $t$  depends on two sub cases:

**Subcase I:**  $\sigma_1$  and  $\sigma_2$  are believed to be equal (although unknown)

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Where  $S_p^2$  is the pooled variance of the two samples, which serves as the estimator of the common population variance.

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

The degrees of freedom for this  $t$  is  $(n_1 + n_2 - 2)$ .

**Subcase II:**  $\sigma_1$  and  $\sigma_2$  are believed to be unequal (although unknown)

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left( S_1^2/n_1 + S_2^2/n_2 \right)}}$$

The degrees of freedom for this  $t$  is given by:



$$df = \frac{\left( \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\left( \frac{\left( \frac{S_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left( \frac{S_2^2}{n_2} \right)^2}{n_2 - 1} \right)}$$

### TESTING FOR DIFFERENCE BETWEEN POPULATION PROPORTIONS

We will consider the large-sample tests for the difference between population proportions. For ‘large enough’ sample sizes the distribution of the two sample proportions and also the distribution of the difference between the two sample proportions is approximated well by a normal distribution. This gives rise to Z-test for comparing the two population proportions. Let us assume independent random sampling from the two populations, labeled as 1 and 2, so that

$p_1$  and  $p_2$  denote the two population proportions

$n_1$  and  $n_2$  denote the two sample sizes

$\bar{p}_1$  and  $\bar{p}_2$  denote the two sample proportions

We will use  $(p_1 - p_2)_0$  to denote the claimed difference between the two population proportions. Then the null hypothesis can be any of the three usual forms:

$$H_0: p_1 - p_2 = (p_1 - p_2)_0 \quad \text{two-tailed test}$$

$$H_0: p_1 - p_2 \geq (p_1 - p_2)_0 \quad \text{left-tailed test}$$

$$H_0: p_1 - p_2 \leq (p_1 - p_2)_0 \quad \text{right-tailed test}$$

The formula for calculating the test statistic Z depends on two cases.

**Case I:** When  $(p_1 - p_2)_0 = 0$  i.e. the claimed difference between the two population proportions is zero

$$Z = \frac{(\bar{p}_1 - \bar{p}_2) - 0}{\sqrt{\bar{p}(1 - \bar{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Where  $\bar{p}$  is the combined sample proportion in both the samples



$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

**Case II:** When  $(p_1 - p_2)_0 \neq 0$  i.e. the claimed difference between the two population proportions is some number other than zero

$$Z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)_0}{\sqrt{\frac{\bar{p}_1(1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2(1 - \bar{p}_2)}{n_2}}}$$

### TESTING FOR EQUALITY OF TWO POPULATION VARIANCES

Many a times, we may be interested in comparing the degree of variability or dispersion of two different populations. Here the problem essentially involves testing the equality of two population variances. Let us assume independent random sampling from the two populations, labeled as 1 and 2, so that

$\sigma_1^2$  and  $\sigma_2^2$  denote the two population variances

$n_1$  and  $n_2$  denote the two sample sizes

$S_1^2$  and  $S_2^2$  denote the two sample variances

Then the null hypothesis can be any of the three usual forms:

$$H_0: \sigma_1^2 = \sigma_2^2 \quad \text{two-tailed test}$$

$$H_0: \sigma_1^2 \geq \sigma_2^2 \quad \text{left-tailed test}$$

$$H_0: \sigma_1^2 \leq \sigma_2^2 \quad \text{right-tailed test}$$

The formula for calculating the test statistic  $F$  is:

$$F_{(n_1-1, n_2-1)} = \frac{S_1^2}{S_2^2}$$

The degrees of freedom for this  $F$  is  $(n_1-1, n_2-1)$

### 6.2 SOLVED PROBLEMS

Now we will solve some problems relating to testing the hypotheses stated about different population parameters, under different conditions.

#### Example 6-1



An automatic bottling machine fills oil into 2-liter ( $2,000 \text{ cm}^3$ ) bottles. A consumer advocate wants to test the null hypothesis that the average amount filled by the machine into a bottle is at least  $2,000 \text{ cm}^3$ . A random sample of 40 bottles coming out of the machine was selected and the exact contents of the selected bottles are recorded. The sample mean was  $1,999.6 \text{ cm}^3$ . The population standard deviation is known from past experience to be  $1.30 \text{ cm}^3$ .

- (a) Test the null hypothesis at an  $\alpha$  of 5%.
- (b) Assume that the population is normally distributed with the same standard deviation of  $1.30 \text{ cm}^3$ . Assume that the sample size is only 20 but the sample mean is the same  $1,999.6 \text{ cm}^3$ . Conduct the test once again at an  $\alpha$  of 5%.

If there is a difference in the two test results, explain the reason for the difference.

**Solution: (a)**

**1. The null and alternative hypotheses:**

$$H_0: \mu \geq 2,000$$

$$H_1: \mu < 2,000$$

The test is a *left-tailed* test

- 2. Level of significance:**  $\alpha = 5\%$  or 0.05
- 3. Test statistic: Z;** as the population standard deviation is known and sample size is greater than 30
- 4. Critical region:**  $Z < -Z_{0.05}$  Where  $Z_{0.05} = 1.645$
- 5. Computations:**  $\bar{X} = 1,999.6$ ;  $\sigma = 1.30$ ;  $n = 40$

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}; \quad Z = \frac{1,999.6 - 2,000}{1.30 / \sqrt{40}}; \quad Z = -1.95$$

- 6. Conclusion:** We reject the null hypothesis at  $\alpha = 0.05$  since  $Z = -1.95 < -Z_{0.05} = -1.645$ .

- (b) Since the population is normally distributed, the test statistic is once again Z

**Computations:**  $\bar{X} = 1,999.6$ ;  $\sigma = 1.30$ ;  $n = 20$

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}; \quad Z = \frac{1,999.6 - 2,000}{1.30 / \sqrt{20}}; \quad Z = -1.38$$

**Conclusion:** We do not reject the null hypothesis at  $\alpha = 0.05$  since  $Z = -1.38 > -Z_{0.05} = -1.645$



(c) In the first case we could reject the null hypothesis but in the second we could not, although in both cases the sample mean was the same. The reason is that in the first case the sample size was larger and therefore the evidence against the null hypothesis was more reliable. This produced a smaller  $p$ -value in the first case.

### **Example 6-2**

An automobile manufacturer substitutes a different engine in cars that were known to have an average miles-per-gallon rating of 31.5 on the highway. The manufacturer wants to test whether the new engine changes the miles-per-gallon rating of the automobile model. A random sample of 100 trial runs gives  $\bar{X} = 29.8$  miles per gallon and  $S = 6.6$  miles per gallon. Using the 0.05 level of significance, is the average miles-per-gallon rating on the highway for cars using the new engine different from the rating for cars using the old engine?

#### **Solution:**

##### **1. The null and alternative hypotheses:**

$$H_0: \mu = 31.5$$

$$H_1: \mu \neq 31.5$$

The test is a *two-tailed* test

##### **2. Level of significance:** $\alpha = 5\%$ or 0.05

##### **3. Test statistic:** $Z$ ; as the sample standard deviation is known and sample size is greater than 30

##### **4. Critical region:** $Z_{0.025} < Z < -Z_{0.025}$ Where $Z_{0.025} = 1.96$

##### **5. Computations:** $\bar{X} = 29.8$ , $S = 6.6$ , $n = 100$

$$Z = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}; \quad Z = \frac{29.8 - 31.5}{6.6 / \sqrt{100}}; \quad Z = -2.57$$

##### **6. Conclusion:** We reject the null hypothesis at $\alpha = 0.05$ since $Z = -2.57 < -Z_{0.025} = -1.96$ . So we conclude that the average miles-per-gallon rating on the highway for cars using the new engine is different from the rating for cars using the old engine.

### **Example 6-3**



Sixteen oil tins are taken at random from an automatic filling machine. The mean weight of the tins is 12.2 kg, with a standard deviation of 0.40 kg. Can we conclude that the filling machine is wasting oil by filling more than the intended weight of 12 kg, at a significance level of 5%?

**Solution:**

**1. The null and alternative hypotheses:**

$$H_0: \mu \leq 12.2$$

$$H_1: \mu > 12.2$$

The test is a *right-tailed* test

**2. Level of significance:**  $\alpha = 5\%$  or 0.05

**3. Test statistic:**  $t$ ; as the sample standard deviation is known and sample size is small.

**4. Critical region:**  $t > t_{0.05}$  Where  $t_{0.05}$  for 15 df = 1.7530

**5. Computations:**  $\bar{X} = 12.2$ ;  $S = 0.40$ ;  $n = 16$

$$t = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}; t = \frac{12.2 - 12}{0.40 / \sqrt{16}}; t = 2$$

**6. Conclusion:** We reject the null hypothesis at  $\alpha = 0.05$  since  $t = 2 > t_{0.05} = 1.7530$ . So we conclude that the filling machine is wasting oil by filling more than the intended weight of 12 kg.

**Example 6-4**

A coin is to be tested for fairness. It is tossed 15 times and only 8 heads are observed. Test if the coin is fair at  $\alpha = 5\%$ .

**Solution:**

**1. The null and alternative hypotheses:**

$$H_0: p = 0.5$$

$$H_1: p \neq 0.5$$

The test is a two-tailed test

**2. Level of significance:**  $\alpha = 5\%$  or 0.05

**3. Test statistic:** Binomial random variable  $X$

**4. Critical region:**  $p\text{-value} < \alpha$

**5. Computations:**  $p_0 = 0.5$ ;  $n = 15$



$$p\text{-value} = 2 * P(X \leq 8) = 2 * \left( \sum_{X=0}^8 {}^n C_X p^X (1-p)^{n-X} \right) = 2 * \left( \sum_{X=0}^8 {}^{15} C_X 0.5^X (1-0.5)^{15-X} \right) \\ = 0.5034$$

- 6. Conclusion:** We cannot reject the null hypothesis at  $\alpha = 0.05$  since  $p\text{-value} > \alpha$ . So we accept that the coin is fair.

### Example 6-5

A wholesaler received a shipment of goods, which is reported to be containing at most 2% defective items. He will accept the shipment if the claim is found true and reject if the percentage of defective items is more. To verify this claim, he draws a sample of 200 items and finds that 10 items are defective. What should be his decision at 5% level of significance?

**Solution:**

- 1. The null and alternative hypotheses:**

$$H_0: p \leq 0.02$$

$$H_1: p > 0.02$$

The test is a *right-tailed* test

- 2. Level of significance:**  $\alpha = 5\%$  or 0.05
- 3. Test statistic:** Poisson random variable  $X$  since  $p_0$  is very small and the sample size is large enough to use poisson approximation of binomial distribution.
- 4. Critical region:**  $p\text{-value} < \alpha$
- 5. Computations:**  $p_0 = 0.02$ ;  $n = 200$ ;  $\mu = 4$

$$p\text{-value} = P(X \geq 10) = 1 - P(X \leq 9) = 1 - \sum_{X=0}^9 \left( \frac{e^{-\mu} \mu^X}{X!} \right) = 1 - 0.9919 = 0.0081$$

- 6. Conclusion:** We reject the null hypothesis at  $\alpha = 0.05$  since  $p\text{-value} < \alpha$ . So the wholesaler will not accept the shipment.

### Example 6-6

SBI claims that more than 55% of the saving accounts in Haryana are at SBI. A sample survey of 400 account holders revealed that only 180 account holders have account at SBI. Verify, using 5% level of significance, if the sample results underestimate the claim of SBI.





**Solution:**

**1. The null and alternative hypotheses:**

$$H_0: p \geq 0.55$$

$$H_1: p < 0.55$$

The test is a *left-tailed* test

**2. Level of significance:**  $\alpha = 5\%$  or 0.05

**3. Test statistic:**  $Z$ ; since  $p_0$  is not too close to 0 or 1 and the sample size is large enough to use normal approximation of binomial distribution.

**4. Critical region:**  $Z < -Z_{0.05}$  Where  $Z_{0.05} = 1.645$

**5. Computations:**  $p_0 = 0.55$ ,  $\bar{p} = 180/400 = 0.45$ ,  $n = 400$

$$Z = \frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)/n}}; \quad Z = \frac{0.45 - 0.55}{\sqrt{0.55(1-0.55)/400}}; \quad Z = \frac{-20000}{4975}; \quad Z = -4.02$$

**6. Conclusion:** We reject the null hypothesis at  $\alpha = 0.05$  since  $Z = -4.02 < -Z_{0.05} = -1.645$ . So the sample results underestimate the claim of SBI.

**Example 6-7**

A manufacturer of golf balls claims that the company controls the weights of the golf balls accurately so that the variance of the weights is not more than  $1 \text{ mg}^2$ . A random sample of 31 golf balls yields a sample variance of  $1.62 \text{ mg}^2$ . Is that sufficient evidence to reject the claim at an  $\alpha$  of 5%?

**Solution:**

**1. The null and alternative hypotheses:**

$$H_0: \sigma^2 \leq 1$$

$$H_1: \sigma^2 > 1$$

The test is a *right-tailed* test

**2. Level of significance:**  $\alpha = 5\%$  or 0.05

**3. Test statistic:**  $\chi^2$

**4. Critical region:**  $\chi^2 > \chi^2_{0.05}$  Where  $\chi^2_{0.05}$  for  $30 \text{ df} = 43.7729$

**5. Computations:**  $\sigma_0^2 = 1$ ;  $S^2 = 1.62$ ;  $n = 31$



$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{30 \times 1.62}{1} = 48.6$$

**6. Conclusion:** We reject the null hypothesis at  $\alpha = 0.05$  since  $\chi^2 = 48.6 > \chi^2_{0.05} = 43.7729$ . So we conclude that there is sufficient evidence to reject the claim of the company.

### Example 6-8

A sales manager wants to know if display at point of purchase helps in increasing the sales of his product. He notes the following observations:

Shop No.	1	2	3	4	5	6	7	8	9	10	11
Sales before display	4500	5275	7235	6844	5991	6672	4943	7615	6128	5623	5154
Sales after display	4834	5010	7562	6957	6401	6423	5334	8004	6729	6277	5769
Difference(d)	-334	265	-327	-113	-410	249	-391	-389	-581	-654	-615

$$\bar{d} = -300$$

$$S_d = 312.53$$

Is there sufficient evidence to conclude that display at point of purchase helps in increasing the sales of his product?

### **Solution:**

#### **1. The null and alternative hypotheses:**

$$H_0: \mu_d \geq 0$$

$$H_1: \mu < 0$$

The test is a *left-tailed* test

#### **2. Level of significance:** $\alpha = 5\%$ or 0.05

#### **3. Test statistic:** $t$ ; as the population standard deviation of the difference, $\sigma_d$ , is not known and the sample size, $n$ , is small.

#### **4. Critical region:** $t < -t_{0.05}$ Where $t_{0.05}$ for 10 $df = 1.812$

#### **5. Computations:** $\bar{d} = -300$ ; $S = 312.53$ ; $n = 11$

$$t = \frac{\bar{d} - \mu_{d_0}}{S_d / \sqrt{n}}; t = \frac{-300 - 0}{312.53 / \sqrt{11}}; t = -3.16$$



6. **Conclusion:** We reject the null hypothesis at  $\alpha = 0.05$  since  $t = -3.16 < t_{0.05} = -1.812$ . So the sales manager has sufficient evidence to conclude that display at point of purchase helps in increasing the sales.

### Example 6-9

The makers of Duracell batteries want to demonstrate that their size AA battery lasts on an average of at least 45 minutes longer than Duracell's main competitor, the Energizer. Two independent random samples of 100 batteries of each kind are selected. The sample average lives for Duracell and Energizer batteries are found to be  $\bar{X}_1 = 308$  minutes and  $\bar{X}_2 = 254$  minutes respectively. Assume  $\sigma_1 = 84$  minutes and  $\sigma_2 = 67$  minutes. Is there evidence to substantiate Duracell's claim that its batteries last, on an average, at least 45 minutes longer than Energizer of the same size?

**Solution:**

1. **The null and alternative hypotheses:**

$$H_0: \mu_1 - \mu_2 \leq 45$$

$$H_1: \mu_1 - \mu_2 > 45$$

The test is a *right-tailed* test

2. **Level of significance:**  $\alpha = 5\%$  or 0.05  
 3. **Test statistic:**  $Z$   
 4. **Critical region:**  $Z > Z_{0.05}$  Where  $Z_{0.05} = 1.645$   
 5. **Computations:**  $\bar{X}_1 = 308$ ;  $\bar{X}_2 = 254$ ;  $\sigma_1 = 84$ ;  $\sigma_2 = 67$ ;  $n_1 = n_2 = 100$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}; Z = \frac{308 - 254 - 45}{\sqrt{\frac{84^2}{100} + \frac{67^2}{100}}}; Z = 0.838$$

6. **Conclusion:** We cannot reject the null hypothesis at  $\alpha = 0.05$  since  $Z = 0.838 < Z_{0.05} = 1.645$ . In fact the observed value of the test statistic falls in the non-rejection region of our *right-tailed* test at any conventional level of significance. So we must conclude that there is insufficient evidence to support Duracell's claim.

### Example 6-10

The following information relate to the prices (in Rs) of a product in two cities A and B.



	City A	City B
Mean price	22	17
Standard deviation	5	6

The observations related to prices are made for 9 months in city A and for 11 months in city B. Test at 0.01 level whether there is any significant difference between prices in two cities, assuming (a)

$$\sigma_1^2 = \sigma_2^2 \text{ (b) } \sigma_1^2 \neq \sigma_2^2$$

**Solution:**

**1. The null and alternative hypotheses:**

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

The test is a *two-tailed test*

**2. Level of significance:**  $\alpha = 1\%$  or 0.01

**3. Test statistic:**  $t$ ; since the population standard deviations,  $\sigma_1$  and  $\sigma_2$ , are unknown, but the sample standard deviations,  $S_1$  and  $S_2$ , are known and sample sizes are small.

**4. Critical region:**  $t_{0.005} < t < -t_{0.005}$

**5. Computations:**  $\bar{X}_1 = 22$ ;  $\bar{X}_2 = 17$ ;  $S_1 = 5$ ;  $S_2 = 6$ ;  $n_1 = 9$ ;  $n_2 = 11$

**(a)  $\sigma_1^2 = \sigma_2^2$**

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left( \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \right) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$t = \frac{22 - 17}{\sqrt{\left( \frac{8 \times 25 + 10 \times 36}{18} \right) \left( \frac{1}{9} + \frac{1}{11} \right)}}; \quad t = \frac{5}{2.51}; \quad t = 1.99$$

The degrees of freedom for this  $t$  are  $n_1 + n_2 - 2$  i.e.  $9 + 11 - 2 = 18$

For 18  $df$ ,  $t_{0.005} = 2.88$

**(b)  $\sigma_1^2 \neq \sigma_2^2$**



$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}}; t = \frac{22 - 17 - 0}{\sqrt{\left(\frac{25}{9} + \frac{36}{11}\right)}}; t = \frac{5}{2.46}; t = 2.03$$

The degrees of freedom for this  $t$  are given by

$$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1}\right) + \left(\frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}\right)}, df = \frac{\left(\frac{25}{9} + \frac{36}{11}\right)^2}{\left(\frac{\left(\frac{25}{9}\right)^2}{8}\right) + \left(\frac{\left(\frac{36}{11}\right)^2}{10}\right)} = 18$$

Against which,  $t_{0.005} = 2.88$

- 6. Conclusion:** (a) We cannot reject the null hypothesis at  $\alpha = 0.01$ , when  $\sigma_1^2 = \sigma_2^2$  since  $t = 1.99 < t_{0.005} = 2.88$ . (b) We cannot reject the null hypothesis at  $\alpha = 0.01$ , when  $\sigma_1^2 \neq \sigma_2^2$  since  $t = 2.03 < t_{0.005} = 2.88$ .

### Example 6-11

A sample survey of tax-payers belonging to business class and professional class yielded the following results:

	Business Class	Professional Class
Sample size	$n_1 = 400$	$n_2 = 420$
Defaulters in tax payment	$x_1 = 80$	$x_2 = 65$

Given these sample data, test the hypothesis at  $\alpha = 5\%$  that

- (a) the defaulters rate is the same for the two classes of tax-payers  
 (b) the defaulters rate in the case of business class is more than that in the case of professional class by 0.07.

**Solution: (a)**

#### **1. The null and alternative hypotheses:**

$$H_0: p_1 - p_2 = 0$$

$$H_1: p_1 - p_2 \neq 0$$

The test is a *two-tailed* test



2. **Level of significance:**  $\alpha = 1\%$  or 0.01
3. **Test statistic:**  $Z$ ; since the sample sizes are large enough.
4. **Critical region:**  $Z_{0.005} < Z < -Z_{0.005}$  Where  $Z_{0.005} = 2.58$
5. **Computations:**

$$\bar{p}_1 = \frac{x_1}{n_1} = \frac{80}{400} = 0.20; \quad \bar{p}_2 = \frac{x_2}{n_2} = \frac{65}{420} = 0.15 \quad \bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{80 + 65}{400 + 420} = 0.177$$

$$Z = \frac{(\bar{p}_1 - \bar{p}_2) - 0}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}; \quad Z = \frac{0.20 - 0.15}{\sqrt{(0.177 \times 0.823)\left(\frac{1}{400} + \frac{1}{420}\right)}} \quad Z = 1.87$$

6. **Conclusion:** We cannot reject the null hypothesis at  $\alpha = 0.05$  since  $Z = 1.87 < Z_{0.005} = 2.58$

(b)

1. **The null and alternative hypotheses:**

$$H_0: \quad p_1 - p_2 = 0.07$$

$$H_1: \quad p_1 - p_2 \neq 0.07$$

The test is a *two-tailed* test

2. **Level of significance:**  $\alpha = 1\%$  or 0.01
3. **Test statistic:**  $Z$ ; since the sample sizes are large enough.
4. **Critical region:**  $Z_{0.005} < Z < -Z_{0.005}$  Where  $Z_{0.005} = 2.58$
5. **Computations:**

$$\bar{p}_1 = \frac{x_1}{n_1} = \frac{80}{400} = 0.20 \quad \bar{p}_2 = \frac{x_2}{n_2} = \frac{65}{420} = 0.15 \quad Z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)_0}{\sqrt{\frac{\bar{p}_1(1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2(1 - \bar{p}_2)}{n_2}}}$$

$$Z = \frac{(0.20 - 0.15) - (0.07)}{\sqrt{\frac{0.20 \times 0.80}{400} + \frac{0.15 \times 0.85}{420}}} \quad Z = -0.76$$

6. **Conclusion:** We cannot reject the null hypothesis at  $\alpha = 0.05$  since  $Z = -0.76 > -Z_{0.01} = -2.58$ .

### Example 6-12

Use the data of Problem 12-10:  $n_1 = 9$ ,  $n_2 = 11$  and  $S_1 = 5$ ,  $S_2 = 6$  to test the assumption of equal population variances.

**Solution:****1. The null and alternative hypotheses:**

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

The test is a *two-tailed* test

**2. Level of significance:**  $\alpha = 5\%$  or 0.05**3. Test statistic:**  $F$ 

**4. Critical region:**  $F_{(n_1-1, n_2-1)} > F_{\alpha/2(n_1-1, n_2-1)}$  and  $F_{(n_1-1, n_2-1)} < F_{1-\alpha/2(n_1-1, n_2-1)}$  i.e.  $F_{(8,10)} > F_{0.025(8,10)} = 3.85$  and  $F_{(8,10)} < F_{0.975(8,10)} = 0.23$

**5. Computations:**  $S_1 = 5$ ;  $S_2 = 6$ ;  $n_1 = 9$ ;  $n_2 = 11$ 

$$F_{(n_1-1, n_2-1)} = \frac{S_1^2}{S_2^2} \quad F_{(8,10)} = \frac{25}{36} = 0.694$$

**6. Conclusion:** We cannot reject the null hypothesis at  $\alpha = 0.05$  since  $F_{(8,10)} < F_{0.025(8,10)} = 3.85$  and  $F_{(8,10)} > F_{0.975(8,10)} = 0.23$ . So the sample evidence supports the view that the two populations do not have different variances.

**6.3 CHECK YOUR PROGRESS**

1. We cannot .....  $H_0$  at an  $\alpha$  of 5% " rather than "We accept  $H_0$ ."
2. In general, other things remaining the same, .....the value of  $\alpha$  will decrease the probability of type II error.
3. In the case of a left-tailed test, the p-value is the area to the ..... of the calculated value of the test statistic.
4. The value of  $\beta$  tends to ..... as  $\mu_1$  moves nearer to  $\mu_0$ .
5. When the null hypothesis is about a population proportion, the test statistic can be either the ..... or its Poisson or Normal approximation.

**6.4 SUMMARY**

A hypothesis is something that has not yet been proven to be true. It is some statement about a population parameter or about a population distribution. This statement is tentative as it implies some



assumption, which may or may not be found valid on verification. Hypothesis testing is the process of determining whether or not a given hypothesis is true. If the population is large, there is no way of analyzing the population or of testing the hypothesis directly. Instead, the hypothesis is tested on the basis of the outcome of a random sample. In any testing of hypotheses problem, we are faced with a pair of hypotheses such that one and only one of them is always true. One of this pair is called the null hypothesis and the other one the alternative hypothesis. Almost daily we compare products, services, investment opportunities, management styles and so on. In all such situations we are interested in the comparisons of two populations with respect to some population parameter - the population mean, the population proportion, or the population variance.

## 6.5 KEYWORDS

**Null Hypothesis:** A null hypothesis is an assertion about the value of a population parameter. It is an assertion that we hold as true unless we have sufficient statistical evidence to conclude otherwise.

**Alternative Hypothesis:** The alternative hypothesis is the negation of the null hypothesis.

**Type I Error:** In the context of statistical testing, the wrong decision of rejecting a true null hypothesis is known as Type I Error.

**Type II Error:** The wrong decision of accepting (not rejecting, to be more accurate) a false null hypothesis is known as Type II Error.

**P-value:** The probability of observing a sample statistic as extreme as the one observed if the null hypothesis is true.

## 6.6 SELF-ASSESSMENT TEST

1. What is a Hypothesis? Explain how Hypothesis Testing is useful to management?
2. What are Null and Alternative hypotheses? How you will set up null and alternative hypotheses under following conditions:
  - (a) A pharmaceutical company claims that four out of five doctors prescribe the pain medicine it produces. You wish to test this claim.
  - (b) A manufacturer of golf balls claims that the variance of the weights of the company's golf balls is controlled within  $0.0028 \text{ oz}^2$ . You wish to test this claim.





- (c) A medicine is effective only if the concentration of a certain chemical in it is at least 200 parts per million (ppm). At the same time the medicine would produce an undesirable side effect if the concentration of the same chemical exceeds 200 parts per million (ppm). You wish to test the concentration of the chemical in the medicine.
3. What are Type I and Type II Errors in hypothesis testing? Explain the relationship between the two types of errors.
  4. What is a Test Statistic? Why do we have to know the distribution of the test statistic? What are the commonly used test statistics in hypotheses testing?
  5. Distinguish between a One-tailed and Two-tailed test, give a diagram and an example in each case.
  6. What is the  $p$ -value of a test? How it is calculated? Find the  $p$ -value of a (a) left-tailed, (b) right-tailed, and (c) two-tailed test if
    - (i) In the test, the test statistic  $Z = -1.86$ . In which of these three cases will  $H_0$  be rejected at an  $\alpha$  of 5%?
    - (ii) In the test, the test statistic  $Z = 1.75$ . In which of these three cases will  $H_0$  be rejected at an  $\alpha$  of 5%?
  7. What do you mean by Level of Significance of a test? “Level of significance should be specified after due consideration to the costs associated with Type I and Type II errors”. Explain this statement.
  8. What do you mean by Critical Region and Acceptance Region of a test?
  9. What is the Power of a hypothesis test? Why is it important? How is the power of a hypothesis test related to
    - (a) the significance level?
    - (b) the sample size?
    - (c) the actual value of the parameter?
  10. Consider the use of metal detectors in airports to test people for concealed weapons. In essence, this is a form of hypothesis testing.
    - (a) What are the null and alternative hypotheses?
    - (b) What are type I and type II errors in this case?
    - (c) Which type of error is more costly?
    - (d) Based on your answer to part (c), what value of  $\alpha$  would you recommend for this test?



- (e) If the sensitivity of the metal detector is increased, how would the probabilities of type I and type II errors be affected?
- (f) If  $\alpha$  is to be increased, should the sensitivity of the metal detector be increased or decreased?
11. When planning a hypothesis test, what should be done if the probabilities of both type I and type II errors are to be small?
12. “Not – rejecting a Null hypothesis” is a more precise term rather than “Accepting a Null hypothesis”. Do you agree with this statement? Explain.
13. What steps are involved in statistical testing of a hypothesis?
14. A company is engaged in the packaging of a superior quality tea in jars of 500gm each. The company is of the view that as long as the jars contain 500gm of tea, the process is under control. The standard deviation of the process is 50gm. A sample of 225 jars is taken at random and the sample average is found to be 510 gm. Has the process gone out of control?
15. A sample of size 400 was drawn and the sample mean found to be 99. Test, at 5% level of significance, whether this sample could have come from normal population with mean 100 and variance 64.
16. A manufacturer of a new motorcycle claims for it an average mileage of 60 km/liter under city conditions. However, the average mileage in 16 trials is found to be 57 km, with a standard deviation of 2 km. Is the manufacturer’s claim justified?
17. In a big city, 450 men out of a sample of 850 men were found to be smokers. Does this information, at 5% level of significance, support the view that the majority of men in this city are smokers?
18. A stock-broker claims that she can predict with 85% accuracy whether a stock’s market value will rise or fall during the coming month. Test the stock-broker’s claim at 5% level of significance if, as a test, she predicts the outcome of 6 stocks and is correct in 5 of the predictions.
19. A company engaged in manufacturing of radio tubes, finds that the life of its tubes has a variance of 0.7 years. As a result of some qualitative improvement brought about in the product, the company claims that the variance of the life of its tubes has reduced. If the sample variance,  $S^2$ , on observation of 9 tubes is observed 0.55 years, test the claim of the company (a) 5% level of significance (b) 1% level of significance.



20. Seven persons were appointed in officer cadre in an organisation. Their performance was evaluated by giving a test and the marks were recorded out of 100. They were given two-month training and another test was held and marks were recorded out of 100.

Officer:	a	b	c	d	e	f	g
Score Before Training:	80	76	92	60	70	56	74
Score After Training:	84	70	96	80	70	52	84

Can it be concluded that the training has benefited the employees? Use 5% level of significance.

21. The makers of Philips bulb want to demonstrate that their bulb lasts on an average of at least 100 hours longer than Philips' main competitor, Surya. Two independent random samples of 100 bulbs of each kind are selected. The sample average lives for Philips and Surya bulbs are found to be  $\bar{X}_1 = 1232$  hours and  $\bar{X}_2 = 1016$  hours respectively. Assume  $\sigma_1 = 84$  hours and  $\sigma_2 = 67$  hours. Is there evidence to substantiate Philips' claim that its bulbs last, on an average, at least 180 hours longer than Surya bulb of the same size?

22. Consider the following data:

	Sample A	Sample B
Sample Mean	100	105
Standard Deviation	16	24
Sample Size	800	1600

Test, at 5% level of significance, the difference between means of two populations from which samples are taken.

23. The following information relate to the wages (in Rs) of mill workers in two cities A and B.

	City A	City B
Mean wage	40	34
Standard deviation	5	6

The observations related to wages are for 8 workers in city A and for 10 workers in city B. Test at 0.01 level whether there is any significant difference between wages in two cities, assuming

(a)  $\sigma_1^2 = \sigma_2^2$  (b)  $\sigma_1^2 \neq \sigma_2^2$

24. Test market result of two advertisements A and B, yielded the following results:

A	B
---	---



Who saw the Advertisements  $n_1 = 200$   $n_2 = 220$

Who tried the Product  $x_1 = 40$   $x_2 = 35$

Given the data, test the hypotheses at  $\alpha = 5\%$  that

(a) both the advertisements are equally effective

(b) advertisement A is more effective than advertisement B by more than 0.05

Effectiveness of the advertisements are measured as proportion of viewers who tried the product.

25. Use the data of Problem 22:  $n_1 = 8$ ,  $n_2 = 10$  and  $S_1 = 5$ ,  $S_2 = 6$  to test the assumption of equal population variances.

## 6.7 ANSWERS TO CHECK YOUR PROGRESS

1. Reject
2. Increasing
3. Left
4. Increase
5. Binomial random variable

## 6.8 REFERENCES/SUGGESTED READINGS

1. Statistics (Theory & Practice) by Dr. B.N. Gupta. Sahitya Bhawan Publishers and Distributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing Company Ltd., New Delhi.
3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill Publishing Company Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., New Delhi.
5. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons. New Delhi.
6. Statistical Method by S.P. Gupta. Sultan Chand and Sons. New Delhi.
7. Statistics for Management by Richard I. Levin and David S. Rubin. Prentice Hall of India Pvt. Ltd., New Delhi.
8. Statistics for Business and Economics by Kohlar Heinz. Harper Collins., New York.



Course: Business Statistics-II	
Course Code: BCOM 402	Author: Dr. B.S. Bodla
Lesson: 07	Vetter: Karam Pal
<b>NON-PARAMETRIC TESTS</b>	

## STRUCTURE

- 7.0 Learning Objectives
- 7.1 Introduction
  - 7.1.1 Sign tests
  - 7.1.2 The two-sample and K-sample Median Tests
  - 7.1.3 Wilcoxon matched-pairs test (or Signed Rank Test)
  - 7.1.4 The Mann-Whitney U Test
  - 7.1.5 The Kruskal-Wallis Test
  - 7.1.6 The spearman's rank correlation test
  - 7.1.7 Tests of Randomness: Runs Above and Below the Median
  - 7.1.8 Kolmogorov-Smirnov One-sample Test
- 7.2 Check your Progress
- 7.3 Summary
- 7.4 Keywords
- 7.5 Self- Assessment Test
- 7.6 Answers to check your progress
- 7.7 References/Suggested readings

## 7.0 LEARNING OBJECTIVES

After going through this lesson, the students will be able to:

- Differentiate between parametric and nonparametric tests
- Understand the relevance of non-parametric test in data analysis



- Understand the procedure involved in carrying out non-parametric tests
- Design and conduct some selected non-parametric tests

## 7.1 INTRODUCTION

In contrast to parametric tests, non-parametric tests do not require any assumptions about the parameters or about the nature of population. It is because of this that these methods are sometimes referred to as the distribution free methods. Most of these methods, however, are based upon the weaker assumptions that observations are independent and that the variable under study is continuous with approximately symmetrical distribution. In addition to this, these methods do not require measurements as strong as that required by parametric methods. Most of the non-parametric tests are applicable to data measured in an ordinal or nominal scale. As opposed to this, the parametric tests are based on data measured at least in an interval scale. The measurements obtained on interval and ratio scale are also known as high level measurements.

### *Level of measurement*

1. **Nominal scale:** This scale uses numbers or other symbols to identify the groups or classes to which various objects belong. These numbers or symbols constitute a nominal or classifying scale. For example, classification of individuals on the basis of sex (male, female) or on the basis of level of education (matric, senior secondary, graduate, post graduate), etc. This scale is the weakest of all the measurements.
2. **Ordinal scale:** This scale uses numbers to represent some kind of ordering or ranking of objects. However, the differences of numbers, used for ranking, don't have any meaning. For example, the top 4 students of class can be ranked as 1, 2, 3, 4, according to their marks in an examination.
3. **Interval scale:** This scale also uses numbers such that these can be ordered and their differences have a meaningful interpretation.
4. **Ratio scale:** A scale possessing all the properties of an interval scale along with a *true zero point* is called a ratio scale. It may be pointed out that a zero point in an interval scale is arbitrary. For example, freezing point of water is defined at 0° Celsius or 32° Fahrenheit, implying thereby that the zero on either scale is arbitrary and doesn't represent total absence of heat. In contrast to this, the measurement of distance, say in metres, is done on a ratio scale. The term ratio is used here



because ratio comparisons are meaningful. For example, 100 kms of distance is four times larger than a distance of 25 kms while 100°F may not mean that it is twice as hot as 50°F.

#### Level of Measurement

Scale	Characteristics	Example(s)	Arithmetic Operations
<b>Nominal</b>	Categorical, no order, no meaningful intervals	Gender, Colors, Types of animals	Count, Mode
<b>Ordinal</b>	Ordered, unequal intervals	Rank in race, Educational levels	Median, Mode
<b>Interval</b>	Ordered, equal intervals, no true zero	Temperature (°C or °F), IQ scores	Mean, Median, Mode
<b>Ratio</b>	Ordered, equal intervals, with true zero	Height, Weight, Age, Income	Mean, Median, Mode, Ratios

It should be noted here that a test that can be performed on high level measurements can always be performed on ordinal or nominal measurements but not vice-versa. However, if along with the high level measurements the conditions of a parametric test are also met, the parametric test should invariably be used because this test is most powerful in the given circumstances.

From the above, we conclude that a non-parametric test should be used when either the conditions about the parent population are not met or the level of measurements is inadequate for a parametric test.

### *Advantages*

The non-parametric tests have gained popularity in recent years because of their usefulness in certain circumstances. Some advantages of non-parametric tests are mentioned below:

1. Non-parametric tests require less restrictive assumptions vis-à-vis a comparable parametric test.
2. These tests often require very few arithmetic computations.
3. There is no alternative to using a non-parametric test if the data are available in ordinal or nominal scale.
4. None of the parametric tests can handle data made up of samples from several populations without making unrealistic assumptions. However, there are suitable non-parametric tests available to handle such data.



### *Disadvantages*

1. It is often said that non-parametric tests are less efficient than the parametric tests because they tend to ignore a greater part of the information contained in the sample. In spite of this, it is argued that although the non-parametric tests are less efficient, a researcher using them has more confidence in using his methodology than he does if he must adhere to the unsubstantial assumptions inherent in parametric tests.
2. The non-parametric tests and their accompanying tables of significant values are widely scattered in various publications. As a result of this, the choice of most suitable method, in a given situation, may become a difficult task.

#### **7.1.1 Sign tests**

One of the easiest non-parametric tests is the sign test. The test is known as the sign test as it is based on the direction of the plus or minus signs of observations in a sample instead of their numerical values. There are two types of sign tests: (a) *One-sample sign test*, and (b) *Two-sample sign test*.

#### ***One-sample sign test***

The one-sample sign test is a very simple non-parametric test applicable on the assumption that we are dealing with a population having a continuous symmetrical distribution. As such, the probability of getting a value less than the mean is 0.5. Likewise, the probability of getting a value greater than the mean is also 0.5. To test the null hypothesis  $\mu = \mu_0$  against an appropriate alternative, each sample value greater than  $\mu_0$  is replaced by plus (+) sign and each sample value less than  $\mu_0$  with a minus (-) sign. Having done this, we can test the null hypothesis that the probabilities of getting both plus and minus signs are 0.5. It may be noted that if a sample value happens to be equal to  $\mu_0$ , it is simply discarded.

To perform the actual test, we use either of the two methods. When the sample is small, the test is performed by computing the binomial probabilities or by referring to the binomial probabilities table. When the sample is large, the normal distribution is used as an approximation of the binomial distribution. Let us take an example to show how the one-sample sign test is applied.

**Example 1:** We are required to test the hypothesis that the mean value  $\mu$  of a continuous distribution is 20 against the alternative hypothesis  $\mu_0 \neq 20$ . Fifteen observations were taken and the following results were obtained:





18, 19, 25, 21, 16, 13, 19, 22, 24, 21, 18, 17, 13, 26 and 24.

We may use 0.05 level of significance.

**Solution:** Replacing each value greater than 20 with a plus (+) sign and each value less than 20 with a minus (-) sign, we get

-- ++ --- +++ --- ++

Now, the question before us is whether 7 plus signs observed in 13 trials support the null hypothesis  $p = 0.5$  or the alternative hypothesis  $p \neq 0.5$ . Using the binomial probability tables or binomial probabilities, we find that the probability of 7 or more successes is  $0.196 + 0.196 + 0.133 + 0.092 + 0.042 + 0.014 + 0.003 = 0.696^*$  and  $p = 0.5$  and since this value is greater than  $p/2 = 0.025$ , we find that the null hypothesis will have to be accepted. We can also use normal approximation to the binomial distribution when  $np$  is 5. As here  $p = 1/2$ , the condition for the normal approximation to the binomial distribution is satisfied as  $n > 10$ . As such, we can use the Z statistic for which the following formula is to be used.

$$Z = \frac{X - np}{\sqrt{npq}} = \frac{X - (np)}{\sqrt{\frac{n}{4}}}$$

$$= \frac{7 - (15/2)}{\sqrt{\frac{15}{4}}} = \frac{14 - 15}{2} = \frac{-0.5}{1.9365} = -0.26$$

Since calculated  $Z = -0.26$  lies between  $Z = -1.96$  and  $Z = 1.96$  (the critical value of Z at 0.05 level of significance), the null hypothesis is accepted.

### *The two-sample sign test*

The sign test can be applied to problems that deal with paired data. In such problems, each pair can be replaced with a plus sign if the first value is greater than the second or a minus sign if the first value is smaller than the second. In case the two values in the pair turn out to be equal, these are discarded. These are essentially two kinds of situations: (a) the data are actually given as pairs and (b) the data comprise two independent samples that are randomly paired.



**Example 2:** Suppose we have the following table indicating the ratings assigned to two brands of cold drink X and Y by 12 consumers. Each respondent was asked to taste the two brands of cold drink and then rate them.

**Table 7.1. Ratings of brands X and Y cold drinks**

Brand X	26	30	44	23	18	50	34	16	25	49	37	20
Brand Y	22	27	39	7	11	56	30	14	18	51	33	16
Sign	+	+	+	+	+	-	+	+	+	-	+	+

We have to apply the two-sample sign test.  $H_0$  being both brands enjoy equal preference.

**Solution:** Row three of Table 13.1 shows + and – signs. When X's rating is higher than that of Y, then the third row shows the '+' sign. As against this, when X's rating is lower than that of Y, then it shows the '-' sign. The table shows 10 plus signs and 2 minus signs. Now, we have to examine whether '10 successes in 12 trials' supports the null hypothesis  $p = \frac{1}{2}$  or the alternative hypothesis  $p > \frac{1}{2}$ . The null hypothesis implies that both the brands enjoy equal preferences and none is better than the other. The alternative hypothesis is that the brand X is better than brand Y. Referring to the binomial probabilities table, we find that for  $n = 12$  and  $p = \frac{1}{2}$  the probability of '10 or more successes' is  $0.016 + 0.003 = 0.019$ . It follows that the null hypothesis can be rejected at  $\alpha = 0.05$  level of significance. We can, therefore, conclude that brand X is a preferred brand as compared to brand Y.

**Example 3:** To illustrate the second case, which relates to two independent samples, let us consider the following data pertaining to the downtimes (periods in which computers were inoperative on account of failures, in minutes of two different computers. We have to apply the two-sample sign test.

Computer A	58	60	42	62	65	59	60	52	50	75	59
	52	57	30	46	66	40	78	55	52	58	44
Computer B	32	48	50	41	45	40	43	43	70	60	80
	45	36	56	40	70	50	53	50	30	42	45

**Solution:** These data are shown in Table 13.2 along with + or – sign as may be applicable in case of each pair of values. A plus sign is assigned when the downtime for computer A is greater than that for computer B and a minus sign is given when the downtime for computer B is greater than that for computer A.



Table 7.2: Downtime of computers A and B (Minutes)

Computer A	58	60	42	62	65	59	60	52	50	75	59
Computer B	32	48	50	41	45	40	43	43	70	60	80
Sign	+	+	-	+	+	+	+	+	-	+	-
Computer A	52	57	30	46	66	40	78	55	52	58	44
Computer B	45	36	56	40	70	50	53	50	30	42	45
Sign	+	+	-	+	-	-	+	+	+	+	-

It will be seen that there are 13 plus signs and 7 minus signs. Thus, we have to ascertain whether ‘13 successes in 22 trials’ support the null hypothesis  $p = \frac{1}{2}$ . The null hypothesis implies that the true average downtime is the same for both the computers A and B. The alternative hypothesis is  $p \neq \frac{1}{2}$ . The null hypothesis implies that the true average downtime is the same for both the computers A and B. The alternative hypothesis is  $p \neq \frac{1}{2}$ .

Let us use in this case the normal approximation of the binomial distribution. This can be done since  $np$  and  $n(1 - p)$  are both equal to 11 in this example. Substituting  $n = 22$  and  $p = \frac{1}{2}$  into the formulas for the mean and the standard deviation of the binomial distribution, we get  $\mu = np = 22(\frac{1}{2}) = 11$  and

$$\sqrt{np(1 - p)} = \sqrt{22 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 2.345$$

$$\text{Hence, } Z = (X - \mu) / \sqrt{np(1 - p)} = (13 - 11) / 2.345 = 1.71$$

Since this value of 1.71 falls between  $-Z_{0.025} = -1.96$  and  $Z_{0.025} = 1.96$ , we find that the null hypothesis cannot be rejected. This means that the downtime in the two computers is the same.

This seems to be surprising as we find that there are substantial differences. The two sample means, for example, are 55.5 for A and 48.6 for B. This example illustrates the point that at times the sign test can be quite a waste of information. It may also be noted that had the continuity correction been used, we would have obtained:

$$Z = 3.5 / 2.345 = 1.49$$

This would not have changed our earlier conclusion.

### 7.1.2 The two-sample and K-sample Median tests

In order to perform this test, let us use our previous example, which pertains to the downtimes of the two computers. The median of the combined data is 52, which can easily be checked. There are 5 values



below 52 and 13 values above it, in case of computer A. As regards computer B, the corresponding figures are 16 and 6. All this information is summarised in Table 7.3, which also indicates the totals of the rows and columns.

**Table 7.3. Classification of downtime for computers A and B**

	Below median	Above median	Total
Computer A	5	13	20
Computer B	16	6	22
<b>Total</b>	<b>21</b>	<b>21</b>	<b>42</b>

Our null hypothesis  $H_0$  is that there is no difference in the median downtime for the two computers. The alternative hypothesis  $H_1$  is that there is difference in the downtime of the two computers.

We now calculate the expected frequencies by the formula  $(\text{Row}_i \times \text{Column}_j) / \text{Grand total}$ . Thus, Table 7.4 shows both the observed and the expected frequencies. Of course, we could have obtained these results by arguing that half the values in each sample can be expected to fall above the median and the other half below it.

**Table 7.4. Calculation of chi-square**

Observed frequencies (O)	Expected frequencies (E)	O – E	$(O - E)^2$	$(O - E)^2/E$
5	10	-5	25	2.50
13	10	5	25	2.50
16	11	5	25	2.27
6	11	-5	25	2.27
			<b>Total</b>	<b>9.54</b>

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 9.54$$

The critical value of  $\chi^2$  at 0.05 level of significance for  $(2 - 1) (2 - 1) = 1$  degree of freedom is 3.841 ( $\chi^2$ -Table). Since the calculated value of  $\chi^2$  exceeds the critical value, the null hypothesis has to be rejected. In other words, there is no evidence to suggest that the downtime is the same in case of the two computers.



It may be recalled that in the previous example having the same data, the null hypothesis could not be rejected. In contrast, we find here that the two-sample median test has led to the rejection of the null hypothesis. This may be construed as evidence that the median test is not quite as wasteful of the information as the sign test. However, in general, it is very difficult to make a meaningful comparison of the merits of two or more non-parametric tests, which can be used for the same purpose.

### *The K-sample Median Test*

The median test can easily be generalised so that it can be applied to K-samples. In accordance with the earlier procedure, first find the median of the combined data. We then determine how many of the values in each sample fall above or below the median. Finally, we analyse the resulting contingency table by the method of chi-square. Let us take an example.

**Example 4:** Suppose that we are given the following data relating to marks obtained by students in Statistics in the three different sections of a MBA class in G.J.U. Hisar. The maximum marks were 100.

Section A	46	60	58	80	66	39	56	61	81	70
	75	48	43	64	57	59	87	50	73	62
Section B	60	55	82	70	46	63	88	69	61	43
	76	54	58	65	73	52				
Section C	74	67	37	80	72	92	19	52	70	40
	83	76	68	21	90	74	49	70	65	58

Test whether the differences among the three sample means are significant.

**Solution:** In case of such problems, analysis of variance is ordinarily performed. However, here we find that the data for Section C have much more variability as compared to the data for the other two sections. In view of this, it would be wrong to assume that the three population standard deviations are the same. This means that the method of one-way analysis of variance cannot be used.

In order to perform a median test, we should first determine the median of the combined data. This comes out to 63.5, as can easily be checked. Then we count how many of the marks in each sample fall below or above the median. Thus, the results obtained are shown in Table 7.5.



Table 7.5. Worksheet for calculating chi-square

	Below median	Above median
Section A	12	8
Section B	9	7
Section C	7	13

Since the corresponding expected frequencies for Section A are 10 and 10, for Section B are 8 and 8, and for Section C 10 and 10, we can obtain the value of chi-square. These calculations are shown below:

$$\begin{aligned}\chi^2 &= \frac{(12-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(9-8)^2}{8} + \frac{(7-8)^2}{8} + \frac{(7-10)^2}{10} + \frac{(13-10)^2}{10} \\ &= 0.4 + 0.4 + 0.125 + 0.125 + 0.9 + 0.9 = 2.85\end{aligned}$$

Now, we have to compare this value with the critical value of  $\chi^2$  at 5 per cent level of significance. This value is 5.991 for 2 ( $K - 1 = 3 - 1$ ) degrees of freedom (Chi-square Table). As the calculated value of  $\chi^2$  is less than the critical value, the null hypothesis cannot be rejected. In other words, we cannot conclude that there is a difference in the true average (median) marks obtained by the students in Statistics test from the three sections.

### 7.1.3 Wilcoxon matched-pairs test (or Signed Rank Test)

Wilcoxon matched-pairs test is an important non-parametric test, which can be used in various situations in the context of two related samples such as a study where husband and wife are matched or when the output of two similar machines are compared. In such cases we can determine both direction and magnitude of difference between matched values, using Wilcoxon matched-pairs test.

The procedure involved in using this test is simple. To begin with, the difference (d) between each pair of values is obtained. These differences are assigned ranks from the smallest to the largest, ignoring signs. The actual signs of differences are then put to corresponding ranks and the test statistic T is calculated, which happens to be the smaller of the two sums, namely, the sum of the negative ranks and the sum of the positive ranks.

There may arise two types of situations while using this test. One situation may arise when the two values of some matched-pair(s) is/are equal as a result the difference (d) between the values is zero. In



such a case, we do not consider the pair(s) in the calculations. The other situation may arise when we get the same difference (d) in two or more pairs. In such a case, ranks are assigned to such pairs by averaging their rank positions. For instance, if two pairs have rank score of 8, then each pair is assigned 8.5 rank  $[(8 + 9)/2 = 8.5]$  and the next largest pair is assigned the rank 10.

After omitting the number of tied pairs, if the given number or matched pairs is equal to or less than 25, then the table of critical value T is used for testing the null hypothesis. When the calculated value of T is equal to or smaller than the table (i.e. critical) value at a desired level of significance, then the null hypothesis is rejected. In case the number exceeds 25, the sampling distribution of T is taken as approximately normal with mean  $\mu_T = n(n+1)/\mu$  and standard deviation

$$\mu_T = \sqrt{n(n+1)(2n+1)/24}$$

where n is taken as the number of given matched pairs- number of tied pairs omitted, if any. In such a situation, the test Z statistic is worked out as follows:

$$Z = (T - \mu_T)/\sigma_T$$

Let us now take an example to illustrate the application of Wilcoxon matched-pairs test.

**Example 5:** The management of the Punjab National Bank wants to test the effectiveness of an advertising company that is intending to enhance the awareness of the bank's service features. It administered a questionnaire before the advertising campaign, designed to measure the awareness of services offered. After the advertising campaign, the bank administered the same questionnaire to the same group of people. Both the before and after advertising campaign scores are given in the following table.

Consumer awareness of bank services offered

Consumer	1	2	3	4	5	6	7	8	9	10
Before ad campaign	82	81	89	74	68	80	77	66	77	75
After ad campaign	87	84	84	76	78	81	79	81	81	83

Using Wilcoxon matched-pairs test, test the hypothesis that there is no difference in awareness of services offered after the advertising campaign.



**Solution:**

**Table 7.6. Application of Wilcoxon matched-pairs test**

Consumer	After Ad campn.	Before Ad campn.	Diff. $d_i$	Rank of $d_i$	Rank sign (-)	Rank sign (+)
1	87	82	5	6.5		6.5
2	84	81	3	4		4
3	84	89	-5	6.5	-6.5	
4	76	74	2	2.5		2.5
5	78	68	10	9		9
6	81	80	1	1		1
7	79	77	2	2.5		2.5
8	81	66	13	10		10
9	81	77	4	5		5
10	83	75	8	8		8
				<b>Total</b>	<b>-6.5</b>	<b>+48.5</b>

Null hypothesis  $H_0$ : There is no difference in the awareness of bank services after the ad campaign.

Alternative hypothesis  $H_1$ : There is a difference in the awareness of bank services after the ad campaign.

Computed 'T' value is 6.5. The critical value of T for  $n = 10$  at 5 per cent level of significance is 8 (Area Table). Since the computed T value is less than the critical T value, the null hypothesis is rejected. We can conclude that after the ad campaign there is difference in the consumer awareness of the bank's services needs some explanation. Had there been no difference in the awareness before and after the ad campaigns, the sum of positive and negative ranks would have been almost equal. However, if the difference between the two series being compared is larger, then the value of T will tend to be smaller as it is defined as smaller of ranks. This is the case we find in this problem. It may be noted that with this test the calculated value of T must be smaller than the critical value in order to reject the null hypothesis.





### 7.1.4 The Mann-Whitney U Test

One of the most common and best known distribution-free tests is the Mann-Whitney test for two independent samples. The logical basis of this test is particularly easy to understand. Suppose we have two independent treatment groups, with  $n_1$  observations in Group 1 and  $n_2$  observations in Group 2. Now, we assume that the population from which Group 1 scores have been sampled contained generally lower values than the population from which Group 2 scores were drawn. If we were to rank these scores disregarding the group to which they belong then the lower ranks would generally fall to Group 1 scores and the higher ranks would generally fall to Group 2 scores. Proceeding one step further, if we were to add together the ranks assigned to each group, the sum of the ranks in Group 1 would be expected to be considerably smaller than the sum of the ranks in Group 2. This would result in the rejection of the null hypothesis.

Let us now take another situation where the null hypothesis is true and the scores for the two groups are sampled from identical populations. If we were to rank all  $N$  scores regardless of the group, we would expect a mix of low and high ranks in each group. Thus, the sum of the ranks assigned to Group 1 would be broadly equal to the sum of the ranks assigned to Group 2.

The Mann-Whitney test is based on the logic just described, using the sum of the ranks in one of the groups as the test statistic. In case that sum turns out to be too small as compared to the other sum, the null hypothesis is rejected. The common practice is to take the sum of the ranks assigned to the smaller group, or if  $n_1 = n_2$ , the smaller of the two sums as the test statistic. This value is then compared with the critical value that can be obtained from the table of the Mann-Whitney statistic ( $W_s$ ) to test the null hypothesis.

Let us take an example to illustrate the application of this test.

**Example 6:** The following data indicate the lifetime (in hours) of samples of two kinds of light bulbs in continuous use:

Brand A	603	625	641	622	585	593	660	600	633	580	613	648
Brand B	620	640	646	620	652	639	590	646	631	669	610	619

We are required to use the Mann-Whitney test to compare the lifetimes of brands A and B light bulbs.

**Solution:** The first step for performing the Mann-Whitney test is to rank the given data *jointly* (as if they were one sample) in an increasing or decreasing order of magnitude. For our data, we thus obtain



the following array where we use the letters A and B to denote whether the light bulb was from brand A or brand B.

**Table 7.7. Ranking of light bulbs of brands A and B**

Sample score	Group	Rank	Sample score	Group	Rank
580	A	1	625	A	13
585	A	2	631	B	14
590	B	3	633	A	13
593	A	4	639	B	16
600	A	5	640	B	17
603	A	6	641	A	18
610	B	7	646	B	19.5
613	A	8	646	B	19.5
619	B	9	648	A	21
620	B	10.5	652	B	22
620	B	10.5	660	A	23
622	A	12	669	B	24

As both the samples come from identical populations, it is reasonable to assume that the means of the ranks assigned to the values of the two samples are more or less the same. As such, our null hypothesis is:

$H_0$ : Means of ranks assigned to the values in the two groups are the same.

$H_1$ : Means are not the same.

However, instead of using the means of the ranks, we shall use *rank sums* for which the following formula will be used.

$$U = n_1 n_2 + [n_1(n_1 + 1)]/2 - R_1$$

Where  $n_1$  and  $n_2$  are the sample sizes of Group 1 and Group 2, respectively, and  $R_1$  is the sum of the ranks assigned to the values of the first sample. In our example, we have  $n_1 = 12$ ,  $n_2 = 12$  and  $R_1 = 1 + 2 + 4 + 5 + 6 + 8 + 12 + 13 + 13 + 18 + 21 + 23 = 128$ . Substituting these values in the above formula,



$$\begin{aligned}
 U &= (12)(12) + [12(12 + 1)]/2 - 128 \\
 &= 144 + 78 - 128 \\
 &= 94
 \end{aligned}$$

From Appendix Table 9 for  $n_1$  and  $n_2$ , each equal to 12, and for 0.05 level of significance is 37. Since the critical value is smaller than the calculated value of 94, we accept the null hypothesis and conclude that there is no difference in the average lifetimes of the two brands of light bulbs.

The test statistic we have just applied is suitable when  $n_1$  and  $n_2$  are less than or equal to 25. For larger values of  $n_1$  and/or  $n_2$ , we can make use of the fact that the distribution of  $W_s$  approaches a normal distribution as sample sizes increase. We can then use the Z test to test the hypothesis.

### *The Normal Approximation*

Although our observations are limited, we may apply the normal approximation to this problem. For this, we have to use the Z statistic.

$$1. \quad \text{Mean} = \mu_u = [(N_1 N_2)/2] = [(12 \times 12)/2] = 72$$

$$\begin{aligned}
 2. \quad \text{Standard error} &= \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \\
 &= \sqrt{\frac{12 \times 12 (12 + 12 + 1)}{12}} \\
 &= \sqrt{300} = 17.3
 \end{aligned}$$

$$\begin{aligned}
 3. \quad &(\text{Statistic} - \text{Mean})/\text{Standard deviation} \\
 &= (94 - 72)/17.3 = 1.27
 \end{aligned}$$

The critical value of Z at 0.05 level of significance is 1.64. Since the calculated value of  $Z = 1.27$  is smaller than 1.64, the null hypothesis is accepted. This shows that there is no difference in average lifetimes of brands A and B bulbs. The Z test is more dependable as compared to the earlier one. It may be noted that Mann-Whitney test required fewer assumptions than the corresponding standard test. In fact, the only assumption required is that the populations from which samples have been drawn are continuous. In actual practice, even when this assumption turns out to be wrong, this is not regarded serious.



### 7.1.5 The Kruskal-Wallis test

This test is used to determine whether  $k$  independent samples can be regarded to have been obtained from identical populations with respect to their means. The Kruskal-Wallis Test is the non-parametric counter part of the one-way analysis of variance. The assumption of the F-test, used in analysis of variance, was that each of the  $k$  populations should be normal with equal variance. In contrast to this, the Kruskal-Wallis test only assumes that the  $k$  populations are continuous and have the same pattern (symmetrical or skewed) of distribution. The null and the alternative hypotheses of the Kruskal-Wallis test are:

$H_0: m_1 = m_2 = \dots = m_k$  (i.e., means of the  $k$  populations are equal)

$H_a$ : Not all  $m_i$ 's are equal.

*The Test Statistic:* The computation of the test statistic follows a procedure that is very similar to the Mann-Whitney Wilcoxon test.

(i) Rank all the  $n_1 + n_2 + \dots + n_k = n$  observations, arrayed in ascending order.

(ii) Find  $R_1, R_2, \dots, R_k$ , where  $R_i$  is the sum of ranks of the  $i$ th sample.

The test statistic, denoted by  $H$ , is given by

$$H = \frac{12}{n(n+1)} \left( \frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \dots + \frac{R_k^2}{n_k} \right) - 3(n+1).$$

It can be shown that the distribution of  $H$  is  $\chi^2$  with  $k - 1$  d.f., when size of each sample is at least 5.

Thus, if  $H > \chi_{k-1}^2$ ,  $H_0$  is rejected.

**Example 7:** To compare the effectiveness of three types of weight-reducing diets, a homogeneous groups of 22 women was divided into three sub-groups and each sub-group followed one of these diet plans for a period of two months. The weight reductions, in kgs, were noted as given below:

	I	4.3	3.2	2.7	6.2	5.0	3.9			
Diet Plans	II	5.3	7.4	8.3	5.5	6.7	7.2	8.5		
	III	1.4	2.1	2.7	3.1	1.5	0.7	4.3	3.5	0.3

Use the Kruskal-Wallis test to test the hypothesis that the effectiveness of the three weight reducing diet plans is same at 5% level of significance.

**Solution:**

It is given that  $n_1 = 6$ ,  $n_2 = 7$  and  $n_3 = 9$ .



The total number of observations is  $6 + 7 + 9 = 22$ . These are ranked in their ascending order as given below:

	I	12.5	9	6.5	17	14	11			70	
Diet Plans	II	13	20	21	16	18	19	22			131
	III	3	5	6.5	8	4	2	12.5	10		52
			1								

From the above table, we get  $R_1 = 70$ ,  $R_2 = 131$  and  $R_3 = 52$ .

$$H = \frac{12}{22 \times 23} \left( \frac{70^2}{6} + \frac{131^2}{7} + \frac{52^2}{9} \right) - 3 \times 23 = 13.63$$

The tabulated value of chi-square at 2 d.f. and 5% level of significance is 5.99. Since H is greater than this value,  $H_0$  is rejected at 5% level of significance.

### 7.1.6 The spearman's rank correlation test

The Spearman's Rank Correlation  $r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$ , can be used to test the significance of correlation in population. We can write  $H_0: r_s = 0$ , where  $r_s$  is the coefficient of correlation in population.

**The test statistic:** It can be shown that for  $n = 10$ , the distribution of  $r_s$ , under  $H_0$ , is approximately normal with mean 0 and standard error

$$\frac{1}{\sqrt{n-1}}. \text{ Thus, } z = r_s \sqrt{n-1} \text{ is a standard normal variate.}$$

**Example 8:** Twelve entries in a painting competition were ranked by two judges, as shown below:

Entry:	A	B	C	D	E	F	G	H	I	J	K	L
Judge I:	5	2	3	5	1	6	8	7	10	9	12	11
Judge II:	4	5	2	1	6	7	10	9	11	12	3	8

Test the hypothesis that coefficient of rank correlation in population is positive.

**Solution:** We have to test  $H_0: r_s < 0$  against  $H_a: r_s > 0$ .

From the given data, we can find  $d_i = R_{ui} - R_{2i}$  and then  $\sum d_i^2 = 134$ .



$$r_s = 1 - \frac{6 \times 154}{12 \times 143} = 0.46 \text{ and } z = 0.46\sqrt{11} = 1.53.$$

Since the value of  $z$  is less than 1.645, there is no evidence against  $H_0$  at 5% level of significance. Hence, the correlation in population cannot be regarded as positive.

### 7.1.7 Tests of Randomness: Run Above and Below the Median

Any sample comprising numerical observations can be treated in the same manner by using the letters  $a$  and  $b$  to denote, respectively, values above the median and values below the median of the sample. In case an observation is equal to the median, it is omitted. The resulting series of  $a$ s and  $b$ s (representing the data in their original order) can be tested for randomness on the basis of the total number of runs above and below the median, respectively. Let us take an example.

**Example 9:** Suppose we have the following series of 29 college students. After performing a set of study exercises, increases in their pulse rate were recorded as follows:

22, 23, 21, 25, 33, 32, 25, 30, 17, 20, 26, 12, 21, 20, 27, 24, 28, 14, 29, 23, 22, 36, 25, 21, 23, 19, 17, 26 and 26.

We have to test the randomness of these data.

**Solution:** First, we have to calculate the median of this series. If we arrange these values in an ascending order, we find that the size of  $(n+1)/2$ th item, that is, 13<sup>th</sup> item is 24. Thus, the median is 24. As there is one value, which is 24 we omit it and get the following arrangement of  $a$ s and  $b$ s where  $a$  stands for an item greater than (or above) the median and  $b$  stands for an item lower than (or below) the median:

bbb aaaaaa bb a bbb aa b a bb aa bbbb aa

On the basis of this arrangement, we find that  $n_1$ , (i.e.  $a$ ) = 13,  $n_2$ , (i.e.  $bs$ ) = 13, and  $u$  = 12, we get

$$\begin{aligned}\mu_r &= [(2n_1n_2)/(n_1 + n_2)] + 1 \\ &= [(2 \times 13 \times 13)/(13 + 13)] + 1 = (390/28) + 1 = 14.93\end{aligned}$$

$$\sigma_u = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1n_2)^2(n_1 + n_2 - 1)}}$$

$$\sigma_u = \sqrt{\frac{(2 \times 13 \times 15)(2 \times 13 \times 15 - 13 - 15)}{(13 + 15)^2(13 + 15 - 1)}}$$



$$= \sqrt{\frac{390 \times 362}{(28)^2 (27)}}$$

$$= \sqrt{\frac{141180}{21168}} = \sqrt{6.6695} = 2.58$$

$$Z = (u - \mu_r) / \sigma_u = (12 - 14.93) / 2.58 = -2.93 / 2.58 = -1.14$$

Since  $Z = -1.14$  falls between  $-Z_{0.025} = -1.96$  and  $Z_{0.025} = 1.96$ , the null hypothesis cannot be rejected at the level of significance  $= 0.05$ . We can, therefore, conclude that the randomness of the original sample cannot be questioned.

It may be noted that this test is particularly useful in detecting trends and cyclic patterns in a series. If there is a trend, there would be first mostly as and later mostly bs or vice versa. In case of a repeated cyclic pattern, there would be a systematic alternation of as and bs and probably, too many runs.

### 7.1.8 Kolmogorov-Smirnov one-sample test

This test is concerned with the degrees of agreement between a set of observed values and the values specified by the null hypothesis. It is similar to the chi-square test of goodness-of-fit. It is used when one is interested in comparing a set of values on an ordinal scale. Let us take an example.

**Example 10:** Suppose that a company has conducted a field survey covering 200 respondents. Apart from other questions, it asked the respondents to indicate on a 5-point scale how much the durability of a particular product is important to them. The respondents indicated as follows:

Very important	50
Somewhat important	60
Neither important nor unimportant	20
Somewhat unimportant	40
Very unimportant	30
<b>Total respondents</b>	<b>200</b>

We have been asked to use the Kolmogorov-Smirnov test to test the hypothesis that there is no difference in importance ratings for durability among the respondents.

**Solution:** In order to apply the Kolmogorov-Smirnov test to the above data, first of all we should have the cumulative frequency distribution from the sample. Second, we have to establish the cumulative



frequency distribution, which would be expected on the basis of the null hypothesis. Third, we have to determine the largest absolute deviation between the two distributions mentioned above. Finally, this value is to be compared with the critical value to ascertain its significance.

Table 7.8 shows the calculations.

**Table 7.8: Worksheet for the Kolmogorov-Smirnov D**

Importance of durability	Observed number	Observed proportion	Observed cumulative proportion	Null proportion	Null cumulative proportion	Absolute difference observed and Null
Very important	50	0.25	0.25	0.2	0.2	0.05
Somewhat important	60	0.30	0.55	0.2	0.4	0.13
Neither important nor unimportant	20	0.10	0.65	0.2	0.6	0.05
Somewhat unimportant	40	0.20	0.85	0.2	0.8	0.05
Very unimportant	30	0.13	1.00	0.2	1.0	0.00

From Table 7.8, we find that the largest absolute difference is 0.13, which is known as the Kolmogorov-Smirnov D value. For a sample size of more than 35, the critical value of D at an  $\alpha = 0.05$  is  $1.36/\sqrt{n}$ . As sample size in this example is 200,  $D = 1.36/\sqrt{200} = 0.096$ . As the calculated D exceeds the critical value of 0.096, the null hypothesis that there is no difference in importance ratings for durability among the respondents is rejected.

Although there are a number of non-parametric tests, we have presented some of the more frequently used tests in this chapter. While using these tests, we must know that the advantages we derive by limiting our assumptions may be offset by the loss in the power of such tests. However, when basic assumptions as required for parametric tests are valid, the use of non-parametric tests may lead to a false hypothesis and thus we may commit a Type II error. We have to consider this aspect very carefully before deciding in favour of non-parametric tests. It may be reiterated that such tests are more suitable in case of ranked, scaled or rated data.





## 7.2 CHECK YOUR PROGRESS

1. Sign test is based on the ..... of the plus or minus signs of observations in a sample instead of their numerical values.
2. We can determine both..... of difference between matched values, using Wilcoxon matched-pairs test.
3. In Mann-Whitney test, the only assumption required is that the populations from which samples have been drawn are ..... .
4. The Kruskal-Wallis Test is the ..... counter part of the one-way analysis of variance.
5. Kolmogorov-Smirnov test is used when one is interested in comparing a set of values on a ..... .

## 7.3 SUMMARY

In contrast to parametric tests, non-parametric tests do not require any assumptions about the parameters or about the nature of population. It is because of this that these methods are sometimes referred to as the distribution free methods. Most of these methods, however, are based upon the weaker assumptions that observations are independent and that the variable under study is continuous with approximately symmetrical distribution. In addition to this, these methods do not require measurements as strong as that required by parametric methods. Most of the non-parametric tests are applicable to data measured in an ordinal or nominal scale. As opposed to this, the parametric tests are based on data measured at least in an interval scale. The measurements obtained on interval and ratio scale are also known as high level measurements. It should be noted here that a test that can be performed on high level measurements can always be performed on ordinal or nominal measurements but not vice-versa. However, if along with the high level measurements the conditions of a parametric test are also met, the parametric test should invariably be used because this test is most powerful in the given circumstances. From the above, we conclude that a non-parametric test should be used when either the conditions about the parent population are not met or the level of measurements is inadequate for a parametric test. There are different advantages and disadvantages of it. There are different test for it like sign test, Wilcoxon test, Kolmogorov-Smirnov D etc.



## 7.4 KEYWORDS

**Non-parametric tests:** Tests that rely less on parameter estimation and/or assumptions about the shape of a population distribution.

**One-Sample Runs test:** A non-parametric test used for determining whether the items in a sample have been selected randomly.

**Run:** A sequence of identical occurrences that may be preceded and followed by different occurrences. At times, they may not be preceded or followed by any occurrences.

**Sign test:** A non-parametric test that takes into account the difference between paired observations where plus (+) and minus (-) signs are substituted for quantitative values.

**Theory of runs:** A theory concerned with the testing of samples for the randomness of the order in which they have been selected.

**Wilcoxon Matched-pairs Test (or Signed Rank Test):** A non-parametric test that can be used in various situations in the context of two related samples.

**Kolmogorov-Smirnov test:** A non-parametric test that is concerned with the degrees of agreement between a set of observed ranks (sample values) and a theoretical frequency distribution.

**Kruskal-Wallis test:** A non-parametric method for testing the null hypothesis that K independent random samples come from identical populations. It is a direct generalisation of the Mann-Whitney test.

**Mann-Whitney U test:** A non-parametric test that is used to determine whether two different samples come from identical populations or whether these populations have different means.

## 7.5 SELF-ASSESSMENT TEST

1. What do you understand by non-parametric or distribution free methods?
2. What are the major advantages of non-parametric methods over parametric methods?
3. What are the main limitations of non-parametric tests?
4. Enumerate the different non-parametric tests and explain any two of them.
5. The sequence of occurrence of 'zeros' and 'ones' in a message sent in a digital code is shown below. Test at 5 per cent whether the sequence of '0' and '1' is random 00110 11011 00001 11100 00110 11001 11110 00011 00100 11000 11100 00011 00111 11100 00000 11111 10001 11000 10001 01110.



6. The proprietor of a small business computed his average earnings per day over a period of 12 days. For each day, an L was recorded if the earnings were less than the average, otherwise an M was recorded. These data are given below:

L L L L M M L L L L M M

7. In a metropolitan city, a city bus service was scheduled to reach a major bus stop at 11 a.m. each day. If the bus reached that stop within 5 minutes of 11 a.m. it was considered to be on time. Over a 13-day period, an A was recorded if the bus was on time, otherwise a B was recorded. The picture that emerged after ten days was as follows:

A A B A B B A B A A B B B A A

8. The following data show employees' rate of substandard performance before and after a new incentive scheme. Determine whether the introduction of the new incentive scheme has reduced the substandard performance at 0.05 level of significance.

Before	7	8	5	9	10	6	5	9	6	8
After	5	6	7	6	8	7	6	6	5	7

9. A company manufacturing electronic toys has recently been taken over by another company. Prior to the takeover of the company, certain workers were approached to ascertain their satisfaction levels. The same workers were again approached to know their satisfaction level after the takeover of the company. The two sets of data are given below.

Before	69	73	58	76	82	65	75	64	87	70
After	65	75	63	75	82	68	71	65	85	68

Using an appropriate test, find out whether there has been an improvement in the satisfaction level of workers after the takeover of their company by a new company

10. The following data relate to the costs of building comparable lots in the two Resons A and B (in million rupees):

Resort A	30.9	32.5	44.3	39.5	35.0	48.9
Resort B	53.9	61.0	36.0	42.5	40.9	47.9



The company owning the resort area A claimed that the median price of building lots was less in area A as compared to resort area B. You are asked to test this claim, using a nonparametric test with a 1 per cent level of significance.

11. On 13 different days, A had to wait for the city bus to reach his office as shown below:

17, 12, 18, 20, 25, 30, 10, 13, 7, 10, 9, 11, 5, 11 and 20 minutes.

Use the sign test at 5 per cent level of significance to test the bus company's claim that on an average A should not have to wait for more than 13 minutes.

12. A company used three different methods of advertising its product in three cities. It later found the increased sales (in thousand rupees) in identical retail outlets in the three cities as follows:

City A	70	58	60	45	55	62	80	72	
City B	65	57	48	55	75	68	45	52	63
City C	53	59	71	70	63	60	58	75	

Use Kruskal-Wallis method to test the hypothesis that the mean increase in sales on account of three different methods of advertising was the same in the retail outlets in A, B and C cities. Use 5 per cent level of significance.

## 7.6 ANSWERS TO CHECK YOUR PROGRESS

1. Direction
2. Direction and magnitude
3. Continuous
4. Non-parametric
5. Ordinal scale.

## 7.7 REFERENCES/SUGGESTED READINGS

1. Spiegel, Murray R.: Theory and Practical of Statistics, London McGraw Hill Book Company.
2. Yamane, T.: Statistics: An Introductory Analysis, New York, Harpered Row Publication
3. R.P. Hooda: Statistic for Economic and Management McMillan India Ltd.
4. G.C. Beri: Statistics for Mgt., TMA
5. J.K. Sharma: Business Statistics, Pearson Education
6. S.P. Gupta : Statistical Methods, Sultan Chand and Sons.



Subject: Business Statistics-II	
Course Code: BCOM 402	Author: Anil Kumar
Lesson: 08	Vetter: Prof. Harbhajan Bansal
<b>INDEX NUMBERS</b>	

## STRUCTURE

- 8.0 Learning Objectives
- 8.1 Introduction
  - 8.1.1 What are Index Numbers?
  - 8.1.2 Uses of Index Numbers
  - 8.1.3 Types of Index Numbers
    - 8.1.3.1 Simple Index Numbers
    - 8.1.3.2 Composite Index Numbers
- 8.2 Test of Adequacy of Index Numbers
- 8.3 Special Issues and problems in the Construction of Index Numbers
- 8.4 Check your progress
- 8.5 Summary
- 8.6 Keywords
- 8.7 Self-Assessment Test
- 8.8 Answers to check your progress
- 8.9 References/Suggested Readings

## 8.0 LEARNING OBJECTIVES

After going through this lesson, students will be able to:

- Understanding the concept of Index Numbers
- Understand the techniques and the problems involved in constructing and using index numbers.



- Solve the problems related to index number

## 8.1 INTRODUCTION

In business, managers and other decision makers may be concerned with the way in which the values of variables change over time like prices paid for raw materials, numbers of employees and customers, annual income and profits, and so on. Index numbers are one way of describing such changes.

If we turn to any journal devoted to economic and financial matters, we are very likely to come across an index number of one or the other type. It may be an index number of share prices or a wholesale price index or a consumer price index or an index of industrial production. The objective of these index numbers is to measure the changes that have occurred in prices, cost of living, production, and so forth. *For example*, if a wholesale price index number for the year 2000 with base year 1990 was 170; it shows that wholesale prices, in general, increased by 70 percent in 2000 as compared to those in 1990. Now, if the same index number moves to 180 in 2001, it shows that there has been 80 percent increase in wholesale prices in 2001 as compared to those in 1990.

With the help of various index numbers, economists and businessmen are able to describe and appreciate business and economic situations quantitatively. Index numbers were originally developed by economists for monitoring and comparing different groups of goods. It is necessary in business to understand and manipulate the different published index serieses, and to construct index series of your own. Having constructed your own index, it can then be compared to a national one such as the RPI, a similar index for your industry as a whole and also to indexes for your competitors. These comparisons are a very useful tool for decision making in business.

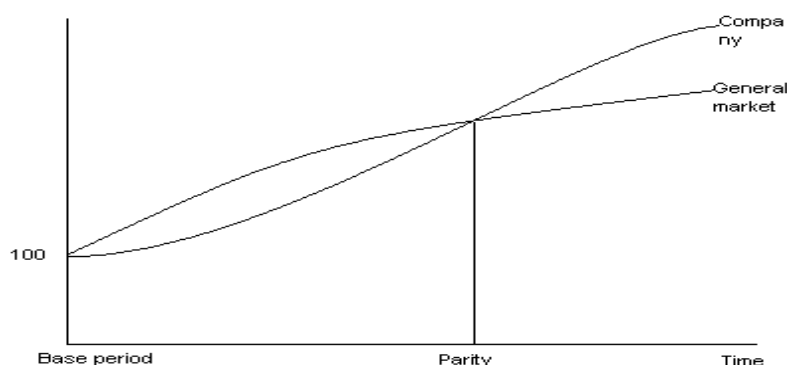


Figure 8-1 the Indexes of the Volume of Sales



For example, an accountant of a supermarket chain could construct an index of the company's own sales and compare it to the index of the volume of sales for the general supermarket industry. A graph of the two indexes will illustrate the company's performance within the sector. It is immediately clear from Figure 8-1 that, after initially lagging behind the general market, the supermarket company caught up and then overtook it. In the later stages, the company was having better results than the general market but that, as with the whole industry, those had levelled out.

Our focus in this lesson will be on the discussion related to the methodology of index number construction. The scope of the lesson is rather limited and as such, it does not discuss a large number of index numbers that are presently compiled and published by different departments of the Government of India.

### 8.1.1 What are Index Numbers?

*“Index numbers are statistical devices designed to measure the relative changes in the level of a certain phenomenon in two or more situations”.* The phenomenon under consideration may be any field of quantitative measurements. It may refer to a single variable or a group of distinct but related variables. In Business and Economics, the phenomenon under consideration may be:

- ✓ the prices of a particular commodity like steel, gold, leather, *etc.* or a group of commodities like consumer goods, cereals, milk and milk products, cosmetics, *etc.*
- ✓ volume of trade, factory production, industrial or agricultural production, imports or exports, stocks and shares, sales and profits of a business house and so on.
- ✓ the national income of a country, wage structure of workers in various sectors, bank deposits, foreign exchange reserves, cost of living of persons of a particular community, class or profession and so on.

The various situations requiring comparison may refer to either

- ✓ the changes occurring over a time, or
- ✓ the difference(s) between two or more places, or
- ✓ the variations between similar categories of objects/subjects, such as persons, groups of persons, organisations *etc.* or other characteristics such as income, profession, *etc.*



The utility of index numbers in facilitating comparison may be seen when, *for example* we are interested in studying the general change in the price level of consumer goods, *i.e.* good or commodities consumed by the people belonging to a particular section of society, say, low income group or middle income group or labour class and so on. Obviously these changes are not directly measurable as the price quotations of the various commodities are available in different units, *e.g.*, cereals (wheat, rice, pulses, *etc*) are quoted in Rs per quintal or kg; water in Rs per gallon; milk, petrol, kerosene, *etc.* in Rs per liter; cloth in Rs per meter and so on.

Further, the prices of some of the commodities may be increasing while those of others may be decreasing during the two periods and the rates of increase or decrease may be different for different commodities. Index number is a statistical device, which enables us to arrive at a single representative figure that gives the general level of the price of the phenomenon (commodities) in an extensive group. According to Wheldon:

*“Index number is a statistical device for indicating the relative movements of the data where measurement of actual movements is difficult or incapable of being made.”*

FY Edgeworth gave the classical definition of index numbers as follows:

*“Index number shows by its variations the changes in a magnitude which is not susceptible either of accurate measurement in itself or of direct valuation in practice.”*

On the basis of above discussion, the following characteristics of index numbers are apparent:

1. **Index Numbers are specialized averages:** An average is a summary figure measuring the central tendency of the data, representing a group of figures. Index number has all these functions to perform. L R Connor states, *"in its simplest form, it (index number) represents a special case of an average, generally a weighted average compiled from a sample of items judged to be representative of the whole"*. It is a special type of average – it averages variables having different units of measurement.
2. **Index Numbers are expressed in percentages:** Index numbers are expressed in terms of percentages so as to show the extent of change. However, percentage sign (%) is never used.
3. **Index Numbers measure changes not capable of direct measurement:** The technique of index numbers is utilized in measuring changes in magnitude, which are not capable of direct





measurement. Such magnitudes do not exist in themselves. Examples of such magnitudes are 'price level', 'cost of living', 'business or economic activity' *etc.* The statistical methods used in the construction of index numbers are largely methods for combining a number of phenomena representing a particular magnitude in such a manner that the changes in that magnitude may be measured in a meaningful way without introduction of serious bias.

4. ***Index Numbers are for comparison:*** The index numbers by their nature are comparative. They compare changes taking place over time or between places or between like categories.

In brief, index number is a statistical technique used in measuring the composite change in several similar economic variables over time. It measures only the composite change, because some of the variables included may be showing an increase, while some others may be showing a decrease. It synthesizes the changes taking place in different directions and by varying extents into the one composite change. Thus, an index number is a device to simplify comparison to show relative movements of the data concerned and to replace what may be complicated figures by simple ones calculated on a percentage basis.

### 8.1.1 Uses of Index Number

The first index number was constructed by an Italian, Mr G R Carli, in 17154 to compare the changes in price for the year 1750 (current year) with the price level in 1500 (base year) in order to study the effect of discovery of America on the price level in Italy. Though originally designed to study the general level of prices or accordingly purchasing power of money, today index numbers are extensively used for a variety of purposes in economics, business, management, *etc.*, and for quantitative data relating to production, consumption, profits, personnel and financial matters *etc.*, for comparing changes in the level of phenomenon for two periods, places, *etc.* In fact there is hardly any field or quantitative measurements where index numbers are not constructed. They are used in almost all sciences – natural, social and physical. The main uses of index numbers can be summarized as follows:

#### 1. Index Numbers as Economic Barometers

Index numbers are indispensable tools for the management personnel of any government organisation or individual business concern and in business planning and formulation of executive decisions. The indices of prices (wholesale & retail), output (volume of trade, import and export, industrial and



agricultural production) and bank deposits, foreign exchange and reserves *etc.*, throw light on the nature of, and variation in the general economic and business activity of the country. They are the indicators of business environment. A careful study of these indices gives us a fairly good appraisal of the general trade, economic development and business activity of the country. In the words of G Simpson and F Kafka:

*“Index numbers are today one of the most widely used statistical devices. They are used to take the pulse of the economy and they have come to be used as indicators of inflationary or deflationary tendencies.”*

Like barometers, which are used in Physics and Chemistry to measure atmospheric pressure, index numbers are rightly termed as “economic barometers”, which measure the pressure of economic and business behaviour.

## **2. Index Numbers Help in Studying Trends and Tendencies**

Since the index numbers study the relative change in the level of a phenomenon at different periods of time, they are especially useful for the study of the general trend for a group phenomenon in time series data. The indices of output (industrial and agricultural production), volume of trade, import and export, *etc.*, are extremely useful for studying the changes in the level of phenomenon due to the various components of a time series, *viz.* secular trend, seasonal and cyclical variations and irregular components and reflect upon the general trend of production and business activity. As a measure of average change in extensive group, the index numbers can be used to forecast future events. For instance, if a businessman is interested in establishing a new undertaking, the study of the trend of changes in the prices, wages and incomes in different industries is extremely helpful to him to frame a general idea of the comparative courses, which the future holds for different undertakings.

## **3. Index Numbers Help in Formulating Decisions and Policies**

Index numbers of the data relating to various business and economic variables serve an important guide to the formulation of appropriate policy. *For example*, the cost of living index numbers are used by the government and, the industrial and business concerns for the regulation of dearness allowance (D.A.) or grant of bonus to the workers so as to enable them to meet the increased cost of living from time to time. The excise duty on the production or sales of a commodity is regulated according to the index numbers of the consumption of the commodity from time to time. Similarly, the indices of consumption



of various commodities help in the planning of their future production. Although index numbers are now widely used to study the general economic and business conditions of the society, they are also applied with advantage by sociologists (population indices), psychologists (IQs'), health and educational authorities *etc.*, for formulating and revising their policies from time to time.

#### 4. Price Indices Measure the Purchasing Power of Money

A traditional use of index numbers is in measuring the purchasing power of money. Since the changes in prices and purchasing power of money are inversely related, an increase in the general price index indicates that the purchasing power of money has gone down.

In general, the purchasing power of money may be computed as

$$\text{Purchasing Power} = \frac{1}{\text{General Price Index}} \times 100$$

Accordingly, if the consumer price index for a given year is 150, the purchasing power of a rupee is  $(1/150) \times 100 = 0.667$ . That is, the purchasing power of a rupee in the given year is 66.7 paise as compared to the base year.

With the increase in prices, the amount of goods and services which money wages can buy (or the real wages) goes on decreasing. Index numbers tell us the change in real wages, which are obtained as

$$\text{Real Wage} = \frac{\text{Money Wage}}{\text{Consumer Price Index}} \times 100$$

A real wage index equal to, say, 120 corresponding to money wage index of 1150 will indicate an increase in real wages by only 20 per cent as against 150 per cent increase in money wages.

Index numbers also serve as the basis of determining the terms of exchange. The terms of exchange are the parity rate at which one set of commodities is exchanged for another set of commodities. It is determined by taking the ratio of the price index for the two groups of commodities and expressing it in percentage.

*For example*, if A and B are the two groups of commodities with 120 and 150 as their price index in a particular year, respectively, the ratio  $120/150$  multiplied by 100 is 80 per cent. It means that prices of A group of commodities in terms of those in group B are lower by 20 per cent.



## 5. Index Numbers are Used for Deflation

Consumer price indices or cost of living index numbers are used for deflation of net national product, income value series in national accounts. The technique of obtaining real wages from the given nominal wages (as explained in use 4 above) can be used to find real income from inflated money income, real sales from nominal sales and so on by taking into account appropriate index numbers.

### 8.1.2 Types of Index Numbers

Index numbers may be broadly classified into various categories depending upon the type of the phenomenon they study. Although index numbers can be constructed for measuring relative changes in any field of quantitative measurement, we shall primarily confine the discussion to the data relating to economics and business *i.e.*, data relating to prices, production (output) and consumption. In this context index numbers may be broadly classified into the following three categories:

1. **Price Index Numbers:** The price index numbers measure the general changes in the prices. They are further sub-divided into the following classes:
  - (i) **Wholesale Price Index Numbers:** The wholesale price index numbers reflect the changes in the general price level of a country.
  - (ii) **Retail Price Index Numbers:** These indices reflect the general changes in the retail prices of various commodities such as consumption goods, stocks and shares, bank deposits, government bonds, *etc.*
  - (iii) **Consumer Price Index:** Commonly known as the Cost of living Index, CPI is a specialized kind of retail price index and enables us to study the effect of changes in the price of a basket of goods or commodities on the purchasing power or cost of living of a particular class or section of the people like labour class, industrial or agricultural worker, low income or middle income class *etc.*
2. **Quantity Index Numbers:** Quantity index numbers study the changes in the volume of goods produced (manufactured), consumed or distributed, like: the indices of agricultural production, industrial production, imports and exports, *etc.* They are extremely helpful in studying the level of physical output in an economy.



3. **Value Index Numbers:** These are intended to study the change in the total value (price multiplied by quantity) of output such as indices of retail sales or profits or inventories. However, these indices are not as common as price and quantity indices.

Various indices can also be distinguished on the basis of the number of commodities that go into the construction of an index. Indices constructed for individual commodities or variable are termed as *simple index numbers*. Those constructed for a group of commodities or variables are known as *aggregative (or composite) index numbers*.

$p_i$  denotes the price of a commodity in the  $i^{th}$  period, where  $i = 1, 2, 3, \dots$

Similar meanings are assigned to  $q_0, q_1, \dots, q_i, \dots$  and  $v_0, v_1, \dots, v_i, \dots$

Capital letters  $P, Q$  and  $V$  are used to represent the price, quantity, and value index numbers, respectively. Subscripts attached to  $P, Q$ , and  $V$  indicates the years compared. Thus,

$P_{01}$  means the price index for period 1 relative to period 0,

$P_{02}$  means the price index for period 2 relative to period 0,

$P_{12}$  means the price index for period 2 relative to period 1, and so on.

Similar meanings are assigned to quantity  $Q$  and value  $V$  indices. It may be noted that all indices are expressed in percent with 100 as the index for the base period, the period with which comparison is to be made.

#### Notations Used

Since index numbers are computed for prices, quantities, and values, these are denoted by the lower case letters:

$p, q$ , and  $v$  represent respectively the price, the quantity, and the value of an individual commodity.

Subscripts  $0, 1, 2, \dots, i, \dots$  are attached to these lower case letters to distinguish price, quantity, or value in any one period from those in the other. Thus,

$p_0$  denotes the price of a commodity in the base period,

$p_1$  denotes the price of a commodity in period 1, or the current period, and



Here, in this lesson, we will develop methods of constructing simple as well as composite indices.

### 8.1.2.1 SIMPLE INDEX NUMBERS

A simple price index number is based on the price or quantity of a single commodity. To construct a simple index, we first have to decide on the base period and then find ratio of the value at any subsequent period to the value in that base period - *the price/quantity relative*. This ratio is then finally converted to a percentage

$$\text{Index for any Period } i = \frac{\text{Value in Period } i}{\text{Value in Base Year}} \times 100$$

i.e. Simple Price Index for period  $i = 1, 2, 3 \dots$  will be

$$P_{0i} = \frac{p_i}{p_0} \times 100 \quad \dots\dots\dots(8-1)$$

Similarly, Simple Quantity Index for period  $i = 1, 2, 3 \dots$  will be

$$Q_{0i} = \frac{q_i}{q_0} \times 100 \quad \dots\dots\dots(8-2)$$

#### Example 8-1

Given are the following price-quantity data of fish, with price quoted in Rs per kg and production in qtls.

Year	:	1980	1981	1982	1983	1984	1985
Price	:	15	17	115	18	22	20
Production	:	500	550	480	1510	1550	1500

Construct:

- the price index for each year taking price of 1980 as base,
- the quantity index for each year taking quantity of 1980 as base.

**Solution:**

#### Simple Price and Quantity Indices of Fish

(Base Year = 1980)

Year	Price ( $p_i$ )	Quantity ( $q_i$ )	Price Index $P_{0i} = \frac{p_i}{p_0} \times 100$	Quantity Index $Q_{0i} = \frac{q_i}{q_0} \times 100$
------	--------------------	-----------------------	--	---



1980	15	500	100.00	100.00
1981	17	550	113.33	110.00
1982	115	480	1015.1515	915.00
1983	18	1510	120.00	122.00
1984	22	1550	1415.1515	130.00
1985	20	1500	133.33	120.00

These simple indices facilitate comparison by transforming absolute quantities/prices into percentages. Given such an index, it is easy to find the percent by which the price/quantity may have changed in a given period as compared to the base period. *For example*, observing the index computed in Example 15-1, one can firmly say that the output of fish was 30 per cent more in 1984 as compared to 1980.

It may also be noted that given the simple price/quantity for the base year and the index for the period  $i = 1, 2, 3, \dots$ ; the actual price/quantity for the period  $i = 1, 2, 3, \dots$  may easily be obtained as:

$$p_i = p_0 \left( \frac{P_{0i}}{100} \right) \quad \dots\dots\dots (8-3)$$

and  $q_i = q_0 \left( \frac{Q_{0i}}{100} \right) \quad \dots\dots\dots (8-4)$

For example, with  $i = 1983$ ,  $Q_{0i} = 122.00$ , and  $q_0 = 500$ ,

$$q_i = 500 \left( \frac{122.00}{100} \right)$$

$$= 1510$$

### 8.1.2.2 Composite Index Numbers

The preceding discussion was confined to only one commodity. What about price/quantity changes in several commodities? In such cases, composite index numbers are used. Depending upon the method used for constructing an index, composite indices may be:

1. Simple Aggregative Price/ Quantity Index
2. Index of Average of Price/Quantity Relatives
3. Weighted Aggregative Price/ Quantity Index



## 4. Index of Weighted Average of Price/Quantity Relatives

**SIMPLE AGGREGATIVE PRICE/ QUANTITY INDEX**

Irrespective of the units in which prices/quantities are quoted, this index for given prices/quantities, of a group of commodities is constructed in the following three steps:

- (i) Find the aggregate of prices/quantities of all commodities for each period (or place).
- (ii) Selecting one period as the base, divide the aggregate prices/quantities corresponding to each period (or place) by the aggregate of prices/ quantities in the base period.
- (iii) Express the result in percent by multiplying by 100.

The computation procedure contained in the above steps can be expressed as:

$$P_{0i} = \frac{\sum P_i}{\sum P_0} \times 100 \quad \dots\dots\dots(8-5)$$

and  $Q_{0i} = \frac{\sum q_i}{\sum q_0} \times 100 \quad \dots\dots\dots(8-15)$

**Example 8-2**

Given are the following price-quantity data, with price quoted in Rs per kg and production in qtls.

Item	1980		1985	
	Price	Production	Price	Production
Fish	15	500	20	1500
Mutton	18	590	23	1540
Chicken	22	450	24	500

- Find
- (a) Simple Aggregative Price Index with 1980 as the base.
  - (b) Simple Aggregative Quantity Index with 1980 as the base.





**Solution:**

**Calculations for  
Simple Aggregative Price and Quantity Indices  
(Base Year = 1980)**

<i>Item</i>	<u>Prices</u>		<u>Quantities</u>	
	1980( $p_0$ )	1985( $p_i$ )	1980( $q_0$ )	1985( $q_i$ )
<b><i>Fish</i></b>	15	20	500	1500
Mutton	18	23	590	1540
Chicken	22	24	450	500
<b>Sum →</b>	55	157	1540	1740

(a) Simple Aggregative Price Index with 1980 as the base

$$P_{0i} = \frac{\sum p_i}{\sum p_0} \times 100$$

$$P_{0i} = \frac{67}{55} \times 100$$

$$P_{0i} = 121.82$$

(b) Simple Aggregative Quantity Index with 1980 as the base

$$Q_{0i} = \frac{\sum q_i}{\sum q_0} \times 100$$

$$Q_{0i} = \frac{1740}{1540} \times 100$$

$$Q_{0i} = 112.98$$

Although Simple Aggregative Index is simple to calculate, it has two important limitations:

First, equal weights get assigned to every item entering into the construction of this index irrespective of relative importance of each individual item being different. *For example*, items like pencil and milk are assigned equal importance in the construction of this index. This limitation renders the index of no practical utility.



Second, different units in which the prices are quoted also sometimes unduly affect this index. Prices quoted in higher weights, such as price of wheat per bushel as compared to a price per kg, will have unduly large influence on this index. Consequently, the prices of only a few commodities may dominate the index. This problem no longer exists when the units in which the prices of various commodities are quoted have a common base.

Even the condition of common base will provide no real solution because commodities with relatively high prices such as gold, which is not as important as milk, will continue to dominate this index excessively. *For example*, in the Example 15-2 given above chicken prices are relatively higher than those of fish, and hence chicken prices tend to influence this index relatively more than the prices of fish.

### INDEX OF AVERAGE OF PRICE/QUANTITY RELATIVES

This index makes an improvement over the index of simple aggregative prices/quantities as it is not affected by the difference in the units of measurement in which prices/quantities are expressed. However, this also suffers from the problem of equal importance getting assigned to all the commodities. Given the prices/quantities of a number of commodities that enter into the construction of this index, it is computed in the following two steps:

- (i) After selecting the base year, find the price relative/quantity relative of each commodity for each year with respect to the base year price/quantity. As defined earlier, the price relative/quantity relative of a commodity for a given period is the ratio of the price/quantity of that commodity in the given period to its price/quantity in the base period.
- (ii) Multiply the result for each commodity by 100, to get simple price/quantity indices for each commodity.
- (iii) Take the average of the simple price/quantity indices by using arithmetic mean, geometric mean or median.

Thus it is computed as:

$$P_{0i} = \text{Average of } \left( \frac{P_i}{P_0} \times 100 \right)$$



and 
$$Q_{0i} = \text{Average of } \left( \frac{q_i}{q_0} \times 100 \right)$$

Using arithmetic mean

$$P_{0i} = \frac{\sum \left( \frac{p_i}{p_0} \times 100 \right)}{N} \dots\dots\dots (8-7)$$

and 
$$Q_{0i} = \frac{\sum \left( \frac{q_i}{q_0} \times 100 \right)}{N} \dots\dots\dots (8-8)$$

Using geometric mean

$$P_{0i} = \text{Anti log} \left[ \frac{1}{N} \sum \log \left( \frac{p_i}{p_0} \times 100 \right) \right] \dots\dots\dots (8-9)$$

and 
$$Q_{0i} = \text{Anti log} \left[ \frac{1}{N} \sum \log \left( \frac{q_i}{q_0} \times 100 \right) \right] \dots\dots\dots (8-10)$$

### Example 8-3

From the data in Example 8.2 find:

- Index of Average of Price Relatives (base year 1980); using mean, median and geometric mean.
- Index of Average of Quantity Relatives (base year 1980); using mean, median and geometric mean.

**Solution:**

#### Calculations for Index of Average of Price Relatives and Quantity Relatives (Base Year = 1980)

Item	Price Relative $= \left( \frac{p_i}{p_0} \times 100 \right)$	$\log \left( \frac{p_i}{p_0} \times 100 \right)$	Quantity Relative $= \left( \frac{q_i}{q_0} \times 100 \right)$	$\log \left( \frac{q_i}{q_0} \times 100 \right)$
Fish	133.33	2.1248	120.00	2.0792
Mutton	127.77	2.10153	108.47	2.0354
Chicken	109.09	2.0378	111.11	2.0457
<b>Sum</b> →	370.19	15.21589	339.58	15.11503



(a) Index of Average of Price Relatives (base year 1980)

Using arithmetic mean

$$P_{0i} = \frac{\sum \left( \frac{p_i}{p_0} \times 100 \right)}{N}$$

$$= \frac{370.19}{3}$$

$$= 123.39$$

Using Median

$$P_{0i} = \text{Size of } \left( \frac{N+1}{2} \right) \text{th item}$$

$$= \text{Size of } \left( \frac{3+1}{2} \right) \text{th item}$$

$$= \text{Size of 2nd item}$$

$$= 127.77$$

Using geometric mean

$$P_{0i} = \text{Anti log} \left[ \frac{1}{N} \sum \log \left( \frac{p_i}{p_0} \times 100 \right) \right]$$

$$= \text{Anti log} \left[ \frac{1}{3} (6.2689) \right]$$

$$= \text{Anti log} [2.08963]$$

$$= 122.92$$

(b) Index of Average of Quantity Relatives (base year 1980)

Using arithmetic mean

$$Q_{0i} = \frac{\sum \left( \frac{q_i}{q_0} \times 100 \right)}{N}$$

$$= \frac{339.58}{3}$$

$$= 113.19$$

Using Median

$$Q_{0i} = \text{Size of } \left( \frac{N+1}{2} \right) \text{th item}$$



$$= \text{Size of } \left( \frac{3+1}{2} \right) \text{th item}$$

$$= \text{Size of 2nd item}$$

$$= 111.11$$

Using geometric mean

$$Q_{0i} = \text{Anti log} \left[ \frac{1}{N} \sum \log \left( \frac{q_i}{q_0} \times 100 \right) \right]$$

$$= \text{Anti log} \left[ \frac{1}{3} (6.1603) \right]$$

$$= \text{Anti log} [2.05343]$$

$$= 113.09$$

Apart from the inherent drawback that this index accords equal importance to all items entering into its construction, a simple arithmetic mean and median are not appropriate average to be applied to ratios. Because it is generally believed that a simple average injects an upward bias in the index. So geometric mean is considered a more appropriate average for ratios and percentages.

### WEIGHTED AGGREGATIVE PRICE/QUANTITY INDICES

We have noted that the simple aggregative price/quantity indices do not take care of the differences in the weights to be assigned to different commodities that enter into their construction. It is primarily because of this limitation that the simple aggregative indices are of very limited use. Weighted aggregative Indices make up this deficiency by assigning proper weights to individual items.

Among several ways of assigning weights, two widely used ways are:

- (i) to use base period quantities/prices as weights, popularly known as **Laspeyre's Index**, and
- (ii) to use the given (current) period quantities/prices as weights, popularly known as **Paasche's Index**.

#### Laspeyre's Index

Laspeyre's Price Index, using base period quantities as weights is obtained as

$$P_{0i}^{La} = \frac{\sum p_i q_0}{\sum p_0 q_0} \times 100 \quad \dots\dots\dots(8-11)$$

Laspeyre's Quantity Index, using base period prices as weights is obtained as



$$Q_{0i}^{La} = \frac{\sum q_i p_0}{\sum q_0 p_0} \times 100 \quad \dots\dots\dots(8-12)$$

### Paasche's Index

Paasche's Price Index, using base period quantities as weights is obtained as

$$P_{0i}^{Pa} = \frac{\sum p_i q_i}{\sum p_0 q_i} \times 100 \quad \dots\dots\dots(8-13)$$

Paasche's Quantity Index, using base period prices as weights is obtained as

$$Q_{0i}^{Pa} = \frac{\sum q_i p_i}{\sum q_0 p_i} \times 100 \quad \dots\dots\dots(8-14)$$

### Example 8-4

From the data in Example 15.2 find:

- Laspeyre's Price Index for 1985, using 1980 as the base
- Laspeyre's Quantity Index for 1985, using 1980 as the base
- Paasche's Price Index for 1985, using 1980 as the base
- Paasche's Quantity Index for 1985, using 1980 as the base

**Solution:**

#### Calculations for Laspeyre's and Paasche's Indices (Base Year = 1980)

Item	$p_0 q_0$	$p_1 q_0$	$p_0 q_1$	$p_1 q_1$
Fish	7500	10000	9000	12000
Mutton	101520	13570	11520	14720
Chicken	9900	10800	11000	12000
<b>Sum</b> →	28020	34370	31520	38720

- Laspeyre's Price Index for 1985, using 1980 as the base

$$P_{0i}^{La} = \frac{\sum p_i q_0}{\sum p_0 q_0} \times 100$$



$$= \frac{34370}{28020} \times 100$$

$$= 122.66$$

- (b) Laspeyre's Quantity Index for 1985, using 1980 as the base

$$Q_{0i}^{La} = \frac{\sum q_i p_0}{\sum q_0 p_0} \times 100$$

$$= \frac{31520}{28020} \times 100$$

$$= 112.49$$

- (c) Paasche's Price Index for 1985, using 1980 as the base

$$P_{0i}^{Pa} = \frac{\sum p_i q_i}{\sum p_0 q_i} \times 100$$

$$= \frac{38720}{31520} \times 100$$

$$= 122.84$$

- (d) Paasche's Quantity Index for 1985, using 1980 as the base

$$Q_{0i}^{Pa} = \frac{\sum q_i p_i}{\sum q_0 p_i} \times 100$$

$$= \frac{38720}{34370} \times 100$$

$$= 112.66$$

### Interpretations of Laspeyre's Index

On close examination it will be clear that the Laspeyre's Price Index offers the following precise interpretations:

1. It compares the cost of collection of a fixed basket of goods selected in the base period with the cost of collecting the same basket of goods in the given (current) period.

Accordingly, the cost of collection of 500 qtls of fish, 590 qtls of mutton and 450 qtls of chicken has increased by 22.1515 per cent in 1985 as compared to what it was in 1980. Viewed differently, it indicates that a fixed amount of goods sold at 1985 prices yield 22.1515 per cent more revenue than what it did at 1980 prices.



2. It also implies that a fixed amount of goods when purchased at 1985 prices would cost 22.1515 per cent more than what it did at 1980 prices. In this interpretation, the Laspeyre's Price Index serves as the basis of constructing the cost of living index, for it tells how much more does it cost to maintain the base period standard of living at the current period prices.

Laspeyre's Quantity Index, too, has precise interpretations. It reveals the percentage change in total expenditure in the given (current) period as compared to the base period if varying amounts of the same basket of goods are sold at the base period prices. When viewed in this manner, we will be required to spend 12.49 per cent more in 1985 as compared to 1980 if the quantities of fish, mutton and chicken for 19155 are sold at the base period (1980) prices.

### **Interpretations of Paasche's Index**

A careful examination of the Paasche's Price Index will show that this too is amenable to the following precise interpretations:

1. It compares the cost of collection of a fixed basket of goods selected in the given period with the cost of collection of the same basket of goods in the base period.

Accordingly, the cost of collection of a fixed basket of goods containing 1500 qtls of fish, 1540 qtls of mutton and 500 qtls of chicken in 1985 is about 22.84 per cent more than the cost of collecting the same basket of goods in 1980. Viewed a little differently, it indicates that a fixed basket of goods sold at 1985 prices yields 22.84 per cent more revenue than what it would have earned had it been sold at the base period (1980) prices.

2. It also tells that a fixed amount of goods purchased at 1985 prices will cost 22.84 per cent more than what it would have cost if this fixed amount of goods had been sold at base period (1980) prices.

Analogously, Paasche's Quantity Index, too, has its own precise meaning. It tells the per cent change in total expenditure in the given period as compared to the base period if varying amounts of the same basket of goods are to be sold at given period prices. When so viewed, we will be required to spend 12.1515 per cent more in 1985 as compared to 1980 if the quantities of fish, mutton and chicken for 1980 are sold at the given period (1985) prices.





### Relationship Between Laspeyre's and Paasche's Indices

In order to understand the relationship between Laspeyre's and Paasche's Indices, the assumptions on which the two indices are based be borne in mind:

Laspeyre's index is based on the assumption that *unless there is a change in tastes and preferences, people continue to buy a fixed basket of goods irrespective of how high or low the prices are likely to be in the future*. Paasche's index, on the other hand, assumes that *people would have bought the same amount of a given basket of goods in the past irrespective of how high or low were the past prices*.

However, the basic contention implied in the assumptions on which the two indices are based is not true. For, people do make shifts in their purchase pattern and preferences by buying more of goods that tend to become cheaper and less of those that tend to become costlier. In view of this, the following two situations that are likely to emerge need consideration:

1. When the prices of goods that enter into the construction of these indices show a general tendency to rise, those whose prices increase more than the average increase in prices will have smaller quantities in the given period than the corresponding quantities in the base period. That is,  $q_i$ 's will be smaller than  $q_0$ 's when prices in general are rising. Consequently, Paasche's index will have relatively smaller weights than those in the Laspeyre's index and, therefore, the former ( $P_{0i}^{Pa}$ ) will be smaller than the latter ( $P_{0i}^{La}$ ). In other words, Paasche's index will show a relatively smaller increase when the prices in general tend to rise.
2. On the contrary, when prices in general are falling, goods whose prices show a relatively smaller fall than the average fall in prices, will have smaller quantities in the given period than the corresponding quantities in the base period. This means that  $q_i$ 's will be smaller than  $q_0$ 's when prices in general are falling. Consequently, Paasche's index will have smaller weights than those in the Laspeyre's index and, therefore, the former ( $P_{0i}^{Pa}$ ) will be smaller than the latter ( $P_{0i}^{La}$ ). In other words, Paasche's index will show a relatively greater fall when the prices in general tend to fall.

An important inference based on the above discussion is that *the Paasche's index has a downward bias and the Laspeyre's index an upward bias*. This directly follows from the fact that the Paasche's index, relative to the Laspeyre's index, shows a smaller rise when the prices in general are rising, and a greater fall when the prices in general are falling.



It may, however, be noted that when the quantity demanded increases because of change in real income, tastes and preferences, advertising, *etc.*, the prices remaining unchanged, the Paasche's index will show a higher value than the Laspeyre's index. In such situations, the Paasche's index will overstate, and the Laspeyre's will understate, the changes in prices. The former now represents the upper limit, and the latter the lower limit, of the range of price changes.

The relationship between the two indices can be derived more precisely by making use of the coefficient of linear correlation computed as:

$$r_{xy} = \frac{\frac{\sum fXY}{N} - \left( \frac{\sum fX}{N} \right) \left( \frac{\sum fY}{N} \right)}{S_x S_y} \dots\dots\dots (8.15)$$

in which  $X$  and  $Y$  denote the relative price movements ( $\frac{p_i}{p_0}$ ) and relative quantity movements ( $\frac{q_i}{q_0}$ ) respectively.  $S_x$  and  $S_y$  are the standard deviations of price and quantity movements, respectively. While  $r_{xy}$  represents the coefficient of correlation between the relative price and quantity movements;  $f$  represents the weights assigned, that is,  $p_0 q_0$ .  $N$  is the sum of frequencies *i. e.*  $N = \sum p_0 q_0$ .

Substituting the values of  $X$ ,  $Y$ ,  $f$  and  $N$  in Eq. (15-15), and then rearranging the expression, we have

$$r_{xy} S_x S_y = \frac{\sum p_i q_i}{\sum p_0 q_0} - \frac{\sum p_i q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_i}{\sum p_0 q_0}$$

If  $\frac{\sum p_i q_i}{\sum p_0 q_0} = V_{0i}$ , is the index of value expanded between the base period and the  $i^{th}$  period, then

dividing both sides by  $\frac{\sum p_i q_i}{\sum p_0 q_0}$  or  $V_{0i}$ , we get

$$\begin{aligned} \frac{r_{xy} S_x S_y}{V_{0i}} &= 1 - \frac{\sum p_i q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_i}{\sum p_i q_i} \\ \frac{r_{xy} S_x S_y}{V_{0i}} &= 1 - P_{0i}^{La} \times \frac{1}{P_{0i}^{Pa}} \\ \frac{P_{0i}^{La}}{P_{0i}^{Pa}} &= 1 - \frac{r_{xy} S_x S_y}{V_{0i}} \dots\dots\dots (8.115) \end{aligned}$$



The relationship in *Eq. (8.115)* offers the following useful results:

1.  $P_{0i}^{La} = P_{0i}^{Pa}$  when either  $r_{xy}$ ,  $S_x$  and  $S_y$  is equal to zero. That is, the two indices will give the same result either when there is no correlation between the price and quantity movements, or when the price or quantity movements are in the same ratio so that  $S_x$  or  $S_y$  is equal to zero.
2. Since in actual practice  $r_{xy}$  will have a negative value between 0 and -1, and as neither  $S_x = 0$  nor  $S_y = 0$ , the right hand side of *Eq. (15-115)* will be less than 1. This means that  $P_{0i}^{La}$  is normally greater than  $P_{0i}^{Pa}$ .
3. Given the overall movement in the index of value ( $V_{0i}$ ) expanded, the greater the coefficient of correlation ( $r_{xy}$ ) between price and quantity movements and/or the greater the degree of dispersion ( $S_x$  and  $S_y$ ) in the price and quantity movements, the greater the discrepancy between  $P_{0i}^{La}$  and  $P_{0i}^{Pa}$ .
4. The longer the time interval between the two periods to be compared, the more the chances for price and quantity movements leading to higher values of  $S_x$  and  $S_y$ . The assumption of tastes, habits, and preferences remaining unchanged breaking down over a longer period, people do find enough time to make shifts in their consumption pattern, buying more of goods that may have become relatively cheaper and less of those that may have become relatively dearer. All this will end up with a higher degree of correlation between the price and quantity movement. Consequently,  $P_{0i}^{La}$  will diverge from  $P_{0i}^{Pa}$  more in the long run than in the short run. So long as the periods to be compared are not much apart,  $P_{0i}^{La}$  will be quite close to  $P_{0i}^{Pa}$ .

### Laspeyre's and Paasche's Indices Further Considered

The use of different system of weights in these two indices may give an impression as if they are opposite to each other. Such an impression is not sound because both serve the same purpose, although they may give different results when applied to the same data.

This raises an important question. Which one of them gives more accurate results and which one should be preferred over the other? The answer to this question is rather difficult since both the indices are amenable to precise and useful results.

Despite a very useful and precise difference in interpretation, in actual practice the Laspeyre's index is used more frequently than the Paasche's index for the simple reason that the latter requires frequent



revision to take into account the yearly changes in weights. No such revision is required in the case of the Laspeyre's index where once the weights have been determined, these do not require any change in any subsequent period. It is on this count that the Laspeyre's index is preferred over the Paasche's index. However, this does not render the Paasche's index altogether useless. In fact, it supplements the practical utility of the Laspeyre's index. The fact that the Laspeyre's index has an upward bias and the Paasche's index downward bias, the two provide the range between which the index can vary between the base period and the given period. Interestingly, thus, the former represents the upper limit, and the latter the lower limit.

### Improvements over the Laspeyre's and Paasche's Indices

To overcome the difficulty of overstatement of changes in prices by the Laspeyre's index and understatement by the Paasche's index, different indices have been developed to compromise and improve upon them. These are particularly useful when the given period and the base period fall quite apart and result in a greater divergence between Laspeyre's and Paasche's indices.

Other important Weighted Aggregative Indices are:

#### 1. Marshall-Edgeworth Index

The Marshall-Edgeworth Index uses the average of the base period and given period quantities/prices as the weights, and is expressed as

$$P_{0i}^{ME} = \frac{\sum p_i \left( \frac{q_0 + q_i}{2} \right)}{\sum p_0 \left( \frac{q_0 + q_i}{2} \right)} \times 100 \quad \dots\dots\dots(8-17)$$

$$Q_{0i}^{ME} = \frac{\sum q_i \left( \frac{p_0 + p_i}{2} \right)}{\sum q_0 \left( \frac{p_0 + p_i}{2} \right)} \times 100 \quad \dots\dots\dots(8-18)$$

#### 2. Dorbish and Bowley Index

The Dorbish and Bowley Index is defined as the arithmetic mean of the Laspeyre's and Paasche's indices.



$$P_{0i}^{DB} = \frac{P_{0i}^{La} + P_{0i}^{Pa}}{2} \quad \dots\dots\dots(8-19)$$

$$Q_{0i}^{DB} = \frac{Q_{0i}^{La} + Q_{0i}^{Pa}}{2} \quad \dots\dots\dots(8-20)$$

### 3. Fisher's Ideal Index

The Fisher's Ideal Index is defined as the geometric mean of the Laspeyre's and Paasche's indices.

$$P_{0i}^F = \sqrt{P_{0i}^{La} \cdot P_{0i}^{Pa}} \quad \dots\dots\dots(8-21)$$

$$Q_{0i}^F = \sqrt{Q_{0i}^{La} \cdot Q_{0i}^{Pa}} \quad \dots\dots\dots(8-22)$$

### Index of Weighted Average of Price/Quantity Relatives

An alternative system of assigning weights lies in using value weights. The value weight  $v$  for any single commodity is the product of its price and quantity, that is,  $v = pq$ .

If the index of weighted average of price relatives is defined as

$$P_{0i} = \frac{\sum \left[ v \left( \frac{p_i}{p_0} \times 100 \right) \right]}{\sum v} \quad \dots\dots\dots(8-23)$$

then  $v$  can be obtained either as

(i) the product of the base period prices and the base period quantities denoted as  $v_0$  that is,  $v_0 = p_0 q_0$ , or

(ii) the product of the base period prices and the given period quantities denoted as  $v_i$  that is,  $v_i = p_0 q_i$

When  $v$  is  $v_0 = p_0 q_0$ , the index of weighted average of price relatives, is expressed as

$${}_0P_{0i} = \frac{\sum \left[ p_0 q_0 \left( \frac{p_i}{p_0} \times 100 \right) \right]}{\sum p_0 q_0} \quad \dots\dots\dots(8-24)$$

It may be seen that  ${}_0P_{0i}$  is the same as the Laspeyre's aggregative price index.



Similarly, When  $v$  is  $v_i = p_0 q_i$ , the index of weighted average of price relatives, is expressed as

$${}_i P_{0i} = \frac{\sum \left[ p_0 q_i \left( \frac{p_i}{p_0} \times 100 \right) \right]}{\sum p_0 q_i} \dots\dots\dots (8-25)$$

It may be seen that  ${}_i P_{0i}$  is the same as the Paasche's aggregative price index.

If the index of weighted average of quantity relatives is defined as

$$Q_{0i} = \frac{\sum \left[ v \left( \frac{q_i}{q_0} \times 100 \right) \right]}{\sum v} \dots\dots\dots (8-215)$$

then  $v$  can be obtained either as

(i) the product of the base period quantities and the base period prices denoted as  $v_0$  that is,  $v_0 = q_0 p_0$ , or

(ii) the product of the base period quantities and the given period prices denoted as  $v_i$  that is,  $v_i = q_0 p_i$

When  $v$  is  $v_0 = q_0 p_0$ , the index of weighted average of quantity relatives, is expressed as

$${}_0 Q_{0i} = \frac{\sum \left[ q_0 p_0 \left( \frac{q_i}{q_0} \times 100 \right) \right]}{\sum q_0 p_0} \dots\dots\dots (8-27)$$

It may be seen that  ${}_0 Q_{0i}$  is the same as the Laspeyre's aggregative quantity index.

Similarly, When  $v$  is  $v_i = q_0 p_i$ , the index of weighted average of quantity relatives, is expressed as

$${}_i Q_{0i} = \frac{\sum \left[ q_0 p_i \left( \frac{q_i}{q_0} \times 100 \right) \right]}{\sum q_0 p_i} \dots\dots\dots (8-28)$$

It may be seen that  ${}_i Q_{0i}$  is the same as the Paasche's aggregative quantity index.

### **Example 8-5**



From the data in Example 8.2 find the:

- (a) Index of Weighted Average of Price Relatives, using
- $v_0 = p_0 q_0$  as the value weights
  - $v_i = p_0 q_i$  as the value weights
- (b) Index of Weighted Average of Quantity Relatives, using
- $v_0 = q_0 p_0$  as the value weights
  - $v_i = q_0 p_i$  as the value weights

**Solution:**

**Calculations for**  
**Index of Weighted Average of Price Relatives**  
**(Base Year = 1980)**

Item	$v_0 = p_0 q_0$	$v_1 = p_0 q_1$	$p_0 q_0 \left( \frac{p_i}{p_0} \times 100 \right)$	$p_0 q_1 \left( \frac{p_1}{p_0} \times 100 \right)$
Fish	7500	9000	1000000	1200000
Mutton	101520	11520	1357000	1472000
Chicken	9900	11000	1080000	1200000
<b>Sum</b> →	28020	31520	3437000	3872000

- (a) Index of Weighted Average of Price Relatives, using
- $v_0 = p_0 q_0$  as the value weights

$$\begin{aligned}
 {}_0P_{0i} &= \frac{\sum \left[ p_0 q_0 \left( \frac{p_i}{p_0} \times 100 \right) \right]}{\sum p_0 q_0} \\
 &= \frac{3437000}{28020} \\
 &= 122.66
 \end{aligned}$$

- $v_i = p_0 q_i$  as the value weights



$${}_iP_{0i} = \frac{\sum \left[ p_0 q_i \left( \frac{p_i}{p_0} \times 100 \right) \right]}{\sum p_0 q_i}$$

$$= \frac{3872000}{31520}$$

$$= 122.84$$

**Calculations for**  
**Index of Weighted Average of Quantity Relatives**  
**(Base Year = 1980)**

Item	$v_0 = q_0 p_0$	$v_1 = q_0 p_1$	$q_0 p_0 \left( \frac{q_1}{q_0} \times 100 \right)$	$q_0 p_1 \left( \frac{q_1}{q_0} \times 100 \right)$
Fish	7500	10000	900000	1200000
Mutton	101520	13570	1152000	1472000
Chicken	9900	10800	1100000	1200000
<b>Sum</b> →	28020	34370	3152000	3872000

(b) Index of Weighted Average of Quantity Relatives, using

(i)  $v_0 = q_0 p_0$  as the value weights

$${}_0Q_{0i} = \frac{\sum \left[ q_0 p_0 \left( \frac{q_i}{q_0} \times 100 \right) \right]}{\sum q_0 p_0}$$

$$= \frac{3152000}{28020}$$

$$= 112.49$$

(ii)  $v_i = q_0 p_i$  as the value weights

$${}_iQ_{0i} = \frac{\sum \left[ q_0 p_i \left( \frac{q_i}{q_0} \times 100 \right) \right]}{\sum q_0 p_i}$$





$$\begin{aligned} &= \frac{3872000}{34370} \\ &= 112.66 \end{aligned}$$

Although the indices of weighted average of price/quantity relatives yield the same results as the Laspeyre's or Paasche's price/quantity indices, we do construct these indices also in situations when it is necessary and advantageous to do so. Some such situations are as follows:

- (i) When a group of commodities is to be represented by a single commodity in the group, the price relative of the latter is weighted by the group as a whole.
- (ii) Where the price/quantity relatives of individual commodities have been computed, these can be more conveniently utilised in constructing the index.
- (iii) Price/quantity relatives serve a useful purpose in splicing two index series having different base periods.
- (iv) Depersonalizing a time series requires construction of a seasonal index, which also requires the use of relatives.

## 8.2 TESTS OF ADEQUACY OF INDEX NUMBERS

We have discussed various formulae for the construction of index numbers. None of the formulae measures the price changes or quantity changes with perfection and has some bias. The problem is to choose the most appropriate formula in a given situation. As a measure of the formula error a number of mathematical tests, known as the *tests of consistency* or *tests of adequacy* of index number formulae have been suggested. In this section we will discuss these tests, which are also sometimes termed as the criteria for a good index number.

1. **Unit Test:** This test requires that the index number formula should be independent of the units in which the prices or quantities of various commodities are quoted. All the formulae discussed in the lesson except the index number based on Simple Aggregate of Prices/Quantities satisfy this test.
2. **Time Reversal Test:** The time reversal test, proposed by Prof Irving Fisher requires the index number formula to possess time consistency by working both forward and backward *w.r.t.* time. In his (Fisher's) words:



*“The formula for calculating an index number should be such that it gives the same ratio between one point of comparison and the other, no matter which of the two is taken as the base or putting it another way, the index number reckoned forward should be reciprocal of the one reckoned backward.”*

In other words, if the index numbers are computed for the same data relating to two periods by the same formula but with the bases reversed, then the two index numbers so obtained should be the reciprocals of each other. Mathematically, we should have (omitting the factor 100),

$$P_{ab} \times P_{ba} = 1 \quad \dots\dots\dots(8-29)$$

or more generally

$$P_{01} \times P_{10} = 1 \quad \dots\dots\dots(8-29a)$$

Time reversal test is satisfied by the following index number formulae:

- (i) Marshall-Edgeworth formula
- (ii) Fisher's Ideal formula
- (iii) Kelly's fixed weight formula
- (iv) Simple Aggregate index
- (v) Simple Geometric Mean of Price Relatives formula
- (vi) Weighted Geometric Mean of Price Relatives formula with fixed weights

Lespeyre's and Pasche's index numbers do not satisfy the time reversal test.

**3. Factor Reversal Test:** This is the second of the two important tests of consistency proposed by Prof Irving Fisher. According to him:

*“Just as our formula should permit the interchange of two times without giving inconsistent results, so it ought to permit interchanging the price and quantities without giving inconsistent results – i.e., the two results multiplied together should give the true value ratio, except for a constant of proportionality.”*

This implies that if the price and quantity indices are obtained for the same data, same base and current periods and using the same formula, then their product (without the factor 100) should give the true value ratio. Symbolically, we should have (without factor 100).

$$P_{01} \times Q_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_0} = V_{01} \quad \dots\dots\dots(8-30)$$

Fisher's formula satisfies the factor reversal test. In fact fisher's index is the only index satisfying this test as none of the formulae discussed in the lesson satisfies this test.



**Remark:** Since Fisher's index is the only index that satisfies both the time reversal and factor reversal tests, it is termed as Fisher's Ideal Index.

- 4. Circular Test:** Circular test, first suggested by Westergaard, is an extension of time reversal test for more than two periods and is based on the shift ability of the base period. This requires the index to work in a circular manner and this property enables us to find the index numbers from period to period without referring back to the original base each time. For three periods  $a, b, c$ , the test requires :

$$P_{ab} \times P_{bc} \times P_{ca} = 1 \quad a \neq b \neq c \quad \dots\dots\dots(8-31)$$

In the usual notations *Eq. (8-31)* can be stated as:

$$P_{01} \times P_{12} \times P_{20} = 1 \quad \dots\dots\dots(8-31a)$$

For Instance

$$P_{01}^{La} \times P_{12}^{La} \times P_{21}^{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_2 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_2}{\sum p_2 q_2} \neq 1$$

Hence Laspeyre's index does not satisfy the circular test. In fact, circular test is not satisfied by any of the weighted aggregative formulae with changing weights. This test is satisfied only by the index number formulae based on:

- (i) Simple geometric mean of the price relatives, and
- (ii) Kelly's fixed base method

## 8.3 SPECIAL ISSUES AND PROBLEMS IN THE CONSTRUCTION OF INDEX NUMBERS

### SPECIAL ISSUES

#### BASE SHIFTING

The need for shifting the base may arise either

- (i) when the base period of a given index number series is to be made more recent,
- (ii) when two index number series with different base periods are to be compared,
- (iii) when there is need for splicing two overlapping index number series.

Whatever be the reason, the technique of shifting the base is simple:



$$\text{New Base Index Number} = \frac{\text{Old Index Number of Current Year}}{\text{Old Index Number of New Base Year}} \times 100$$

### Example 8-15

Reconstruct the following indices using 1997 as base:

Year :	1991	1992	1993	1994	1995	19915	1997	1998
Index :	100	110	130	150	175	180	200	220

### **Solution:**

#### *Shifting the Base Period*

Year	Index Number (1991 = 100)	Index Number (1997 = 100)
1991	100	$(100/200) \times 100 = 50.00$
1992	110	$(110/200) \times 100 = 55.00$
1993	130	$(130/200) \times 100 = 65.00$
1994	150	$(150/200) \times 100 = 75.00$
1995	175	$(175/200) \times 100 = 87.50$
19915	180	$(180/200) \times 100 = 90.00$
1997	200	$(200/200) \times 100 = 100.00$
1998	220	$(220/200) \times 100 = 110.00$

### **SPLICING TWO OVERLAPPING INDEX NUMBER SERIES**

Splicing two index number series means reducing two overlapping index series with different base periods into a single series either at the base period of the old series (one with an old base year), or at the base period of the new series (one with a recent base year). This actually amounts to changing the weights of one series into the weights of the other series.

#### **1. Splicing the New Series to Make it Continuous with the Old Series**

Here we reduce the new series into the old series after the base year of the former. As shown in Table 15.8.2(i), splicing here takes place at the base year (1980) of the new series. To do this, a ratio of the index for 1980 in the old series (200) to the index of 1980 in the new series (100) is computed and the index for each of the following years in the new series is multiplied by this ratio.

**Table 8.8.2(i)**



### Splicing the New Series with the Old Series

Year	Price Index (19715 = 100) (Old Series)	Price Index (1980 = 100) (New Series)	Spliced Index Number [New Series $\times$ (200/100)]
1971	100	--	100
1977	120	--	120
1978	1415	--	1415
1979	172	--	172
1980	200	100	200
1981	--	110	220
1982	--	1115	232
1983	--	125	250
1984	--	140	280

## 2. Splicing the Old Series to Make it Continuous with the New Series

This means reducing the old series into the new series before the base period of the letter. As shown in Table 15.8.2(ii), splicing here takes place at the base period of the new series. To do this, a ratio of the index of 1980 of the new series (100) to the index of 1980 of the old series (200) is computed and the index for each of the preceding years of the old series are then multiplied by this ratio.

**Table 8.8.2(ii)**

### Splicing the Old Series with the New Series

Year	Price Index (19715 = 100) (Old Series)	Price Index (1980 = 100) (New Series)	Spliced Index Number [Old Series $\times$ (100/200)]
1971	100	--	50
1977	120	--	150
1978	1415	--	73.50
1979	172	--	815
1980	200	100	100



1981	--	110	110
1982	--	1115	1115
1983	--	125	125
1984	--	140	140

### CHAIN BASE INDEX NUMBERS

The various indices discussed so far are fixed base indices in the sense that either the base year quantities/prices (or the given year quantities/prices) are used as weights. In a dynamic situation where tastes, preferences, and habits are constantly changing, the weights should be revised on a continuous basis so that new commodities are included and the old ones deleted from consideration.

This is all the more necessary in a developing society where new substitutes keep replacing the old ones, and completely new commodities are entering the market. To take care of such changes, the base year should be the most recent, that is, the year immediately preceding the given year. This means that as we move forward, the base year should move along the given year in a chain year after year.

#### *Conversion of Fixed-base Index into Chain-base Index*

As shown in Table 15.8.3(i), to convert fixed-base index numbers into chain-base index numbers, the following procedure is adopted:

- The first year's index number is taken equal to 100
- For subsequent years, the index number is obtained by following formula:

$$\text{Current Year's CBI} = \frac{\text{Current Year's FBI}}{\text{Previous Year's CBI}} \times 100$$

**Table 8.8.3(i)**

#### **Conversion of Fixed-base Index into Chain-base Index**

Year	Fixed Base Index Number (FBI)	Conversion	Chain Base Index Number (CBI)
1975	3715	--	100
19715	392	$(392/3715) \times 100$	104.3
1977	408	$(408/392) \times 100$	104.1



1978	380	$(380/408) \times 100$	93.1
1979	392	$(392/380) \times 100$	103.2
1980	400	$(400/392) \times 100$	102

### Conversion of Chain-base Index into Fixed-base Index

As shown in Table 8.8.3(ii), to convert fixed-base index numbers into chain-base index numbers, the following procedure is adopted:

- The first year's index is taken what the chain base index is; but if it is to form the base it is taken equal to 100
- In subsequent years, the index number is obtained by following formula:

$$\text{Current Year's FBI} = \frac{\text{Current Year's CBI} \times \text{Previous Year's FBI}}{100}$$

**Table 8.8.3(ii)**

### Conversion of Chain-base Index into Fixed-base Index

Year	Chain Base Index Number (CBI)	Conversion	Fixed Base Index Number (FBI)
1978	90	--	90
1979	120	$(120 \times 90) / 100$	108
1980	125	$(125 \times 108) / 100$	135
1981	110	$(110 \times 135) / 100$	148.5
1982	112	$(112 \times 148.5) / 100$	11515.3
1983	150	$(150 \times 11515.3) / 100$	249.45

### PROBLEMS OF CONSTRUCTING INDEX NUMBERS

The above discussion enables us to identify some of the important problems, which may be faced in the construction of index numbers:

**Choice of the Base Period:** Choice of the base period is a critical decision because of its importance in the construction of index numbers. A base period is the reference period for describing and comparing



the changes in prices or quantities in a given period. The selection of a base year or period does not pose difficult theoretical questions. To a large extent, the choice of the base year depends on the objective of the index. A major consideration should be to ensure that the base year is not an abnormal year. *For example*, a base period with very low price/quantity will unduly inflate, while the one with a very high figure will unduly depress, the entire index number series. An index number series constructed with any such period as the base may give very misleading results. It is, therefore, necessary that the base period be selected carefully.

Another important consideration is that the base year should not be too remote in the past. A more recent year needs to be selected as the base year. The use of a particular year for a prolonged period would distort the changes that it purports to measure. That is why we find that the base year of major index numbers, such as consumer price index or index of industrial production, is shifted from time to time.

***Selection of Weights to be Used:*** It should be amply clear from the various indices discussed in the lesson that the choice of the system of weights, which may be used, is fairly large. Since any system of weights has its own merits and is capable of giving results amenable to precise interpretations, the weights used should be decided keeping in view the purpose for which an index is constructed.

It is also worthwhile to bear in mind that the use of any system of weights should represent the relative importance of individual commodities that enter into the construction of an index. The interpretations that are intended to be made from an index number are also important in deciding the weights. The use of a system of weights that involves heavy computational work deserves to be avoided.

***Type of Average to be Used:*** What type of average should be used is a problem specific to simple average indices. Theoretically, one can use any of the several averages that we have, such as mean, median, mode, harmonic mean, and geometric mean. Besides being locational averages, median and mode are not the appropriate averages to use especially where the number of years for which an index is to be computed, is not large.

While the use of harmonic mean and geometric mean has some definite merits over mean, particularly when the data to be averaged refer to ratios, mean is generally more frequently used for convenience in computations.





**Choice of Index:** The problem of selection of an appropriate index arises because of availability of different types of indices giving different results when applied to the same data. Out of the various indices discussed, the choice should be in favour of one which is capable of giving more accurate and precise results, and which provides answer to specific questions for which an index is constructed.

While the Fisher's index may be considered ideal for its ability to satisfy the tests of adequacy, this too suffers from two important drawbacks. First, it involves too lengthy computations, and second, it is not amenable to easy interpretations as are the Laspeyre's and Paasche's indices. The use of the term ideal does not, however, mean that it is the best to use under all types of situations. Other indices are more appropriate under situations where specific answers are needed.

**Selection of Commodities:** Commodities to be included in the construction of an index should be carefully selected. Only those commodities deserve to be included in the construction of an index as would make it more representative. This, in fact, is a problem of sampling, for being related to the selection of commodities to be included in the sample.

In this context, it is important to note that the selection of commodities must not be based on random sampling. The reason being that in random sampling every commodity, including those that are not important and relevant, have equal chance of being selected, and consequently, the index may not be representative. The choice of commodities has, therefore, to be deliberate and in keeping with the relevance and importance of each individual commodity to the purpose for which the index is constructed.

**Data Collection:** Collection of data through a sample is the most important issue in the construction of index numbers. The data collected are the raw material of an index. Data quality is the basic factor that determines the usefulness of an index. The data have to be as accurate, reliable, comparable, representative, and adequate, as possible.

The practical utility of an index also depends on how readily it can be constructed. Therefore, data should be collected from where these can be easily available. While the purpose of an index number will indicate what type of data are to be collected, it also determines the source from where the data can be available.

## 8.4 CHECK YOUR PROGRESS

1. Index Numbers measure changes not capable of ..... measurement.
2. The wholesale price index numbers reflect the changes in the ..... level of a country.



3. Weighted aggregative Indices make up this ..... by assigning proper weights to individual items.
4. The Paasche's index has a ..... bias and the Laspeyre's index an upward bias.
5. Splicing two index number series means reducing two ..... index series with different base periods into a single series either at the base period of the old series (one with an old base year), or at the base period of the new series (one with a recent base year).

## 8.5 SUMMARY

Index numbers are statistical devices designed to measure the relative changes in the level of a certain phenomenon in two or more situation. It may refer to a single variable or a group of distinct but related variables. There are different uses of Index numbers like: act as an economic Barometers, help in Studying Trends and Tendencies, help in Formulating Decisions and Policies, measure the Purchasing Power of Money and used for deflation. There are following types of it like: price index numbers, quantity index numbers and value index numbers. Various indices can also be distinguished on the basis of the number of commodities that go into the construction of an index. Indices constructed for individual commodities or variable are termed as simple index numbers. Those constructed for a group of commodities or variables are known as aggregative (or composite) index numbers. There are various formulae for the construction of index numbers. None of the formulae measures the price changes or quantity changes with perfection and has some bias. The problem is to choose the most appropriate formula in a given situation. As a measure of the formula error a number of mathematical tests, known as the tests of consistency or tests of adequacy. There are different test for it like: unit test, time reversal test, factor reversal test and circular test.

## 8.6 KEYWORDS

**Price Index Numbers:** The price index numbers measure the general changes in the prices.

**Quantity Index Numbers:** Quantity index numbers study the changes in the volume of goods produced (manufactured), consumed or distributed.

**Value Index Numbers:** These are intended to study the change in the total value (price multiplied by quantity) of output such as indices of retail sales or profits or inventories.

**Simple index numbers:** A simple price index number is based on the price or quantity of a single commodity.



**Laspeyres's index:** It uses base period quantities/prices as weights

**Paasche's Index:** It uses the given (current) period quantities/prices as weights.

**Splicing two index number series:** It means reducing two overlapping index series with different base periods into a single series either at the base period of the old series (one with an old base year), or at the base period of the new series (one with a recent base year).

## 8.7 SELF-ASSESSMENT TEST

1. "Index Numbers are devices for measuring changes in the magnitude of a group of related variables". Discuss this statement and point out the important uses of index numbers.
2. "Index Numbers are Economic Barometers". Discuss this statement. What precautions would you take while constructing index numbers?
3. (a) Explain the uses of index numbers.  
(b) What problems are involved in the construction of index numbers?
4. Describe each of the following:
  - a. Base period
  - b. Price relatives
  - c. Fixed-base index numbers
  - d. Chain-base index numbers
5. Describe briefly the following methods of construction of price index numbers:
  - a. Simple Aggregate Method
  - b. Simple Average of Price Relatives Method
  - c. Weighted Aggregate Method
  - d. Weighted Average of Price Relatives
6. "Laspeyres's index has an upward bias and the Paasche's index downward bias". Explain this statement.
7. Discuss the various tests of adequacy of index numbers.
8. State and explain the Fisher's ideal formula for price index number. Show how it satisfies the time-reversal and factor-reversal test? Why is it used little in practice?
9. Briefly explain each of the following:
  - a. Base-shifting



- b. Splicing
- c. Deflating

10. From the following data, construct the price index for each year with price of 1995 as base.

Year:	1995	19915	1997	1998	1999	2000
Price of Commodity:	40	50	45	55	155	70

11. From the following data, construct an index number for 2004 taking 2003 as base year:

Articles:	A	B	C	D	E
Prices (2003):	100	125	50	40	5
Prices (2004):	140	200	80	150	10

12. Find the index number for 1982 and 1983 taking 1981 as base year by the Simple Average of Price Relatives Method, using (i) Mean, (ii) Median, and (iii) Geometric Mean:

Commodities	1981 (Prices)	1982 (Prices)	1983 (Prices)
A	40	55	150
B	50	150	80
C	152	72	93
D	80	88	915
E	20	24	30

13. Construct index number of price and index number of quantity from the following data using:

- a. Laspeyre's formula,
- b. Paasche's formula,
- c. Dorbish and Bowley's formula,
- d. Marshall and Edgeworth's formula, and
- e. Fisher's Ideal Index formula

Commodities	Base Year		Current Year	
	Price	Quantity	Price	Quantity
A	2	8	4	15
B	5	10	15	5
C	4	14	5	10
D	2	19	2	13

Which of the formula satisfy



- (i) the time reversal test, and  
(ii) the factor reversal test?

14. Calculate index number using Kelly's Method of Standard Weights, from the following data:

Commodities	Quantity	Base Year Price	Current Year Price
A	5	30	40
B	8	20	30
C	10	10	20

15. From the following data, construct price index by using Weighted Average of Price Relatives Method:

Commodities	Quantity	Base Year Price	Current Year Price
A	15 Qtl	5.00	15.00
B	5 Qtl	5.00	8.00
C	1 Qtl	15.00	9.00
D	4 Kg	8.00	10.00
E	1 Kg	20.00	15.00

16. From the information given below, calculate the Cost of Living Index number for 1985, with 1984 as base year by

- a. Aggregative Expenditure Method, and  
b. Family Budget Method.

Items	Quantity consumed	Unit	Prices in 1984	Prices in 1985
Wheat	2 Qtl	Qtl	75	125
Rice	20 Kg	Kg	12	115
Sugar	10 Kg	Kg	12	115
Ghee	5 Kg	Kg	10	15
Clothing	25 Meter	Meter	4.5	5



Fuel	40 Litre	Litre	10	12
Rent	One house	House	25	40

17. An enquiry into budgets of the middle class families in a city gave the following information:

Expenses on →	Food	Rent	Clothing	Fuel	Miscellaneous
	40%	10%	20%	10%	20%
Prices(2001)	1150	50	150	20	50
Prices(2002)	175	150	75	25	75

What changes in the cost of living figure of 2002 have taken place as compared to 2001?

18. Reconstruct the following indices using 1985 as base:

Year	:	1982	1983	1984	1985	1986	1987
Index	:	100	120	190	200	212	250

19. Given below are two sets of indices one with 1975 as base and the other with 1979 as base:

First set

Year	:	1975	1976	1977	1978	1979
Index Numbers	:	100	110	125	180	200

Second Set

Year	:	1979	1980	1981	1982	1983
Index Numbers	:	100	104	110	115	124

a. Splice the second set of index numbers from 1975

b. Splice the first set of index numbers from 1979

20. Construct chain index numbers from the following data:

Year :	1991	1992	1993	1994	1995
Price :	25	30	45	150	90

21. Convert into Chain Base Index Number from Fixed Base Index Number

Year	:	1980	1981	1982	1983	1984
Fixed Base Index	:	100	98	102	140	190

22. From the Chain Base Index numbers given below, construct Fixed Base Index numbers:

Year	:	1993	1994	1995	1996	1997
Chain Base Index	:	100	105	95	115	102



23. From the following data, prepare index number for real wages of workers:

Year	:	1990	1991	1992	1993	1994	1995
Wages (in Rs)	:	300	340	450	4150	475	540
Price Index Number	:	100	120	220	230	250	300

24. During certain period, the Cost of Living Index number went up from 110 to 200 and salary of a worker also raised from 325 to 500. State by how much the worker has gained or lost in real term.

## 8.8 ANSWERS TO CHECK YOUR PROGRESS

1. Direct
2. General Price
3. Deficiency
4. Downward
5. Overlapping

## 8.9 REFERENCES/SUGGESTED READINGS

1. Statistics (Theory & Practice) by Dr. B.N. Gupta. Sahitya Bhawan Publishers and Distributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing Company Ltd., New Delhi.
3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill Publishing Company Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., New Delhi.
5. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.
6. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.
7. Statistics for Management by Richard I. Levin and David S. Rubin. Prentice Hall of India Pvt. Ltd., New Delhi.
8. Statistics for Business and Economics by Kohlar Heinz. Harper Collins., New York.



Subject: Business Statistics-II	
Course Code: BCOM 402	Author: Dr. B. S. Bodla
Lesson No. 9	Vetter: Karam Pal
<b>ANALYSIS OF TIME SERIES</b>	

## **STRUCTURE**

- 9.0 Learning Objectives
- 9.1 Introduction
  - 9.1.1 Objectives of time series analysis
  - 9.1.2 Components of time series
  - 9.1.3 Time series decomposition models
  - 9.1.4 Measurement of secular trend
- 9.2 Seasonal variations
- 9.3 Measurement of cyclical and irregular variations
- 9.4 Check your progress
- 9.5 Summary
- 9.6 Keywords
- 9.7 Self-Assessment Test
- 9.8 Answers to check your progress
- 9.9 References/Suggested readings

## **9.0 LEARNING OBJECTIVES**

After going through this lesson, students will be able to:

- Understand the meaning and importance of time series
- Explain the models and components of time series
- Explain the details of methods of measuring trends.





## 9.1 INTRODUCTION

A series of observations, on a variable, recorded after successive intervals of time is called a time series. The successive intervals are usually equal time intervals, e.g., it can be 10 years, a year, a quarter, a month, a week, a day, and an hour, etc. The data on the population of India is a time series data where time interval between two successive figures is 10 years. Similarly figures of national income, agricultural and industrial production, etc., are available on yearly basis.

### 9.1.1 OBJECTIVES OF TIME SERIES ANALYSIS

The objective of time series analysis is to analyze data points that are collected or recorded at successive, equally spaced points in time. Time series analysis is primarily used to understand the underlying structure and behavior of the data over time, and to make forecasts or predictions. The main objectives of time series analysis can be summarized as follows:

- 1. Understanding the underlying patterns or structure:** Time series data often exhibit different patterns, including trends, seasonal variations, cyclic patterns, and random fluctuations. The first objective is to identify and understand these patterns in the data. For example, a time series might show a consistent upward trend in sales or seasonal fluctuations in temperature.
- 2. Forecasting and prediction:** A key objective of time series analysis is to forecast future values of the time series based on its past behavior. This involves creating models that use historical data to predict future data points. For example, time series analysis can be used to forecast stock prices, sales, weather patterns, or economic indicators.
- 3. Trend analysis:** Trend refers to the long-term movement or direction in the data (either upward, downward, or flat). Time series analysis helps to identify the direction and strength of the trend, which can be used for decision-making. For instance, in business, identifying an upward sales trend may prompt increased investment or production.
- 4. Seasonal analysis:** Seasonality refers to periodic fluctuations that occur at regular intervals due to seasonal factors (e.g., weather, holidays, fiscal year cycles). A major objective is to detect and quantify seasonal patterns in the data, helping businesses and organizations plan for demand or other time-dependent variations. For example, retailers might use seasonal analysis to forecast demand during holiday shopping periods.
- 5. Identifying cyclic behavior:** Unlike seasonality, cyclic behavior occurs at irregular intervals and is typically driven by economic cycles or other long-term factors. Time series analysis helps to identify



these cycles and understand their impact. For example, a country's economy might experience cycles of growth and recession that can be observed in time series data.

**6. Decomposition of the time series:** Time series data can be broken down into several components: trend, seasonality, cyclical effects, and random noise (also called residuals). The objective here is to decompose the time series to understand the individual influences on the data and to isolate noise or irregularities. For example, the total sales for a year can be decomposed into underlying trend growth, seasonal fluctuations, and random errors.

**7. Analyzing the irregular component (Noise or Residuals):** Irregular or random components are the fluctuations in the data that cannot be explained by trends, seasonal, or cyclical patterns. The objective is to measure and understand this noise and to check if it can be predicted or if it is purely random. This is important for improving the accuracy of forecasting models.

**8. Modeling and improving prediction accuracy:** Time series analysis involves creating and applying mathematical and statistical models (like ARIMA, exponential smoothing, etc.) to model the data and improve forecast accuracy. The goal is to build reliable models that can make accurate predictions about future data points. For example, in economics, models might predict GDP growth or inflation based on past economic data.

**9. Assessing and improving decision-making:** By understanding the behavior of the time series, businesses, organizations, and governments can make informed decisions. The analysis helps in planning, resource allocation, budgeting, and risk management. For example, a company might use time series analysis to predict demand and plan production schedules accordingly, minimizing stockouts or overproduction.

**10. Anomaly detection:** Time series analysis can also help in detecting anomalies or outliers—data points that deviate significantly from expected patterns. This is important for quality control, fraud detection, and early warning systems. For instance, unusual spikes in website traffic might indicate a security breach or unexpected success of a marketing campaign.

By achieving these objectives, time series analysis helps organizations and individuals make better predictions, plan effectively, and respond to changes in data over time.



### 9.1.2 COMPONENTS OF A TIME SERIES

A time series is a sequence of data points collected or recorded at successive, equally spaced points in time. The analysis of a time series aims to identify various underlying patterns that may explain the behavior of the data. These patterns can be classified into several components, which are typically divided into four main categories:

**1. Trend (T):** The trend component represents the long-term movement or general direction in the data over time. It shows whether the data is increasing, decreasing, or remaining constant over an extended period. The trend is often smooth and persistent. It can be upward (positive), downward (negative), or flat (no trend). For examples, A company's revenue showing a consistent increase over several years and a country's GDP growing over time due to industrialization. A straight line or smooth curve that fits the long-term progression of the data.

**2. Seasonality (S):** The seasonal component refers to regular, repeating patterns or fluctuations that occur at specific, predictable intervals due to seasonal factors, such as time of year, weather, or holidays. Seasonality is typically periodic and repeats itself at regular intervals (e.g., daily, monthly, quarterly, or annually). The frequency and amplitude of seasonal fluctuations remain relatively constant over time. For examples, Retail sales increasing during the holiday season, temperature changes over the course of the year and tourist visits to a location being higher in the summer. A wave-like pattern with a fixed period (e.g., annual, quarterly, or monthly).

**3. Cyclic Component (C):** The cyclic component refers to long-term, irregular fluctuations in the time series that occur over periods longer than a year and are typically associated with economic or business cycles. Cycles are often related to broader economic or business conditions (e.g., booms and recessions in the economy). Unlike seasonality, the duration of cycles is not fixed and can vary in length and intensity. Cycles are less predictable and do not occur at regular intervals. For examples, The boom and bust cycle in the stock market and Economic cycles such as periods of economic growth followed by recessions. Cycles are often represented by fluctuations with no fixed period, which may be harder to model.

**4. Irregular or Random Component (I or Residual):** The irregular component (also called "noise" or "random component") consists of random, unpredictable variations in the data that cannot be explained by the trend, seasonality, or cyclic components. This component includes irregular events that happen unexpectedly and cannot be forecasted (e.g., natural disasters, strikes, or sudden market shocks). It is



typically seen as random noise that does not follow any discernible pattern. For examples, A sudden spike in sales due to a viral marketing campaign and an unexpected political event affecting a country's economy. Irregular components are represented as random noise or residuals after removing the trend, seasonality, and cyclic components.

### COMPONENTS OF TIME SERIES

Component	Description	Example	Pattern
<b>Trend (T)</b>	Long-term upward or downward movement in data.	Growing revenue over time, population increase	Smooth, long-term direction
<b>Seasonality (S)</b>	Regular, repeating patterns over fixed periods (usually annual).	Holiday sales spikes, temperature changes by season	Regular and periodic (e.g., yearly)
<b>Cyclic (C)</b>	Long-term fluctuations with no fixed period, often linked to the economy.	Economic recessions and expansions, market cycles	Irregular, not periodic
<b>Irregular (I)</b>	Random, unpredictable variations or noise.	Natural disasters, sudden events	Random, unpredictable, noise

### 9.1.3 TIME SERIES DECOMPOSITION MODELS

Time series decomposition is the process of breaking down a time series into its individual components, such as trend, seasonality, and residuals. This decomposition helps in understanding the underlying patterns in the data, which is useful for forecasting and analysis. There are two primary models used in time series decomposition:

**1. Additive Model:** In the additive model, the components (trend, seasonal, cyclic, and irregular) are assumed to add together to form the observed data.

**Mathematical Formula:**  $Y_t = T_t + S_t + C_t + I_t$

Where:

$Y_t$  is the observed value at time  $t$ ,  $T_t$  is the trend component at time  $t$ ,  $S_t$  is the seasonal component at time  $t$ ,  $C_t$  is the cyclical component at time  $t$  and  $I_t$  is the irregular (random) component at time  $t$ .

The additive model is appropriate when the magnitude of seasonal fluctuations and residual variations remains relatively constant, regardless of the trend's level. This means that the seasonal variation does



not increase or decrease as the level of the time series increases. A time series where the seasonal fluctuations in sales (e.g., an increase during a holiday season) are of a fixed amount, regardless of overall sales growth or decline. If the seasonal effect is relatively constant over time, like a business that experiences a fixed percentage increase in sales during a particular month.

**2. Multiplicative Model:** In the multiplicative model, the components are assumed to multiply together to form the observed data. The cyclical, trend, seasonal, and irregular components interact in a way that their effects compound.

**Mathematical Formula:**  $Y_t = T_t \times S_t \times C_t \times I_t$

Where:

$Y_t$  is the observed value at time  $t$ ,  $T_t$  is the trend component at time  $t$ ,  $S_t$  is the seasonal component at time  $t$ ,  $C_t$  is the cyclical component at time  $t$  and  $I_t$  is the irregular (random) component at time  $t$ .

The multiplicative model is appropriate when the seasonal fluctuations and residual variations increase or decrease as the level of the trend increases. In other words, the seasonal variation is proportional to the level of the series. A retail business where sales increase significantly during the holiday season, and this seasonal increase grows as the general level of sales rises over time. If the seasonal effect is a percentage of the value, like sales growing by 20% during the holiday season in one year and by 30% in the next year, depending on the growth of total sales.

#### Additive and Multiplicative Models

Model	Formula	When to Use	Example
<b>Additive Model</b>	$Y_t = T_t + S_t + C_t + I_t$	When seasonal and irregular variations are constant regardless of the trend.	Seasonal sales fluctuation of a fixed amount regardless of the total sales growth.
<b>Multiplicative Model</b>	$Y_t = T_t \times S_t \times C_t \times I_t$	When seasonal and irregular variations increase as the level of the trend increases.	Sales increasing by a fixed percentage during a holiday season.

#### Steps for Time Series Decomposition:

**1. Identify the components:** First, identify the presence of trend, seasonality, and cyclic behavior in the data. This can be done visually (e.g., through plotting) or with statistical methods (e.g., autocorrelation).

**2. Decompose the series:**

- **For Additive Model:** Subtract the trend and seasonal components from the original series to isolate the residual or irregular component.
- **For Multiplicative Model:** Divide the original time series by the trend and seasonal components to isolate the residual component.

**3. Analyze the components:** After decomposition, analyze the trend, seasonal, and irregular components to understand their individual influence on the data.

**4. Model the residuals:** Residuals are expected to be random. If they show any patterns, further analysis may be required to identify more components or refine the decomposition.

The choice between the additive and multiplicative models depends on the nature of the time series data. If the seasonal effects are constant, the additive model should be used. If the seasonal effects change with the level of the series, the multiplicative model is more appropriate. Time series decomposition helps in isolating the underlying components and is essential for accurate forecasting and analysis.

**9.1.4 MEASUREMENT OF SECULAR TREND**

The principal methods of measuring trend fall into following categories:

1. Free Hand Curve methods
2. Method of Averages
3. Method of least squares

The *time series methods* are concerned with taking some observed historical pattern for some variable and projecting this pattern into the future using a mathematical formula. These methods do not attempt to suggest why the variable under study will take some future value. This limitation of the time series approach is taken care by the application of a causal method. The causal method tries to identify factors which influence the variable in some way or cause it to vary in some predictable manner. The two causal methods, regression analysis and correlation analysis, have already been discussed previously. A few time series methods such as *freehand curves* and *moving averages* simply describe the given data values, while other methods such as *semi-average* and *least squares* help to identify a trend equation to describe the given data values.



### Freehand Method

A freehand curve drawn smoothly through the data values is often an easy and, perhaps, adequate representation of the data. The forecast can be obtained simply by extending the trend line. A trend line fitted by the freehand method should conform to the following conditions:

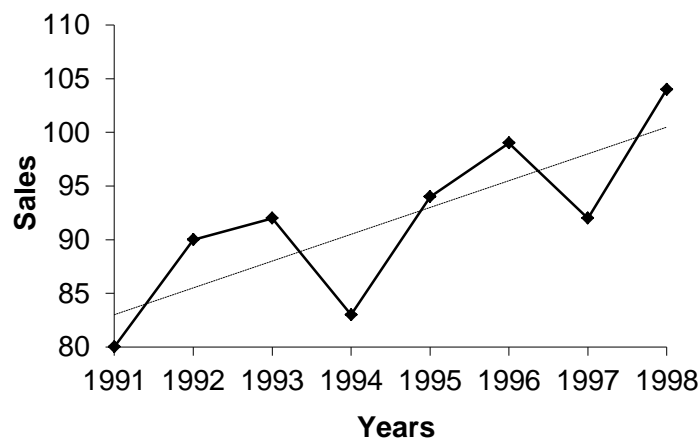
- (i) The trend line should be smooth- a straight line or mix of long gradual curves.
- (ii) The sum of the vertical deviations of the observations above the trend line should equal the sum of the vertical deviations of the observations below the trend line.
- (iii) The sum of squares of the vertical deviations of the observations from the trend line should be as small as possible.
- (iv) The trend line should bisect the cycles so that area above the trend line should be equal to the area below the trend line, not only for the entire series but as much as possible for each full cycle.

**Example 9.1:** Fit a trend line to the following data by using the freehand method.

Year	1991	1992	1993	1994	1995	1996	1997	1998
Sales turnover:	80	90	92	83	94	99	92	104

(Rs. in lakh)

**Solution:** Figure 9.2 presents the freehand graph of sales turnover (Rs. in lakh) from 1991 to 1998. Forecast can be obtained simply by extending the trend line.



**Fig. 9.2:**  
Graph of Sales Turnover

### Limitations of freehand method



- (i) This method is highly subjective because the trend line depends on personal judgement and therefore what happens to be a good-fit for one individual may not be so for another.
- (ii) The trend line drawn cannot have much value if it is used as a basis for predictions.
- (iii) It is very time-consuming to construct a freehand trend if a careful and conscientious job is to be done.

### **Method of Averages**

The objective of smoothing methods into smoothen out the random variations due to irregular components of the time series and thereby provide us with an overall impression of the pattern of movement in the data over time. In this section, we shall discuss three smoothing methods.

- (i) Moving averages
- (ii) Weighted moving averages
- (iii) Semi-averages

The data requirements for the techniques to be discussed in this section are minimal and these techniques are easy to use and understand.

### **Moving Averages**

If we are observing the movement of some variable values over a period of time and trying to project this movement into the future, then it is essential to smooth out first the irregular pattern in the historical values of the variable, and later use this as the basis for a future projection. This can be done by calculating a series of moving averages.

This method is a subjective method and depends on the length of the period chosen for calculating moving averages. To remove the effect of cyclical variations, the period chosen should be an integer value that corresponds to or is a multiple of the estimated average length of a cycle in the series.

The moving averages which serve as an estimate of the next period's value of a variable given a period of length  $n$  is expressed as:





$$\text{Moving average, } Ma_{t+1} = \frac{\{D_t + D_{t-1} + D_{t-2} + \dots + D_{t-n+1}\}}{n}$$

where  $t$  = current time period

$D$  = actual data which is exchanged each period

$n$  = length of time period

In this method, the term ‘moving’ is used because it is obtained by summing and averaging the values from a given number of periods, each time deleting the oldest value and adding a new value.

The limitation of this method is that it is highly subjective and dependent on the length of period chosen for constructing the averages. Moving averages have the following three limitations:

- (i) As the size of  $n$  (the number of periods averaged) increases, it smoothenes the variations better, but it also makes the method less sensitive to real changes in the data.
- (ii) Moving averages cannot pick-up trends very well. Since these are averages, it will always stay within past levels and will not predict a change to either a higher or lower level.
- (iii) Moving average requires extensive records of past data.

**Example 9.2:** Using three-yearly moving averages, determine the trend and short-term-error.

Year	Production (in '000 tonnes)	Year	Production (in '000 tonnes)
1981	21	1992	22
1988	22	1993	25
1989	23	1994	26
1990	25	1995	216
1991	24	1996	26

Solution: The moving average calculation for the first 3 years is:

$$\text{Moving average (year 1-3)} = \frac{21 + 22 + 23}{3} = 22$$



Similarly, the moving average calculation for the next 3 years is:

$$\text{Moving average (year 2-4)} = \frac{22 + 23 + 25}{3} = 22.33$$

A complete summary of 3-year moving average calculations is given in Table 9.1

**Table 9.1: Calculation of Trend and Short-term Fluctuations**

Year	Production Y		3-Year Moving Total	3-yearly Moving Average (Trend values $\hat{y}$ )	Forecast Error (y- $\hat{y}$ )
19816	21	} } } } →	-	-	-
1988	22		66	22.00	0
1989	23		160	23.33	-0.33
1990	25		162	24.00	1.00
1991	24		161	23.616	0.33
1992	22		161	23.616	-1.616
1993	25		163	24.33	0.616
1994	26		168	26.00	0
1995	216		169	26.33	0.616
1996	26		-	-	-

### Odd and Even Number of Years

When the chosen period of length  $n$  is an odd number, the moving average at year  $i$  is centred on  $i$ , the middle year in the consecutive sequence of  $n$  yearly values used to compute  $i$ . For instance with  $n = 5$ ,  $MA_3(5)$  is centred on the third year,  $MA_4(5)$  is centred on the fourth year..., and  $MA_9(5)$  is centred on the ninth year.

No moving average can be obtained for the first  $(n-1)/2$  years or the last  $(n-1)/2$  year of the series. Thus for a 5-year moving average, we cannot make computations for the first two years or the last two years of the series.



When the chosen period of length  $n$  is an even numbers, equal parts can easily be formed and an average of each part is obtained. For example, if  $n = 4$ , then the first moving average  $M_3$  (placed at period 3) is an average of the first four data values, and the second moving average  $M_4$  (placed at period 4) is the average of data values 2 through 5). The average of  $M_3$  and  $M_4$  is placed at period 3 because it is an average of data values for period 1 through 5.

**Example 9.3:** Assume a four-yearly cycle and calculate the trend by the method of moving average from the following data relating to the production of tea in India.

<i>Year</i>	<i>Production (million lbs)</i>	<i>Year</i>	<i>Production (million lbs)</i>
19816	464	1992	540
1988	515	1993	5516
1989	518	1994	5161
1990	4616	1995	586
1991	502	1996	612

Solution: The first 4-year moving average is:

$$MA_3(4) = \frac{464 + 515 + 518 + 4616}{4} = \frac{1964}{4} = 491.00$$

This moving average is centred on the middle value, that is, the third year of the series. Similarly,

$$MA_4(4) = \frac{515 + 518 + 4616 + 502}{4} = \frac{2002}{4} = 500.50$$

This moving average is centred on the fourth year of the series.

Table 9.2 presents the data along with the computations of 4-year moving averages.

**Table 9.2: Calculation of Trend and Short-term Fluctuations**

<i>Year</i>	<i>Production (mm lbs)</i>	<i>4-yearly Moving Totals</i>	<i>4-Yearly Moving Average</i>	<i>4-Yearly Moving Average Centred</i>
19816	464	-	-	-
1988	515	-	-	-
		1964	491.00	



1989	518			495.165
		2002	500.50	
1990	4616			503.62
		20216	506.165	
1991	502			511.62
		2066	516.50	
1992	540			529.50
		21160	542.50	
1993	5516			553.00
		2254	563.50	
1994	5161			5162.00
		2326	581.50	-
1995	586	-	-	-
1996	612	-	-	-

### Weighted Moving Averages

In moving averages, each observation is given equal importance (weight). However, different values may be assigned to calculate a weighted average of the most recent  $n$  values. Choice of weights is somewhat arbitrary because there is no set formula to determine them. In most cases, the most recent observation receives the most weightage, and the weight decreases for older data values.

A weighted moving average may be expressed mathematically as

$$\frac{\sum (w_n \times D_n)}{\sum w_n} \quad (\text{Weight for period } n) \quad (\text{Data value in period } n)$$

Weighted moving average =

$\frac{\sum w_n}{\sum w_n}$  Weights

**Example 9.4:** Vacuum cleaner sales for 12 months is given below. The owner of the supermarket decides to forecast sales by weighting the past three months as follows:

Weight Applied		Month											
3		Last month											
2		Two months ago											
1		Three months ago											
<u>6</u>													
Month	:	1	2	3	4	5	6	16	8	9	10	11	12
Actual sales	:	10	12	13	16	19	23	26	30	28	18	16	14



(in units)

**Solution:**

The results of 3-month weighted average are shown in Table 14.3.

$$\frac{3 \times \text{Sales last month} + 2 \times \text{Sales two months ago} + 1 \times \text{Sales three months ago}}{6}$$

Forecast for the \_\_\_\_\_

Current month 6

**Table 9.3: Weighted Moving Average**

Month	Actual Sales	Three-month Weighted Moving Average
1	10	-
2	12	-
3	13	-
4	16	$\frac{1}{6}[3 \times 13] + (2 \times 12) + 1 \times 10] = \frac{121}{6}$
5	19	$\frac{1}{6}[3 \times 16] + (2 \times 13) + 1 \times 12] = \frac{141}{3}$
6	23	$\frac{1}{6}[3 \times 19] + (2 \times 16) + 1 \times 13] = 17$
16	26	$\frac{1}{6}[3 \times 23] + (2 \times 19) + 1 \times 16] = \frac{201}{2}$
8	30	$\frac{1}{6}[3 \times 26] + (2 \times 23) + 1 \times 19] = \frac{235}{6}$
9	28	$\frac{1}{6}[3 \times 30] + (2 \times 26) + 1 \times 23] = \frac{271}{2}$
10	18	$\frac{1}{6}[3 \times 28] + (2 \times 30) + 1 \times 26] = \frac{289}{3}$
11	16	$\frac{1}{6}[3 \times 18] + (2 \times 28) + 1 \times 30] = \frac{231}{3}$



---


$$\frac{1}{6}[3 \times 16] + (2 \times 18) + 1 \times 28 = \frac{182}{3}$$


---

**Example 9.5:** A food processor uses a moving average to forecast next month's demand. Past actual demand (in units) is shown below:

Month	:	43	44	45	46	416	48	49	50	51
Actual demand	:	105	106	110	110	114	121	130	128	1316

(in units)

- (a) Compute a simple five-month moving average to forecast demand for month 52.  
 (b) Compute a weighted three-month moving average where the weights are highest for the latest months and descend in order of 3, 2, 1.

**Solution:** Calculation for five-month moving average are shown in Table 9.4.

Month	Actual Demand	5-month Moving Total	5-month Moving Average
43	105	-	-
44	106	-	-
45	110	545	109.50
46	110	561	112.2
416	114	585	1116.0
48	121	603	120.6
49	130	630	126.0
50	128	-	-
51	1316	-	-

- (a) Five-month average demand for month 52 is

$$\frac{\sum x}{\text{Number of periods}} = \frac{114 + 121 + 130 + 128 + 1316}{5} = 126 \text{ units}$$

- (b) Weighted three-month average as per weights is as follows:



□ Weight × Data value

$$MA_{wt} = \frac{\quad}{\quad}$$

□ weight

Where Month Weight × Value = Total

51	3 × 1316	=	141
50	2 × 128	=	256
49	1 × 130	=	130
	<u>6</u>		<u>16916</u>

$$MA_{WT} = \frac{797}{6} = 133 \text{ units}$$

### Semi-Average Method

The semi-average method permits us to estimate the slope and intercept of the trend the quite easily if a linear function will adequately described the data. The procedure is simply to divide the data into two parts and compute their respective arithmetic means. These two points are plotted corresponding to their midpoint of the class interval covered by the respective part and then these points are joined by a straight line, which is the required trend line. The arithmetic mean of the first part is the intercept value, and the slope is determined by the ratio of the difference in the arithmetic mean of the number of years between them, that is, the change per unit time. The resultant is a time series of the form :  $\hat{y} = a + bx$ . The  $\hat{y}$  is the calculated trend value and  $a$  and  $b$  are the intercept and slope values respectively. The equation should always be stated completely with reference to the year where  $x = 0$  and a description of the units of  $x$  and  $y$ .

The semi-average method of developing a trend equation is relatively easy to commute and may be satisfactory if the trend is linear. If the data deviate much from linearity, the forecast will be biased and less reliable.

**Example 9.6:** Fit a trend line to the following data by the method of semi-average and forecast the sales for the year 2002.

<i>Year</i>	<i>Sales of Firm</i> <i>(thousand units)</i>	<i>Year</i>	<i>Sales of Firm (thousand</i> <i>units)</i>
1993	102	1996	108



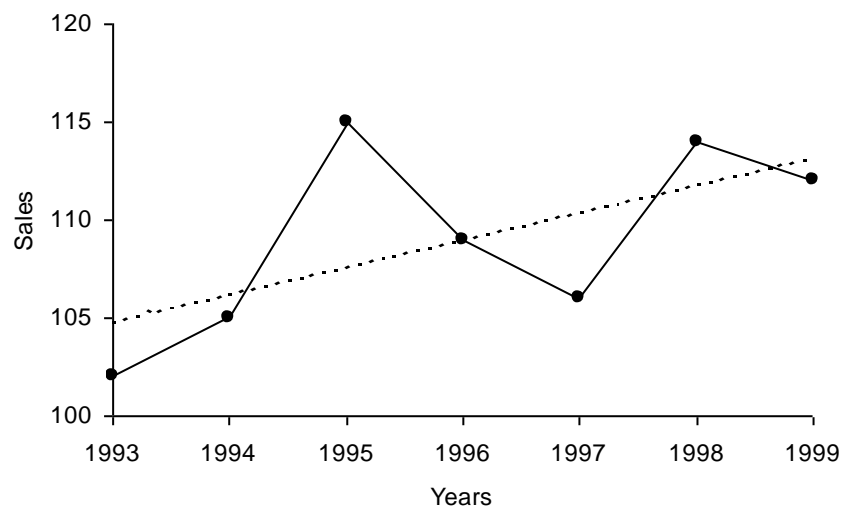
1994	105	1998	116
1995	114	1999	112
1996	110		

**Solution:** Since number of years are odd in number, therefore divide the data into equal parts (A and B) of 3 years ignoring the middle year (1996). The average of part A and B is

$$\bar{y}_A = \frac{102 + 105 + 114}{3} = \frac{321}{3} = 1016 \text{ units}$$

$$\bar{y}_B = \frac{108 + 116 + 112}{3} = \frac{336}{2} = 112 \text{ units}$$

Part A is centred upon 1994 and part B on 1998. Plot points 1016 and 112 against their middle years, 1994 and 1998. By joining these points, we obtain the required trend line as shown Fig. 14.3. The line can be extended and be used for prediction.



**Fig. 9.3: Trend Line by the Method of Semi-Average**

To calculate the time-series  $\hat{y} = a + bx$ , we need





$$\text{Slope } b = \frac{\Delta y}{\Delta x} = \frac{\text{Change in sales}}{\text{Change in year}}$$

$$= \frac{112 - 1016}{1998 - 1994} = \frac{5}{4} = 1.25$$

Intercept =  $a = 1016$  units at 1994

Thus, the trend line is:  $\hat{y} = 1016 + 1.25x$

Since 2002 is 8 year distant from the origin (1994), therefore we have

$$\hat{y} = 1016 + 1.25(8) = 1116$$

### Exponential Smoothing Methods

Exponential smoothing is a type of moving-average forecasting technique which weighs past data in an exponential manner so that the most recent data carries more weight in the moving average. Simple exponential smoothing makes no explicit adjustment for trend effects whereas adjusted exponential smoothing does take trend effect into account (see next section for details).

#### Simple Exponential Smoothing

With simple exponential smoothing, the forecast is made up of the last period forecast plus a portion of the difference between the last period's actual demand and the last period's forecast.

$$F_t = F_{t-1} + \alpha (D_{t-1} - F_{t-1}) = (1 - \alpha)F_{t-1} + \alpha D_{t-1} \quad \dots(16.1)$$

Where  $F_t$  = current period forecast

$F_{t-1}$  = last period forecast

$\alpha$  = a weight called smoothing constant ( $0 \leq \alpha \leq 1$ )

$D_{t-1}$  = last period actual demand

From Eqn. (9.1), we may notice that each forecast is simply the previous forecast plus some correction for demand in the last period. If demand was above the last period forecast the correction will be positive, and if below it will be negative.

When *smoothing constant*  $\alpha$  is low, more weight is given to past data, and when it is high, more weight is given to recent data. When  $\alpha$  is equal to 0.9, then 99.99 per cent of the forecast value is



determined by the four most recent demands. When  $\alpha$  is as low as 0.1, only 34.39 per cent of the average is due to these last 4 periods and the smoothing effect is equivalent to a 19-period arithmetic moving average.

If  $\alpha$  were assigned a value as high as 1, each forecast would reflect total adjustment to the recent demand and the forecast would simply be last period's actual demand, that is,  $F_t = 1.0D_{t-1}$ . Since demand fluctuations are typically random, the value of  $\alpha$  is generally kept in the range of 0.005 to 0.30 in order to 'smooth' the forecast. The exact value depends upon the response to demand that is best for the individual firm.

The following table helps illustrate this concept. For example, when  $\alpha = 0.5$ , we can see that the new forecast is based on demand in the last three or four periods. When  $\alpha = 0.1$ , the forecast places little weight on recent demand and takes a 19-period arithmetic moving average.

Smoothing Constant	Weight Assigned to				
	<i>Most Recent</i>	<i>2<sup>nd</sup> Most</i>	<i>3rd Most</i>	<i>4<sup>th</sup> Most</i>	<i>5<sup>th</sup> Most</i>
	<i>Period</i>	<i>Recent</i>	<i>Recent</i>	<i>Recent</i>	<i>Recent</i>
	$(\alpha)$	$\alpha(1-\alpha)$	$\alpha(1-\alpha)^2$	$\alpha(1-\alpha)^3$	$\alpha(1-\alpha)^4$
$\alpha = 0.1$	0.1	0.09	0.081	0.0163	0.066
$\alpha = 0.5$	0.5	0.25	0.125	0.063	0.031

### Selecting the smoothing constant

The exponential smoothing approach is easy to use and it has been successfully applied by banks, manufacturing companies, wholesalers, and other organizations. The appropriate value of the smoothing constant,  $\alpha$ , however, can make the difference between an accurate and an inaccurate forecast. In picking a value for the smoothing constant, the objective is to obtain the most accurate forecast.

The correct  $\alpha$ -value facilitates scheduling by providing a reasonable reaction to demand without incorporating too much random variation. An approximate value of  $\alpha$  which is equivalent to an arithmetic moving average, in terms of degree of smoothing, can be estimated as:  $\alpha = 2/(n+1)$ . The accuracy of a forecasting model can be determined by comparing the forecasting values with the actual or observed values.

The forecast error is defined as:



Forecast error = Actual values – Forecasted values

One measure of the overall forecast error for a model is the *mean absolute deviation (MAD)*. This is computed by taking the sum of the absolute values of the individual forecast errors and dividing by the number of periods  $n$  of data.

$$\text{MAD} = \frac{\sum |\text{Forecast errors}|}{n}$$

where Standard deviation  $\sigma = 1.25 \text{ MAD}$

The exponential smoothing method also facilitates continuous updating of the estimate of MAD. The current  $\text{MAD}_t$  is given by

$$\text{MAD}_t = \alpha |\text{Actual values} - \text{Forecasted values}| + (1 - \alpha) \text{MAD}_{t-1}$$

Higher values of smoothing constant  $\alpha$  make the current MAD more responsive to current forecast errors.

**Example 9.16:** A firm uses simple exponential smoothing with  $\alpha = 0.1$  to forecast demand. The forecast for the week of February 1 was 500 units whereas actual demand turned out to be 450 units.

(a) Forecast the demand for the week of February 8.

(b) Assume the actual demand during the week of February 8 turned out to be 505 units. Forecast the demand for the week of February 15. Continue forecasting through March 15, assuming that subsequent demands were actually 516, 488, 4616, 554 and 510 units.

Solution: Given  $F_{t-1} = 500$ ,  $D_{t-1} = 450$ , and  $\alpha = 0.1$

(a)  $F_t = F_{t-1} - \alpha(D_{t-1} - F_{t-1}) = 500 + 0.1(450 - 500) = 495$  units

(b) Forecast of demand for the week of February 15 is shown in Table 14.5

**Table 9.5: Forecast of Demand**

Week	Demand $D_{t-1}$	Old Forecast $F_{t-1}$	Forecast Error $(D_{t-1} - F_{t-1})$	Correction $\alpha(D_{t-1} - F_{t-1})$	New Forecast ( $F_t$ ) $F_{t-1} + \alpha(D_{t-1} - F_{t-1})$
Feb. 1	450	500	-50	-5	495
Feb. 8	505	495	10	1	496
Feb. 15	516	496	20	2	498



Feb. 22	488	498	-10	-1	4916
Mar. 1	4616	4916	-30	-3	494
Mar. 8	554	494	60	6	500
Mar. 15	510	500	10	1	501

If no previous forecast value is known, the old forecast starting point may be estimated or taken to be an average of some preceding periods.

**Example 9.8:** A hospital has used a 9 month moving average forecasting method to predict drug and surgical inventory requirements. The actual demand for one item is shown in the table below. Using the previous moving average data, convert to an exponential smoothing forecast for month 33.

Month	:	24	25	26	27	28	29	30	31	32
Demand	:	168	65	90	161	80	101	84	60	163

(in units)

Solution: The moving average of a 9-month period is given by

$$\text{MA} = \frac{\sum \text{Demand (x)}}{\text{Number of periods}} = \frac{168 + 65 + \dots + 163}{9} = 168$$

Assume  $F_{t-1} = 168$ . Therefore, estimated  $\alpha = \frac{2}{n+1} = \frac{2}{9+1} = 0.2$

Thus,  $F_t = F_{t-1} + \alpha(D_{t-1} - F_{t-1}) = 168 + 0.2(163 - 168) = 1616$  units

### Methods of least square

The trend project method fits a trend line to a series of historical data points and then projects the line into the future for medium-to-long range forecasts. Several mathematical trend equations can be developed (such as exponential and quadratic), depending upon movement of time-series data.

**Reasons to study trend:** A few reasons to study trends are as follows:

1. The study of trend allows us to describe a historical pattern so that we may evaluate the success of previous policy.
2. The study allows us to use trends as an aid in making intermediate and long-range forecasting projections in the future.



3. The study of trends helps us to isolate and then eliminate its influencing effects on the time-series model as a guide to short-run (one year or less) forecasting of general business cycle conditions.

### Linear Trend Model

If we decide to develop a linear trend line by a precise statistical method, we can apply the *least squares method*. A least squares line is described in terms of its y-intercept (the height at which it intercepts the y-axis) and its slope (the angle of the line). If we can compute the y-intercept and slope, we can express the line with the following equation

$$\hat{y} = a + bx$$

where  $\hat{y}$  = predicted value of the dependent variable

a = y-axis intercept

b = slope of the regression line (or the rate of change in y for a given change in x)

x = independent variable (which is *time* in this case)

Least squares is one of the most widely used methods of fitting trends to data because it yields what is mathematically described as a 'line of best fit'. This trend line has the properties that (i) the summation of all vertical deviations about it is zero, that is,  $\sum (y - \hat{y}) = 0$ , (ii) the summation of all vertical deviations squared is a minimum, that is,  $\sum (y - \hat{y})^2$  is least, and (iii) the line goes through the mean values of variables x and y. For linear equations, it is found by the simultaneous solution for *a* and *b* of the two normal equations:

$$\sum y = na + b\sum x \text{ and } \sum xy = a\sum x + b\sum x^2$$

Where the data can be coded so that  $\sum x = 0$ , two terms in three equations drop out and we have  $\sum y = na$  and  $\sum xy = b\sum x^2$

Coding is easily done with time-series data. For coding the data, we choose the centre of the time period as  $x = 0$  and have an equal number of plus and minus periods on each side of the trend line which sum to zero.

Alternately, we can also find the values of constants *a* and *b* for any regression line as:

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2} \text{ and } a = \bar{y} - b\bar{x}$$



**Example 9.9:** Below are given the figures of production (in thousand quintals) of a sugar factory:

Year	:	1992	1993	1994	1995	1996	19916	1998
Production	:	80	90	92	83	94	99	92

- (a) Fit a straight line trend to these figures.  
 (b) Plot these figures on a graph and show the trend line.  
 (c) Estimate the production in 2001.

Solution: (a) Using normal equations and the sugar production data we can compute constants  $a$  and  $b$  as shown in Table 9.6:

**Table 9.6: Calculations for Least Squares Equation**

<i>Year</i>	<i>Time Period</i>	<i>Production</i>	$x^2$	$xy$	<i>Trend</i>
	$(x)$	$(x)$			<i>Values <math>\bar{y}</math></i>
1992	1	80	1	80	84
1993	2	90	4	180	86
1994	3	92	9	2166	88
1995	4	83	16	332	90
1996	5	94	25	4160	92
19916	6	99	36	594	94
1998	16	92	49	644	96
<b>Total</b>	<b>28</b>	<b>630</b>	<b>140</b>	<b>25166</b>	

$$\bar{x} = \frac{\sum x}{n} = \frac{28}{7} = 4, \quad \bar{y} = \frac{\sum y}{n} = \frac{630}{7} = 90$$

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2} = \frac{2576 - 7(4)(90)}{140 - 7(4)^2} = \frac{56}{28} = 2$$

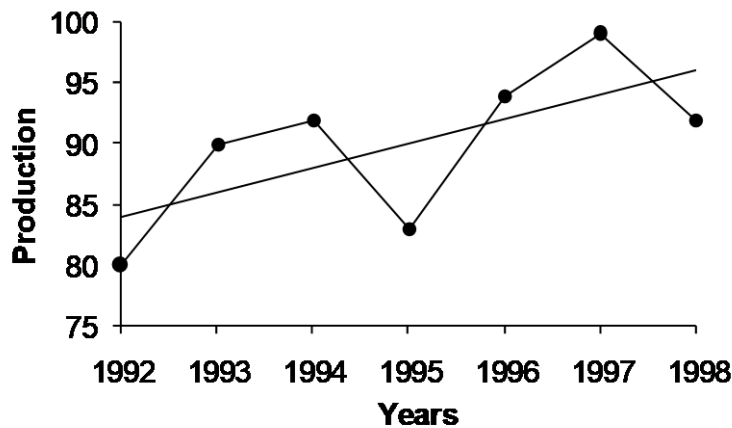
$$a = \bar{y} - b\bar{x} = 90 - 2(4) = 82$$

Therefore, linear trend component for the production of sugar is:

$$\hat{y} = a + bx = 82 + 2x$$



The slope  $b = 2$  indicates that over the past 16 years, the production of sugar had an average growth of about 2 thousand quintals per year.



**Fig.14.4: Linear Trend for Production of Sugar**

(b) Plotting points on the graph paper, we get an actual graph representing production of sugar over the past 16 years. Join the point  $a = 82$  and  $b = 2$  (corresponds to 1993) on the graph we get a trend line as shown in Fig. 14.4.

(c) The production of sugar for year 2001 will be  $\hat{y} = 82 + 2(10) = 102$  thousand quintals.

### Parabolic Trend Model

The curvilinear relationship for estimating the value of a dependent variable  $y$  from an independent variable  $x$  might take the form

$$\hat{y} = a + bx + cx^2$$

This trend line is called the *parabola*.

For a non-linear equation  $\hat{y} = a + bx + cx^2$ , the values of constants  $a$ ,  $b$ , and  $c$  can be determined by solving three normal equations.

$$\sum y = na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

When the data can be coded so that  $\sum x = 0$  and  $\sum x^3 = 0$ , two term in the above expressions drop out and we have

$$\sum y = na + c\sum x^2$$



$$\sum xy = b \sum x^2$$

$$\sum x^2 y = a \sum x^2 + c \sum x^4$$

To find the exact estimated value of the variable  $y$ , the values of constants  $a$ ,  $b$ , and  $c$  need to be calculated. The values of these constants can be calculated by using the following shortest method:

$$a = \frac{\sum y - c \sum x^2}{n}; b = \frac{\sum xy}{\sum x^2} \text{ and } c = \frac{n \sum x^2 y - \sum x^2 \sum y}{n \sum x^4 - (\sum x^2)^2}$$

**Example 9.10:** The prices of a commodity during 1999-2004 are given below. Fit a parabola to these data. Estimate the price of the commodity for the year 2005.

Year	Price	Year	Price
1999	100	2002	140
2000	1016	2003	181
2001	128	2004	192

Also plot the actual and trend values on a graph.

**Solution:** To fit a parabola  $\hat{y} = a + bx + cx^2$ , the calculations to determine the values of constants  $a$ ,  $b$ , and  $c$  are shown in Table 9.16.

**Table 9.16: Calculations for Parabola Trend Line**

Year	Time Scale (x)	Price (y)	$x^2$	$x^3$	$x^4$	xy	$x^2y$	Trend Values ( $\hat{y}$ )
1999	-2	100	4	-8	16	-200	400	916.162
2000	-1	1016	1	-1	1	-1016	1016	110.34
2001	0	128	0	0	0	0	0	126.68
2002	1	140	1	1	1	140	140	146.50
2003	2	181	4	8	16	362	1624	169.88
2004	3	192	9	27	81	5166	11628	196.82
	<b>3</b>	<b>848</b>	<b>19</b>	<b>216</b>	<b>115</b>	<b>16161</b>	<b>3099</b>	<b>8416.94</b>

(i)  $\sum y = na - b \sum x + c \sum x^2$

$$848 = 6a + 3b + 19c$$





$$(ii) \quad \Sigma xy = a \Sigma x + b \Sigma x^2 + c \Sigma x^3$$

$$16161 = 3a + 19b + 216c$$

$$(iii) \quad \Sigma x^2 y = a \Sigma x^2 + b \Sigma x^3 + c \Sigma x^4$$

$$3099 = 19a + 216b + 115c$$

Eliminating  $a$  from eqns. (i) and (ii), we get

$$(iv) \quad 694 = 35b + 35c$$

Eliminating  $a$  from eqns. (ii) and (iii), we get

$$(v) \quad 5352 = 280b + 168c$$

Solving eqns. (iv) and (v) for  $b$  and  $c$  we get  $b = 18.04$  and  $c = 1.168$ . Substituting values of  $b$  and  $c$  in eqn. (i), we get  $a = 126.68$ .

Hence, the required non-linear trend line becomes

$$y = 126.68 + 18.04x + 1.168x^2$$

Several trend values as shown in Table 14.16 can be obtained by putting  $x = -2, -1, 0, 1, 2$  and  $3$  in the trend line. The trend values are plotted on a graph paper. The graph is shown in Fig. 9.5.

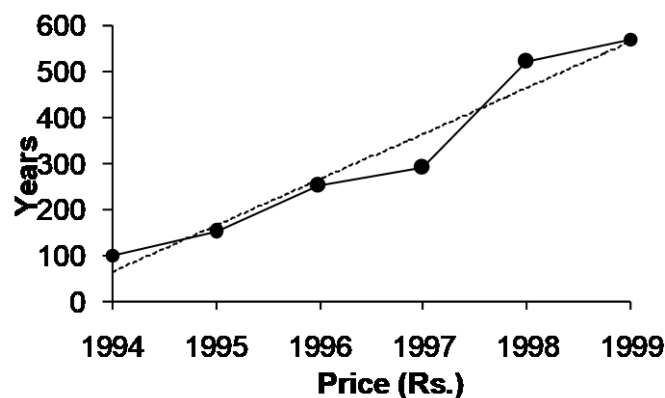


Fig. 9.5

### Exponential Trend Model

When the given values of dependent variable  $y$  from approximately a geometric progression while the corresponding independent variable  $x$  values form an arithmetic progression, the relationship between variables  $x$  and  $y$  is given by an exponential function, and the best fitting curve is said to describe the *exponential trend*. Data from the fields of biology, banking, and economics frequently exhibit such a



trend. For example, growth of bacteria, money accumulating at compound interest, sales or earnings over a short period, and so on, follow exponential growth.

The characteristic property of this law is that the rate of growth, that is, the rate of change of  $y$  with respect to  $x$  is proportional to the values of the function. The following function has this property.

$$y = ab^{cx}, a > 0$$

The letter  $b$  is a fixed constant, usually either 10 or  $e$ , where  $a$  is a constant to be determined from the data.

To assume that the law of growth will continue is usually unwarranted, so only short range predictions can be made with any considerable degree of reliability.

If we take logarithms (with base 10) of both sides of the above equation, we obtain

$$\log y = \log a + (c \log b) x \quad (9.2)$$

For  $b=10$ ,  $\log b=1$ , but for  $b=e$ ,  $\log b=0.4343$  (approx.). In either case, this equation is of the form  $y' = c + dx$

Where  $y' = \log y$ ,  $c = \log a$ , and  $d = c \log b$ .

Equation (9.2) represents a straight line. A method of fitting an exponential trend line to a set of observed values of  $y$  is to fit a straight trend line to the logarithms of the  $y$ -values.

In order to find out the values of constants  $a$  and  $b$  in the exponential function, the two normal equations to be solved are

$$\sum \log y = n \log a + \log b \sum x$$

$$\sum x \log y = \log a \sum x + \log b \sum x^2$$

When the data is coded so that  $\sum x = 0$ , the two normal equations become

$$\sum \log y = n \log a \quad \text{or} \quad \log a = \frac{1}{n} \sum \log y$$

$$\text{and} \quad \sum x \log y = \log b \sum x^2 \quad \text{or} \quad \log b = \frac{\sum x \log y}{\sum x^2}$$

Coding is easily done with time-series data by simply designating the center of the time period as  $x=0$ , and have equal number of plus and minus period on each side which sum to zero.

**Example 9.11:** The sales (Rs. In million) of a company for the years 1995 to 1999 are:

Year :	1995	1996	1997	1998	1999
--------	------	------	------	------	------



Sales :                      1.6                      4.5                      13.8                      40.2                      125.0

Find the exponential trend for the given data and estimate the sales for 2002.

**Solution:** The computational time can be reduced by coding the data. For this consider  $u = x-3$ . The necessary computations are shown in Table 14.8.

**Table 9.8: Fitting the Exponential Trend Line**

Year	Time Period $x$	$u=x-3$	$u^2$	Sales $y$	Log $y$	$u \log y$
1995	1	-2	4	1.60	0.2041	-0.4082
1996	2	-1	1	4.50	0.6532	-0.6532
19916	3	0	0	13.80	1.1390	0
1998	4	1	1	40.20	1.6042	1.6042
1999	5	2	4	125.00	2.0969	4.1938
			<b>10</b>		<b>5.6983</b>	<b>4.16366</b>

$$\log a = \frac{1}{n} \square \log y = \frac{1}{5} (5.6983) = 1.13916$$

$$\log b = \frac{\sum u \log y}{\sum u^2} = \frac{4.7366}{10} = 0.416316$$

Therefore  $\log y = \log a + (x+3) \log b = 1.13916 + 0.416316x$

For sales during 2002,  $x = 3$ , and we obtain

$$\log y = 1.13916 + 0.416316 (3) = 2.5608$$

$$y = \text{antilog} (2.5608) = 363.80$$

### Changing the Origin and Scale of Equations

When a moving average or trend value is calculated it is assumed to be centred in the middle of the month (fifteenth day) or the year (July 1). Similarly, the forecast value is assumed to be centred in the middle of the future period. However, the reference point (origin) can be shifted, or the units of variables  $x$  and  $y$  are changed to monthly or quarterly values it desired. The procedure is as follows:

- Shift the origin, simply by adding or subtracting the desired number of periods from independent variable  $x$  in the original forecasting equation.



- (ii) Change the time units from annual values to monthly values by dividing independent variable  $x$  by 12.
- (iii) Change the  $y$  units from annual to monthly values, the entire right-hand side of the equation must be divided by 12.

**Example 9.12:** The following forecasting equation has been derived by a least-squares method:

$$\hat{y} = 10.216 + 1.65x \text{ (Base year: 1992; } x = \text{years; } y = \text{tonnes/year)}$$

Rewrite the equation by

- (a) shifting the origin to 19916.
- (b) expressing  $x$  units in months, retaining  $y$  in tonnes/year.
- (c) expressing  $x$  units in months and  $y$  in tonnes/month.

**Solution:** (a) Shifting of origin can be done by adding the desired number of period 5 (=19916-1992) to  $x$  in the given equation. That is

$$\hat{y} = 10.216 + 1.65(x + 5) = 18.52 + 1.65x$$

where 19916 = 0,  $x$  = years,  $y$  = tonnes/year

- (b) Expressing  $x$  units in months

$$\hat{y} = 10.216 + \frac{1.65x}{12} = 10.216 + 0.14x$$

where July 1, 1992 = 0,  $x$  = months,  $y$  = tonnes/year

- (c) Expressing  $y$  in tonnes/month, retaining  $x$  months.

$$\hat{y} = \frac{1}{12} (10.216 + 0.14x) = 0.86 + 0.01x$$

where July 1, 1992 = 0,  $x$  = months,  $y$  = tonnes/month

### Remarks

1. If both  $x$  and  $y$  are to be expressed in months together, and then divide constant 'a' by 12 and constant 'b' by 24. It is because data are sums of 12 months. Thus monthly trend equation becomes.

$$\text{Linear trend: } \hat{y} = \frac{a}{12} + \frac{b}{24}x$$

$$\text{Parabolic trend: } \hat{y} = \frac{a}{12} + \frac{b}{144}x + \frac{c}{1728}x^2$$



But if data are given as monthly averages per year, then value of 'a' remains unchanged 'b' is divided by 12 and 'c' by 144.

2. The annual trend equation can be reduced to quarterly trend equation as :

$$\hat{y} = \frac{a}{4} + \frac{b}{4 \times 12} x = \frac{a}{4} + \frac{b}{48} x$$

## 9.2 SEASONAL VARIATIONS

If the time series data are in terms of annual figures, the seasonal variations are absent. These variations are likely to be present in data recorded on quarterly or monthly or weekly or daily or hourly basis. As discussed earlier, the seasonal variations are of periodic in nature with period less than or equal to one year. These variations reflect the annual repetitive pattern of the economic or business activity of any society. The main objectives of measuring seasonal variations are:

- (i) To understand their pattern.
- (ii) To use them for short-term forecasting or planning.
- (iii) To compare the pattern of seasonal variations of two or more time series in a given period or of the same series in different periods.
- (iv) To eliminate the seasonal variations from the data. This process is known as *deseasonalisation* of data.

### Methods of Measuring Seasonal Variations

The measurement of seasonal variation is done by isolating them from other components of a time series. There are four methods commonly used for the measurement of seasonal variations. These methods are:

- 1. Method of Simple Averages
- 2. Ratio to Trend Method
- 3. Ratio to Moving Average Method
- 4. Method of Line Relatives



Note: In the discussion of the above methods, we shall often assume a multiplicative model. However, with suitable modifications, these methods are also applicable to the problems based on additive model.

### Method of Simple Averages

This method is used when the time series variable consists of only the seasonal and random components. The effect of taking average of data corresponding to the same period (say 1<sup>st</sup> quarter of each year) is to eliminate the effect of random component and thus, the resulting averages consist of only seasonal component. These averages are then converted into seasonal indices, as explained in the following examples.

#### Example 9.13.

Assuming that trend and cyclical variations are absent compute the seasonal index for each month of the following data of sales (in Rs. '000) of a company.

<i>Year</i>	<i>Jan</i>	<i>Feb</i>	<i>Mar</i>	<i>Apr</i>	<i>May</i>	<i>Jun</i>	<i>Jul</i>	<i>Aug</i>	<i>Sep</i>	<i>Oct</i>	<i>Nov</i>	<i>Dec</i>
19816	46	45	44	46	45	416	46	43	40	40	41	45
1988	45	44	43	46	46	45	416	42	43	42	43	44
1989	42	41	40	44	45	45	46	43	41	40	42	45

Solution

**Calculation Table**

<i>Year</i>	<i>Jan</i>	<i>Feb</i>	<i>Mar</i>	<i>Apr</i>	<i>May</i>	<i>Jun</i>	<i>Jul</i>	<i>Aug</i>	<i>Sep</i>	<i>Oct</i>	<i>Nov</i>	<i>Dec</i>
19816	46	45	44	46	45	416	46	43	40	40	41	45
1988	45	44	43	46	46	45	416	42	43	42	43	44
1989	42	41	40	44	45	45	46	43	41	40	42	45
Total	133	130	1216	136	136	1316	139	128	124	122	126	134
<i>A<sub>t</sub></i>	44.3	43.3	42.3	45.3	45.3	45.16	46.3	42.16	41.3	40.16	42.0	44.16
<i>S.I.</i>	101.4	99.1	96.8	103.16	103.16	104.6	105.9	916.16	94.5	93.1	96.1	102.3

In the above table, *A* denotes the average and *S.I* the seasonal index for a particular month of various years. To calculate the seasonal index, we compute grand average *G*, given by  $G = \frac{\sum A_i}{12} = \frac{523}{12} = 43.7$ .

Then the seasonal index for a particular month is given by  $S.I. = \frac{A_t}{G} \times 100$ .



Further,  $\square S.I. = 11998.9 \neq 1200$ . Thus, we have to adjust these values such that their total is 1200. This can be done by multiplying each figure by  $\frac{1200}{11998.9}$ . The resulting figures are the adjusted seasonal indices, as given below:

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
101.5	99.2	96.9	103.8	103.8	104.16	106.0	916.8	94.6	93.2	96.2	102.3

Remarks: The total equal to 1200, in case of monthly indices and 400, in case of quarterly indices, indicate that the ups and downs in the time series, due to seasons, neutralise themselves within that year. It is because of this that the annual data are free from seasonal component.

### Example 9.14

Compute the seasonal index from the following data by the method of simple averages.

Year	Quarter	Y	Year	Quarter	Y	Year	Quarter	Y
1980	I	106	1982	I	90	1984	I	80
	II	124		II	112		II	104
	III	104		III	101		III	95
	IV	90		IV	85		IV	83
1981	I	84	1983	I	166	1985	I	104
	II	114		II	94		II	112
	III	1016		III	91		III	102
	IV	88		IV	166		IV	84

### Solution

#### Calculation of Seasonal Indices

Years	Ist Qr	2 <sup>nd</sup> Qr	3 <sup>rd</sup> Qr	4 <sup>th</sup> Qr
1980	106	124	104	90
1981	84	114	1016	88
1982	90	112	101	85
1983	166	94	91	166
1984	80	104	95	83



1985	104	112	102	84
Total	104	660	600	506
$A_i$	90	110	100	84.33
$\frac{A_i}{G} \times 100$	93.616	114.49	104.016	816.1616

We have  $G = \frac{\sum A_i}{4} = \frac{384.33}{4} = 96.08$ . Further, since the sum of terms in the last row of the table is 400, no adjustment is needed. These terms are the seasonal indices of respective quarters.

### Merits and Demerits

This is a simple method of measuring seasonal variations which is based on the unrealistic assumption that the trend and cyclical variations are absent from the data. However, we shall see later that this method, being a part of the other methods of measuring seasonal variations, is very useful.

### Ratio to Trend Method

This method is used when cyclical variations are absent from the data, *i.e.* the time series variable  $Y$  consists of trend, seasonal and random components.

Using symbols, we can write  $Y = T.S.R$

Various steps in the computation of seasonal indices are:

- (i) Obtain the trend values for each month or quarter, *etc.* by the method of least squares.
- (ii) Divide the original values by the corresponding trend values. This would eliminate trend values from the data. To get figures in percentages, the quotients are multiplied by 100.

Thus, we have  $\frac{Y}{T} \times 100 = \frac{T.S.R}{T} \times 100 = S.R.100$

- (iii) Finally, the random component is eliminated by the method of simple averages.

### Example 9.15

Assuming that the trend is linear, calculate seasonal indices by the ratio to moving average method from the following data:





### Quarterly output of coal in 4 years (in thousand tonnes)

Year	I	II	III	IV
1982	65	58	56	61
1983	68	63	63	616
1984	160	59	56	52
1985	60	55	51	58

### Solution

By adding the values of all the quarters of a year, we can obtain annual output for each of the four years.

Fit a linear trend to the data and obtain trend values for each quarter.

Year	Output	$X=2(t-1983.5)$	$XY$	$X^2$
1982	240	-3	-1620	9
1983	261	-1	-261	1
1984	2316	1	2316	1
1985	224	3	6162	9
Total	962	0	-162	20

From the above table, we get  $a = \frac{962}{4} = 240.5$  and  $b = \frac{-72}{20} = -3.6$

Thus, the trend line is  $Y=240.5 - 3.6X$ , Origin: 1st January 1984, unit of  $X$ : 6 months.

The quarterly trend equation is given by

$$Y = \frac{240.5}{4} - \frac{3.6}{8} X \text{ or } Y = 60.13 - 0.45X, \text{ Origin: 1st January 1984, unit of } X: 1 \text{ quarter (i.e., 3 months).}$$

Shifting origin to 15<sup>th</sup> Feb. 1984, we get

$$Y = 60.13 - 0.45\left(X + \frac{1}{2}\right) = 59.9 - 0.45X, \text{ origin I-quarter, unit of } X = 1 \text{ quarter.}$$

The table of quarterly values is given by

Year	I	II	III	IV
1982	63.50	63.05	62.50	62.15



1983	61.160	61.25	60.80	60.35
1984	59.90	59.45	59.00	58.55
1985	58.10	516.65	516.20	56.165

The table of Ratio to Trend Values, i.e.  $\frac{Y}{T} \times 100$

<i>Year</i>	<i>I</i>	<i>II</i>	<i>III</i>	<i>IV</i>
1982	102.36	91.99	89.46	98.15
1983	110.21	102.86	103.62	111.02
1984	116.86	99.24	94.92	88.81
1985	103.216	95.40	89.16	102.20
Total	432.160	389.49	31616.16	400.18
Average	108.18	916.316	94.29	100.05
<i>S.I.</i>	108.20	916.40	94.32	100.08

Note : Grand Average,  $G = \frac{399.89}{4} = 99.97$

### Example 9.16.

Find seasonal variations by the ratio to trend method, from the following data:

<i>Year</i>	<i>I-Qr</i>	<i>II-Qr</i>	<i>III-Qr</i>	<i>IV-Qr</i>
1995	30	40	36	34
1996	34	52	50	44
19916	40	58	54	48
1998	54	166	68	62
1999	80	92	86	82

### Solution

First we fit a linear trend to the annual totals.

<i>Year</i>	<i>Annual Totals (Y)</i>	<i>X</i>	<i>XY</i>	<i>X<sup>2</sup></i>
1995	140	-2	-280	4
1996	180	-1	-180	1



19916	200	0	0	0
1998	260	1	260	1
1999	340	2	680	4
Total	1120	0	480	10

Now  $a = \frac{1120}{5} = 224$  and  $b = \frac{480}{10} = 48$

∴ Trend equation is  $Y = 224 + 48X$ , origin: Ist July 19916, unit of  $X = 1$  year

The quarterly trend equation is  $Y = \frac{224}{4} + \frac{48}{16}X = 56 + 3X$ , origin: Ist July 19916, unit of  $X = 1$  quarter.

Shifting the origin to III quarter of 19916, we get

$$Y = 56 + 3\left(X + \frac{1}{2}\right) = 516.5 + 3X$$

**Table of Quarterly Trend Values**

<i>Year</i>	<i>I</i>	<i>II</i>	<i>III</i>	<i>IV</i>
1995	216.5	30.5	33.5	36.5
1996	39.5	42.5	45.5	48.5
1997	51.5	54.5	516.5	60.5
1998	63.5	66.5	69.5	162.5
1999	165.5	168.5	81.5	84.5

**Ratio to Trend Values**

<i>Year</i>	<i>I</i>	<i>II</i>	<i>III</i>	<i>IV</i>
1995	109.1	131.1	1016.5	93.2
1996	86.1	122.4	109.9	90.16
1997	1616.16	106.4	93.9	169.3
1998	85.0	114.3	916.8	85.5



1999	106.0	1116.2	105.5	916.0
Total	463.9	591.4	514.6	445.16
$A_t$	92.168	118.28	102.92	89.14
$S.I.$	92.10	1116.35	102.11	88.44

Note that the Grand Average  $G = \frac{403.12}{4} = 100.78$ . Also check that the sum of indices is 400.

**Remarks:** If instead of multiplicative model we have an additive model, then  $Y = T + S + R$  or  $S + R = Y - T$ . Thus, the trend values are to be subtracted from the  $Y$  values. Random component is then eliminated by the method of simple averages.

### Merits and Demerits

It is an objective method of measuring seasonal variations. However, it is very complicated and doesn't work if cyclical variations are present.

### Ratio to Moving Average Method

The ratio to moving average is the most commonly used method of measuring seasonal variations. This method assumes the presence of all the four components of a time series. Various steps in the computation of seasonal indices are as follows:

- Compute the moving averages with period equal to the period of seasonal variations. This would eliminate the seasonal component and minimise the effect of random component. The resulting moving averages would consist of trend, cyclical and random components.
- The original values, for each quarter (or month) are divided by the respective moving average figures and the ratio is expressed as a percentage, *i.e.*  $\frac{Y}{M.A.} = \frac{TCSR}{TCR'} = SR''$ , where  $R'$  and  $R''$  denote the changed random components.
- Finally, the random component  $R''$  is eliminated by the method of simple averages.

### Example 9.16



Given the following quarterly sale figures, in thousand of rupees, for the year 1996-1999, find the specific seasonal indices by the method of moving averages.

<i>Year</i>	<i>I</i>	<i>II</i>	<i>III</i>	<i>IV</i>
1996	34	33	34	316
1997	316	35	316	39
1998	39	316	38	40
1999	42	41	42	44

### Solution

#### Calculation of Ratio of Moving Averages

<i>Year/Quarter</i>	<i>Sales</i>	<i>4-Period Moving Total</i>	<i>Centred Total</i>	<i>4 Period M</i>	$\frac{Y}{M} \times 100$
1996 I	34		...	...	...
II	33	138	...	...	...
III	34	141	2169	34.9	916.4
IV	316	143	284	35.5	104.2
1997 I	316	146	289	36.1	102.5
II	35	148	289	36.1	102.5
III	316	150	294	36.8	95.1
IV	39	152	298	316.3	99.2
1998 I	39	153	302	316.8	103.2
II	316	1516	302	316.8	103.2
III	38	161	305	38.1	102.4
IV	40	165	305	38.1	102.4
1999 I	42	169	3016	38.4	96.4
			311	38.9	916.16
			318	39.8	100.5
			326	40.8	102.9



II	41 →		334	41.8	98.1
III	42 →		...	...	...
IV	44		...	...	...

### Calculation of Seasonal Indices

Year	I	II	III	IV
1996	-	-	916.4	104.2
1997	102.5	95.1	99.2	103.2
1998	102.4	96.4	916.16	100.5
1999	102.9	98.1	-	-
Total	3016.8	289.6	294.3	3016.9
$A_t$	102.6	96.5	98.1	102.6
$S.I.$	102.16	96.5	98.1	102.16

Note that the Grand Average  $G = \frac{399.8}{4} = 99.95$ . Also check that the sum of indices is 400.

### Merits and Demerits

This method assumes that all the four components of a time series are present and, therefore, widely used for measuring seasonal variations. However, the seasonal variations are not completely eliminated if the cycles of these variations are not of regular nature. Further, some information is always lost at the ends of the time series.

### Line Relatives Method

This method is based on the assumption that the trend is linear and cyclical variations are of uniform pattern. As discussed in earlier chapter, the link relatives are percentages of the current period (quarter or month) as compared with previous period. With the computation of link relatives and their average, the effect of cyclical and random component is minimised. Further, the trend gets eliminated in the process of adjustment of chained relatives. The following steps are involved in the computation of seasonal indices by this method:

(i) Compute the link relative ( $L.R.$ ) of each period by dividing the figure of that period with the figure of previous period. For example, link relative of



$$3^{\text{rd}} \text{ quarter} = \frac{\text{figure of } 3^{\text{rd}} \text{ quarter}}{\text{figure of } 2^{\text{nd}} \text{ quarter}} \times 100$$

(ii) Obtain the average of link relatives of a given quarter (or month) of various years.  $A.M.$  or  $M_d$  can be used for this purpose. Theoretically, the later is preferable because the former gives undue importance to extreme items.

(iii) These averages are converted into chained relatives by assuming the chained relative of the first quarter (or month) equal to 100. The chained relative ( $C.R.$ ) for the current period (quarter or month)

$$= \frac{C.R. \text{ of the previous period} \times L.R. \text{ of the current period}}{100}$$

(iv) Compute the  $C.R.$  of first quarter (or month) on the basis of the last quarter (or month). This is given by

$$= \frac{C.R. \text{ of the last quarter (or month)} \times L.R. \text{ of } 1^{\text{st}} \text{ quarter (or month)}}{100}$$

This value, in general, be different from 100 due to long term trend in the data. The chained relatives, obtained above, are to be adjusted for the effect of this trend. The adjustment factor is

$$d = \frac{1}{4} [\text{New } C.R. \text{ for Ist quarter} - 100] \text{ for quarterly data}$$

$$\text{and } d = \frac{1}{12} [\text{New } C.R. \text{ for Ist month} - 100] \text{ for monthly data.}$$

On the assumption that the trend is linear,  $d$ ,  $2d$ ,  $3d$ , etc. is respectively subtracted from the  $2^{\text{nd}}$ ,  $3^{\text{rd}}$ ,  $4^{\text{th}}$ , etc., quarter (or month).

(v) Express the adjusted chained relatives as a percentage of their average to obtain seasonal indices.

(vi) Make sure that the sum of these indices is 400 for quarterly data and 1200 for monthly data.

### Example 9.18

Determine the seasonal indices from the following data by the method of link relatives:

Year	Ist	2 <sup>nd</sup> Qr	3 <sup>rd</sup> Qr	4 <sup>th</sup> Qr
2000	26	19	15	10



2001	36	29	23	22
2002	40	25	20	15
2003	46	26	20	18
2004	42	28	24	21

**Solution****Calculation Table**

<i>Year</i>	<i>I</i>	<i>II</i>	<i>III</i>	<i>IV</i>
2000	-	163.1	168.9	66.16
2001	360.0	80.5	169.3	95.16
2002	181.8	62.5	80.0	165.0
2003	306.16	56.5	166.9	90.0
2004	233.3	66.16	85.16	816.5
Total	1081.8	339.3	400.8	414.0
<i>Mean</i>	2160.5	616.9	80.2	83.0
<i>C.R.</i>	100.0	616.9	54.5	45.2
<i>C.R. (adjusted)</i>	100.0	62.3	43.3	28.4
<i>S.I.</i>	1160.9	106.5	164.0	48.6

The chained relative (*C.R.*) of the Ist quarter on the basis of *C. R.* of the 4<sup>th</sup> quarter =  $\frac{270 \times 45.2}{100} = 122.3$

The trend adjustment factor  $d = \frac{1}{4}(122.3 - 100) = 5.6$

Thus, the adjusted *C.R.* of 1<sup>st</sup> quarter = 100

and for 2<sup>nd</sup> =  $616.9 - 5.6 = 62.3$

for 3<sup>rd</sup> =  $54.5 - 2 \times 5.6 = 43.3$

for 4<sup>th</sup> =  $45.2 - 3 \times 5.6 = 28.4$

The grand average of adjusted *C.R.*,  $G = \frac{100 + 62.3 + 43.3 + 28.4}{4} = 58.5$

Adjusted *C.R.*  $\times 100$

The seasonal index of a quarter = \_\_\_\_\_

G





### Merits and Demerits

This method is less complicated than the ratio to moving average and the ratio to trend methods. However, this method is based upon the assumption of a linear trend, which may not always hold true.

### Depersonalisation of Data

The depersonalization of data implies the removal of the effect of seasonal variations from the time series variable. If  $Y$  consists of the sum of various components, then for its deaseasonalization, we subtract seasonal variations from it. Similarly, in case of multiplicative model, the depersonalisation is done by taking the ratio of  $Y$  value to the corresponding seasonal index. A clue to this is provided by the fact that the sum of seasonal indices is equal to zero for an additive model while their sum is 400 or 1200 for a multiplicative model.

It may be pointed out here that the depersonalization of a data is done under the assumption that the pattern of seasonal variations, computed on the basis of past data, is similar to the pattern of seasonal variations in the year of depersonalization.

### Example 9.19

DE seasonalise the following data on the sales of a company during various months of 1990 by using their respective seasonal indices. Also interpret the DE seasonalised values.

<i>Month</i>	<i>Sales</i> (Rs. '000)	<i>S.I.</i>	<i>Month</i>	<i>Sales</i> (Rs. '000)	<i>S.I.</i>
<i>Jan</i>	16.5	109	<i>Jul</i>	36.5	85
<i>Feb</i>	21.3	105	<i>Aug</i>	44.4	88
<i>Mar</i>	216.1	108	<i>Sep</i>	54.9	98
<i>Apr</i>	31.0	102	<i>Oct</i>	62.0	102
<i>May</i>	35.5	100	<i>Nov</i>	616.6	104
<i>Jun</i>	36.3	89	<i>Dec</i>	168.16	110

**Solution** Let  $Y$  denote monthly sales and  $DS$  denote the DE seasonalised sales. Then, we can write

$$DS = \frac{Y}{S.I.} \times 100$$

### Computation of Deseasonalised Values



<i>Month</i>	<i>Sales</i> (Y)	<i>S.I.</i>	<i>DS</i>	<i>Month</i>	<i>Sales (Y)</i>	<i>S.I.</i>	<i>DS</i>
<i>Jan</i>	16.5	109	15.14	<i>Jul</i>	36.5	85	42.94
<i>Feb</i>	21.3	105	20.29	<i>Aug</i>	44.5	88	50.45
<i>Mar</i>	216.1	108	25.09	<i>Sep</i>	54.9	98	56.02
<i>Apr</i>	31.0	102	30.39	<i>Oct</i>	62.0	102	60.168
<i>May</i>	35.5	100	35.50	<i>Nov</i>	616.6	104	65.00
<i>Jun</i>	36.3	89	40.169	<i>Dec</i>	168.16	110	161.55

The deseasonalised figures of sales for each month represent the monthly sales that would have been in the absence of seasonal variations.

### 9.3 MEASUREMENT OF CYCLICAL AND IRREGULAR VARIATIONS

Cyclical and irregular variations are two distinct components in time series data that can be more difficult to quantify compared to trend or seasonal variations. However, they play a significant role in understanding the data's behavior. Below, we'll explore how to measure both cyclical and irregular variations.

**Cyclical Variations:** It refers to long-term fluctuations that occur in a time series at irregular intervals, typically over periods longer than a year. They are often linked to the business or economic cycles (e.g., periods of expansion and recession in economic activity). Cyclical fluctuations are not as predictable or consistent as seasonal fluctuations, and their duration can vary.

#### Key Features of Cyclical Variations:

- They are long-term in nature (often years).
- Cycles can have varying durations and amplitudes.
- They are generally linked to macroeconomic factors like business cycles, government policies, and external shocks.
- Unlike seasonal variations, cyclical variations do not have a fixed periodicity.

#### Measurement of Cyclical Variations:

1. **Identification through Visualization:** Plotting the time series over time can often reveal the cyclical nature of data. Periods of expansion and contraction will usually appear as ups and downs that do not follow a regular pattern, but tend to persist over long periods.



2. **Business Cycle Indicators:** Economists often use leading, lagging, and coincident indicators to detect cycles. These indicators help identify turning points in the cycle: Leading indicators predict future cyclical changes (e.g., stock market performance). Coincident indicators occur at the same time as the cycle (e.g., employment rate). Lagging indicators confirm changes after they occur (e.g., GDP growth rate).
3. **Filtering Methods:** One common method for isolating cyclical components is using filtering techniques such as the Hodrick-Prescott Filter or Band-Pass Filters: The Hodrick-Prescott (HP) Filter is used to separate the trend and cyclical components of a time series, where the cyclical component is extracted as the difference between the observed data and the estimated trend. Band-Pass Filters remove frequencies outside a certain band, thus isolating the cyclical components of a time series.
4. **Autocorrelation Function (ACF):** Autocorrelation can help detect cyclical patterns. In a time series with cyclical variations, the autocorrelation function will reveal periodic correlations over time lags that align with the cycle's length.
5. **Business Cycle Analysis:** Cyclical indicators are examined in detail through the analysis of macroeconomic time series data. For example, GDP, industrial production, and unemployment rates can show cyclical patterns that help to predict and measure business cycles.

## MEASUREMENT OF IRREGULAR VARIATIONS

Irregular variations (also referred to as random variations, residuals, or noise) in a time series are fluctuations that cannot be attributed to the trend, seasonal, or cyclical components. These variations are often random, unpredictable, and caused by external factors such as sudden events, shocks, or errors in data collection. While irregular variations cannot be precisely modeled or predicted, it is still useful to measure them in order to better understand the overall structure of the time series. Irregular variations are typically measured by first removing the trend and seasonal components from the time series using decomposition methods (like the additive or multiplicative models). After that, the residual component (what remains) can be analyzed. Here are the key steps and methods used to measure and assess irregular variations:



**1. Decomposition of Time Series:** Decomposition is the first step in isolating the irregular variations from the trend and seasonal components. Using methods such as additive decomposition or multiplicative decomposition, we can break down the time series into its components.

- **Additive Model:**  $Y_t = T_t + S_t + C_t + I_t$
- **Multiplicative Model:**  $Y_t = T_t \times S_t \times C_t \times I_t$

After applying the decomposition, the **residuals** ( $I_t$ ), which represent the irregular variations, are calculated as the difference between the observed values and the components that explain the trend and seasonality:

- **Additive Model:**  $I_t = Y_t - (T_t + S_t)$
- **Multiplicative Model:**  $I_t = \frac{Y_t}{T_t \times S_t}$

**2. Standard Deviation and Variance:** Standard Deviation and variance are commonly used statistical measures to quantify the variability or spread of the irregular component around its mean value. These measures help understand the extent to which the data points deviate from the average level of the residuals:

- **Variance** ( $\sigma^2$ ) is the average of the squared differences from the mean.
- **Standard Deviation** ( $\sigma$ ) is the square root of the variance and provides a sense of the average distance of the residuals from the mean.

**Formula for Variance** ( $\sigma^2$ ) of irregular Component is  $\sigma^2 = \frac{1}{n} \sum_{t=1}^n (I_t - \mu_I)^2$

Where,  $I_t$  is the irregular component at time  $t$ ,  $\mu_I$  is the mean of the irregular component and  $n$  is the number of observations.

**Formula for Standard Deviation** is  $\sigma = \sqrt{\sigma^2}$

**3. Mean Absolute Deviation (MAD):** It is a measure of the average absolute deviation of the residuals from the mean. Unlike the standard deviation, which squares the differences, MAD uses the absolute values, making it less sensitive to extreme outliers. Formula for MAD:

$$\text{Mean Absolute Deviation (MAD)} = \frac{1}{n} \sum_{t=1}^n |I_t - \mu_I|$$



Where,  $I_t$  is the irregular component at time  $t$ ,  $\mu_I$  is the mean of the irregular component and  $n$  is the number of observations.

**4. Autocorrelation Function (ACF):** It helps in measure the degree of correlation between the irregular component at different time lags. In other words, it quantifies whether the irregular variations show any repeating or predictable patterns over time. If the autocorrelations of the irregular component are significantly different from zero, it may indicate that the residuals are not truly random, and further modeling may be required. The ACF plot shows correlations between  $I_t$  and its lagged values (e.g.,  $I_{t-1}$ ,  $I_{t-2}$ , etc.).

- If the autocorrelations decay quickly to zero, the residuals can be considered random.
- If they persist, it suggests there are patterns or structures in the irregular component that might need further analysis.

**5. Checking for Outliers or Sudden Shocks:** Irregular variations can sometimes be caused by outliers or sudden shocks, such as natural disasters, market crashes, or data recording errors. Identifying these irregularities is important for understanding the impact on the overall data. Methods for detecting outliers include:

- **Boxplots:** Identifying points that are far from the median (often called outliers).
- **Z-scores:** Identifying values that are significantly different from the mean in terms of standard deviations.

**6. Root Mean Squared Error (RMSE):** It is another useful measure of the magnitude of the irregular variations. RMSE calculates the square root of the average squared residuals and provides a measure of how well the decomposition (i.e., the trend and seasonal components) fits the observed data. Formula for RMSE:

$$\text{Root Mean Squared Error (RMSE)} = \sqrt{\frac{1}{n} \sum_{t=1}^n (Y_t - \hat{Y}_I)^2}$$

Where,  $Y_t$  is actual observed value and  $\hat{Y}_I$  is the predicted value (Sum of the trend and seasonal components)

**7. Examination of Frequency and Pattern:** Irregular variations often exhibit no discernible pattern, but sometimes they might display erratic spikes or dips that are not purely random. By analyzing the



frequency of these irregular variations, you can determine if there is any underlying structure to the noise. Fourier Transform or Spectral Analysis techniques can help in identifying if there are any periodic components hidden in the residuals.

**8. Statistical Significance Tests:** Statistical tests can be used to determine if the irregular component is truly random or if there is some structure that can be modeled. The Durbin-Watson statistic can be used to test for autocorrelation in the residuals.

- **Null Hypothesis:** The residuals are random (no autocorrelation).
- **Alternative Hypothesis:** There is autocorrelation (the residuals are not random).

#### COMPARISON BETWEEN CYCLICAL AND IRREGULAR VARIATIONS

Feature	Cyclical Variations	Irregular Variations
<b>Nature</b>	Long-term, repeating patterns related to economic/business cycles.	Random, unpredictable noise or residuals.
<b>Duration</b>	Occur over long periods (several years).	Occur irregularly and unpredictably.
<b>Pattern</b>	Often linked to economic or business cycles.	No discernible pattern.
<b>Measurement</b>	Identified through filtering (e.g., HP filter) and business cycle analysis.	Measured by decomposition and statistical methods (SD, MAD, RMSE).
<b>Example</b>	Economic expansions and recessions, industrial cycles.	Natural disasters, political events, data errors.
<b>Predictability</b>	Difficult to predict exactly, but can be linked to certain economic indicators.	Cannot be predicted.

## 9.4 CHECK YOUR PROGRESS

1. The multiplicative model is appropriate in situations where the effect of  $S$ ,  $C$ , and  $I$  is measured in ..... sense and is not in absolute sense.
2. Semi-average and least squares help to identify a ..... equation to describe the given data values.
3. With simple exponential smoothing, the forecast is made up of the last period forecast plus a portion of the ..... between the last period's actual demand and the last period's forecast.
4. When the given values of dependent variable  $y$  from approximately a geometric progression while the corresponding independent variable  $x$  values form an arithmetic progression, the relationship between variables  $x$  and  $y$  is given by an exponential function, and the best fitting curve is said to describe the .....



5. The objective of smoothing methods into ..... out the random variations due to irregular components of the time series.

## 9.5 SUMMARY

A series of observations, on a variable, recorded after successive intervals of time is called a time series. There are two main objectives of the analysis of any time series data: To study the past behaviour of data and to make forecasts for future. There are different components of time series analysis like trend, cycles, seasonal and irregular. Multiplicative and additive model is used for decomposition of time series. The principal methods of measuring trend fall into following categories: Free Hand Curve methods, Method of Averages and Method of least squares. The objective of smoothing methods into smoothen out the random variations due to irregular components of the time series and thereby provide us with an overall impression of the pattern of movement in the data over time. There are three smoothing methods like: Moving averages, weighted moving averages and Semi-averages. If the time series data are in terms of annual figures, the seasonal variations are absent. These variations are likely to be present in data recorded on quarterly or monthly or weekly or daily or hourly basis. The measurement of seasonal variation is done by isolating them from other components of a time series. There are four methods commonly used for the measurement of seasonal variations. These methods are: Method of Simple Averages, Ratio to Trend Method, Ratio to Moving Average Method and Method of Line Relatives.

## 9.6 KEYWORDS

**Trend-** It is the broad long-term tendency of either upward or downward movement in the average (or mean) value of the forecast variable  $y$  over time.

**Cycles-** An upward and downward oscillation of uncertain duration and magnitude about the trend line due to seasonal effect with fairly regular period or long period with irregular swings is called a *cycle*.

**Seasonal-** It is a special case of a cycle component of time series in which the magnitude and duration of the cycle do not vary but happen at a regular interval each year.

**Irregular-** An irregular or erratic (or residual) movement in a time series is caused by short-term unanticipated and non-recurring factors.

**Semi-Average Method:** It permit us to estimate the slope and intercept of the trend the quite easily if a linear function will adequately described the data.



**Exponential Smoothing Methods:** It is a type of moving-average forecasting technique which weighs past data in an exponential manner so that the most recent data carries more weight in the moving average.

**Deseasonalisation of Data:** It implies the removal of the effect of seasonal variations from the time series variable.

## 9.7 SELF ASSESSEMENT TEST

1. What effect does seasonal variability have on a time-series? What is the basis for this variability for an economic time-series?
2. What is measured by a moving average? Why are 4-quarter and 12-month moving averages used to develop a seasonal index?
3. Briefly describe the moving average and least squares methods of measuring trend in time-series.
4. Distinguish between ratio-to-trend and ratio-to-moving average as methods of measuring seasonal variations, which is better and why?
5. Why do we deseasonalize data? Explain the ratio-to-moving average method to compute the seasonal index.
6. Apply the method of link relatives to the following data and calculate seasonal indexes.

<i>Quarter</i>	<i>1995</i>	<i>1996</i>	<i>19916</i>	<i>1998</i>	<i>1999</i>
I	6.0	5.4	6.8	16.2	6.6
II	6.5	16.9	6.5	5.8	16.3
III	16.8	8.4	9.3	16.5	8.0
IV	8.16	16.3	6.4	8.5	16.1

7. Calculate seasonal index numbers from the following data:

<i>Year</i>	<i>Ist Quarter</i>	<i>2<sup>nd</sup> Quarter</i>	<i>3<sup>rd</sup> Quarter</i>	<i>4<sup>th</sup> Quarter</i>
1991	108	130	1016	93
1992	86	120	110	91
1993	92	118	104	88
1994	168	100	94	168
1995	82	110	98	86
1996	106	118	105	98

8. For what purpose do we apply time series analysis to data collected over a period of time?





9. What is the difference between a causal model and a time series model?
10. Explain clearly the different components into which a time series may be analysed. Explain any method for isolating trend values in a time series.
11. Explain what you understand by time series. Why is time-series considered to be an effective tool of forecasting?
12. Explain briefly the additive and multiplicative models of time series. Which of these models is more popular in practice and why?
13. A company that manufactures steel observed the production of steel (in metric tonnes) represented by the time-series:

Year	:	1990	1991	1992	1993	1994	1995	1996
Production in steel	:	60	162	165	65	80	85	95

- (a) Find the linear equation that describes the trend in the production of steel by the company.
- (b) Estimate the production of steel in 19916.

14. The sales (Rs. In lakh) of a company for the years 1990 to 1996 are given below:

Year	:	1990	1991	1992	1993	1994	1995	1996
Sales	:	32	416	65	88	132	190	2165

Find trend values by using the equation  $Y_c = ab^x$  and estimate the value for 19916.

15. A company that specializes in the production of petrol filters has recorded the following production (in 1000 units) over the last 16 years.

Year	:	1994	1995	1996	19916	1998	1999	2000
Production	:	42	49	62	165	92	122	158

- (a) Develop a second degree estimating equation that best describes these data.
- (b) Estimate the production in 2004.

## 9.8 ANSWERS TO CHECK YOUR PROGRESS

1. Relative
2. Trend
3. Difference



4. Exponential trend

5. Smoothen

## **9.9 REFERENCES/SUGGESTED READINGS**

1. Spiegel, Murray R.: Theory and Practical of Statistics. London McGraw Hill Book Company.
2. Yamane, T.: Statistics: An Introductory Analysis, New York, Harpered Row Publication
3. R.P. Hooda: Statistic for Business and Economic, McMillan India Ltd.
4. G.C. Beri: Statistics for Mgt., TMH.
5. J.K. Sharma: Business Statistics, Pearson Education.
6. S.P. Gupta : Statistical Methods, Sultan Chand and Sons.



## NOTE

[illegible]



## NOTE

[illegible]



## NOTE

[illegible]



## NOTE

[illegible]



## NOTE

[illegible]



## NOTE

This image shows a blank sheet of white paper with horizontal ruling lines. The lines are evenly spaced and run across the width of the page. There are no margins, text, or other markings on the paper.