2025

Introduction to Statistics

Dr. Babasaheb Ambedkar Open University

Expert Committee

Prof. (Dr.) Nilesh Modi Professor and Director, School of Computer Science, Dr. Babasaheb Ambedkar Open University, Ahmedabad	**(Chairman)**
Prof. (Dr.) Ajay Parikh Professor and Head, Department of Computer Science Gujarat Vidyapith, Ahmedabad	**(Member)**
Prof. (Dr.) Satyen Parikh Dean, School of Computer Science and Application Ganpat University, Kherva, Mahesana	**(Member)**
Prof. M. T. Savaliya Associate Professor and Head, Computer Engineering Department Vishwakarma Engineering College, Ahmedabad	**(Member)**
Dr. Himanshu Patel Assistant Professor, School of Computer Science, Dr. Babasaheb Ambedkar Open University, Ahmedabad	**(Member Secretary)**

Course Writer

Units	Contributors
1	**Dr. Padmalochan Hazarika, Gauhati University**
2, 12	**Ajanta Majumdar, Gauhati Commerce College**
3, 4	**Dr. Joydeep Baruah, O.K.D. Institute of Social Change and Development**
5, 13	**Dr. Tarakeswar Chaudhary, Retired Professor, Cotton College**
6	**Harekrishna Deka, KKHSOU**
7, 8, 9, 10	**Dr. Rijusmita Sarma, HoD, Department of Statistics, L.C.B. College, Gauhati**
11	**Dr. Rupam Barman, Tezpur University**
14	**Dr. Pranjal Sarma, L.C.B. College, Gauhati & Dr. Bhaskar Sarmah, KKHSOU**

Content Reviewer & Editor

Prof. (Dr.) Nilesh Modi
Professor and Director, School of Computer Science,
Dr. Babasaheb Ambedkar Open University, Ahmedabad

Acknowledgement: The content in this book is modifications based on the work created and shared by the Krishna Kanta Handiqui State Open University (KKHSOU) for the subject Descriptive Statistics used according to terms described in Creative Commons Attribution-Share Alike 4.0 International (CC BY-SA 4.0)



This publication is made available under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International
(CC BY-NC-SA 4.0) https://creativecommons.org/licenses/by-nc-sa/4.0/

ISBN:

Printed and published by: Dr. Babasaheb Ambedkar Open University, Ahmedabad

While all efforts have been made by editors to check accuracy of the content, the representation of facts, principles, descriptions and methods are that of the respective module writers. Views expressed in the publication are that of the authors, and do not necessarily reflect the views of Dr. Babasaheb Ambedkar Open University. All products and services mentioned are owned by their respective copyrights holders, and mere presentation in the publication does not mean endorsement by Dr. Babasaheb Ambedkar Open University. Every effort has been made to acknowledge and attribute all sources of information used in preparation of this learning material. Readers are requested to kindly
notify missing attribution, if any.

Dr. Babasaheb Ambedkar Open University

BSCCS-204

Introduction to Statistics

Block-1:

Unit-1: Measures of Central Tendency 07

Unit-2: Measures of Dispersion 33

Unit-3: Skewness, Moments and Kurtosis 57

Unit-4: Correlation 81

Unit-5: Regression 95

Block-2:

Unit-6: Fundamentals of Probability 104

Unit-7: Conditional Probability 129

Unit-8: Random Variables and its Probability Distribution 150

Block-3:

Unit-9: Theoretical Probability Distributions (Discrete Variable – I) 163

Unit-10: Theoretical Probability Distributions (Discrete Variable – II) 181

Unit-11: Theoretical Distributions (Continuous Variable) 191

Block-4:

Unit-12: Index Numbers 204

Unit-13: Time Series 230

Unit-14: Measurement of Economic Inequality 248

UNIT 1: MEASURES OF CENTRAL TENDENCY

UNIT STRUCTURE

1.1 Learning Objectives

1.2 Introduction

1.3 Measures of Central Tendency

- 1.3.1 Definition
- 1.3.2 Characterstics of a Good Average
- 1.3.3 'Σ' Symbol
- 1.3.4 Different Types of Measures of Central Tendency

1.4 Let Us Sum Up

1.5 Further Reading

1.6 Answers to Check Your Progress

1.7 Model Questions

1.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- explain the meaning of measures of central tendency or averages
- describe the various types of averages along with their advantages, limitations and uses
- explain the meaning of weighted mean.

1.2 INTRODUCTION

We need a numerical expression which summaries the characteristics of the whole set of data under study. 'Measures of Central Location' or 'Measures of Central Tendency' popularly known as 'Averages' serve this purpose. A figure which represents a series of values should obviously be greater than the lowest value and less than the highest value. It should be a value somewhere between these two limits, possibly at the centre where most of the values of the series center. Such a figure is called a Measure of Central Tendency.

Measures of central tendency, very often, are not fully representative

of a given set of data. This happens when the extent of variation of individual values in relation to the average, or in relation to the other values is large. As an illustration, let us observe the following three series:

Series A	40	40	40	40	40
Series B	35	39	41	42	43
Series C	10	18	35	57	80

In the first series the arithmetic mean or simply mean (sum of the values divided by the number of values) is 40 and the values of all the items are identical each being equal to 40. The mean fully represents the series in general and the individual items in particular. The data (items or observations) are not at all scattered. In the second series, although the mean is 40, all the observations are not very much scattered as the minimum value of the series is 35 and the maximum value is 43. Hence in case of the second series also the mean is a good representation of the series. Although none of the observations of the series is equal to the mean of the series yet the discrepancy between the mean and any other observation is not so significant. In case of the third series, we observe that all the items or observations of the series are different. This series also has the same mean 40. In case of this series the observations are widely scattered. Clearly, in case of this series the mean neither satisfactorily represents the entire series in general nor the individual items of the series in particular. Thus we have observed that although all the above three series have the same average (arithmetic mean is a method of measuring average) yet they widely differ from one another in terms of their formation. When the extent of variation (deviation or scatteredness) of the individual values (items or observations) of a distribution or series in relation to their average or in relation to the other values is large then measures of central tendency or averages cannot be representative of the distribution. Hence it is important for any investigation not only to know the average of any type (mean, median or mode) but also the scatteredness of the various observations of a distribution.

1.3 MEASURES OF CENTRAL TENDENCY OR AVERAGES

1.3.1 Definition

An average of a distribution (i.e. a distribution of the values of a variable like height, weight, income etc.) is a representative value of that distribution. This representative value usually lies at the central part of distribution. "A measure of central tendency or an average of a certain distribution is nothing but a representative value of that distribution which enables us to comprehend in a single effort the significance of the whole."

Unit of average: The unit of average of a distribution is the unit of that distribution.

1.3.2 Characteristics of a Good Average

An average will be termed as a good if it possesses the following characteristics:

i) It should be easy to understand and calculate.

ii) It should be rigidly defined which means that the definition should be so clear that its interpretation does not differ from person to person.

iii) The average should be based on all the values of the variable.

iv) The average should have sampling stability. This means that the value of an average calculated from various independent random samples of the same size from a given population should not vary much from another.

v) It should be capable of further algebric treatment.

vi) The average should not be unduly affected by extreme values.

1.3.3 'Σ' Symbol

In order to denote sum (i.e., a total of certain quantities) the Greek letter 'Σ' (capital sigma) is used. For example, if variable x

takes the values $x_1, x_2, x_3, \ldots, \ldots, ..., x_n$ then the sum of these values of the variable i.e. $(x_1, x_2, x_3, \ldots, \ldots, ..., x_n)$ is denoted by $\sum_{i=1}^{n} x_i$ or $\sum x$. The symbol $\sum_{i=1}^{n} x_i$ means that the lower limit of i is 1 and the symbol Σ means that all the values of x_i for i = 1, 2, 3, …, …, ..., n are to added. Thus $\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots \ldots ... + x_n$. Again, the symbol $\sum x$ implies 'sum of the values of x'.

1.3.4 Different Types of Measures of Central Tendency

The following three types of measures of central tendency or averages are in use.

(a) Mean, (b) Median, and (c) Mode.

a) MEAN: There are three types of mean, namely,

(i) Arithmetic Mean (A.M.), (ii) Geometric Mean (G.M.) and

(iii) Harmonic Mean (H.M.)

So far as mean is concerned, we shall discuss arithmetic mean only since this is the most popular technique among the different types of mean.

In unit two, we have already discussed about arithmetic mean and geometric mean. Here we will discuss some special case related to Arithmetic mean.

The A.M. of a variable x is generally denoted by $\bar{x}$ and is defined as the sum of the values of divided by the total number of values of x.

Thus, if $x_1, x_2, x_3, ..., x_n$ be n values of x then A.M ($\bar{x}$) will be given by,

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum x}{n}$$

In case of ungrouped frequency distribution table:

$x : x_1 \quad x_2 \quad x_3 \quad \ldots \quad x_n$

$f : f_1 \quad f_2 \quad f_3 \quad \ldots \quad f_n$

the A.M. ($\bar{x}$) will be given by,

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \ldots + f_nx_n}{f_1 + f_2 + f_3 + \ldots + f_n}$$

$$= \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

$$= \frac{\sum fx}{\sum f} = \frac{\sum fx}{N}$$

where, $N = f_1 + f_2 + f_3 + \ldots + f_n$

- **Arithmetic Mean (A.M.) of grouped frequency distributions:**

In order to determine the A.M. of a grouped frequency distribution an assumption is made that the observations included in a class represented by a class interval are concentrated around the centre of that class interval. For example, if in a distribution of marks of some students the frequency of the class interval 50-60 is 8 (say) then we assume that each of 8 students is getting marks around 55. Consequently we may approximately take that each of the 8 students is getting 55 marks which is the mid value of the class interval 50-60 and we say '8 is the frequency of 55'. To obtain the A.M. of a grouped frequency distribution, we take the frequencies of the different class intervals to be the frequencies of the mid-values of the corresponding classes. This converts a grouped frequency distribution to a discrete (or ungrouped) frequency distribution. Here by applying the arithmetic mean formula for a discrete or ungrouped frequency distribution we can find the arithmetic mean of a grouped frequency distribution.

Thus, in case of grouped frequency distribution A.M. ($\bar{x}$) is defined as $\bar{x} = \frac{\sum fx}{N}$

where, $N = f_1 + f_2 + f_3 + ... + f_n = \sum f$ and x is the mid value of a class.

Example 1: Find A.M. of the following frequency distribution:

x	1	2	3	4	5	6	7	8	9
y	7	11	16	17	26	31	11	1	1

Solution: First of all we shall prepare the following frequency table:

x	f	fx
1	7	7
2	11	22
3	16	48
4	17	68
5	26	130
6	31	186
7	11	77
8	1	8
9	1	9
	N = 121	Σfx = 555

A.M. $(\bar{x}) = \frac{\sum fx}{N} = \frac{555}{121}$ = 4.59 (Approx.)

Example 2: Determine mean of the following distribution:

Daily wages (in Rs.):	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of workers:	6	5	8	15	7	6	3

Solution: Since there are three methods of obtaining mean, namely arithmetic mean (A.M.) method, geometric mean (G.M.) method and harmonic mean (H.M.) method, we shall apply A.M. method to find mean of the given distribution. (We shall see subsequently that A.M. is the most popularly used measure of central tendency.)

To find mean by applying arithmetic mean technique we form the following table:

Wages	Mid value ($\bar{x}$) = $\frac{l_1 + l_2}{2}$	No. of workers (f)	fx
0-10	5	6	30
10-20	15	5	75
20-30	25	8	200
30-40	35	15	525
40-50	45	7	315
50-60	55	6	330
60-70	65	3	195
		N = 50	Σfx = 1670

Mean ($\bar{x}$) = $\frac{\sum fx}{N} = \frac{1670}{N} = 33.4$

i.e.; The required mean (i.e. average) wage = **Rs. 33.40**

Note: l_1 and l_2 imply respectively the lower limit and the upper limit of a class interval.

ALTERNATIVE METHODS OF FINDING A.M.:

a) Assumed Mean Method or Short-Cut Method: In this method A.M. $\bar{x}$ is calculated by using the following formula:

$$\bar{x} = A + \frac{\sum fd}{N}$$

where A = Assumed mean of x, d = x – A, N = total frequency. This formula is used when the values of 'd' do not have a common factor.

b) Step-deviation Method: If the values of (d = x – A) in the above formula have one or more common factors then this method is applied. The formula is:

$$\bar{x} = A + \frac{\sum fd'}{N} \times h$$

where $d' = \frac{d}{h}$, h = Highest common factor of 'd'.

In case of grouped frequency distributions with equal class intervals, h is equal to the length of the class intervals.

A note on assumed mean: Any value of a variable x can be considered to be the assumed mean of x. However, the assumed mean should be taken from the central part of the values

Example 3: Determine arithmetic mean of the following distribution:

Height (in cm)	130-134	135-139	140-144	145-149
Frequency	5	15	28	24
Height (in cm)	150-154	15-159	160-164	
Frequency	17	10	1	

Solution: We denote height by the variable x and we take the assumed mean A of x to be 147.

Class Interval	Mid value (x)	Frequency (f)	d = x – A A = 147	$d' = \frac{d}{h}$ h = 5	fd'
130-134	132	5	–15	–3	–15
135-139	137	15	–10	–2	–30
140-144	142	28	–5	–1	–28
145-149	147	24	0	0	0
150-154	152	17	5	1	17
155-159	157	10	10	2	20
160-164	162	1	15	3	3
		N = 100			$\Sigma fd' = -33$

Now, A.M. $(\bar{x}) = A + \frac{\sum fd'}{N} \times h$

$= 147 + \frac{-33}{100} \times 5$

$= 147 + \frac{-33}{20}$

$= 147 - 1.65$

$= 145.35$

i.e. the required A.M. = **145.35 cm**

PROPERTIES OF A.M.:

Property 1: If each value of a variable is increased (decreased) by a constant c, then the A.M. of the new values is increased (decreased) by c. if each value of a variable is multiplied by a constant c then the A.M. of the new values is c times the A.M. of the original values. Again, if each value of a variable is divided by a constant c ($c \neq 0$) then the A.M. of the new values is equal to the A.M. of the original values divided by c.

Note that if $u = \frac{x \pm a}{b}$, then $\bar{u} = \frac{\bar{x} \pm a}{b}$ where a and b are constants.

Property 2: The sum of the deviations measured from the mean is zero.

In case of individual series $\Sigma(x - \bar{x}) = 0$, where $\bar{x} = \Sigma x/n$ and in case of frequency distributions $\Sigma f(x - \bar{x}) = 0$,

where $\bar{x} = \Sigma fx/N$, $N = \Sigma f$.

Property 3: If the A.M. of a distribution having n_1 value is $\bar{x}_1$ and the A.M. of another distribution having n_2 values is $\bar{x}_2$ then the A.M. $\bar{x}$ of the combined distribution is:

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

This property can be extended for any number of distributions.

Example 4 (a): If $\bar{u} = \frac{x - 55}{10}$ and $\bar{x} = 59$, what is the value of $\bar{u}$.

b) The A.M. of the values of a variable x is 25.

i) If each value is increased by 5, what will be the new A.M.?

ii) If each value is decreased by 7, what will be the new A.M.?

iii) If each value is multiplied by 2.5 what will be the new A.M.?

iv) If each value is divided by 4, what will be the new A.M.?

c) Two series with 28 and 36 observations have means 2.9 and 5.6 respectively. Find the mean of the combined series.

Solution (a) : $u = \frac{x - 55}{10}$

$$\Rightarrow \bar{u} = \frac{\bar{x} - 55}{10}$$

When $\bar{x} = 59$, $\bar{u} = \frac{59 - 55}{10} = \frac{4}{10} = 0.4$

Solution (b) :

i) The A.M. of x i.e. $\bar{x} = 25$, Let $u = x + 5$.

(u is the variable each of whose values is more than the corresponding value of x by 5.)

$\therefore \bar{u} = \bar{x} + 5 = 25 + 5 = 30$

The new A.M. = 30

ii) Let u be the variable each of whose values is less than the corresponding value of x by 7.

$\therefore u = x - 7$

$\Rightarrow \bar{u} = \bar{x} - 7 = 25 - 7 = 18$

i.e. the new A.M. = 18

iii) If u is the variable each of whose values is 2.5 times the corresponding value x then $u = 2.5x$

$\Rightarrow \bar{u} = 2.5\ \bar{x} = 2.5 \times 2.5 = 6.25$

Thus the new A.M. = 6.25

iv) Let $u = \frac{x}{4}$

$$\Rightarrow \bar{u} = \frac{\bar{x}}{4} = \frac{25}{4} = 6.25$$

Thus the new A.M. = 6.25

Solution (c): Let $\bar{x}$ be the mean of the combined series.

Then $\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

where $n_1 = 28$, $n_2 = 36$, $\bar{x}_1 = 2.9$, $\bar{x}_2 = 5.6$.

Now, $\bar{x} = \frac{28 \times 2.9 + 36 \times 5.6}{28 + 36}$

$= \frac{81.2 + 201.6}{64}$

$= \frac{282.8}{64}$

$\therefore$ A.M = 4.42

Determination of Arithmetic Mean in case of Cumulative Frequency Distributions: We shall illustrate with the following example how arithmetic mean of cumulative frequency distributions can be determined.

Example 5: The following are the marks obtained by the students of class XII of a certain Higher Secondary School. Find the average marks using arithmetic mean technique.

Marks	No. of students
Less than 10	5
Less than 20	17
Less than 30	31
Less than 40	41
Less than 50	49

Solution: The above cumulative frequency distribution should first be converted into a simple frequency distribution as under:

We convert the given cumulative frequency distribution to the simple frequency distribution as follows:

Marks	No. of students
0 - 10	5
10 - 20	17 - 5 = 12
20 - 30	31 - 17 = 14
30 - 40	41 - 31 = 10
40 - 50	59 - 41 = 8

Now arithmetic mean of the data can be obtained by using the Direct Method as under:

Calculation of A.M.:

Marks	Mid value (x) = $\frac{l_1 + l_2}{2}$	No. of workers (f)	fx
0 - 10	5	5	25
10 - 20	15	12	180
20 - 30	25	14	350
30 - 40	35	10	350
40 - 50	45	8	360
		N = 49	fx = 1265

$$\text{A.M.} = \left(\bar{x}\right) = \frac{\sum fx}{N} = \frac{1265}{49} = 25.81$$

Note: Student will solve the problem by using step-deviation method.

Example 6: Calculate Arithmatic Mean from the following data using short cut method.

Marks (out of 50):	1–10	11–20	21–30	31–40	41–50
No. of students:	5	7	10	6	2

Solution: Calculation of Arithmatic Mean

Here A = 25.5 (Assumed mean)

Marks	f	Mid-value x	fx	d = x – A	fd
1-10	5	5.5	27.5	–20	–100
11-20	7	15.5	108.5	–10	–70
21-30	10	25.5	255.0	0	0
31-40	6	35.5	213.0	+10	60
41-50	2	45.5	91.0	+ 20	40
Total	N = 30		Σfx = 695.0		Σfd = –70

By short cut method

$$\bar{x} = A + \frac{\sum fd}{N}$$

$$= 25.5 + \frac{-70}{30}$$

$$= 25.5 - 2.33$$

$$= 23.17 \text{ marks.}$$

Example 7: Calculate mean from the following data.

Value	frequency
Less than 10	4
Less than 20	10
Less than 30	15
Less than 40	25
Less than 50	30
Less than 60	35
Less than 70	45
Less than 80	65

Solution: In the problem cumulative frequencies and classes are given. We will first convert the data in simple series from the given cumulative frequencies. After this, we will calculate the mean.

Here Assumed mean, A = 35.

Calculation of Mean

Value	Individual frequency	Mid-value (xi)	$d' = \frac{x_i - 35}{10}$	fd'
0–10	4	5	–3	–12
10–20	10 – 4 = 6	15	–2	–12
20–30	15 – 10 = 5	25	–1	–5
30–40	25 – 15 = 10	35	0	0
40–50	30 – 25 = 5	45	1	5
50–60	35 – 30 = 5	55	2	10
60–70	45 – 35 = 10	65	3	30
70–80	65 – 45 = 20	75	4	80
	N = 65			$\Sigma fd' = 96$

$$\therefore \bar{x} = a + \frac{\sum fd'}{N} \times H$$

$$\therefore A = 35, \Sigma fd' = 96, N = 65, h = 10$$

$$\therefore \bar{x} = 35 + \frac{96}{65} \times 10 = 35 + 14.77 = 49.77$$



LET US KNOW

Arithmetic mean is the most popularly used measure of averages. Most people are familiar with this technique. The basic advantages of this measure over other measures of central tendency are: (i) It is comparatively easier to understand and calculate; (ii) It provides a good basis for comparison; (iii) It can be determined for all values namely positive, negative or zero'; (iv) It is amenable for further mathematical treatment. Mainly because of its easiness in understanding and calculating, it has become most popular and hence it is called an ideal average. Usually mean or average means arithmetic mean. It is widely used in social, economic and business problems.

Weighted Arithmetic Mean : When we determine the arithmetic mean of series by assigning weights to the different quantities of the series depending upon their relative importance then this arithmetic mean will be called the weighted arithmetic mean. For example, if we want to know the change in the cost of living of a particular community over a period of time then we must assign appropriate weights to the quantities of the different items of consumption of that community while constructing the index. Since all the commodities consumed by them are not of equal importance hence simple mean of the prices of the commodities consumed by them will not reflect the true cost.

Let the weights attached to the quantities $x_1, x_2, ..., x_n$ be $w_1, w_2, ..., w_n$ respectively. The weighted mean of these quantities is denoted by $\bar{x}_w$ and is given by:

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + \ldots + w_nx_n}{w_1 + w_2 + \ldots w_n} = \frac{\sum wx}{\sum w}$$

Note: (i) This formula is similar to the formula for obtaining A.M. of a frequency distribution. Here instead of the frequencies f_1, f_2, ..., ..., ..., f_n of x_1, x_2, ..., ..., ..., x_n respectively we take their respective weights w_1, w_2, ..., ..., ..., w_n.

USES OF WEIGHTED MEAN:

i) In order to determine the mean of the quantities whose weights are not equal the formula for weighted mean is applied.

ii) In order to determine the mean of the sub-series of a series, the formula for weighted mean is used.

Example 6: The marks obtained by three students A, B and C in physics, Chemistry and Mathematics out of 100 in each subject in a certain entrance test are:

	Physics	Chemistry	Mathematics
A	60	65	70
B	75	60	50
C	55	60	65

i) Rank the three students on the basis of their performance if equal weights are given to the subjects.

ii) Rank the students if weights are given as below:
Physics: 30%, Chemistry: 20%, Mathematics: 50%

Solution: Let x, y and z denote the marks obtained by A, B, and C respectively is Physics, Chemistry and Mathematics.

Let $\bar{x}(\bar{x}_w)$, $\bar{y}(\bar{y}_w)$ and $\bar{z}(\bar{z}_w)$ denote the simple (weighted) average of marks obtained by A, B and C respectively.

[Recall that when equal weights are given to all observations then it is the case of simple average or average and when different weights are assigned, then it is the case of weighted average. Again, average is usually obtained by arithmetic mean technique.]

i) $\bar{x} = \frac{60+65+70}{3} = \frac{195}{3} = 65$

ii) $\bar{y} = \frac{75+60+50}{3} = \frac{185}{3} = 61.67$

iii) $\bar{z} = \frac{55+60+65}{3} = \frac{180}{3} = 60$

From the average marks obtained by A, B and C we find that the ranking positions of A, B and C are 1st, 2nd and 3rd respectively. Again,

$$\bar{x}_w = \frac{60\times 30\% + 65\times 20\% + 70\times 50\%}{30\% + 20\% + 50\%}$$

$$= \frac{60\times 0.3 + 65\times 0.2 + 70\times 0.5}{0.3+0.2+0.5} \left[\Theta\ 30\% = \frac{30}{100} = 0.3 \text{etc.}\right]$$

$$= \frac{18+13+35}{1} = 66$$

$$\bar{y}_w = \frac{75\times 0.3 + 60\times 0.2 + 50\times 0.5}{0.3+0.2+0.5}$$

$$= \frac{22.5+12+25}{1} = 59.5$$

$$\bar{z}_w = \frac{55\times 0.3 + 60\times 0.2 + 65\times 0.5}{0.3+0.2+0.5}$$

$$= \frac{16.5+12+32.5}{1} = 61$$

We find from above that when the weights are assigned as given to the marks obtained in Physics, Chemistry and Mathematics, the ranking positions of A, B and C becomes 1st, 2nd and 3rd respectively.

b) MEDIAN: The median of a series or distribution in ascending or descending order is that observation of the distribution which divides the distribution into two equal parts. Thus there are equal number of observations on the right and on the left of the median value.

In order to determine the median of an individual series, first of all we have to observe whether the values (observations) are

in a define order or not i.e. whether the values are in ascending or in descending order or not. If the values are not in a define order, then these values are to be arranged either in ascending or in descending order. If there are odd number of values in the series, then the $\left(\frac{n+1}{2}\right)$th value from the beginning (and also from the end) will be the median. If the number of values is even then the arithmetic mean of the $\frac{n}{2}$th value and the $\left(\frac{n}{2}+1\right)$th value will be the median.

Example 7: Determine median for the following series:

i) 77, 73, 72, 70, 75, 79, 78

ii) 94, 33, 86, 68, 32, 80, 48, 70

Solution:

i) Arranging the values of the series in ascending order, we get: 70, 72, 73, 75, 78, 79

No. of terms in the series = 7 = An odd number

The required median = $\frac{7+1}{2}$th term i.e; 4th term = 75

ii) Arranging the data (values or observations) in ascending order, we get: 32, 33, 48, 68, 70, 80, 86, 94

No. of term in the series = 8 = An even number

Now, $\frac{n}{2}$th term = $\frac{8}{2}$th term = 4th term and $\left(\frac{n}{2}+1\right)$th term = 5th term

$\therefore$ The required median = $\frac{68+70}{2}$ = 69.

MEDIAN OF AN UNGROUPED FREQUENCY DISTRIBUTION:

In order to determine the median of an ungrouped frequency distribution which is in ascending order we have to, first of all, form a cumulative frequency table. If the number of observations is odd then the $\frac{N=1}{2}$th (N being the total frequency) term

(observation) will be the median. If we find that $\frac{N=1}{2}$ is greater then x but less than or equal to y where x and y are two consecutive values in the cumulative frequency column, then the observation whose cumulative frequency is y will be the median. Again, if the number of observations is even (i.e., if the total frequency is even) then as in case of individual series we are to determine the $\frac{n}{2}$th and the $\left(\frac{n}{2}+1\right)$th terms (observations). The A.M. of these two terms will be the median. Very often these two terms are the same and consequently when these two terms are equal, the $\frac{n}{2}$th term can be taken as the median.

Example 8: Determine median for the following distribution:

Wages (Rs.):	20	21	22	23	24	25	26	27	28
No. of workers:	8	10	11	16	20	25	19	9	6

Solution:

Wages (Rs.)	No. of workers (f)	Cumulative frequency (f)
20	8	8
21	10	18
22	11	29
23	16	45
24	20	65
25	25	90
26	19	109
27	9	118
28	6	124
	N = 124	

Here total frequency (i.e. total no. of observations) = 124 which is even. Hence A.M. of the $\frac{n}{2}$th and the $\left(\frac{n}{2}+1\right)$th terms will be the median.

Now $\frac{n}{2} = \frac{124}{2} = 62$ and $\frac{N=1}{2} = 63$. We find from the cumulative frequency column that 62 and 63 lie between 45 and 65. Since 65 is the cumulative frequency of 24 hence each of the 62th and the 63th terms will be 24.

Hence the required median = Rs. 24.

Note: The 62th term = Rs. 24, and the 63th term = Rs. 24 and their A.M.= $\frac{Rs.24 + Rs.24}{2} = \frac{Rs.48}{2}$ = Rs. 24.

MEDIAN OF A GROUPED FREQUENCY DISTRIBUTION: In case of grouped frequency distributions, one may consider the $\frac{n}{2}$th observation as the median if N is even. When N is odd, $\left(\frac{n}{2}+1\right)$th observation will be the median.

Median (Me) = $L + \frac{\frac{N}{2} - f_c}{f} \times 1$

Where L = Lower class limit (lower class boundary) in case of exclusive (inclusive) classification.

f = Frequency i.e. simple frequency of the median class

f_c = Cumulative frequency of the class preceeding the median class.

N = Total frequency

I = length of the median class

Note: This formula for obtaining median of a grouped frequency distribution holds only when the distribution is in ascending order. If the distribution is in descending order then it is to be first of all arranged in ascending order in order to apply the above formula.

Example 9: Determine median for the following distribution:

Weekly wages (Rs.)	50-55	55-60	60-65	65-70	70-75
No. of workers	6	10	22	30	16
Weekly wages (Rs.)	75-80	80-85			
No. of workers	12	15			

Solution:

Weekly wages (Rs.)	No. of workers (f)	Cumulative frequency (fc)
50-55	6	6
55-60	10	16
60-65	22	38
65-70	30	68
70-75	16	84
75-80	12	96
80-85	15	111
	N = 111	

Here N = 111, which is odd.

Now, $\frac{N=1}{2}$ th term = $\frac{111+1}{2}$ th term = 56th term. From the cumulative frequency table we find that the 56th term lies in the class 65 - 70. 65 - 70 is the median class.

Now, median = $L+\frac{\frac{N}{2}-fc}{f}\times 1$

Here L = 65, f = 30, fc = 38, N = 111, I = 5

Median = $65+\frac{\frac{111}{2}-38}{30}\times 5$

$= 65+\frac{55.5-38}{30}\times 5$

$= 65+\frac{17.5}{6}$

= 65 + 2.92

= 67.92

i.e. the required median = Rs. 67.92

Advantages and Limitations of Median:

- **Advantages:**
  - i) Extreme values do not affect median.
  - ii) Median is easy to understand; it is easy also to determine.
  - iii) Median can also be determined graphically

iv) Median can be determined for distributions having open end class intervals.

➢ **Limitations:**

i) In order to determine the median of a distribution, the distribution must be arranged in a define order if it is not in an order. This is not needed in other measures of central tendency.

ii) Median of a distribution is not based on all the observations of the distribution.

Uses of median: In order to determine the average of distribution having open-end class intervals median is the best measure. In case of income distribution the use of median gives better results.

c) MODE: The mode of a distribution is that observation of the distribution whose frequency is the maximum. It is to be noted that mode is not unique which means that a distribution may have more than one modes.

Clearly an individual or single series (distribution) does not have a mode. In many cases of ungrouped frequency distributions, mode can be detected simply by observation. Mode of a grouped frequency distribution is obtained by using the following formula.

$$\text{Mode } (M_o) = L + \frac{f_1 = f_0}{2f_1 - f_0 - f_2} \times I$$

Where L = Lower limit/lower boundary of the model class

f_o = Frequency of the class preceeding the model class

f_1 = Frequency of the model class

f_2 = Frequency of the class succeeding the model class

And I = Length of the model class

➢ **Note:** i) The class (specified by a class interval) whose frequency is the maximum is called the model class.

ii) The above formula is used when all the classes are of equal length.

Example 10: Determine mode for the following distribution:

Marks	1-5	6-10	11-15	16-20	21-25	26-30
No. of students	7	10	16	32	24	18
Marks	31-35	36-40	41-45			
No. of students	10	5	1			

Solution: Since the frequency of the class 16-20 is the maximum, hence this class is the model class. The class intervals of the given distribution are as per the inclusive method of classification and hence in determining mode we must take the lower boundary of the model class.

Now, Mode $= L + \frac{f_1 = f_0}{2f_1 - f_0 - f_2} \times I$

$= 15.5 + \frac{32 - 16}{2 \times 32 - 16 - 24} \times 5$

$= 15.5 + \frac{16}{64 - 40} \times 5$

$= 15.5 + \frac{16}{24} \times 5$

= 15.5 + 3.33 = **18.33 marks**

Example 11: Determine mode/modes, if any, of the following series:

i) 3, 4, 5, 2, 3, 4, 1, 6, 4;

ii) 7, 9, 11, 7, 6,5, 9, 13;

iii) 3, 5, 6, 7, 9, 12, 3, 6, 5, 9, 12, 7

Solution:

i) The number 4 is repeated the maximum of number times (thrice). Hence 4 is the mode of the given series.

ii) We find that the frequency of 7 and 9 is the maximum each being equal to 2. Hence the two modes of the series are 7 and 9.

iii) Since the frequency of each observation is the same (being 2 in each case), hence the given series has no mode.

Advantages and Limitations of Mode:

- **Advantages:**
  - i) In most cases the mode/modes of an ungrouped frequency distribution can be determined simply by observation.
  - ii) Mode is not affected by extreme values
  - iii) Mode is easy to understand
  - iv) Mode can be determined graphically.
- **Limitations:**
  - i) Mode is not based on all the observations
  - ii) It is not suitable for further mathematical treatment.
  - iii) Like arithmetic mean we cannot know the sum of the observations of a distribution if we know the mode and the number of observations of the distribution.

Uses of Mode: The concept of mode is used by manufacturers, businessmen, agriculturists, etc. For example, a manufacture of shoes is interested in the model size of shoes and manufactures them in large quantities. It is useful in industry and business. Weather forecasts are based on mode. The concept of mode is used in socio-economic surveys besides being used in business and commerce.

RELATIONSHIP AMONG MEAN, MEDIAN AND MODE: Prof. Kari Pearson has established an empirical (experimental) relationship among mean, median and mode. For any distribution the following relationship known as empirical relationship approximately holds:

Mean – Mode = 3 (Mean – Median)



CHECK YOUR PROGRESS

Q.1: What do you mean by measures of central tendency or averages?

Q.2: Define weighted arithmetic mean.

Q.3: Find mean and median from the following distribution:

Age (year):	20	19	18	17	16	15	14	13	12	11
No. of students:	1	2	4	8	11	10	7	4	2	1

Q.4: Determine mean, median and mode from the following data:

Weekly wages (Rs.):	15	16	17	18	19	20
No. of workers:	6	12	23	30	9	1

Q.5: Determine mean, median and mode from the following statistical distribution.

Class interval:	6.5-7.5	7.5-8.5	8.5-9.5	9.5-10.5
Frequency:	5	12	25	48
Class interval:	10.5-11.5	11.5-12.5	12.5-13.5	
Frequency:	32	6	1	

Q.6: The weekly expenditures of a few families are given below. Calculate arithmetic mean and median.

Weekly expenditures (in Rs.):	110-120	120-130	130-140	140-150	150-160
No. of families:	6	15	38	62	106
Weekly expenditures (in Rs.):	160-170	170-180	180-190	190-200	
No. of families:	50	18	12	3	

Q.7: Calculate arithmetic mean and median of the distribution given below. Calculate mode using the empirical relation among the three.

Class limit:	130-134	135-139	140-144	145-149
Frequency:	7	12	15	6



1.4 LET US SUM UP

In this unit we have learnt the meaning of measure of central tendency (mean, media and mode). Then we discussed some properties of arithmetic mean. Here we also discussed the relationship between mean, medium and mode.



1.5 FURTHER READING

1) Agarwal, D. R. (2006). *Business Statistics*. Delhi: Vrinda Publications.
2) Gupta, S. C. (1994). *Fundamentals of Statistics*. New Delhi: Himalayan Publishing House.
3) Rajagopalan, S. P. & Sattanthan, R. (2009). *Business Statistics and Operations Research*. New Delhi: Tata McGraw-Hill.
4) Sharma, J. K. (2007). *Business Statistics*. New Delhi: Pearson Education Ltd.
5) Verma, A. P. (2007). *Business Statistics*. Guwahat: Asian Books Private Limited.



1.6 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: A measure of central tendency or an average of a certain distribution is nothing but a representative value of that distribution which enables us to comprehend in a single effort the significance of the whole.

Ans. to Q. No. 2: When we determine the arithmetic mean of series by assigning weights to the different quantities of the series depending upon their relative importance then this arithmetic mean will be called the weighted arithmetic mean.

Ans. to Q. No. 3: Mean (A.M.) = 15.54 years; Median = 16 years

Ans. to Q. No. 4: Men (A.M.) = Rs. 17.33; Median = Rs. 17; Mode = Rs. 18

Ans. to Q. No. 5: Mean (A.M.) = 9.98; Median = 9.97; Mode = 10.17

Ans. to Q. No. 6: Mean (A.M.) = Rs. 152.64; Median = Rs. 153.67

Ans. to Q. No. 7: 139.5, 139.63, 139.89



1.7 MODEL QUESTIONS

Q.1: Define median of a distribution. Discuss its advantages and limitations.

Q.2: Define mode. What are its advantages and limitations?

Q.3: Write down the empirical relationship among mean, median and mode.

Q.4: What are the properties of A.M?

*** ***** ***

UNIT 2: MEASURES OF DISPERSION

UNIT STRUCTURE

2.1 Learning Objectives

2.2 Introduction

2.3 Dispersion

- 2.3.1 Objectives of Studying Dispersion
- 2.3.2 Essentials of a Good Measure of Dispersion
- 2.3.3 Different Measures of Dispersion

2.4 Let Us Sum Up

2.5 Further Reading

2.6 Answer To Check Your Progress

2.7 Model Questions

2.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- explain the meaning of dispersion or variation
- describe the various methods of obtaining dispersion for various types of distribution
- describe the concept of skewness and types of skewness
- describe the methods of measuring skewness.

2.2 INTRODUCTION

In this unit, we are going to discuss about dispersion or variation. In the previous unit we discussed about one single figure that represents the entire data. Here we shall discuss about the degree to which numerical data tend to spread about an average value. Dispersion measures the extent to which the items vary from central value. Thus through this unit you will also able to explain the concept of skewness and types of skewness.

2.3 DISPERSION

Scatteredness of data around an average is termed as dispersion or variation. To quite Spiegel, "The degree to which numerical data tend to spread about an average value is called variation or dispersion".

2.3.1 Objectives of Studying Dispersion

The objectives of studying Dispersion are as follows:

i) To study the reliability of average.
ii) To control the variation.
iii) To make the comparison among series of observations with regards to variations.
iv) To make further statistical analysis.

2.3.2 Essentials of a Good Measure of Dispersion

i) A measure of dispersion should be based on all the observations of a series.
ii) It should be rigidly defined.
iii) It should be easy to understand and calculate.
iv) It should not extremely affected by extreme values.
v) It should not be affected much by fluctuations of sampling.
vi) It should be usable for further statistical analysis.

2.3.3 Different Measures of Dispersion

The following are the various measures of dispersion: (a) Range (b) Interquartile Range and Quartile Deviation (c) Mean Deviation (d) Standard Deviation and Variance (e) Coefficient of variation. While measures (a), (b), (c) and (d) are called absoluate measures, measure (e) is called a relative measure of dispersion. An absolute measure is that which possesses the same unit of the distribution for which the measure is obtained. On the other hand a relative measure is a ratio or percentage and as such it is a pure number. It is to be mentioned that all measures of central tendency are absolute measures.

a) Range: The range of a distribution is the difference between the largest and the smallest observations of the distribution. Thus if L denotes the largest observation and S denotes the smallest observation of a distribution then the range R of the distribution will be: R = L – S

Example 1: If the marks obtained by six students are 24, 12, 16, 11, 40 and 42, find the range of these marks.

Solution: Here L = 42, S = 11

Range R = L – S = 42 – 11 = 31 (marks)

Example 2: Determine range of the following distribution:

Weights (in kg):	40	47	56	62	70
No. of students:	4	7	11	3	1

Solution: Here L = 70kg, S = 40kg

Range R = L – S = (70 – 40)kg = 30kg

Example 3. The following distribution is a distribution of height. Determine the range of the distribution:

Height (in cm.):	120-129	130-139	140-149	150-159
No. of students:	10	17	23	8

Solution: Range R = (Upper limit of 150-159) – (Lower limit of 120-129) = (159-120) cm = 39 cm.

Note: In case of a grouped frequency distribution with inclusive classification of data like the above distribution, we may also obtain range as below:

R = (Upper class boundary of 150-159) – (Lower class boundary of 120-129) = (159.5 - 119.5) cm = 40cm.

b) Interquartile Range and Quartile Deviation (Q.D.): The interquartile range of a distribution is the difference between the third quartile Q_3 and the first quartile Q_1 of the distribution. Again, half the interquartile range of a distribution is called the quartile deviation (Q.D.) of the distribution. Thus,

Interquartile range = $Q_3 - Q_1$, and Q.D. = $\frac{Q_3 - Q_1}{2}$

Formulas to Calculate Interquartile Range and Quartile Deviation:

For Individual Servies:

Inrerquartile Range = $Q_3 - Q_1$

Q.D. = $\frac{Q_3 - Q_1}{2}$

Where Q_1 = size of $\frac{n+1^{th}}{4}$ item

n = number of observations

Q_3 = size of $\left(\frac{3(n+1)}{4}\right)^{th}$ item

n = number of observations

For Discrete Series:

Interquartile Range = $Q_3 - Q_1$

Q.D. = $\frac{Q_3 - Q_1}{2}$

Where Q_1 = Locate the size of $\frac{N+1^{th}}{4}$ item in the cumulative frequency column and corresponding x value is Q_1.

N = Sum of frequencies.

Q_3 = Locate the size of $\left(\frac{3(N+1)}{4}\right)^{th}$ item in the cumulative frequency column and corresponding x value is Q_3.

N = Sum of frequencies.

For Continuous Series:

Interquartile Rang = $Q_3 - Q_1$

Q.D. = $\frac{Q_3 - Q_1}{2}$

Here we first locate the size of $\frac{N^{th}}{4}$ item in c.f column and the value of Q_1 will lie in the corresponding class interval. (i.e. Q_1 class)

Where $Q_1 = l_1 + \frac{\frac{N}{4} - c.f.}{f} \times i$

l_1 = lower limit of class interval.

N = Sum of frequencies.

c.f. = Cumulative frequencies of the class preceding the Q_1 class.

i = Class interval.

First, we locate the size of $\frac{3N}{4}^{th}$ item in c.f column and the value of Q_3 will lie in the corresponding class interval.

$$Q_3 = l_1 + \frac{3\left(\frac{N}{4}\right) - c.f.}{f} \times i$$

Where l_1 = lower limit of class interval.

N = Sum of frequencies.

c.f. = Cumulative frequencies of the class preceding the Q_3 class.

i = Class interval.

Example 1: Calculate Range and Q.D. of the following observation:

20 25 29 30 35 39 41 48 51 60 and 70

Solution: Range is 70 – 20 = 50

For Q. D., we need to calculate values of Q_3 and Q_1.

Q_1 is the size of $\frac{N+1}{4}^{th}$ value

Here N being 11. Q_1 is the size of 3rd value.

As the values are already arranged, in ascending order, it can be seen that Q_1, the 3rd value is 29.

Similarly. Q_3 is size of $\frac{3(N+1)}{4}^{th}$ value : i.e. 9th value which is 51

Here Q_3= 51

$$Q.D. = \frac{Q_3 - Q_1}{2} = \frac{51 - 29}{2} = 11$$

Example 2: For the following distribution of marks scored by a class of 40 students. Calculate the range and Q.D.

Class interval	No of students (f)
0–10	5
10–20	8
20–40	16
40–60	7
60–90	4
	40

Solution: Range is just the difference between the upper limit of the highest class and the lower limit of the lowest class. So, Range is 90 – 0 = 90

For Q.D., first calculate cumulative frequencies as follows:

(Class Interval) (c.i.)	(Frequency) (f.)	(Cumulative Frequency) (c.f.)
0–10	5	05
10–20	8	13
20–40	16	29
40–60	7	36
60–90	4	40
	N = 40	

Q_1 is the size of $\frac{N}{4}^{th}$ value in a continuous series. Thus, it is the size of the 10^{th} value. The class containing the 10^{th} value is 10-20. Here Q_1 lies in class 10-20. Now, to calculate the exact value of Q_1, the following formula is used:

$$Q_1 = L + \frac{\frac{N}{4} - c.f.}{f} \times i$$

Where L = 10 (Lower limit of Quartile class)

c.f. = 5 (value of c.f. for the class preceding the Quartile class)

i = 10 (interval of the Quartile class) and

f = 8 (frequency of the Quartile class)

Thus, $Q_1 = 10 + \frac{10-5}{8} \times 10 = 16.25$

Similarly, Q_3 is the size of $\frac{3N^{th}}{4}$ value i.e., 30^{th} value which lies in class 40 – 60. Now using the formula for Q_3 its value can be calculated as follows:

$$Q_1 = L + \frac{\frac{3N}{4} - c.f.}{f} \times i$$

$$= 40 + \frac{30-29}{7} \times 20$$

$$= 42.87$$

$$\therefore Q.D. = \frac{42.87 - 16.25}{2}$$

$$= 13.31$$

Example 3: Calculate Quartile deviation from the following.

Age in years	No. of members
20	3
30	61
40	132
50	153
60	140
70	51
80	3

Solution:

Age in years	No. of members	c.f.
20	3	3
30	61	64
40	132	196
50	153	349
60	140	489
70	51	540
80	3	543

Q_1 = Value of $\left(\frac{N+1}{4}\right)^{th}$ item

= Value of $\left(\frac{543+1}{4}\right)^{th}$ item

= Value of $\frac{544}{4}^{th}$ item

= Value of 136^{th} item

= 40 years

Q_3 = Value of $3\left(\frac{N+1}{4}\right)^{th}$ item

= Value of $3\left(\frac{543+1}{4}\right)^{th}$ item

= Value of 3×136^{th} item

= 408^{th} item which is 60 years

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

$$= \frac{60 - 20}{2}$$

$$= \frac{20}{2}$$

= 10 years

Example 4: Calculate the range and Quartile deviation

Wages (Rs.)	Labouress
30–32	12
32–34	18
34–36	16
36–38	14
38–40	12
40–42	8
42–44	6

Solution: Range = L – S

= 44 – 30 = 14 years

x	f	c.f.
30-32	12	12
32-34	18	30
34–36	16	46
36–38	14	60
38–40	12	72
40–42	8	80
42–44	6	86

Q_1 = Size of $\frac{N}{4}^{th}$ item

$= \frac{86}{4} = 21.5$

$\therefore Q_1$, lies in the group 32 – 34

$$Q_3 = L + \frac{\frac{3N}{4} - c.f.}{f} \times i$$

$= 32 + \frac{21.5 - 12}{18} \times 2$

$= 32 + \frac{19}{18}$

$= 32 + 1.06 = 33.06$

Q_3 = Size of $\frac{3N}{4}^{th}$ item

$= \frac{3 \times 86}{4}$

$= 64.5^{th}$ item

Q_3, lies in the group 38 – 40

$$Q_3 = L + \frac{\frac{3N}{4} - c.f.}{f} \times i$$

$= 38 + \frac{64.5 - 60}{12} \times 2$

$= 38 + 0.75$

$= 38.75$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{38.75 - 33.06}{2}$$

$$= \frac{5.69}{2} = 2.85$$

Note: When a distribution is in increasing order then the three quantities which divide the distribution into four equal parts are called the quartiles of the distribution. In increasing order these quartiles are denoted by Q_1, Q_2 and Q_3. Q_1 is called the first or lower quartile Q_2 is called the second or middle quartile and Q_3 is called the third or upper quartile. The second quartile Q_2 and the median (Me) of a distribution are the same.

c) Mean Deviation (M.D.): The arithmetic mean of the absolute deviations of the observations of a distribution from its mean, median or mode is known as mean deviation.

If a variable x takes n values then

$$\text{M.D.} = \frac{|x_1 - A| + |x_2 - A| + ... + |x_n - A|}{n}$$

$$\text{or} \quad \text{M.D.} = \frac{\sum |x - A|}{n} = \frac{\sum |d|}{n}$$

Here A = mean, median or mode of x and d = x – A

Again, if the frequencies of $x_1, x_2, ..., ..., ..., x_n$ are $f_1, f_2, ..., ..., ..., f_n$ respectively then

$$\text{M.D.} = \frac{f_1|x_1 - A| + f_2|x_2 - A| + ... + f_n|x_n - A|}{N}, \quad N = \Sigma f$$

$$\text{or} \quad \text{M.D.} = \frac{\sum f|x - A|}{N} = \frac{\sum f|d|}{N}$$

Here A = mean, median or mode of the distribution, d = x – A

If A = mean, then we get mean deviation from mean. Likewise, we get mean deviation from median and mean deviation from mode. Mean deviation from mean ($\bar{x}$) can be denoted by the symbol M.D. ($\bar{x}$). Likewise, mean deviations from median and mode can be denoted by M.D. (Me) and M.D. (Mo) respectively.

Note:

i) In case of mean deviation from mean, 'A' may be A.M., G.M. or H.M. But usually A is taken as A.M.

ii) $|x - A|$ is called the absolute value of the deviation. By absolute value we mean the magnitude of the value without considering the sign. Thus $x - A$ may be positive or negative but is $|x - A|$ always positive. Thus $|5 - 2| = |3| = 3$, $|7 - 9| = |-2| = 2$ etc.

iii) Since the sum of the deviations measured from the mean is zero, hence in case of mean deviation we always take absolute deviations.

Advantages and Limitations of Mean Deviations:

Advantages:

i) It is based on all the observations.

ii) It is less affected by extreme values in comparison to standard deviation.

iii) Since deviations are taken from average (mean, median or mode), therefore mean deviation is considered to be a good measure for comparing the variability among two or more distributions.

Limitations:

i) In mean deviation actual signs of deviations are discarded by taking absolute values of the deviations.

ii) Mean deviation from mode is not considered to be a good measure of dispersion.

iii) One cannot determine mean deviation for a grouped frequency distribution containing open-end class interval.

Example 5: For the following distribution determine mean deviation (M.D.) from mean:

x	:	10	11	12	13	14
f	:	3	12	18	12	3

Solution: First of all, we form the following table.

x	f	fx	$\|d\| = \|x - \bar{x}\|$	f|d|
10	3	30	2	6
11	12	132	1	12
12	18	216	0	0
13	12	156	1	12
14	3	42	2	6
	N = 48	$\Sigma fx = 576$		$\Sigma f\|d\| = 36$

A.M. $(\bar{x}) = \frac{\sum fx}{N} = \frac{576}{48} = 12$

(Usually mean implies arithmetic mean.)

Now, M.D. from mean = $\frac{\sum f|d|}{N} = \frac{36}{48} = 0.75$

d) Standard Deviation (S.D.): The positive square root of the arithmetic mean of the squares of the deviations of the values of a variable from its arithmetic mean is called the standard deviation of that variable. The symbol σ (a Greek letter pronounced as 'sigma') is used to denote standard deviation.

If a variable x takes n values $x_1, x_2, ..., ..., ..., x_n$ and if $\bar{x}$ be the arithmetic mean of these values, then

$$\sigma = \sqrt{\frac{f_1(x_1 - \bar{x})^2 + f_2(x_2 - \bar{x})^2 + ... + f_n(x_n - \bar{x})^2}{n}}$$

i.e. $\sigma = \sqrt{\frac{\sum(x_1 - \bar{x})^2}{n}}$, $\bar{x} = \frac{\sum x}{n}$

Again, in case of frequency distribution:

x	x_1	x_2	.....	x_n
f	f_1	f_2	.....	f_n

Standard deviation will be:

$$\sigma = \sqrt{\frac{f_1(x_1 - \bar{x})^2 + f_2(x_2 - \bar{x})^2 + ... + f_n(x_n - \bar{x})^2}{N}}, \quad N = \Sigma f$$

i.e. $\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^2}{N}}$, $\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N}$

Instead of the symbol σ we may use the symbol σ_x to clearly signify the standard deviation of the variable x.

Alternative methods of obtaining standard deviation:

a) Assumed Mean Method: In case of individual series we have the following formula:

$$\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}, \quad d = x - A$$

And A = Assumed mean of x.

In case of frequency distribution,

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

b) Step deviation method: If the deviations from assumed mean A denoted by d have some common factors then these deviations are divided by the highest common factor h (say) and the step deviations denoted by d' are obtained. Thus $d' = d/h$, where h is the H.C.F. of the deviations d. In this method, we have the following formula:

For individual series, $\sigma = \sqrt{\frac{\sum f'^2}{n} - \left(\frac{\sum f'}{n}\right)^2} \times h$

For frequency distribution $\sigma = \sqrt{\frac{\sum fd'^2}{n} - \left(\frac{\sum fd'}{n}\right)^2} \times h$

$d' = \frac{d}{h} = \frac{x - A}{h}$, A = Assumed mean of the variable x whose values are the various values of the distribution.

Empirical relationship among Q.D., M.D. and S.D.: It can be verified empirically that for a distribution the following relationship among Q.D., M.D. and S.D. approximately holds. However, in case of normal distribution, this relationship is exact.

$$\text{Q.D.} = \frac{5}{6}, \text{M.D.} = \frac{2}{3}\text{S.D.}$$

or, 6Q.D. = 5 M.D. = 4 S.D.

Advantages and Limitations of S.D.: Among all the measures of dispersion standard deviation is considered to be the best measure. Standard deviation possesses almost all the qualities that a good measure of dispersion should have. It is rigidly defined. It is based on all the observations. The sampling stability of this measure is the maximum in comparison to all other measures of dispersion. While mean deviation ignores the algebraic signs of deviations, standard deviation is free from this demerit as in standard deviations squares of the deviations are taken. The formula for standard deviation is amenable for further mathematical treatment. Normal curve can be analysed with the help of standard deviation. Standard deviation plays a significant role in sampling theory and correlation analysis.

The main limitation of standard deviation is that it is difficult to calculate in comparison to other measures of dispersion. Moreover, in comparison to mean deviation it is more affected by extreme values. If we make a comparative study of the advantages and limitations of various measures of dispersion, we find that standard deviation is the best of all the measures of dispersion. Hence it is widely used as a measure of dispersion. It is called an ideal measure of dispersion.

An Important Property of Standard Deviation (S.D.): The S.D. of a series remains unaltered if a constant is added to or subtracted from each value of the series. For example, if the standard deviation of the series $x_1, x_2, ..., ..., ..., x_n$ is σ then the S.D. of the series $x_1 \pm c, x_2 \pm c, ..., ..., ..., x_n \pm c$, will also be where c ($c \neq 0$) is a constant. On the other hand, if each value of a series is multiplied by a constant c then the S.D. of the new series will be equal to c times the S.D. of the original series. Again, if each value of a series is divided by a constant c then the S.D. of the new series will be equal to the S.D. of the original series divided by c.

e) Coefficient of Variation (C.V.): It is a relative measure of dispersion. It is expressed as a percentage and is useful in comparing the variability of two or more sets of data especially when they are expressed in different units of measurement. According to Karl Pearson, ''Coefficient of variation is the percentage variation in mean, standard deviation being considered as the total variation in the mean.'' Thus coefficient of variation abbreviated as C.V. is defined as below:

$$C.V. = \frac{\sigma}{\bar{x}} \times 100\%$$

Although we have several relative measures of dispersion yet the coefficient of variation is the most commonly used relative measure of dispersion. Other relative measures are rarely used in practice. While comparing the variability between distributions, the distribution having minimum C.V. is said to be less variable, more stable, more uniform, more consistent or more homogeneous. On the other hand, the other distribution is said to be more variable, less stable, less uniform, less consistent or less homogeneous.

Note: Suppose the scores of a cricket player A in three matches are 0, 10 and 80. And the scores of another cricket player B in these matches be 28, 30 and 32. Although the mean scores of both the players are same both being 30, we say that B is more consistent then A. Even people without having the knowledge of variability or dispersion say like this. If we calculate the C.V. for both the series of scores we shall find that the C.V. of scores of B is less than the C.V. of scores of A. In fact, the magnitude of any measure of dispersion will be less in case of scores attained by B than the scores attained by A.

Example 6: Find standard deviation of the following observations: 8,10, 12, 14, 16, 18, 20, 22, 24, 26

Solution: We shall solve this problem by Direct Method as well as by Step-Deviation Method.

Direct Method:

Value x	Deviation from the mean $x = \bar{x}$, ($\bar{x} = 17$)	$(x-\bar{x})^2$
8	–9	81
10	–7	49
12	–5	25
14	–3	9
16	–1	1
18	1	1
20	3	9
22	5	25
24	7	49
26	9	81
$\Sigma x = 170$		$\Sigma(x-\bar{x})^2 = 330$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{170}{10} = 17$$

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

$$= \sqrt{\frac{330}{10}}$$

$$= \sqrt{33}$$

$$= 5.74$$

Note: When the mean $\bar{x}$ is an integer the formula $\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$ should be used to find S.D. When $\bar{x}$ is not an integer then the $\sigma = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$ or $\sigma = \sqrt{\frac{\Sigma x^2}{n} - \bar{x}^2}$ formula should be used. Because when $\bar{x}$ is not an integer determination of the values of $(x - \bar{x})^2$ becomes vary time consuming.

Step-Deviation Method: Let the given values be the values of the variable x and let the assumed mean of x be 18.

Calculation of S.D.:

Value	d = x – 18	$d' = \frac{d}{2}$	d'^2
8	–10	–5	25
10	–8	–4	16
12	–6	–3	9
14	–4	–2	4
16	–2	–1	1
18	0	0	0
20	2	1	1
22	4	2	4
24	6	3	9
26	8	4	16
Total		$\Sigma d' = -5$	$\Sigma d'^2 = 85$

$$\Theta\ \sigma = \sqrt{\frac{\sum fd'^2}{n} - \left(\frac{\sum fd'}{n}\right)^2} \times h$$

$$= \sqrt{\frac{85}{10} - \left(\frac{-5}{10}\right)^2} \times 2$$

$$= \sqrt{8.50 - 0.25} \times 2$$

$$= \sqrt{8.25} \times 2$$

$$= 2.87 \times 2$$

$$= 5.74$$

Note: If we use assumed mean method then we will have to apply the following formula:

$$\sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd^2}{n}\right)^2}$$

But if the values of d have a common factor then we shall have to divide the values of d by that common factor and the values obtained thereby are to be assumed to be the values of the variable d'. Then we get the formula for step-deviation method. If the values of d have two or more common factors

then we are to divide the values of d by its H.C.F. and the values obtained thereby will be considered to be the values of d. If the values of d have only one common factor then that will be the highest common factor.

Example 7: Find S.D. from the following record of number of car accidents in a street.

No. of accidents (x):	1	2	4	5	6
No. of days:	2	3	3	1	1

Solution: We shall offer the solution by the Assumed Mean Method.

x	f	d = x – 4	fd	fd^2
1	2	–3	–6	18
2	3	–2	–6	–12
4	3	0	0	0
5	1	1	1	1
6	1	2	2	4
	N = 10		$\sqrt{\sum fd'}$ = –9	$\sqrt{\sum fd'^2}$ = 35

$$\sigma = \sqrt{\frac{\sum fd'}{N} - \left(\frac{\sum fd'}{N}\right)^2}$$

$$= \sqrt{\frac{35}{10} - \left(\frac{-9}{10}\right)^2}$$

$$= \sqrt{3.5 - 0.81}$$

$$= \sqrt{2.69}$$

$$= 1.64$$

Example 8: Calculate mean and standard deviation from the following distribution:

Age (year):	20-25	25-30	30-35	35-40	40-45	45-50
No. of Persons:	170	110	80	45	40	35

Solution:

Age (year)	Mid value (x)	No. of persons (f)	d = x – 32.5	$d' = \frac{d}{2}$	fd'	fd'^2
20-25	22.5	170	–10	–2	–340	680
25-30	22.7	110	–5	–1	–110	110
30-35	32.5	80	0	0	0	0
35-40	37.5	45	5	1	45	45
40-45	42.5	49	10	2	80	160
45-50	47.5	35	15	3	105	315
		N = 480		$\Sigma fd'$ = –220	$\Sigma fd'^2$ =1310	

Here, assumed mean A = 32.5

Now, A.M.$(\bar{x})$ = A + $\frac{\sum fd'}{N} \times h$

= 32.5 + $\frac{-220}{480} \times 5$

= 32.5 – 2.99

= 29.51

The required A.M. = 29.51 years.

S.D. (σ) = $\sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times h$

= $\sqrt{\frac{1310}{480} - \left(\frac{220}{480}\right)^2} \times 5$

= $\sqrt{2.7292 - 0.2101} \times 5$

= $\sqrt{2.5191} \times 5$

= 1.5871 × 5

= 7.936

∴ The required S.D. = 7.936 years.

Example 9: Calculate mean and standard deviation from the following data:

Value (Rs.):	90-99	80-89	70-79	60-69	50-59	40-49	30-39
No. of Persons:	2	12	22	20	14	4	1

Solution:

Determine of A.M. and S.D.:

Class interval	Mid value (x)	Frequency (f)	d = x – 64.5	$d' = \frac{d}{10}$	fd'	fd'^2
90-99	94.5	2	30	3	6	18
80-89	84.5	12	20	2	24	48
70-79	74.5	22	10	1	22	22
60-69	64.5	20	0	0	0	0
50-59	54.5	14	–10	–1	–14	14
40-49	44.5	4	–20	–2	–8	16
30-39	34.5	1	–30	–3	–3	9
	N = 75				$\Sigma fd'$ = 27	$\Sigma fd'^2$ = 127

Now, A.M. $(\bar{x}) = A + \frac{\Sigma fd'}{N} \times h$

$= 64.7 + \frac{27}{75} \times 10$

$= 64.5 + 3.6$

$= 68.10$

$\therefore$ The required A.M. = Rs. 68.10

S.D. $(\sigma) = \sqrt{\frac{\Sigma fd'}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times h$

$= \sqrt{\frac{127}{75} - \left(\frac{27}{75}\right)^2} \times 10$

$= \sqrt{1.6933 - 0.1296} \times 10$

$= \sqrt{1.5637} \times 10$

$= 1.2505 \times 10$

$= 12.505$

$= 12.51$

$\therefore$ The required S.D. = Rs. 12.51

Example 10: Calculate standard deviation and coefficient of variation from the following data:

Age (year):	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of Persons:	3	61	132	153	140	51	2

Solution:

Calculation of S.D. and C.V.:

Age	Mid value (x)	Frequency (f)	d = x–55	$d' = \frac{d}{10}$	fd'	fd'^2
20-30	25	3	–30	–3	–9	27
30-40	35	61	–20	–2	–122	244
40-50	45	132	–10	–1	–132	132
50-60	55	153	0	0	0	0
60-70	65	140	10	1	140	140
70-80	75	51	20	2	102	204
80-90	85	2	30	3	6	18
N = 542					$\Sigma fd' = -15$	$\Sigma fd'^2 = 765$

$$\text{S.D. } (\sigma) = \sqrt{\frac{765}{542} - \left(\frac{-15}{542}\right)^2} \times 10$$

$$= \sqrt{1.4114 - 0.0001} \times 10$$

$$= \sqrt{1.4105} \times 10$$

$$= 11.876$$

i.e. S.D. = **11.876 years.**

$$\text{Again, C.V.} = \frac{\sigma}{\bar{x}} \times 100\%$$

$$\text{Now, } \bar{x} = A + \frac{\Sigma fd'}{N} \times h$$

$$= 55 + \frac{-15}{542} \times 10$$

$$= 55 - 0.277 = 54.723$$

i.e. A.M. = $\bar{x}$ = 54.723 years

$$\therefore \text{ C.V.} = \frac{11.876}{54.723} \times 100\% = \textbf{21.76\%}$$



CHECK YOUR PROGRESS

Q.1: Determine standard deviation and coefficient of variation for the following distribution:

Age (year):	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of Persons:	6	42	86	122	78	30	5

Q.2: Below are given yearly profits of a few small companies. Determine arithmetic mean and standard deviation:

Profit (in thousand Rs.):	20-30	30-40	40-50	50-60	60-70
No. of Companies:	30	58	62	85	112
Profit (in thousand Rs.):	70-80	80-90	90-100		
No. of Companies:	112	57	26		

Q.3: The following distribution is the distribution of ages of 100 students. Find the standard deviation of the distribution.

Age (in years):	16-17	17-18	18-19	19-20	20-21
No. of Students:	4	14	18	28	20
Age (in years):	21-22	22-23			
No. of Students:	12	4			

Q.4: i) What is the S.D. of the following series:

5, 5, 5, 5, 5, 5, 5

ii) Suppose the S.D. of the series a_1, a_2, ..., ..., ..., a_n is σ. What is the S.D. of the series $a_1 + 5$, $a_2 + 5$, ..., ..., ..., $a_n + 5$?

iii) What is the relationship between the S.D.s of the following two series?

10, 20, 30, 70, 100, 110

And 1, 2, 3, 7, 10, 11

Q.5: Chose the correct answer:

i) Which of the following is a unitless measure of dispersion?

a) Standard derivation b) Mean deviation

c) Coefficient of variation d) Range

ii) The relationship between mean deviation and standard deviation is:

a) 3 M.D. = 2 S.D. b) 6 M.D. = 5 S.D.

c) 5 M.D. = 4 S.D. d) M.D. = S.D.

iii) If the minimum value in a series is 20 and its range is 47, the maximum value of the series is:

a) 67 b) 57

c) 48 d) None of the above.



2.4 LET US SUM UP

- In this unit we have learnt the meaning of measure of dispersion.
- Then we have discussed different measures of range, interquartile range and quartile deviation, mean deviation, standard deviation.
- The unit also encompass the concept of variance and coefficient of variation.



2.5 FURTHER READING

1) Agarwal, D. R. (2006). *Business Statistics*. Delhi: Vrinda Publications.

2) Gupta S. C. (1994). *Fundamentals of Statistics*. New Delhi: Himalayan Publishing House.

3) Rajagopalan, S. P. & Sattanthan R. (2009). *Business Statistics and Operations Research*. New Delhi: Tata McGraw-Hill

4) Sharma, J. K. (2007). *Business Statistics*. New Delhi: Pearson Education Ltd.

5) Verma, A. P. (2007). *Business Statistics*. Guwahat: Asian Books Private Limited.



2.6 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: S.D. = 119, C.V. = 21.74;

Ans. to Q. No. 2: A.M. = Rs. 63.32, S.D. = Rs. 18.69;

Ans. to Q. No. 3: 1.47 years

Ans. to Q. No. 4: (i) 0; (ii) σ; (iii) S.D. of the 2nd series = $\frac{1}{10}$ x S.D. of the first series

Ans. to Q. No. 5: (i) (c), (ii) (c), (iii) (a).



2.7 MODEL QUESTIONS

Q.1: What do you mean by dispersion? What are the various measures of dispersion?

Q.2: Distinguish between absolute and relative measures of dispersion.

Q.3: Define standard deviation. Why is standard deviation most widely used as a measure of dispersion?

Q.4: Write a note on coefficient of variation.

Class:	0-10	10-20	20-30	30-40	40-50
Frequency:	6	12	22	48	56
Class:	50-60	60-70	70-80		
Frequency:	32	18	6		

*** ***** ***

UNIT 3 : SKEWNESS, MOMENTS AND KURTOSIS

UNIT STRUCTURE

3.1 Learning Objectives

3.2 Introduction

3.3 Skewness

3.4 Measures of Skewness

- 3.4.1 Karl Pearson's Co-efficient of Skewness
- 3.4.2 Boweley's Co-efficient of Skewness

3.5 Moments

- 3.5.1 Moments About Mean
- 3.5.2 Moments About Arbitrary Point A
- 3.5.3 Relation Between Central and Raw Moments

3.6 Karl Pearson's Beta (β) and Gamma (γ) Coefficient

- 3.6.1 Coefficient of Skewness Based on Moments

3.7 Kurtosis

- 3.7.1 Measures of Kurtosis

3.8 Let Us Sum Up

3.9 Further Reading

3.10 Answers To Check Your Progress

3.11 Model Questions

3.1 LEARNING OBJECTIVES

After going through this unit, you will able to:

- Know the complementary relationship of skewness with measures of central tendency and dispersion in describing a set of data.
- understand 'moments' as a convenient and unifying method for summerizing several Introduction to statistical measures.

3.2 INTRODUCTION

In order to make proper comparison between two or more distributions and to reveal clearly the salient features of a frequency distribution, we need

to study certain statistical measures.The measures of central tendency tells us about the concentration of the observations about the middle of the distribution and the measure of dispersion give us an idea about the speead and scatter of the observation about some measure of central tendency. We may come across frequency distributions which differ very widely in their nature and composition and yet may have the same central tendency and dispersion.

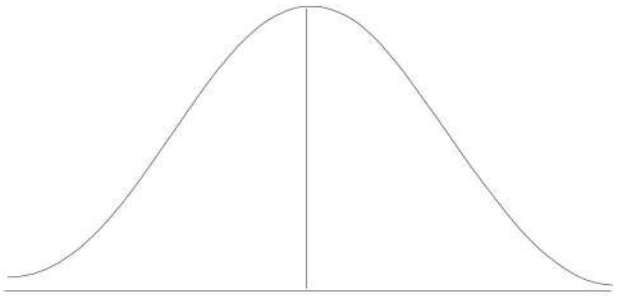
Thus these two measures viz. central tendency and dispersion are inadequate to characterise a distribution completely and these must be supported and supplemented by two more measures viz skewness and Kurtosis. **Skewness** helps us to study the shape i.e. symmetry or asymmetry of the distribution. While **Kurtosis** refers to the flatness or peakedness of the curve which may be drawn with the help of the given data. These four measures viz central tendency, dispersion, skewness and kurtosis are sufficient to describe a frequency distribution completely.

3.3 SKEWNESS

Literal meaning of skewness is lack of symmetry. We study skewness to have an idea about the shape of the curve which we can draw with the help of the given frequency distribution. It helps us to determine the nature and extent of the concentration of the observations towards the higher or lower values of the variable. In a symmetrical frequency distribution, which is unimodal, if the frequency curve is folded about the ordinate at the mean, the two halves so obtained will coincide with each other.

A frequency distribution, which is not symmetrical is called **asymmetrical** or **skewed**. In a skewed distribution, extreme values in a data set move towards one side or tail of the distribution, thereby lergthening that tail. A frequency distribution. For which the curve has a longer tail towards the right is said to be **positively skewed** (Figure 5.2) and if the longer tail lies towards the left, it is said to be negatively skewed (Figure 5.3).

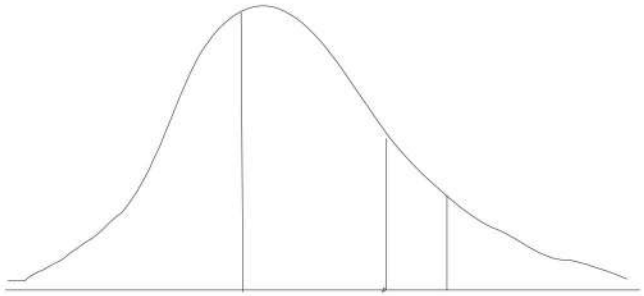
Note : 1) For a symmetrical distribution, a Bell-shaped (normal curve) is obtained for which Mean = Median = Mode i.e. these three values fall at the same point. (Figure 5.1)



Mean = Mode = Median

(Fig. 5.1)

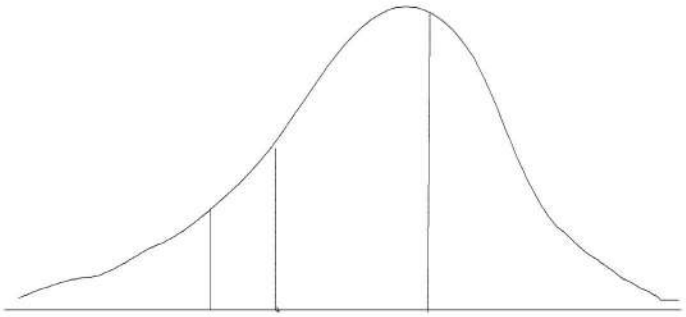
(Symmetrical Distribution)



Mode = Median = Mean

(Fig. 5.2)

(Positively Skewed Distribution)



$\bar{x}$ Median Mode

(Fig. 5.3)

(Negatively Skewed Distribution)

2) For a positively skewed distribution A.M > Median > Mode, and for a negatively skewed distribution AM < Median < Mode.

3) For a skewed distribution, Quartiles Q_1 and Q_3 are not equidistant from the Median i.e.

Q_3 – Median # Median – Q_1

3.4 MEASURES OF SKEWNESS

For a skewed distribution, the difference between AM and Mode i.e (AM – Mode) can also be taken as a measure of skewness as for a positively skewed distribution AM > Mode and for a negatively skewed distribution AM < Mode. But the statistician observed that this measure may not be desirable as the diffence between mean and mode is expressed in the same unit as the distribution and therefore can not be used for comparing skewness of two or more distributions having different units of measurement.

In order to overcome this shortcoming, relative measure of skewness (free from units of measurement), also commonly known as Co-efficient of skewness which is pure number suggested by Karl pearson is preferred for measuring skewness.

3.4.1 Karl Pearson's Co-efficient of Skewness

Karl Pearson's Co-efficient of Skewness is given by the Formula:

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}} = \frac{\bar{x} - M_o}{\sigma}$$

But quite often, Mode is ill-defined. In such a situation, we use the empieical relation between Mean, Mode and Median for moderately skewed distribution: Mean – Mode = 3 (Mean – Median)

$$\text{In that case, } Sk_p = \frac{3(\text{Mean} - \text{Median})}{\text{S.D.}}$$

Remarks:

1) Theoretically, Karl Pearson's Co-efficient of Skewness lies between ± 3. But in practice these limits are rasely attained. For a moderately skewed distribution, SK_p lies between ± 1.
2) For a symmetrical distribution Mean = Mode = Median. Therefore, $SK_p = 0$. In other words, for a symmetrical distribution $SK_p = 0$.
3) SK > 0 if $\bar{x}$ > Med > Mo

   $\therefore$ For a positively skewed distribution, the value of Mean is the greatest of the three measures and value of Mode is the leasl. On the other hand, SK < 0 if Mo > Med > $\bar{x}$

   $\therefore$ For a negatively skewed distribution the value of Mode is the greatest of the three measures and value of Mean is the leasl.

3.4.2 Boweley's Co-efficient of Skewness

There is another method suggested by Prof. Bowely known as **Bowleys'** Co-efficient of Skewness. It is given by

$$Sk_B = \frac{Q_3 + Q_1 - 2\text{Med}}{Q_3 - Q_1}$$

Remarks:

1) The values of Sk_3 lies between ± 1.
2) This method of measuring skewness is quite useful when the distribution has open end classes.

Note: It is important to note that–

1) These two measures of Skewness namely Karl Pearson's and Bowley's co-efficient of Skewness can not be compared. On certain occasions it is possible that one of them gives a possible that one of them gives a positive value while the other gives a negative value.
2) Out of the two measures, SK_p is more reliable due to the fact that sometime it may happen that $SK_B = 0$ but the distribution is not perfectly symmetrical.

Example 1: Given for a distribution, pearson's measure of skewness is 0.4, AM = 30 and standard deviation = 8. Find Median and Mode.

Solution: Given, $\bar{x} = 30$, $\sigma = 8$ and $SK_p = 0.4$

We know that, $SK_p = \dfrac{\bar{x} - \text{Mode}}{\sigma}$

$$\Rightarrow 0.4 = \frac{30 - \text{Mode}}{8}$$

$$\Rightarrow 3.2 = 30 - \text{Mode}$$

$$\Rightarrow \text{Mode} = 30 - 3.2 = 26.8$$

Also, $SK_p = \dfrac{3(\bar{x} - \text{Median})}{\sigma}$

$$\Rightarrow 0.4 = \frac{3(30 - \text{Median})}{8}$$

$$\Rightarrow 3.2 = 90 - 3\ \text{Median}$$

$$\Rightarrow 3\ \text{Median} = 86.8$$

$$\Rightarrow \text{Median} = 28.93$$

Example 2: From the data given below, calculate the co-efficient of variation–

AM = 86, Median = 80, Pearsonian measure of skewness = 0.42.

Solution: Given $\bar{x}$ = 86, Median = 80, SK_p = 0.42.

We know that, $SK_p = \frac{3(\bar{x} - \text{Median})}{\sigma}$

$$\Rightarrow 0.42 = \frac{3(86-80)}{\sigma}$$

$$\Rightarrow \sigma = \frac{18}{0.42} = 42.86$$

$\therefore$ Co-efficient of variation = $\frac{\sigma}{\bar{x}} \times 100$

$$= \frac{42.86}{86} \times 100$$

= 49.84%

Example 3: Given below are the Arithmetic mean, the Median and Standard Deviation of two distributions, Determine which distribution is more skewed.

i) AM = 22, Median = 24, SD = 10

ii) AM = 22, Median = 25, SD = 12

Solution: Here Pearson's Co-efficient of Skewness (SK_p) is applicable.

$$SK_p = \frac{3(\text{AM} - \text{Median})}{\text{SD.}}$$

For distribution (i), $SK_p = \frac{3(22-24)}{10} = -0.6$

For distribution (ii), $SK_p = \frac{3(22-25)}{12} = -0.75$

$\therefore$ For (i) $|SK_p| = 0.6$

For (ii) $|SK_p| = 0.75$

Since, $|SK_p|$ for (ii) > $|SK_p|$ for (i), hence distribution (ii) is more skewed.

Example 4: For a group 10 observations, sum of the observations is 452 and sum of squares of the observations is 24,270 and Mode is 43.7. Find the Pearson's co-efficient of Skewness.

Solution: Given n = 10, $\Sigma x = 452$, $\Sigma x^2 = 452$, $\Sigma x^2 = 24{,}270$ and Mode = 43.7

$\therefore$ Mean $(\bar{x}) = \frac{\Sigma x}{n} = \frac{452}{10} = 45.2$

$$S.D.(\sigma) = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2} = \sqrt{\frac{24270}{10} - (45.2)^2} = 19.59$$

$$\therefore\ SK_p = \frac{\bar{x} - Mode}{\sigma} = \frac{45.2 - 43.7}{19.59} = -0.08$$

Example 5: In a frequency distribution, the co-efficient of Skewness lased on quartiles is 0.6. If the sum of upper and lower quaitiles is 100 and Median is 38, find Q_3.

Solution : Given $SK_B = 0.6$, $Q_3 + Q_1 = 100$, Median = 38

We know that, $SK_B = \frac{Q_3 + Q_1 - 2Median}{Q_3 - Q_1}$

$$\Rightarrow 0.6 = \frac{100 - 76}{Q_3 - Q_1}$$

$$\Rightarrow Q_3 - Q_1 = \frac{24}{0.6} = 40$$

Given $Q_3 + Q_1 = 100$

$\therefore (Q_3 - Q_1) + (Q_3 + Q_1) = 40 + 100$

$\Rightarrow 2Q_3 = 140$

$\Rightarrow Q_3 = 70$

Example 6: Calculate Karl Pearson's Co-efficient of Skewness from the data given below.

Wt (Kg):	10	20	30	40	50	60	70
Frequency:	1	5	12	22	17	9	4

Solution: Table for calculation of co-efficient of Skewness

Wt (kg) (x)	Frequency (f)	d = x – A	$d' = d/10$	fd'	fd'^2
10	1	–30	–3	–3	9
20	5	–20	–2	–10	20
30	12	–10	–1	–12	12
40	22	0	0	0	0
50	17	10	1	17	17
60	9	20	2	18	36
70	4	30	3	12	36
	70 = N			$\Sigma fd' = 22$	$\Sigma fd'^2 = 130$

Let assumed mean = A = 40

Here i = 10

$$\therefore \bar{x} = A + \frac{\Sigma fd'}{N} \times i$$

$$= 40 + \frac{22}{70} \times 10$$

= 43.14 kg.

Mode = 40 (By inspection)

$$SD(\sigma) = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times i$$

$$= \sqrt{\frac{130}{70} - \left(\frac{22}{70}\right)^2} \times 10 = 13.27$$

$$\therefore \ SK_P = \frac{\bar{x} - Mode}{\sigma} = \frac{43.1 - 40}{13.26} = 0.23$$

Example 7: From the data given below, calculate Karl Pearson's co-efficient of Skewness.

Ht (cm):	130-134	135-139	140-144	145-149	150-154	155-159	160-164
f:	3	12	21	28	19	12	5

Solution: Table for calculating Karl Pearson's co-efficient of Skewness.

Ht (cm) value(x)	(f)	Mid	d = x–A	$d'=d/5$	fd^1 Boundarie	fd^{12}	Class
130-134	3	132	–15	–3	–9	27	129.5–134.5
135–139	12	137	–10	–2	–24	48	134.5–139.5
140–144	21	142	–5	–1	–21	21	139.5–144.5
145–149	28	14.7	0	0	0	0	144.5–154.5
150–154	19	152	5	1	19	19	149.5–154.5
155–159	12	157	10	2	24	48	154.5–159.5
160–164	5	162	15	3	15	45	159.5–164.5
	100 = N				4 = $\Sigma fd'$	208 = Σfd^{12}	

Let A = 147 Here i = 5

$$\therefore \bar{x} = A + \frac{\Sigma fd'}{N} \times i = 147 + \frac{4 \times 5}{100} = 147.2 \text{ cm}$$

$$\sigma = \sqrt{\frac{\Sigma fd^{12}}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$$

$$= \sqrt{\frac{208}{100} - \left(\frac{4}{100}\right)^2} \times 5$$

$$= \sqrt{2.08 - .0016} \times 5 = 7.21 \text{ cm}$$

Here Modal Class is 144.5 – 149.5

$$\therefore \text{ Mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$$= 144.5 + \frac{28 - 21}{56 - 21 - 19} \times 5$$

$$= 144.5 + \frac{35}{16} = 146.69 \text{ cm}$$

$$\therefore SK_P = \frac{\bar{x} - \text{Mode}}{\sigma} = \frac{147.2 - 146.69}{7.21} = 0.07$$

Example 8: Fine the most suitable measure of Skewness from the data given below.

Age (yrs):	Below 20	20–25	25–30	30–35	35–40	40–45	45–50
No. of employee 50 and above 21	13	29	46	60	112	94	45

Solution: This is an open-ent series. So. the most suitable measure of Skewness is Bowley's Co-efficient of Skewness.

Age (yrs)	No. employee (f)	c.f.
Below 20	13	13
20 – 25	29	42
25 – 30	46	88
30 – 35	60	148
35 – 40	112	260
40 – 45	94	354
45 – 50	45	399
50 and above	21	420
	420 = N	

We know that, Bowley's coefficiant of Skewness

$$= SK_B = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

Here N = 420

$$\therefore\ Q_1 = \text{Value of } \left(\frac{N}{4}\right)^{th} = \frac{420}{4} = 105^{th} \text{ item}$$

$$\text{Median} = \text{Value of } \left(\frac{N}{2}\right)^{th} = \frac{420}{2} = 210^{th} \text{ item}$$

$$Q_3 = \text{Value of } \left(\frac{3N}{4}\right)^{th} = \frac{3 \times 420}{4} = 315^{th} \text{ item}$$

$\therefore\ Q_1$ class is 30 – 35

Median class is 35 – 40

Q_3 class is 40 – 45

$$\therefore\ Q_1 = l + \frac{N/4 - cf}{f} \times i$$

$$= 30 + \frac{105 - 88}{60} \times 5$$

= 30 + 1.42

= 31.42 yrs

$$\text{Median} = l + \frac{N/2 - cf}{f} \times i$$

$$= 35 + \frac{210 - 148}{112} \times 5$$

= 34 + 2.77

= 37.77 yrs

$$\text{and } Q_3 = l + \frac{3N/4 - cf}{f} \times i$$

$$= 40 + \frac{315 - 260}{94} \times 5$$

= 42 + 2.93 = 42.93 yrs

$$\therefore SK_B = \frac{42.93 + 31.42 - 2 \times 37.77}{42.93 - 31.42}$$

$$= \frac{-1.19}{11.51} = -0.103$$

3.5 MOMENTS

In statistics, 'Moments' are used to describe the various characteristics of a frequency distribution viz central tendency, dispersion, Skwness and Kurtosis.

3.5.1 Moments About Mean

Let a variable x takes the values $x_1, x_2, ..., ..., x_n$. Then $\bar{x} = \frac{\Sigma x}{n}$ is its arithmetic mean.

The r^{th} moment of x about the mean $\bar{x}$, is usually denoted by μr [where μ is the letter (mu) of the Greek alphabet] is defined as

$$\mu_r = \frac{1}{n}\Sigma(x-\bar{x})^r \quad = r = 1, 2, 3, 4 \qquad 5.5.1\ (a)$$

Similarly, for a frequency distribution of a variable x

x	x_1 x_2..... x_n	
y	f_1 f_2.....f_n	$\Sigma f = N$

Then, $\bar{x} = \frac{\Sigma fx}{N}$ is the arithmetic mean

Then r^{th} moment about $\bar{x}$ is defined as

$$\mu_r = \frac{1}{N}\Sigma f(x-\bar{x})^r, \; r = 1, 2, 3, 4 \qquad 5.5.1\ (b)$$

Putting r = 1 in 5.5.1 (a) and 5.5.1 (b) we get

$$\mu_1 = \frac{1}{n}\Sigma(x-\bar{x})^r = 0 \quad \rightarrow \text{For individual series}$$

$$\mu_1 = \frac{1}{N}\Sigma f(x-\bar{x})^r = 0 \quad \rightarrow \text{For frequency distribution}$$

[Because, the algebric sum of deviations of a given set of observations from their AM is zero]

Thus, the first moment about mean is always zero.

Putting r = 2.

$$\mu_2 = \frac{1}{n}\Sigma(x-\bar{x})^2 = \sigma x^2 \quad \rightarrow \text{For individual series.}$$

$$\mu_2 = \frac{1}{N}\Sigma f(x-\bar{x})^2 = \sigma x^2 \quad \rightarrow \text{For frequency distribution.}$$

Thus, the second moment about mean is the variance of the distribution.

Also,
$$\left.\begin{aligned}\mu_3 &= \frac{1}{n}\Sigma(x-\bar{x})^3\\ \mu_4 &= \frac{1}{n}\Sigma(x-\bar{x})^4\end{aligned}\right\}$$ For individual series

and
$$\left.\begin{aligned}\mu_3 &= \frac{1}{N}\Sigma f(x-\bar{x})^3\\ \mu_4 &= \frac{1}{N}\Sigma f(x-\bar{x})^4\end{aligned}\right\}$$ For frequency distribution

3.5.2 Moments About Arbitrary Point A

When actual mean is in fractions, moments are first calculated about an assumed mean A (arbitrary point) and then are converted about the actual mean.

The r^{th} moment of a variable x about an arbitrary point A is given by

$$\mu'_r = \frac{\Sigma(x-A)^r}{n} \rightarrow \text{For individual series}$$

and $$\mu'_r = \frac{\Sigma f(x-A)^r}{N} \rightarrow \text{For frequency distribution.}$$

Now, $$\mu'_1 = \frac{1}{n}\Sigma(x-A)^2 = \frac{1}{n}\Sigma x - \frac{1}{n}\Sigma A$$

$$= \bar{x} - \frac{1}{n}\times nA = \bar{x} - A$$

Also, $$\mu'_1 = \frac{1}{N}\Sigma f(x-A)$$

$$= \frac{1}{N}\Sigma fz - \frac{1}{N}\Sigma fA$$

$$= \bar{x} - \frac{A}{N}\Sigma f = \bar{x} - \frac{A}{N}\times N = \bar{x} - A$$

$\therefore$ For both individual series and frequency distribution.

$$\boxed{\mu'_1 = \bar{x} - A}$$

Note : 1) μ_r , the r^{th} moment about the mean; r = 1, 2, 3, 4, ..., ... are also called **Central** moments and μ'_r , the r^{th} moment about any arbitrary point A are called **raw moments** or **non-central moments.**

2) The moments (raw or central) of higher order other than fourth order are seldom used.

3.5.3 Relation Between Central and Raw Moments

Central Moment $= \mu_r = \frac{1}{n}\Sigma(x-\bar{x})^r$

Raw Moments about an arbitrary point A

$$\mu'_r = \frac{1}{n}\Sigma(x-A)^r$$

Now, $\mu_r = \frac{1}{n}\Sigma(x-\bar{x})^r$

$$= \frac{1}{n}\Sigma\{x-A-(\bar{x}-A)\}^r$$

$$= \frac{1}{n}\Sigma\left(x-A-\mu'_1\right)^r$$

$$= \frac{1}{n}\left[\Sigma(x-A)^r - {}^rC_1\,\mu'_1\Sigma(x-A)^{r-1} + {}^rC_2(\mu'_1)^2\,\Sigma(x-A)^{r-2} + \ldots\ldots + (-1)^r(\mu'_1)^r\right]$$

$$= \frac{1}{n}\left[\Sigma(x-A)^r - {}^rC_1\,\mu'_1\frac{1}{n}\Sigma(x-A)^{r-1} + {}^rC_2(\mu'_1)^2\,\Sigma(x-A)^{r-2} + \ldots\ldots + (-1)^r(\mu'_1)^r\right]$$

$$= \mu'_r - {}^rC_1\,\mu'_1\mu'_{r-1} + {}^rC_2(\mu'_1)^2\mu'_{r-2} + \ldots\ldots + (-1)^r(\mu'_1)^r$$

From the above relation, for various values of r we have

For r = 1, $\mu_1 = \mu'_1$

For r = 2, $\mu_2 = \mu_2' - {}^2C_1\mu'^2_1 + {}^2C_2\mu_1'^2\mu_0'$

$$\Rightarrow \boxed{\mu_2 = \mu_2' - (\mu_1')^2}$$

$$\mu_0 = \frac{1}{n}\Sigma(x-x)^0$$

$$= \frac{1}{n}\Sigma 1$$

$$= \frac{n}{n} = 1$$

$$\mu_0' = \frac{1}{n}\Sigma(x-A)^0$$

$$= \frac{1}{n}\Sigma 1$$

$$= \frac{n}{n} = 1$$

For r = 3, $\mu_3 = \mu_3' - {}^3C_1\mu_1'\mu_2' + {}^3C_2(\mu_1')^2\mu_1' - {}^3C_3(\mu_1')^3\mu_0'$

$$\Rightarrow \boxed{\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3}$$

Similarly,

For r = 4, $\boxed{\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6\mu_2'(\mu_1')^2 - (\mu_1')^4}$

Example 9: The first two moments of a distribution about the value 5 of the variable are 2 and 20 respectively. Find the mean and variance of the distribution.

Solution: given A = 5, $\mu_1' = 2, \mu_2' = 20$

We know that mean $(\bar{x}) = A + \mu_1' = 5 + 2 - 7$

Variance $(\sigma^2) = \mu_2 = \mu_2' - (\mu_1')^2$ = 20 – $(2)^2$ = 16

Example 10: Find the first four moment about the mean for the set of numbers 2, 4, 6 and 8.

Solution: $\mu_r = r^{th}$ Central moment $= \frac{\Sigma(x-\bar{x})^r}{n}$, r = 1, 2, 3, 4

Table for Calculation of First Four Moments

x	$x-\bar{x} = x-5$	$(x-\bar{x})^2$	$(x-\bar{x})^3$	$(x-\bar{x})^4$
2	–3	9	–27	81
4	–1	1	–1	1
6	1	1	1	1
8	3	9	27	81
$20 = \Sigma x$	$0 = \Sigma(x-\bar{x})$	$20 = \Sigma(x-\bar{x})^2$	$0 = \Sigma(x-\bar{x})^3$	$164 = \Sigma(x-\bar{x})^4$

$$\therefore \bar{x} = \frac{\Sigma x}{n} = \frac{20}{4} = 5$$

$$\therefore \mu_1 = \frac{\Sigma(x-\bar{x})}{4} = 0, \quad \mu_2 = \frac{\Sigma(x-\bar{x})^2}{4} = \frac{20}{4} = 5$$

$$\mu_3 = \frac{\Sigma(x-\bar{x})^3}{4} = 0, \quad \mu_4 = \frac{\Sigma(x-\bar{x})^4}{4} = \frac{164}{4} = 41$$

Example 11: The first four moments of a distribution about the value 4 are– 1.5, 17, –30 and 108. Find the moments abour mean.

Solution: given A = 4

$$\mu_1' = -1.5, \mu_2' = 17, \mu_3' = -30, \mu_4' = 108$$

$\therefore$ First moment about mean = μ_1

$$= A + \mu_1'$$

$$= 4 - 1.5 = 2.5$$

Second moment about mean = μ_2

$$= \mu_2' - (\mu_1')^2$$

$$= 17 - (-1.5)^2 = 14.75$$

Third moment about mean = μ_3

$$= \mu_3' - 3\mu_3'\mu_3' + 2(\mu_3')^3$$

$$= -30 - 3\ (-1.5) \times (17) + 2\ (-1.5)^3$$

$$= -30 + 76.5 - 6.75 = 39.75$$

Fourth central moment about mean

$$= \mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6(\mu_1')^2 - 3(\mu_1')^4$$

$$= 108 - 4 \times (-1.5)\ (-30) + 6\ (-1.5)^2 \times 17 - 3\ (-1.5)^4$$

$$= 108 - 180 + 229.5 - 15.1875$$

$$= 142.3125$$

3.6 KARL PEARSON'S BETA (β) AND GAMMA (γ) COEFFICIENT

Prof Karl Pearson defined the following four co-efficients based on the 1st four central moment.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \qquad \rightarrow 5.6.1$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4} \qquad \rightarrow 5.6.2$$

$$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3} \quad \left[\Theta\ \mu_2 = \sigma^2\right] \qquad \rightarrow 5.6.3$$

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3 \qquad \rightarrow 5.6.4$$

It may be stated here that these co-efficients are pure numbers independent of units of measurement and as such can be conveniently used for comparative studies. In practice, they are used as measure of Skewness and Kurtosis as discussed below.

3.6.1 Co-efficient of Skewness Based on Moments

The co-efficient of Skewness, based on the moments is given by–

$$SK = \frac{\beta_2 + 3\sqrt{\beta_1}}{2(5\beta^2 - 6\beta_1 - 9)}$$

$\therefore$ Sk = 0 if $\beta_1 = 0$ or $\beta_2 + 3 = 0 \Rightarrow \beta_2 = -3$

But $\mu_4 = \frac{1}{n}\Sigma(x - \bar{x})^4 > 0$

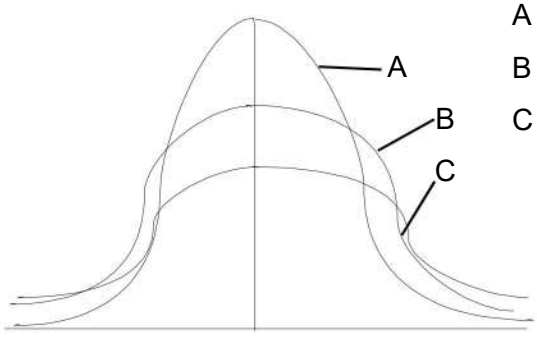
and $\mu_2 = \frac{1}{n}\Sigma(x - \bar{x})^2 > 0$

$\therefore \beta_2 = \frac{\mu_4}{\mu_2^2} > 0$

Since β_2 can not be negative, SK = 0 if $\beta_1 = 0$.

3.7 KURTOSIS

So far we have studied three measure viz central tendency, dispersion and skewness to describe the characteristics of a frequency distribution. However, even if we know all these three measures, we all not in a position to charcterise a distribution completely. The diagram given below will clarify the situation.



A – Lapto Kurtic
B – Meso Kurtic
C – Platy Kurtic

All these three curves are symmetrical about the mean and have same variation (range). But the shape and nature of the middle part of the curve are different i.e the peakedness of the three curves are not same. So. to identify a distribution completaly, we need one more measure which is called **'Kurtosis'**.

The word **Kurtosis** comes from a Greek word meaning 'humped'. In statistics, **Kurtosis refers to the degree of flatness or peakedness of the frequency curve.**

Curve of type B which is neither flat peaked is known as normal curveand shape of its hump is accepted as a standard one. This curve is termed as **mesokurtic.** The curves of type A, which is more peaked than the normal curve are known as **leptokurtic** and the curves of type C, which are flatter than the normal curve are called **platykurtic**.

3.7.1 Measures of Kurtosis

As a measure of Kurtosis, Karl Pearson gave the coefficient Beta two (β_2) or its derivative Gramma two (δ_2) defined as follows:

$$\beta_2 = \frac{\mu_4}{\mu_2^{\,2}} = \frac{\mu_4}{6^4} \qquad \rightarrow 5.7.1 \text{ (a)}$$

$$\alpha_2 = \beta_2 - 3 = \frac{\mu_4}{6^4} - 3 \quad \rightarrow 5.7.2 \text{ (b)}$$

For a normal curve (mesakurtic curve), $\beta_2 = 3$ or $r_2 = 0$. For a lepto kurtic curve, $\beta_2 > 3$ or $r_2 > 0$ and for a platy kurtic curve, $\beta_2 < 3$ or $r_2 < 0$.

Note : A British statistician W. S. Gosset explained the use of the terms **platykurtic** and **leptokurtic** as: **'platykurtic curves like platyplus, are squat with short tails; leptokurtic curves are high with long tails like the kangaroos noted for leaping.**

Example 12: The first four moments of a distribution about the value 5 and 2, 20, 40 and 50. Do you think that the distribution is platykurtic?

Solution: Here A = 5, $\mu_1' = 2$, $\mu_2' = 20$, $\mu_3' = 40$, $\mu_4' = 50$

$\therefore$ Mean $(\bar{x}) = \mu_1' + A = 2 + 5 = 7$

Now, $\mu_2 = \mu_2' - (\mu_1')^2 = 20 - (2)^2 = 16$

$\therefore$ Variance $\mu_2 = 16$

$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3$

= 40 – 3 x 2 x 20 + 2 $(2)^3$ = – 64..7

$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$

= 50 – 4 x 2 x 40 + 6 x 20 x $(2)^2$ – 3$(2)^4$ = 162

$$= \beta_2 = \frac{\mu_4}{\mu_2^{\ 2}} = \frac{162}{(16)^2} = 0.63$$

Since $\beta_2 < 3$, the distribution is platykurtic.

Example 13: The first four moments of a distribution about the origin are 1, 4, 10 and 46 respectively. Calculate the central moments and indicates the nature of the distribution.

Solution: given A = 0, $\mu_1' = 1, \mu_2' = 4, \mu_3' = 10, \mu_4' = 46$

$\therefore$ Mean $(\bar{x}) = A + \mu_1' = 0 + 1 = 1$

Variance $(\sigma)^2 = \mu_2 = \mu_2' - (\mu_1')^2 = 4 - (1)^2 = 3$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$

= 10 – 3 x 4 x 1 + 2$(1)^3$

= 0

$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$

= 46 – 4 x 10 x 1 + 6 x 4 x 1 – 3 = 27

$\therefore$ Karl Pearson's moment co-efficient of Skewness is given by

$$\delta_1 = \frac{\mu_3}{\mu_2^{\ 3/2}} = 0, \ \beta_1 = \frac{\mu_3^{\ 2}}{\mu_2^{\ 3}} = 0$$

Since, $\beta_1 = 0$ the given distribution is symmetrical (normal).

Again, Karl Pearson's measure of Kurtosis is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^{\ 2}} = \frac{27}{(3)^2} = 3$$

and $\delta_2 = \beta_2 - 3 = 0$

Since, $\beta_2 = 3$, it is mesokurtic (normal) distribution with

Mean $(\bar{x}) = 1$

and S.D. $= \sqrt{3} = 1.732$

Example 14: If $\beta_1 = +1, \beta_2 = 4$ and variance = 9, find the values of μ_3 and μ_4 and comment upon the nature of Kurtosis.

Solution: Given $\beta_1 = +1, \beta_2 = 4$

Variance $= \sigma^2 = \mu_2 = 9$

Now, $\beta_1 = +1 \Rightarrow \frac{\mu_3^2}{\mu_2^3} = 1 \Rightarrow \mu_3^2 = \mu_2^3 = 9 \times 9 \times 9 = (3 \times 3 \times 3)^2$

$\Rightarrow \mu_3^2 = (27)^2$

$\Rightarrow \mu_3 = \pm 27^2$

Also $\beta_2 = 4$

$\Rightarrow \frac{\mu_4}{\mu_2^2} = 4$

$\Rightarrow \mu_4 = 4 \times 9 \times 9 = 324$

$\therefore \mu_3 = \pm 27$ and $\mu_4 = 324$

$\mu_2 = 4 > 3$

Hence the given distribution is leptokurtic.



CHECK YOUR PROGRESS

Q.1: Explain meaning of skewness using sketches of frequency curves. How does skewness different from dispersion.

Q.2: Distinguish between Karl Pearson's and Bowley's measure of skewness. Which one of these would you prefer and why?

Q.3: How is Kurtosis measured ?

Q.4: Fill in the blanks:

i) For a symmetrical distribution mean, median and mode are .....

ii) If $\mu_2 > 3$, the curve is called .....

iii) If $\mu_1 = 0$, the distribution is .....

iv) If Mode > Mean, the distribution is ..... skewed.

v) Kurtosis measures ..... of the frequency curve.

vi) Bowley's co-efficient of skewness lies between .....

vii) Relative measure of Kurtosis is .....

Q.5: Say whether the following statements are true or false.

i) For a symmetrical distribution $\beta_1 = 0$.

ii) Bowely's co-efficient of skewness lie between ± 3.

iii) For a negatively skewed distribution mean > mode.

iv) The first moment about the origin L mean.

v) Variance = μ_2

vi) If $\beta_2 > 3$, the curve is platykurtic.

Q.6: Consider the following distribution–

	Distribution A	**Distribution B**
Mean	100	90
Median	90	80
Variance	100	100

i) Both the distribution have same degree of skewness.

ii) Distribution A has the same degree variation as B.

–Say whether the following statements are true or false. Justify your answer.

Q.7: For a distribution $\bar{x} = 30$, $\sigma = 8$ and SK_p= 0.4. Find Median and Mode.

Q.8: For a distribution, Mean = 50, CU = 40%, SK_p = –0.4. Find SD, Median and Mode.

Q.9: For a group of 10 items, $\Sigma x = 452$ and Mode = 43.7. Find the pearsonion co-efficient of Skewness.

Q.10: The measure of skewness for a distribution is – 0.8. If the upper and lower quartiles are respectively 56.6 and 44.1, find Median.

Q.11: Calculate Karl Pearson's co-efficient of skewness from the data given below.

Marks	No. of students
21 – 25	5
26 – 30	15
31 – 35	28
36 – 40	42
41 – 45	15
46 – 50	12
51 – 55	3

Q.12: Calculate Bowley's co-efficient of skewness

Wt (kg):	0–15	15–30	30–45	45–60	60–75	75–90	90 and above
f:	20	30	30	35	45	15	5

Q.13: a) Find the first and second moments about the mean for the numbers 3, 4, 5 and 8.

b) Find the first, second, third moments of the numbers 1, 3, 6, 7, 8 about the value 4.

Q.14: a) The first three moments of a distribution about the value 1 of the variable are 2, 25 and 80. Find the mean, standard deviation and the moment measure of skewness.

b) The first four moments of a distribution about x = 2 are 1, 2.5, 5.5 and 16. Calculate the four moments about ($\bar{x}$) and about zero.

Q.15: a) The central moments of a distribution are given by μ_2 = 140, μ_3 = 148 and μ_4 = 6030, calculate the moment measure of skewness and kurtosis and comment on the shape of the distribution.

b) From the data given below do you think that the distribution is platykurtic?

N = 100, $\Sigma fd = 50$, $\Sigma fd^2 = 1970$, $\Sigma fd^3 = 2948$ and Σfd^4 = 86, 752; where d = x – 48.



3.8 LET US SUM UP

In this unit we have learnt about the following–

- Meaning of symmetrical and asymmetrical distribution.
- Meaning of Skewness and types of Skewness.
- Different measures of skewness.
- Meaning of moments, central moments, Raw moments.
- Use of moments for describing various characteristics of frequency distribution.
- Meaning of kurtosts and types of Skewness.
- To measure Skewness and Kurtosis using moments.



3.9 FURTHER READING

1) Agarwal, D. R. (2006). *Business Statistics*. Delhi: Vrinda Publication.
2) Arora, P. N., Arora, Sumeet & Arora, S. *Comprehensive Statistical Methods.* S. Chand and Company.
3) Ghosh, R. K. and Saha, S. *Business Mathematics and Statistice.* New Central Book Agency (P) Ltd.
4) Gupra, S. C. *Fundamental of Statistics.* New Delhi: Himalaya Publishing House.
5) Gupta, S. C. and Gupta, Indira. *Business Statistics.* New Delhi: Himalaya Publishing House.
6) Sharma, J. K. *Business Statistics.* New Delhi: Pearson Education Ltd.



3.10 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 4: (i) Equal, (ii) Laptakurtic, (iii) symmetrical, (iv) negatively, (v) peakedness/flatness, (vi) ± 1, (vii) β_2 or r_2.

Ans. to Q. No. 5: (i) True, (ii) False, (iii) False, (iv) True, (v) True, (vi) False.

Ans. to Q. No. 6: (i) True, (ii) False

Ans. to Q. No. 7: Median = 28.93, Mode = 26.8

Ans. to Q. No. 8: SD = 20, Mode = 58, Median = 52.67

Ans. to Q. No. 9: $SK_p = 0.077$

Ans. to Q. No. 10: Median = 55.35

Ans. to Q. No. 11: $SK_p = -0.109$

Ans. to Q. No. 12: $SK_B = -0.1126$

Ans. to Q. No. 13: a) $\mu_1 = 0,\ \mu_2 = 3.5$

b) $\mu_1' = 1,\ \mu_2' = 7.8,\ \mu_3' = 14.2$

Ans. to Q. No. 14: a) mean = 3, SD = 4.58

$\delta_1 = \sqrt{\beta_1} = -0.561$

b) Moments about mean.

$\mu_2 = 1.5,\ \mu_3 = 0,\ \mu_4 = 6$

Moments about zero

$\mu_1' = 3,\ \mu_2' = 10.5,\ \mu_3' = 40.5,\ \mu_4' = 168$

Ans. to Q. No. 15: a) $\delta_1 = 0.0893,\ \beta_2 = 0.3076$

The distribution is approximately symmetrical and platykurtic

b) $\beta_2 = 2.2.14,\ \beta_2 < 3$, distribution is platykurtic.



3.11 MODEL QUESTIONS

Q.1: Averages dispersion, skewness and kurtosis are complementary to one another in understanding a frequency distribution.

Q.2: Define moments establish the relationship between the moments about mean and moments about any orbitrary point.

Q.3: The first two moments of a distribution about the value 5 are 2 and 20. Find the mean and variance.

Q.4: Calculate Karl Peason's Co-efficients of Skewness for the given distribution.

Marks less than:	20	40	60	80	100
No. of students:	18	40	70	90	100

Q.5: Calculate Bowley's Co-efficient of Skewness.

Marks:	0–10	10–20	20–30	30–40	40–50
No. of students:	5	9	12	8	6

Q.6: The arithmetic mean of a certain distribution is 5. The second and the third moments about the mean are 20 and 140 respectively. Find the third moment of the distribution about zero.

*** ***** ***

UNIT 4: CORRELATION

UNIT STRUCTURE

4.1 Learning Objectives

4.2 Introduction

4.3 Correlation: Meaning and Concept

4.4 Types of Correlation

4.5 Methods of Measuring Correlation

4.6 Scatter Diagram and Correlation

4.7 Correlation Coefficient

4.8 Underlying Assumptions

4.9 Interpretation of Correlation Coefficient

4.10 Rank Correlation Coefficient

4.11 Let Us Sum Up

4.12 Further Reading

4.13 Answers To Check Your Progress

4.14 Model Questions

4.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- explain the concept of correlation coefficients of correlation and their properties
- discuss different types of correlation and regression
- explain the essential features of correlation and regression analysis
- explain and compare the various aspects of correlation.

4.2 INTRODUCTION

Depending on the number of variables that we study, we may have univariate, bivariate or multivariate series or distribution. Univariate series or distribution is one when we deal with only one variable. For example, if we consider, say, only age of different persons, then we will have a univariate series, i.e. values of only age. Bivariate distribution, on the other hand,

involves two variables (note that 'bi' means two). If, for instance, we take both income and expenditure of persons then distribution thus obtained will constitute a bivariate distribution. Multivariate distribution will, similarly consist of more than two variables. The two concepts-correlation and regression that we will study in this unit are associated with bivariate and/or multivariate series or distributions.



LET US KNOW

A **variable** is an entity that can take any value. Age, income, height, weight, expenditure, marks etc. are examples of variables. Note that for different persons these will take different values. Contrary to it, **constant** is an entity that can assume only fixed value.



ACTIVITY 4.1

Try to list some variables and constants that you find in common day to day life.

.....

.....

4.3 CORRELATION: MEANING AND CONCEPT

In the simplest term, correlation means relation between two variables (note that 'co' simply means two things together). In case of bivariate distributions, many a times, we may be interested in looking at how the two variables are related among themselves. Obviously, when two variables are related, change in any one variable is responded by the other variable showing some changes in it. Correlation analysis tries to measure the 'extent' of such responses among the variables. Thus, correlation is an analysis of co-variation between two (or even more) variables. Two variables are said to be correlated if the change in any one of them results in a corresponding change in the other.

4.4 TYPES OF CORRELATION

Correlation may be classified into the following three types:

a) Positive correlation
b) Negative correlation and
c) Zero Correlation.

a) Positive Correlation: When increase (or decrease) in one variable leads to a corresponding increase (or decrease) in the other variable then it is called positive correlation. The basic idea of positive correlation is that both the variables move in the same direction. For example, there exists positive correlation between the following pairs of variables:

i) Income and expenditures of a group of families
ii) Amount of rainfall and yield of crops
iii) Price and supply of a company
iv) Temperature and sale of ice-creams on different days of a month in summer.

b) Negative Correlation: Negative correlation is, on the other hand, is a situation when increase (or decrease) in one variable is accompanied by a decrease (or increase) in the other variable. Unlike the positive correlation, the fundamental element in negative correlation is that the variables move in opposite directions. For example, following are the situations of negative correlations:

i) Price of a commodity and demand for it.
ii) Sale of woolen garments and the day temperature.
iii) Number of workers and the time required to complete work.

c) Zero correlation: In case, variables are found as uncorrelated i.e., change in one of the variables fail to produce any corresponding change in the other, and then it is called zero correlation. For example, one should expect zero correlation between the heights of workers and the income earned by them, or between price of rice and demand for sugar.

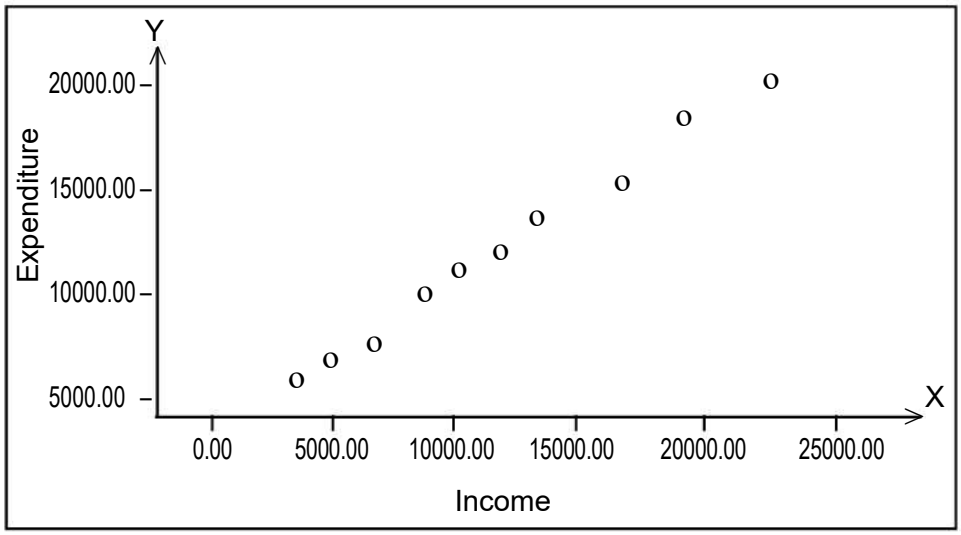
4.5 METHODS OF MEASURING CORRELATION

The following are the methods used to measure the correlation between two variables:

i) Scatter Diagram Method

ii) Karl Pearson's Correlation Method and

iii) Spearman's Rank Correlation Method.



4.6 SCATTER DIAGRAM AND CORRELATION

How can one find whether any two variables are correlated or not? One of the easiest ways for identifying correlation between two variables is scatter diagram. Scatter diagram is a very simple graphical technique of identification of correlation. When we consider respective pairs of values as points and diagram them against X-Y axes then we get the scatter plot of the values. To understand this, let us consider the following bivariate series of income and expenditure of say, ten persons.

Person (No.)	Income (Rs.)	Expenditure (Rs.)
1	9200	6700
2	6860	4256
3	12825	9875
4	17980	12900
5	22125	19700

6	7892	4780
7	4645	3200
8	10925	8645
9	13820	10900
10	24780	21200

Taking income in X-axis and expenditure in Y-axis, the values of the above table can now be plotted in a graph to see whether these values are correlated. The scatter thus obtained will be somewhat like this

It may be seen from the above scatter that points show an upward-rising trend. If the scatter does indicate some sort of trend or pattern then it implies existence of correlation among the variables. If, for example, pattern is such that there is a visible upward trend from left to right, then variables are said to have positive correlation. In case, there is no such identifiable pattern then the variables are said to be uncorrelated.



ACTIVITY 4.2

Note down height and weight of 10 of friends of yours. Then try to draw a scatter plot of the values. What can you say about their possible correlation?

.....

.....



CHECK YOUR PROGRESS

Q.1: Choose the correct answer from the following:

i) Correlation and regression are associated with–

a) Univariate series b) Bivariate series

c) Both bivariate and multivariate series

d) Multivariate series

.....

ii) Scatter plot technique is used to detect–

a) Correlation b) Regression

.....

iii) When variables move in opposite directions then it is–

a) Positive Correlation b) Negative Correlation

.....

4.7 CORRELATION COEFFICIENT

Scatter plots are particularly useful to identify whether the variables are correlated and in case they are then what is the type of such relation. It, however, can't quantify or quantitatively measure the 'amount', 'extent', 'magnitude' or 'degree' of relation that exists between the variables. British statistician Karl Pearson (1867-1936) introduced a simple technique for measuring the magnitude of such correlation. His measure is called Pearson's Product-Moment Correlation Coefficient or simply Pearson' Correlation Coefficient and usually denoted by 'r' (standing for 'relation') which is given as:

$$r_{x,y} = \frac{Cov(x,y)}{\sigma_x.\sigma_y}$$

Here, cov (x, y) means "covariance of x and y" and σx and σy are standard deviations of x and y respectively. Formulas for obtaining these are as bellow:

$$Cov(x, y) = \frac{1}{n}\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)$$

$$\sigma_x = \sqrt{\frac{1}{n}\sum\left(x-\bar{x}\right)^2}$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum\left(y-\bar{y}\right)^2}$$

Putting these in the original formula and simplifying, we will have,

$$r = \frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sqrt{\sum\left(x-\bar{x}\right)^2\left(y-\bar{y}\right)^2}}$$

This can be further simplified and can be written as:

$$r = \frac{n\sum xy - \sum x\sum y}{\sqrt{\left[n\sum x^2-\left(\sum y\right)^2\right]\left[n\sum y^2-\left(\sum y\right)^2\right]}}$$

This formula is relatively easier to calculate and most commonly used to obtain the Pearson's coefficient of correlation; 'n' here means total number of observations.

Here is procedure of finding the 'r' for a series involving two variables x and y:

- Obtain total of x i.e. Σx
- Obtain total of y i.e. Σy
- Get product of x and y for all pairs and then get the total of xy, i.e. Σxy
- Get x^2 and y^2 for all values of x and y and then get their totals, i.e. $\Sigma x^2 y^2$
- Put these values to the formula and get the r.



LET US KNOW



Carl Pearson, later known as Karl Pearson (1857-1936) was born to William Pearson and Fanny Smith, was also an accomplished historian and Germanist. Karl Pearson was educated privately at University College School, after which he went to King's College, Cambridge in 1876 to study mathematics. He then spent part of 1879 and 1880 studying medieval and 16th century German literature at the universities of Berlin and Heidelberg. He graduated from Cambridge University in 1879 as Third Wrangler in the Mathematical Tripos. In 1911 he founded the world's first university statistics department at University College London. Besides, correlation coefficient, chi-square is another significant tool contributed by Pearson.



EXERCISE 4.1

A) Given the following pairs of values:

Capital employed (lakhs of Rs.)	12	13	15	16	18	29	32
Profits (lakhs of Rs.)	6	5	7	8	12	11	17

a) Draw a scatter diagram.

b) Do you see any correlation between capital employed and volume of profits? What can be concluded about their correlation if it exists?

B) Calculate Pearson's coefficient of correlation for following data:

X:	6	8	12	15	18	20	24	28	31
Y:	10	12	15	15	18	25	22	26	28

4.8 UNDERLYING ASSUMPTIONS

The most fundamental assumption of Pearson's correlation coefficient is that it supposes that the variables are quantitative. Further, it assumes a linear relation between the two variables. Therefore, the value of 'r' indicates extent of linearity that exists between the two variables. In case, the value of 'r' is found to be equal to 'zero', it may be best understood as 'absence of linear relations' among the variables. It may, however, be possible that the variables have other non-linear types of relations.

4.9 INTERPRETATION OF CORRELATION COEFFICIENT

Once the value of 'r' is obtained, it is very important to correctly interpret the result. Interpretation of the value of 'r' depends on two things– sign and magnitude. When 'sign' is positive correlation is said to be positive, and when 'sign' happens to be negative correlation is also said to be negative.

On the other hand, the magnitude of the value of correlation coefficient can not exceed the value of 1. The value of 'r', therefore, will always lie in between –1 and +1. More closer is the value to 1, greater is the

extent of linear correlation. Conversely, further away is the value from 1 greater is amount of deviation from (linear) correlation. Many scholars provide ranges of values of 'r' for meaningful interpretation. Most commonly it is held that–

When the value of 'r' is 1 it is called perfect correlation. For value ranging from 0.8 to 1, variables are said to be "highly correlated". When the value lies in between 0.6 to 0.8, variables are called "fairly correlated". For value of 'r' in between 0.3 to 0.6, it is called moderate correlation. For values of 'r' less than 0.3 it is called low correlation. When the value of 'r' is equal to 0, the variables are said to be linearly uncorrelated or linearly independent.



CHECK YOUR PROGRESS

Q.2: Choose the correct answer from the following:

i) Correlation analysis is based on the assumption that–

a) Variables are quantitative

b) Variables are qualitative

.....

ii) Interpretation of r depends on–

a) Only sign b) Only magnitude

c) Both sign and magnitude

.....

iii) When value of r falls within the range of 0.6 to 0.8 it is called–

a) Fair Correlation b) Moderate Correlation

c) Low Correlation c) Perfect Correlation

.....

iv) In case of perfectly positive correlation, the value of r is

a) –1 b) +1 c) 0.8 d) 0

.....

4.10 RANK CORRELATION COEFFICIENT

Many a times, we come across variables, which can not be quantitatively measured. However, these can be 'ranked' and accordingly ‘ordered’. For example, beauty, intelligence, honesty etc. In fact, most of the qualitative attributes are of this type. It is difficult to measure exact amount of honesty that a person has, or exact amount of intelligence that he possesses. It is, nevertheless, very much possible to find out which person is more honest, or intelligent or beautiful so on and so forth. When we try to do such comparisons, we in fact, attach some ranks to individual and then try to arrange in some order depending on the ranks. Such variables which can’t be directly measured but can be ordered are called ordinal variables. When we have a bivariate distribution involving two such ordinal variables then how do we find correlation? In this case Pearson’s correlation coefficient can not be applied as it is based on the fundamental assumption of quantitative variable.

Charles Edward Spearman, a British psychologist, developed a formula in 1904 for obtaining a coefficient of correlation of ranks of two attributes. To obtain the correlation of ranks the two attributes are to be ordered and then their differences of ranks are to be calculated. If say ‘d’ is the difference of ranks between the two attributes then the rank correlation coefficient is given by–

$$\sigma = 1 - \frac{6\sum d^2}{n(n^2-1)}$$ n is the number of observation

Procedure to obtain the rank correlation coefficient is fairly simple and easy:

- Arrange the two attributed in same order by their ranks.
- Subtract each pair of ranks to get their differences, i.e. d and then get d^2.
- Get the total of d^2 i.e. Σd^2.
- Put the values thus obtained in the formula and get the coefficient.

Just like the value of r, the value of also lies in between –1 and +1. Interpretation is also same as r except the foundational assumption of linearity.



LET US KNOW



Charles Edward Spearman, (September 10, 1863 - September 17, 1945) was an English psychologist known for work in statistics. He was a pioneer of factor analysis. He also did seminal work on models for human intelligence. Although Spearman achieved most recognition in his day for his statistical work, he regarded this work as subordinate to his quest for the fundamental laws of psychology, and he is now similarly renowned for both.



CHECK YOUR PROGRESS

Q.3: Choose the correct answer from the following:

i) Rank Correlation analysis is used when

a) Variables are quantitative

b) Variables are qualitative

.....

ii) Spearman's rank correlation does not consider linearity

a) True b) False

.....

iii) Just like the value of r, Rank correlation also falls within +1 and –1

a) True b) False

.....



ACTIVITY 4.3

Ranking of 10 students by two teachers A and B are shown below. Obtain the rank correlation coefficient for the data and figure out the problem in decision making

A:	3	5	8	4	7	10	2	1	6	9
B:	6	4	9	8	1	2	3	10	5	7

.....

.....



4.11 LET US SUM UP

In this unit, we have discussed the following:

- Depending on the number of variables, we have univariate, bivariate or multivariate distributions.
- Correlation and regression are associated with bivariate and/or multivariate distributions
- Correlation is the relation between two variables.
- Scatter diagram is a simple graphical method of identifying correlation. For correlated variables scatter plot exhibits some observable pattern or trend.
- Correlation can be of two types– positive when variables move in the same direction and negative when the variables move in the opposite directions.
- Correlation coefficient is based on assumptions of that variables are quantitative and linearly related.
- Interpretation of the correlation coefficient depends on both sign and magnitude of the value of r.
- Rank correlation is applied for qualitative variables which are not measurable but ordinal.
- Like Pearson's correlation coefficient, Spearman's rank correlation coefficient also lies in between +1 and –1.



4.12 FURTHER READING

1) Agarwal, D. R. (2006). *Business Statistics*. Delhi: Vrinda Publications.
2) Gupta S. C. (1994). *Fundamentals of Statistics*. New Delhi: Himalayan Publishing House.
3) Rajagopalan, S. P. & Sattanthan R. (2009). *Business Statistics and Operations Research*. New Delhi: Tata McGraw-Hill
4) Sharma, J. K. (2007). *Business Statistics*. New Delhi: Pearson Education Ltd.
5) Verma, A. P. (2007). *Business Statistics*. Guwahat: Asian Books Private Limited.



4.13 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: i) c) Both bivariate and multivariate series
ii) a) Correlation
iii) b) Negative Correlation

Ans. to Q. No. 2: i) a) Variables are quantitative
ii) c) both sign and magnitude
iii) a) Fair Correlation
iv) b) +1

Ans. to Q. No. 3: i) a) Variables are quantitative
ii) a) True
iii) a) True



4.14 MODEL QUESTIONS

Q.1: What is meant by correlation? What are the properties o the coefficient of correlation?

Q.2: What are the methods of calculating coefficient of correlation?

Q.3: Calculate Karl pearson's coefficient correlation from the data given below

X :	2	4	6	8	10
Y:	20	18	16	14	12

Q.4: Find Karl Pearson's coefficient of correlation between x and y from following data giving test scores of 10 candidates in mathematics and statistics and interpret:

Scores in Mathematics:	98	70	40	20	85	75	95	80	10	5
Scores in Statistics:	85	65	32	30	80	60	61	55	54	65

Q.5: Find the coefficient of correlation from the following data give below:

X:	12	20	15	18	33	24	30	12	15	22
Y:	30	35	28	36	29	39	30	25	30	38

*** ***** ***

UNIT 5: REGRESSION

UNIT STRUCTURE

5.1 Learning Objectives

5.2 Introduction

5.3 Meaning and Concept of Regression

5.4 Linear Regression

5.5 Line of Regression and Regression Equation

5.6 Coefficient of Regression

5.7 Relation between Correlation and Regression Coefficients

5.8 Correlation and Regression Analysis: A Comparison

5.9 Let Us Sum Up

5.10 Further Reading

5.11 Answers To Check Your Progress

5.12 Model Questions

5.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- explain the concept of regression and regression, lines and equation of regression, and their properties
- discuss different types of regression
- explain the essential features of correlation and regression analysis
- explain and compare the various aspects of regression analysis.

5.2 INTRODUCTION

In the previous unit, we introduced the concept correlation between two variables where it was discussed whether the variables move in the same direction or in the opposite direction and the extent of association between the two variables under study. On the other hand regression analysis is a statistical technique that expresses the relationship between two or more variables in the form of an equation to estimate the value of a variable, based on the given value of another variable. In business, several times it

becomes necessary to have some forecast so that it can take a decision regarding a product or a particular course of action. In order to make a forecast, one has to ascertain some relationship between two or more variables relevant to a particular situation. For example, a company is interested to know how far the demand for television sets will increase in the next five years, keeping in mind the growth of population in a certain town. Here, it clearly assumes that the increase in population will lead to an increased demand for television sets.

5.3 MEANING AND CONCEPT OF REGRESSION

Regression is one of the most important and powerful statistical techniques widely used in both science and social science studies. The term regression was introduced by Francis Galton. In a famous scientific paper ‘‘Family Likeness in Stature’’ published in Proceedings of Royal Society, London in 1886, he first talked about ‘‘regression to mediocrity’’.

Today, broadly speaking, regression analysis is referred to the study of the dependence of one variable on one or more other variables, with a view to estimating and/or predicting the average value of the former in terms of fixed or known values of the latter.

To understand consider this - suppose x and y are two variables such that values of y depends on values of x. Then regression analysis will try to estimate or predict the average value of y that it will take for any given value of x.

5.4 LINEAR REGRESSION

We have understood that regression is the technique of estimating average value of the dependent variable corresponding to any given value of the independent variable when the two variables are dependent upon each other i.e. the two variables are ‘related’ to each other. Now we know that relation may be either linear or non-linear. Linear regression is the technique of regression when variables are linearly related i.e. for each successive change in the independent variable there is a ‘constant’ change in the dependent variable.



LET US KNOW

A dependent variable is a variable that depends on value of some other variable. Independent variable is a variable that does not depend on values of the other variable.

Linearity on the other hand means that when independent variable changes the dependent variable changes by constant rate. Such changes produce a straight line when plotted.



ACTIVITY 5.1

Try to list five dependent variables and five independent variables. Also find out which depends on what.

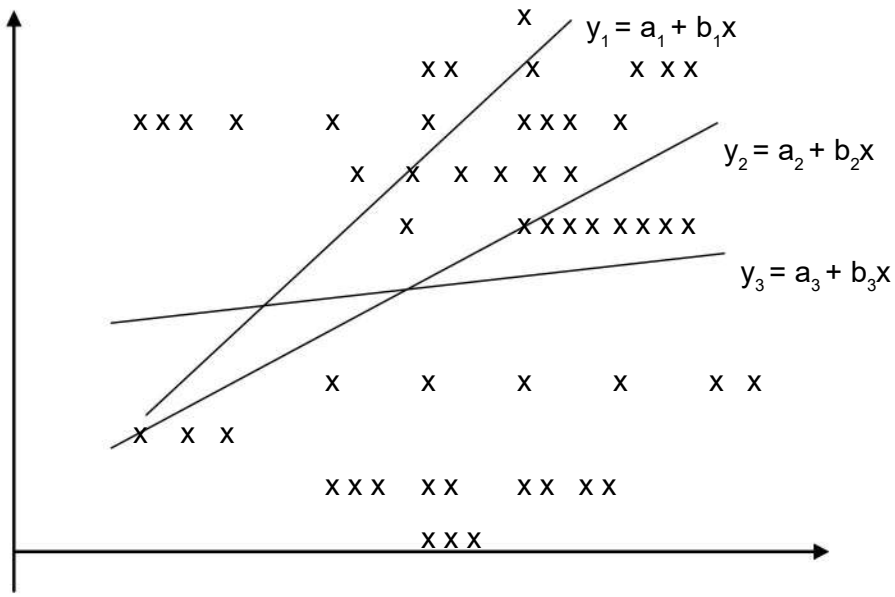
.....

.....

5.5 LINE OF REGRESSION AND REGRESSION EQUATION

Let us consider two variables *x* and *y*, which are linearly related. Relation is such that value of *y* depends on value of *x*. The relation may be expressed as y = a + bx, where *a* and *b* are two constants usually called parameters.

Note that value of *y* now depends on values of *a* and *b* given the value of *x*. For different values of the parameters *a* and *b*, we will have different lines that will pass through the points showing pairs *x* and *y*. Look at the following figure:



In the figure you will notice that three lines passing through the points produces three different values of *y* depending on three different values *a* and *b*. Now since all three values of *y* are 'estimated' value therefore there will be an 'error' of estimation. Error is simply the difference between the 'actual' and 'estimated' values of *y*. That is:

$e = y - \hat{y}$; where $\hat{y}$ is the estimated value of y.

Ideally it would be better if 'error' *r* be zero or minimum. Now since $\hat{y}$ depends on values of *a* and *b* we need to estimate *a* and *b* in such a way that *e* becomes minimum. Suppose, we estimate values of *a* and *b* as $\hat{a}$ and $\hat{b}$ so that we have $\hat{y}$ as:

$\hat{y} = \hat{a} + \hat{b}x$ and resultant *e* is minimum.

Then this equation is called the regression equation and the line that this equation produces is called the line of regression or simply regression line. In other words regression equation is the equation that estimates dependent variable in such a way that the difference between the estimated and actual value of *y* or error in short, is minimum.

The line of regression often interpreted as line of "best fit". Among the all possible lines (look at the figure above showing three such lines) the line of regression fits i.e. touches most of the points (and that is why error is minimum).

When dependent and independent variables are clearly identified then we have single line of regression. When it is not clear will have two lines of regression one showing *y* as dependent and other showing *x* as dependent.

It is useful to remember that in case the variables are perfectly related i.e. if correlation coefficient is either –1 or +1 both the lines will coincide. When the variables are linearly independent regression lines will be perpendicular to each other. Also, both the lines of regression pass through the mean values of the variables.

5.6 COEFFICIENT OF REGRESSION

Consider the regression equation:

$$\hat{y} = \hat{a} + \hat{b}x$$

In this equation $\hat{a}$ is called 'constant' term of the regression equation and $\hat{b}$ is called the 'coefficient' of regression. How can one obtain the $\hat{b}$? In fact, $\hat{b}$ is simply given by

$$\hat{b} = \frac{Cov(x,y)}{\sigma_x}$$

Formal derivation of $\hat{b}$ involves partial differentiation. However, the following formula is used to estimate the regression coefficient–

$$\hat{b} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

Once the $\hat{b}$ is obtained then the constant can be obtained by solving the following equation:

$$\bar{y} = \hat{a} + \hat{b}\bar{x}$$

Where $\bar{y}$ and $\bar{x}$ are means of *y* and *x* respectively. The regression coefficient indicates the amount of variation in dependent variable explained by the independent variable.

5.7 RELATION BETWEEN CORRELATION AND REGRESSION COEFFICIENTS

It may be seen that both coefficient of regression and coefficient of correlation are related to each other. We know that:

$$r = \frac{Cov(x,y)}{\sigma_x \sigma_y}; \text{ and}$$

$$\hat{b} = \frac{Cov(x,y)}{\sigma_x}$$

It is thus clear that–

$$r = \frac{\hat{b}}{\sigma_y}$$

This relation is practically very useful for analysis. Remember that, in case we don't know clearly which one is dependent and which one is independent variable then we will have two lines of regression and hence two coefficients of regression. Let us assume that the two possible coefficients of regression are $\hat{b}$ and $\tilde{b}$ respectively. Then they satisfy the following property–

$$r^2 = \hat{b} \times \tilde{b}$$

This follows from the above property that both regression coefficients have same sign and if value of any one is greater than unity value of the other must be less than unity.

5.8 CORRELATION AND REGRESSION ANALYSIS: A COMPARISON

Correlation analysis simply detects whether two variables are having any relations or not. Such detection simply looks for possible movements in variables and then some identifiable patterns in the movements. On the other hand, regression analysis tries to estimate or predict average value of one variable depending on given values of the other variable. It also tries to explain the extent of variation caused by a variable. In this sense regression analysis has greater explanatory abilities than correlation analysis.



CHECK YOUR PROGRESS

Q.1: Choose/Tick the correct answer from the following:

A) Regression analysis tries to estimate ..... values.

a) Average b) Median

c) Mode d) Range

.....

B) The line of regression is called line of ..... fit.

a) Worst b) Highest

c) Best d) Lowest

.....

C) Regression equation estimates the values such that error is–

a) Minimum b) Maximum

c) Normal d) Abnormal

.....

D) When correlation is 0, then the angle between the lines of regressions–

a) 90^0 b) 30^0 c) 45^0 d) 0^0

.....

E) Lines of regression pass through ..... values of the variables.

a) Maximum b) Minimum

c) Mean d) Median

.....

F) Regression coefficients are of opposite sign.

a) True b) False

.....

G) Both the regression coefficients can be greater than unity.

a) True b) False

.....



5.9 LET US SUM UP

In this unit, we have discussed the following:

- Regression is a method of estimating average value of dependent variable given the value of independent variable.
- Line of regression gives the line of best fit and regression equation defines the line of regression. The line of regression passes through the means of the variables.
- The regression coefficient indicates the amount of variation in dependent variable explained by the independent variable.
- Correlation analysis simply detects whether two variables are having any relations or not. On the hand regression analysis tries to estimate or predict average value of one variable depending on given values of the other variable.



5.10 FURTHER READING

1) Agarwal, D. R. (2006). *Business Statistics*. Delhi: Vrinda Publications.
2) Gupta S. C. (1994). *Fundamentals of Statistics*. New Delhi: Himalayan Publishing House.
3) Rajagopalan, S. P. & Sattanthan R. (2009). *Business Statistics and Operations Research*. New Delhi: Tata McGraw-Hill
4) Sharma, J. K. (2007). *Business Statistics*. New Delhi: Pearson Education Ltd.
5) Verma, A. P. (2007). *Business Statistics*. Guwahat: Asian Books Private Limited.



5.11 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: A) (a) Average, B) (c) Best, C) (a) Minimum, D) (a) 90^0, E) (c) Mean, F) (b) False, G) (b) False



5.12 MODEL QUESTIONS

Q.1: Explain the concept of regression and bring out its utility in business and commerce.

Q.2: What is a regression line? Why can we have two lines of regression? Explain the relationship between lines of regression and correlation coefficient.

Q.3: Why the regression line is called line of best fit? What do you understand by line of best fit?

Q.4: What is regression coefficient? What does it indicate?

Q.5: Fit a line of regression to following data, taking Y as dependent variable and estimate Y for X is equal to 20.

Y:	1	2	4	5	7	8	9
X:	1	3	4	8	9	11	14

Q.6: If followings are the two lines of regression find out the correlation coefficient and means of x and y

$8x - 10y + 66 = 0$ and $40x - 18y = 214$

Q.7: Distinguish between correlation and regression analyses.

*** ***** ***

UNIT 6: FUNDAMENTALS OF PROBABILITY

UNIT STRUCTURE

6.1 Learning Objectives

6.2 Introduction

6.3 Random Experiment

6.4 Classical Definition of Probability

6.5 Set

6.5.1 Subset

6.5.2 Sample Space

6.6 Axiomatic Definition of Probability

6.7 Elementary Theorems on Probability

6.8 Let Us Sum Up

6.9 Further Reading

6.10 Answer to Check Your Progress

6.11 Model Questions

6.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- explain the concept of random experiment
- discuss the concept of events
- define probability
- explain the concept of sample space
- explain limitations of classical and statistical definition and need for axiomatic definition
- calculate probabilities of events
- define some elementary theorems on probability
- explain the concept of independent events
- evaluate probabilities of independent events.

6.2 INTRODUCTION

The subject probability began in the 17th century through efforts of some mathematician (to quote Fermat and Pascal) to answer questions

concerning games of chance. It was not untill the 20th century that a rigorous mathematical theory based on axioms, definitions and theorems was developed.

All phenomena in nature and society are either deterministic or probabilistic. Till the 18th century only deterministic phenomena were amenable to scientific treatment, but during the last two centuries a great deal of research has been carried out on phenomena involving elements of chance.

As time progresses, theory of probability found its way into many applications, not only in mathematical sciences also in wider fields like agriculture, business, medical science etc. All these applications contributed to further development of the theory.

The term probability or chance we use in our daily talk to indicate the degree of personal belief. We often say “there is no probability of raining today”. The statement just reflects the speaker’s belief without any base on experiment or trial. But in statistics the term is used in different way which is more reliable than the former one.

The theory of probability can be discussed in two different stages. The first one is classical approach. In this approach, probability is defined as the ratio of number of favourable cases when the cases are exhaustive, equally likely and mutually exclusive. The second one is modern approach. It is based on concept of sample space, which is again based on set.

6.3 RANDOM EXPERIMENT

If we perform certain experiment under identical conditions we expect to arrive at results that are essentially same provided we can control the value of the variable that may effect the outcome of the experiment. But in many cases we are not in a position to control the value of some variables. As a result, though under identical conditions experiments are performed, the result will vary from one experiment to the others. These experiments are called random experiments.

Examples of random experiments:

1) Let the experiment be tossing of a coin. The result will be either head (H) or tail (T). But we cannot exactly predict what the result will be. The result will depend on chance. There are various factors that will influence the result and all these factors cannot be controlled. This is an example of random experiment. If coin is tossed twice, the result will be {(H, H), (H, T), (T, H), (T, T)}. This is also example of random experiment.
2) Let us consider the random experiment of throwing a six faced cubical die. As the die is perfect, we are sure that one of the faces will come up with one of the numbers, 1, 2, 3, 4, 5, 6.

Event: An event is a collection of possible outcome of a random experiment. An event may consist of a single outcome or a group of outcomes taken together. If it consists of single outcome then it is called elementary event. An event will be denoted by capital letter of English alphabet.

A series of events $A_1, A_2 \ldots, \ldots, \ldots, A_K$ will be called exhaustive if at least one of them is sure to happen in any trial of random experiment. For example, when a coin is tossed, either head (H) or tail (T) must occur. Hence the event is exhaustive.

Two events A and B are said to be mutually exclusive, if the occurrence of one precludes the occurrence of the other. In other words the two events cannot occur simultaneously. For example in casting a die if 6 occurs then other numbers 1, 2, 3, 4 and 5 cannot occur. Hence it is an example of mutually exclusive events.

A series of events $A_1, A_2 \ldots, \ldots, \ldots, A_K$ are said to be equally likely, if one of them cannot be expected to occur in preference to the others in a single trial of random experiment. For example, in tossing of a coin each face is equally likely to occur.

6.4 CLASSICAL DEFINITION OF PROBABILITY

Consistent with the conditions of a random experiment let ‘n’ be exhaustive, mutually exclusive and equally likely cases and ‘m’ of them are

favourable to an event A, then probability of A is defined by 'm/n' and is denoted by 'P (A)' or 'p' so that–

$$P(A) = \frac{m}{n}$$

Since $0 \leq m \leq n$, P(A) lies between 0 and 1.

1) When there is no elementary event favourable to A, then P(A) = 0, since m = 0.
2) When all events are favourable to A then P (A) = 1.

Example 1: Let a six faced die be cast. Find the probability that (i) the number shown on the die is odd (ii) number shown on the die is divisible by 2 and 3 (iii) number shown on the die is divisible by 2 and 5.

Solution: When die is cast there are six cases 1, 2, 3, 4, 5 or 6. These are exhaustive, mutually exclusive and equally likely.

i) Of the cases the favourable cases are 1, 3, 5. Let the event be A.

$$P(A) = \frac{3}{6} = \frac{1}{2}$$

ii) In the second case let the event be B then there is only one favourable case viz. 6.

$$P(B) = \frac{1}{6}$$

iii) In this case let the event be C. Out of the numbers 1, 2, 3, 4, 5, 6 no one is divisible by 2 and 5 simultaneously. Hence no case is favourable to C.

$$P(C) = 0$$

Example 2: A card is drawn at random from a well shuffled pack of cards. Find the probability that it is (i) a king (ii) a queen of spade (iii) a heart.

In a pack of cards there are 52 cards and any card can be drawn and hence total number of cases is 52.

i) Let A denote the first event. There are 4 kings. So number of favourable cases for drawing a king is 4.

$$P(A) = \frac{4}{52} = \frac{1}{13}$$

ii) Let B denote the second event. There are only one queen of spade.

$$P(B) = \frac{1}{52}$$

iii) Let C denote the third event. There are 13 hearts. So favourable cases to C is 13.

$$P(C) = \frac{13}{52} = \frac{1}{4}$$

Limitations of classical definitions: The classical definition requires that n is finite. But there are instances where it may be infinite Secondly we assume elementary events to be equally likely. This is also need not be true always. For example, in tossing of a coin, there are two events head and tail. Unless the coin is unbiased they may not be equally likely. To get a perfectly unbiased coin is a very difficult condition. Hence probability of obtaining a head with any coin cannot be obtained from classical definition.

Statistical definition: To overcome the difficulties encountered in classical definition, a large number of trials of random experiment is considered. Let m be the number of occurrences of an event A associated with n independent trials of random experiment. The probability of the event A is defined by–

$$P(A) = \lim_{n \to \infty} \frac{m}{n}$$

Example: In tossing of a coin 1000 times head comes up 537 times. The probabiity of head coming up is–

$$\frac{537}{1000} = 0.537$$

This definition of probability also has drawback, because the large number involved in definition is vague.

As both the definitions suffer from certain defects another approach has been put forward to overcome all defects. This approach is known as axiomatic approach of probability.

Before considering this approach we shall introduce a concept, which is as Set. We shall briefly state the necessary results here.

6.5 SET

Set is a concept. It has no definition. Inituitively, a set is any well-defined collection of objects. The objects in a set can be anything. The

numbers, letters, people etc. These objects are called elements or members of the Set.

Examples of Sets:

1) The numbers 19, 23, 29 and 31.
2) The vowels of English alphabet : a, e, i, o and u.
3) The states of North-East India.
4) The girl students Anjana, Binita, Chandana, Deepika and Lina.
5) The district head quarters of Assam.

Sets are normally denoted by capital letters and their elements by small letters.

If A denotes the Set given in example 1 then we write,

$$A = \{19, 23, 29, 31\}$$

Elements of a Set are separated by commas and enclosed in { }. This form of representing a Set is called tabular form. This is also called Roster method.

There is another way of representing a Set. Look at the example 3. Here we have not written the names of the states. Let B denote this Set. Then if x be a state.

$$B = \{x : x \text{ is a state of North-East India}\}$$

This form is Set builder form. This is also called Rule method.

If an object x is a member of a Set A, then we write $x \in A$ [Read as x belongs to A]. If A does not contain x then we write $x \notin A$.

Defining set: A Set is finite if it consists of specific number of different elements, otherwise set is infinite. For example Set of months in a year is finite set. It consists of 12 elements. Set of stars in the sky is infinite it cannot be counted. Set of rivers on the earth is **finite**. Though it is difficult to count still it can be counted. Set of natural number is **infinite Set**.

Universal sets: All Sets under investigation will likely be subsets of a fixed set. This fixed set is universal set and denoted by U.

Equality of sets: Two sets A and B are said to be equal if they have the same members. In other words, every elements which belongs to A also belongs to B and every element which belongs to B also belongs to A, then A and B are equal (A = B).

A set does not change if its elements are repeated.

Let A = {5, 6, 7, 5, 8} and B = {7, 5, 7, 6, 7, 8}. Then A = B.

Null Set: It is mathematically convenient to introduce the concept of null set (Empty Set). A Set which contains no elements is a null set. It is denoted by the symbol $\varnothing$.

6.5.1 Subsets

If every element of a Set A is also an element of another Set B, then A is a subset of B. Symbolically it is denoted by $A \subset B$.

If $x \in A \Rightarrow x \in B$, then $A \subset B$.

Let A = {a, b, c} and B = {a, b}

Then $B \subset A$ and A will be a super set of B.

From definitions of subsets and equality of sets we see that if $A \subset B$ and $B \subset A$ then A = B.

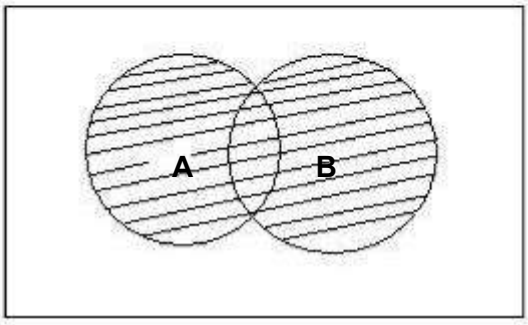
Basic Set operations:

1) **Union of sets:** The union of sets A and B is the set of all elements which belong to A or to B or to both and is denoted by $A \cup B$.

Let A = {1, 2, 3, 4}, B = {2, 4, 6, 8} and C = {1, 3, 5, 7}

Then $A \cup B$ = {1, 2, 3, 4, 6, 8}, $B \cup C$ = {1, 2, 3, 4, 5, 6, 7, 8}

In Venn diagram notation:



$A \cup B$ is the shaded portion.

$A \cup B = \{x : x \in A \text{ or } x \in B\}$

A and B are always subsets of $A \cup B$.

$A \subset A \cup B$ and $B \subset A \cup B$



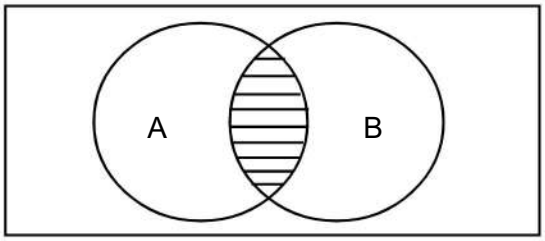
LET US KNOW

Venn diagrams or set diagrams were invented by British mathematician John Venn in 1880. These diagrams show all possible logical relationship between and among different sets that share something in common by using circle.

2) **Intersection of Sets :** The intersection of sets A and B is the set of elements common to A and B and is denoted by $A \cap B$. In the previous example A = {1, 2, 3, 4}, B = (2, 4, 6, 8} and C = {1, 3, 5, 7}.

$A \cap B = \{2, 4\}$, $A \cap C = \{1, 3\}$ and $B \cap C = \varnothing$

In Venn diagram notation,



The shaded portion is $A \cap B$.

$A \cap B$ is subsets of both A and B.

$A \cap B \subset A$ and $A \cap B \subset B$

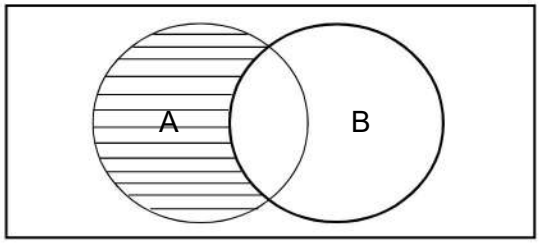
$A \cap B = \{x : x \in A \text{ and } x \in B\}$

If A and B have no common elements then $A \cap B = \varnothing$

In that case A and B are said to be disjoint.

3) **Difference of sets:** The difference of sets A and B is the set of elements which belong to A but not to B and is denoted by A – B.

In the previous example, A – B = {1, 3}



In Venn diagram notation,

Shaded portion is A – B.

Similar definition for B – A

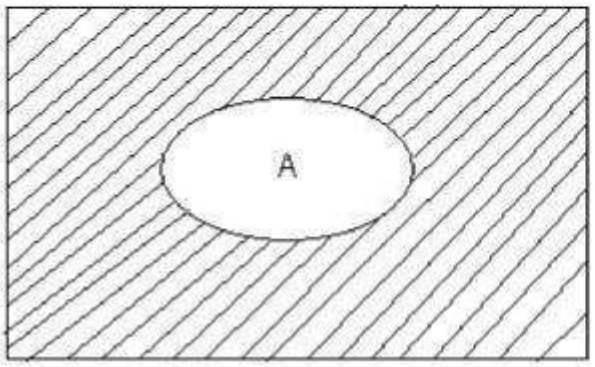
$A - B = \{x : x \in A \text{ and } x \notin B\}$

$A - B \subset A$

4) Complement of a Set: The complement of a Set A is the set of elements which do not belong to A, ie difference of universal set U and A and denoted by A^C.

$$A^C = U - A$$

In Venn diagram notation:



Shaded portion is A^C.

With this background of set we are in a position to define which is known as sample space, which plays significant role in modern development of theory of probability.

Illustrative examples on sets :

1) Let us consider the sets

   $A = \{1, 2, 3, 4\}$ and $B = \{3, 1, 4, 2\}$

   Each elements 1, 2, 3, 4 of A belongs to B

   $$A \subset B$$

   Again each elements 3, 1, 4, 2 of B belongs to A

   $$B \subset A$$

   Hence A = B

2) $A = \{x : 2x = 6\}$ and b=3. Does b belongs to A?

   $A = \{x : 2x = 6\} \Rightarrow A = \{b\}$ when x = 3

   So, $b \in A$

3) Write true and false of the following statements

   i) $8 \in \{1, 3, 5, 7, 9\}$

   ii) If $A = \{1, 4, 6\}$ and $B = \{1, 8\}$ here $A \cap B = \{1\}$

iii) $A \cup A = A$

iv) Set of even numbers is finite

v) The set of letters in the word "follow" and the set of letters in the word "wolf" are different.

Answers: (i) False, (ii) True, (iii) True, (iv) False, (v) False

4) Find union, intersection and diference of the sets

A = {1, b, c, d} and B = {f, b, d, g}

Solution: $A \cup B = \{a, b, c, d, f, g\}$, $A \cap B = \{b, d\}$,

$A - B = \{a, c\}$

5) Let U = {1, 2, 3 ..., ..., ..., 8, 9} and A = {2, 4, 6, 8}

Solution: $A^c = \{1, 3, 5, 7, 9\}$



CHECK YOUR PROGRESS

Q.1: A = {1, 2, 3, 4} and B = {1, 4, 9, 16}.

Fill up the following by $\in$ or $\notin$.

i) 3 ..... A ii) 5 ..... B

iii) 8 ..... B iv) 4 ..... B

Q.2: Which of the following sets are equal : {1, 2, 3}, {3, 2, 1, 3}, (2, 3, 2, 1} and {2, 1, 2}

Q.3: Find all the subsets of {0, 1, 2} and {x, y, z}

Q.4: Find union, intersection and difference of A and B.

Where A = {1, 2, 3, 4, 5} and B = {2, 5, 7}

Q.5: Write True or False of the following:

i) Months of the year in a finite set.

ii) A null set is the set {0}

iii) x = {a, b}, Y = {a, b, c} and Z = {a, b, d}. Then

a) $Z \subset X$ b) $X = Z$ c) $Y \subset X$

iv) A = {1, 0}, then $\{0) \subset A$

v) A = {1, 4, 6} and B = {1, 8} then

a) $A \cup B = \{1, 4, 8\}$ b) $A \cap B = \{1\}$

c) $A - B = \{4, 6, 8\}$

6.5.2 Sample Space

A set S that consists of all possible outcome of a random experiment is called a sample space. Each outcome is a sample point.

Example: Let us consider the result of casting a six-faced die. The outcome will be the appearance of one and only one of the numbers 1, 2, 3, 4, 5, 6, on the upper most face of the die. In other words the six possibilities are such that no two or more of them can occur simultaneously and at least one of them must occur. The space S is {1, 2, 3, 4, 5, 6}. If A denote the event "occurrences of even number" then A = {2, 4, 6} which is a subset of S.

Let us generalise the concept.

Let all possible outcomes of some particular experiment be e_1, e_2, ..., ..., ..., e_n which are such that no two or more of them can occur simultaneously and exactly one of them must occur when the experiment is performed. Any set associated with an experiment which satisfies the above mentioned two properties is called a sample space. The elements or points of the sample space associated with an experiment are called elementary events of the experiment.

If a sample space has finite number of points then it is a finite sample space. If it has many points as there are natural numbers then it is a countably infinite sample space. If it has as many points as there are in $0 \leq x \leq 1$ then it is an infinite sample space.

Examples:

1) If a coin is tossed twice, the four possible results are {HH, HT, TH, TT} which is the sample space. The event that only one head comes up will be {HT, TH} which is a subset of in sample space.
2) From an urn containing 4 balls of different colours red (R), Blue (B), White (W) and Green (G) draw two balls simultaneously. The sample space for this experiment is {RB, RW, RG, BW, BG, WG}.

Note: 1) RB and BR represent the same outcome because we draw balls simultaneously.

Instead of balls drawing simultaneously, let us now draw balls in successions with replacement. The sample space will be {RR, RB, RW, RG, BR, BB, BW, BG, WR, WB, WW, WG, GR, GB, GW, GG}.

2) In this case RB and BR are not same.

6.6 AXIOMATIC DEFINITION OF PROBABILITY

Let S be a sample space. To each event A in a class C of events, a real number P(A) is associated.

P (A) is called probability of the event A if the following axioms are satisfied.

Axiom 1. For any event A in the class C, $P(A) \geq 0$.

Axiom 2. For certain event S in class C, P(S) = 1.

Axiom 3. For any number of mutually exclusive events $A_1, A_2, ..., ..., ..., A_n$ in the class C,

$P(A_1 \cup A_2 \cup ... ...) = P(A_1) + P(A_2) + ...$

In particular for two events, A_1 and A_2

$$P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

Let S be a sample space $\{e_1, e_2, ..., ..., ..., e_n\}$ and $P(\{e_i\}) = p_i$,

i = 1, 2, ..., ..., ..., n such that $0 \leq P_1 \leq 1$ and $\sum_{i=1}^{n} P_i = 1$

Now $P(\{e_i\}) = \frac{1}{n}$

If E be any event consisting of $r(1 \leq r \leq n)$ elementary events then

$P(E) = \frac{1}{n} + \frac{1}{n} + ...$ r times

$= \frac{r}{n}$

$P(E) = \frac{n(E)}{n(S)}$, where n(E) = number of events in E and

n(S) = number of events in S.

6.7 ELEMENTARY THEOREMS ON PROBABILITY

Some elementary theorems on probability :

Let $S = \{e_1, e_2, ..., ..., ..., e_n\}$ be a sample space of an experiment.

Theorem 1: $P(S) = 1$.

Proof: By definition, $P(S) = P(\{e_1\}) + P(\{e_2\}) + ... + P(\{e_n\})$

$= p_1 + p_2 + ... + p_n$

$= 1$

This theorem is very trivial.

Theorem 2: If A and B are two events such that $A \subseteq B$ then $P(A) \leq P(B)$

Proof: $A \subseteq B \Rightarrow$ B contains all the elementary events of A and possibly a few more.

So, the probability of B is equal to the probability of A plus the sum of the probabilities attached to those elementary events which belongs to B but not to A i.e. to B – A. Since the probability is a non-negative number,

$$P(A) \leq P(B)$$

Theorem 3: For any event A, $0 \leq P(A) \leq 1$

By axiom 1, $P(A) \geq 0$ ... ... (i)

For any event A, $A \subseteq S$

By theorem 2, $P(A) \leq P(S) = 1$ (by theorem 1) ... ... (ii)

$\therefore$ From (i) and (ii), $0 \leq P(A) \leq 1$

Theorem 4: Probability of impossible event is zero.

Let $\varnothing$ denote the impossible event

We know, $S = \cup \varnothing$

$\Rightarrow P(S) = P(S \cup \varnothing) = P(S) + P(\varnothing)$ (by axiom 2)

$\Rightarrow P(\varnothing) = 0$

Let us examine the converse of theorem 1 and theorem 4.

For this let us consider the sample space.

$$S = \{HH, HT, TH, TT\}$$

Let us consider, $P(TH) = \frac{1}{2} = P(HH)$

$$P(HT) = 0\ P(TT)$$

If $A = \{HT, TT\}$ then $P(A) = 0$, but A is not impossible event.

Hence converse of theorem 4 is not true.

Let B = {TH, HH} then P(B) = 1, but B is not sure event.

Hence converse of theorem 1 is not true.

Theorem 5: $P(A_2 - A_1) = P(A_2) - P(A_1)$

We know, $A_2 = A_1 \cup (A_2 - A_1)$, where A_1 and A_2 are mutually exclusive.

$$P(A_2) = P[A_1 \cup (A_2 - A_1)]$$
$$= P(A_1) + P(A_2 - A_1)$$
$$\Rightarrow P(A_2 - A_1) = P(A_2) - P(A_1)$$

Theorem 6: If A^C is the complement of A then $P(A^C) = 1 - P(A)$

We know, $A \cup A^C = S$ and $A \cap A^C = \varnothing$

$A \cap A^C = \varnothing \Rightarrow$ A and A^C are mutually exclusive

Hence, $A \cup A^c = S \Rightarrow P(A \cup A^c) = P(S)$

$$\Rightarrow P(A) + P(A^c) = 1$$
$$\Rightarrow P(A^C) = 1 - P(A)$$

Independent Events: If the probability of B occuring is not affected by the occurence or non-occurence of A then A and B are said to be independent events.

In this case $P(A \cap B) = P(A).P(B)$

Conversely of $P(A \cap B) = P(A).P(B)$ then A and B are independent.

Calculation of probabilities: All these probability theories are used in business to evaluate financial and decision making risk. Probability is used to improve business performnance.

Example 1: In a single cast with two dice find the probability of throwing (i) two aces (ii) doublets (iii) five-six

Solution: Any face of either die may turn up. So sample space consists of 6 × 6 = 36 sample points.

i) Two aces means (6, 6) only one point in sample space.

Probability for the aces = $\frac{1}{36}$

ii) For doublets, the point are (1, 1), (2, 2), (3, 3), (4, 4), (5, 5) and (6, 6), which are six in numbers.

Probability for the doublets = $\frac{6}{36} = \frac{1}{6}$

iii) For five-six, we have two points(5, 6) and (6, 5).

Probability for five-six = $\frac{2}{36} = \frac{1}{18}$

Example 2: Two dice one red and other blue are thrown. Find the probability of the event E where E is–

i) Sum of the number shown by the two dice is 7.

ii) Number on the second dice is greater than the number on the 1st.

Solution: The sample space consists of 6x6=36 elementary events.

i) E_1 = {(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)}

$P(E_1) = \frac{6}{36} = \frac{1}{6}$

ii) E_2 = {(1,2), (1,3), (1,4), (1,5), (1,6), (2,3), (2,4), (2,5), (2,6), (3,4), (3,5), (3,6), (4,5), (4,6), (5,6)}

$P(E_2) = \frac{15}{36} = \frac{5}{12}$

Example 3: A coin is tossed three times in succession. Find the probability of the event that have

i) two or more heads

ii) number of heads equal to the number of tail

iii) exactly one head and two tails.

Solution: The sample space is {(HHH), (HHT), (HTH), (THH), (HTT), (THT), (TTH), (TTT)}

i) E_1 = {(HHH), (HHT), (HTH), (THH)}

$P(E_1) = \frac{4}{8} = \frac{1}{2}$

ii) E_2 = {(HHH), (TTT)}

$P(E_2) = \frac{2}{8} = \frac{1}{4}$

iii) E_3 = {(HTT), (THT), (TTH)}

$P(E_3) = \frac{3}{8}$

Example 4: From a pack of usual playing cards, a card is drawn at random and is noted. Calculate the probabilities of the following events:

i) the drawn card is either a spade or a club.

ii) the drawn card is a picture card (ace is include).

iii) the drawn card is of denomination less than 10 and greater than5.

Solution: The sample space consists of 52 elementary events corresponding to each of the 52 cards that can be drawn from the pack.

i) Let, E_1 be the event that the drawn card is a spade or a club. Then E consists of (13 + 13) = 26 elementary events.

$$P(E_1) = \frac{26}{52} = \frac{1}{2}$$

ii) E_2 be the event that drawn card is a picture card. Here aces are taken as picture card. Then E_2 contains 4 × 4 = 16 elementary events.

$$P(E_2) = \frac{16}{52} = \frac{4}{13}$$

iii) E_3 be the event such that denomination lies between 5 and 10. E_3 consists of 4 × 4 = 16 elementary events.

$$P(E_3) = \frac{16}{52} = \frac{4}{13}$$

Example 5: Two drawings, each of 3 balls, are made from a bag containing 5 white and 8 black balls, the balls not being replaced before the second trial. Find the chance that the first drawing will give 3 white and the second 3 black balls.

Solution: Let E_1 be the event of drawing 3 white balls and E_2 be the event of drawing 3 black balls.

The sample space consists of ${}^{13}C_3$ elementary events for E_1.

$$P(E_1) = \frac{{}^{5}C_3}{{}^{13}C_3}$$

$$= \frac{\frac{5.4}{2.1}}{\frac{13.12.11}{3.2.1}}$$

$$= \frac{\frac{5.4}{2.1}}{\frac{3.2.1}{13.12.11}}$$

$$= \frac{5}{143}$$

For E_2 the sample space consists of 10_{C_3} elementary events.

$$P(E_2) = \frac{8_{C_3}}{10_{C_3}}$$

$$= \frac{\frac{8.7.6}{3.2.1}}{\frac{10.9.8}{3.2.1}}$$

$$= \frac{8.7.6}{3.2.1} \times \frac{3.2.1}{10.9.8}$$

$$= \frac{7}{15}$$

Since E_1 and E_2 are independent,

$$P(E_1E_2) = P(E_1).P(E_2)$$

$$= \frac{5}{143} . \frac{7}{15}$$

$$= \frac{7}{429}$$

Example 6: Two drawings, each of 3 balls, are made from a bag containing 5 white and 8 black balls, the balls being replaced before the second trial find the chance that the first drawing will give 3 white, and the second 3 black balls.

Solution: Let, E_1 be the event of drawing 3 white balls and E_2 be the event of drawing 3 black balls.

Since balls are replaced these sample space for both the events will contained same number of elementary events.

E_1 contains 5_{C_3} elementary events and E_2 contains 8_{C_3} elementary events.

$$P(E_1) = \frac{5_{C_3}}{13_{C_3}}$$

$$= \frac{\frac{5.4}{2.1}}{\frac{13.12.11}{3.2.1}}$$

$$= \frac{5}{143}$$

$$P(E_2) = \frac{8_{C_3}}{13_{C_3}}$$

$$= \frac{\frac{8.7.6}{3.2.1}}{\frac{13.12.11}{3.2.1}}$$

$$= \frac{8.7.6}{13.12.11}$$

$$= \frac{28}{143}$$

Since E_1 and E_2 are independent,

$P(E_1 \cap E_2) = P(E_1).P(E_2)$

$$= \frac{5}{143}.\frac{28}{143}$$

$$= \frac{140}{20449}$$

Example 7: A bag contains 5 white and 7 black balls; if two balls are drawn what is the chance that one is white and the other black?

Solution: The sample space consists of $12_{C_2} = \frac{12.11}{2.1} = 66$ elementary events.

Let E_1 be the event of drawing first white and second black balls.

$$P(E_1) = \frac{5}{12}.\frac{7}{11} = \frac{35}{132}$$

Let, E_2 be the event of drawing first black and second white balls.

$$P(E_2) = \frac{7}{12}.\frac{5}{11} = \frac{35}{132}$$

E_1 and E_2 are mutually exclusive. Therefore,

$P(E_1 \cup E_2) = P(E_1) + P(E_2)$

$$= \frac{35}{132} + \frac{35}{132}$$

$$= \frac{70}{132}$$

$$= \frac{35}{66}$$

Example 8: Four cards are drawn from a pack of 52 cards. What is the chance that one of each suit is drawn ?

The sample space consists of 15_{C_4} elementary events.

Let E be the event of drawing one card from each suit.

$$P(E) = \frac{13_{C_1} \times 13_{C_1} \times 13_{C_1} \times 13_{C_1}}{52_{C_4}}$$

$$= \frac{13 \times 13 \times 13 \times 13}{\frac{52.51.50.49}{4.3.2.1}}$$

$$= \frac{28561}{270725}$$

Example 9: From an ordinary pack of cards a card is drawn at random. What is the probability that it is–

i) 3 of clubs or 6 of diamonds.

ii) any suit except heart.

iii) a jack or a queen.

iv) either a spade, or a heart or not a picture card.

Solution: i) Let, A denote 3 of clubs, and B denote 6 of diamonds.

Now $A \cap B = \varnothing$

P(A B) = P(A) + P(B)

$$= \frac{1}{52} + \frac{1}{52}$$

$$= \frac{1}{26}$$

ii) If H denote heart that H/ will denote any suit except heart.

$$P(H) = \frac{13}{52} = \frac{1}{4}$$

$P(H/) = 1 - P(H)$

$$= 1 - \frac{1}{4}$$

$$= \frac{3}{4}$$

iii) Let, A denote that the card is jack & B that is queen. Now

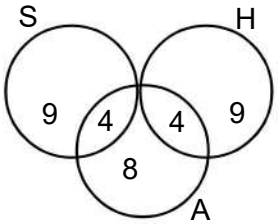
$A \cap B = \varnothing$

$P(A \cup B) = P(A) + P(B)$

$= \frac{4}{13} + \frac{4}{13}$

$= \frac{2}{13}$

iv) Let, S denotes spade, H denotes heart and A denotes picture card. S and H are mutually exclusive but S and A, H and A are not mutually exclusive.



$P(S) = \frac{13}{52}$, $P(H) = \frac{13}{52}$ and $P(A) = \frac{16}{52}$

$P(A') = \frac{36}{52}$, $P(S \cap A') = \frac{9}{52}$

$P(H \cap A') = \frac{9}{52}$

$P(H \cap A') - P(S \cap H) + P(S \cap H \cap A')$

$P(S \cup H \cup A) = \frac{13}{52} + \frac{13}{52} + \frac{36}{52} - \frac{9}{52} - \frac{9}{52}$

$[\because P(S \cap H) = 0, P(S \cap H \cap A') = 0]$

$= \frac{62}{52} - \frac{18}{52}$

$= \frac{44}{52}$

Example 10: Through certain election 3 persons are to be chosen out of five candidates A, B, C, D, E with equal chance being elected. Find the probability that

i) A and B are elected or B and C are elected.

ii) E is elected or C and D are elected.

Solution: Let E_1 be the event that A and B are elected and E_2 be the events that B and C are elected.

S = {ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE}

i) E_1 = {ABC, ABD, ABE}

E_2 = {ABC, BCD, BCE}

$E_1 \cap E_2$ = {ABC}

$P(E_1) = \frac{3}{10}$, $P(E_2) = \frac{3}{10}$, $P(E_1 \cap E_2) = \frac{1}{10}$

$\therefore P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

$P(E_1 \ E_2) = \frac{3}{10} + \frac{3}{10} - \frac{1}{10} = \frac{5}{10} = \frac{1}{2}$

Let E_3 be the event that E is elected and E_4 be the event that B and C are elected.

ii) E_3 = {ABE, ACE, ADE, BCE, BDE, CDE}

E_4 = {ACD, BCD, CDE}

$E_3 \cap E_4$ = {CDE}

$P(E_3) = \frac{6}{10}$, $P(E_4) = \frac{3}{10}$, $P(E_3 \cap E_4) = \frac{1}{10}$

$\therefore P(E_3EE_4) = P(E_3) + P(E_4) - P(E_3CE_4)$

$P(E_3 \ E_4) = \frac{6}{10} + \frac{3}{10} - \frac{1}{10} = \frac{8}{10} = \frac{4}{5}$

Odds for and odds against: Let E be an event. Sometimes we say that the odds for E are a : b. By this we mean

$$P(E) = \frac{a}{a+b}$$

Thus given the odds for an event, the probability of is event can be calculated. Conversely given the probability of an event, the odds for the event can determined. For example

if P(E) = .7 then, $P(E) = \frac{7}{10} = \frac{7}{7+3}$

Hence odds for E are 7 : 3.

If odds for E is a : b then odds against E is b : a.

Composite Event: When an event, comprises of more than one simple event then the event is called a composite event.

The event $A \cup B$ comprises of the sample points in the whole region bounded by A and B. $A \cup B$ is also denoted by A + B. Similarly the event $A \cap B$ is denoted by AB.

Example 11: From a bag containing 4 white and 5 black balls a man draws 3 balls at random; what are the odds against these being all black?

Solution: The total number of ways in which 3 balls can be drawn is 9_{C_3}, and the number of ways of drawing 3 black balls is 5_{C_3}; therefore the chance of drawing 3 black balls

$$= \frac{5_{C_3}}{9_{C_3}} = \frac{5.4.3}{9.8.7} = \frac{5}{42}$$

Thus the odds against the event are 37 to 5.

Example 12: A party of ten take their seats at a round table. What are the odds against two specified persons (A, B) sitting together?

Solution: A having taken his place, B has a choice of 9 places, 2 of which are next to A, hence the odds against B sitting next to A are as 7 : 2.



CHECK YOUR PROGRESS

Q.6: In a single throw with two dice, find the chances of throwing (i) five, (ii) six.

.....

Q.7: From a pack of 52 cards two are drawn at random; find the chance that one is a king and the other a queen.

.....

Q.8: A bag contains 5 white, 7 black and 4 red balls; find the chance that three balls drawn at random are all white.

.....

Q.9: If four coins are tossed, find the chance that, there should be two heads and two tails.

.....

Q.10: One of two events must happen : given that the chance of the one is two-thirds that of the other, find the odds in favour of the other.

.....



ACTIVITY 8.1

You may perform some random experiments like tossing of coin, casting of a die, drawing cards or some other practical field you come across and calculate probabilities. You may verify the accuracy of the result by repeating the experiment.

.....

.....



6.8 LET US SUM UP

In this unit we have discussed the concept of rondom experiment. In certain experiments, value of the variables cannot be controlled. Therefore, the results of these experiments are not the same. These are known as random experiments. Then we have discussed the concept of probability. Another important topic that we have discussed is the set. Set is a well defined collection of objects which may be numbers, letters, people, animal etc.

Ultimately we have given axiomatic definition. We have discussed the concept of probability and their computational procedure. The discussion is extended to composite events. All are discussed through rules and numerical computation.



6.9 FURTHER READING

1) Gupta, S. K. (1968). *Elements of Probability*. NCERT.
2) Medhi, J. (2005). *Statistical Method*. New Age International Publishers.
3) Mukherjee, Kalyan Kumar (1993). *Probability and Statistics*. New Central Book Agency.
4) Murray, R., Spiegel, Jhon, Schiller, R., Srinivasan, Alu (2005). *Probability and Statistics*. Schaum's Outline Series.



6.10 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: i) $\in$, ii) $\notin$, iii) $\notin$, iv) $\in$

Ans. to Q. No. 2: First three sets are equal.

Ans. to Q. No. 3: {0}, {1}, {2}, {0,1}, {0,2}, {1,2}, {0,1,2}, $\varnothing$

{x}, {y}, {z}, {x,y}, {x, z}, [y,z], {x, y, z}, $\varnothing$

Ans. to Q. No. 4: $A \cup B = \{1, 2, 3, 4, 5, 7\}$, $A \cap B = \{2, 5\}$,

$A - B = \{1, 3, 4\}$ $B - A = \{7\}$

Ans. to Q. No. 5: i) True, ii) False, iii) (a) False, (b) False, (c) False,

iv) True, v) (a) False, (b) True, (c) False

Ans. to Q. No. 6: $\frac{1}{9}$

Ans. to Q. No. 7: $\frac{5}{36}$

Ans. to Q. No. 8: $\frac{1}{56}$

Ans. to Q. No. 9: $\frac{3}{8}$

Ans. to Q. No. 10: 2 to 3



6.11 MODEL QUESTIONS

Q.1: There are three events A, B, C one of which must, and only one can, happen; the odds are 8 to 3 against A, 5 to 2 against B; find the odds against C.

Q.2: If two balls are drawn from a bag containing 2 white, 4 red and 5 black balls, what is the chance that

i) they are both red?

ii) one is red and the other black?

Q.3: Two coins are tossed simultaneously. Find the probability of getting–

i) exactly two heads or head is the first coin or both,

ii) getting head in first and load is the 2nd coin or both.

Q.4: A card is drawn at random from a pack of playing cards. Find the probability that it is

i) a ten or a spade.

ii) neither a 4 nor a club.

iii) either a black card or an ace or both.

iv) either a spade or a picture card.

Q.5: Two dice are thrown simultaneously. Find the probability that the sum is

i) divisible by either 2 or 5 but not by both.

ii) does not exceed 6.

iii) greater than 10.

iv) divisible by 2 and 5 both.

Q.6: If A and B are any events prove that

$P(A \cap B) \leq P(A) \leq P(A \cup B) \leq P(A) + P(B)$

when $P(A \cup B) = P(A) + P(B)$

Q.7: If $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{4}$, $P(A \cup B) = \frac{3}{8}$, find P(AB).

Q.8: There are three events A, B, C one of which must and only one can happen. The odds are 8 to 3 against A, 5 to 2 against B. Find the odds against C.

Q.9: A experiment consists of drawing 3 cards in succession from a pack of cards. A_1 is the event king on first draw, A_2 king on the 2nd and A_3 king is the 3rd. State the meaning of

i) $P(A_1 \cap A^{/})$ ii) $P(A_1 \cup A_2)$ iii) $P(A_1^{/} \cup A_2^{/})$

iv) $P(A^{/} \cap A_2^{/} \cap A_3^{/})$ v) $P[(A_1 \cap A_2) \cup (A_2^{/} \cap A_3)]$

*** ***** ***

UNIT 7: CONDITIONAL PROBABILITY

UNIT STRUCTURE

7.1 Learning Objectives

7.2 Introduction

7.3 Conditional Probability

7.4 Multiplication Theorem on Probability

7.5 Total Probability

7.6 Baye's Theorem

7.7 Let Us Sum Up

7.8 Further Reading

7.9 Answers to Check Your Progress

7.10 Model Questions

7.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- define conditional probability
- know multiplication theorem of probability
- define law of probability
- know Baye's theorem.

7.2 INTRODUCTION

In unit 7, we discussed about probbility and some important theorems related to probability. In this unit, we will discuss the concept of conditional probability. We will also discuss dependent and independent events. Multiplication theorems for dependent and independent events will also be discussed in this unit. Finally, we will disuss the law of probability and baye's theorem.

7.3 CONDITIONAL PROBABILITY

In unit 7, we have discussed the methods of finding the probability of events. Supose, we have two events from the same sample space in a

random experiment, does the occurrence of one of the events affect the probability of the other event? Let us see this situation with the following examples

1) Consider a random experiment of throwing a die.

Let S be the sample space and A be the event of of getting an odd numbers.

$\therefore$ S = {1, 2, 3, 4, 5, 6} and A = {1, 3, 5}

$$P(A) = \frac{3}{6} = \frac{1}{2}$$

Suppose we consider another event B such that B be the event of getting the numbers greater than 3. Then B = {4, 5, 6}.

When we say that B has already occurred,the event of getting 'numbers is less than or equal to 3' is not possible. Now, the number of elements of of the sample space is reduced to 3 i.e. {4, 5, 6}. Out of these three points, two points are even number {4, 6}.

Thus, given that if the event B has already occurred, then $P(A) = \frac{2}{3}$.

Hence, the probability of A given that B has already occurred is denoted by $P\left(\frac{A}{B}\right)$.

2) Consider the experiment of tossing three fair coins. The sample space of the experiment is S = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}.

Let E be the event 'at least two tails appear' and F be the event 'first coin shows head'.

Now, suppose we are given that the first coin shows head, i.e. F occurs, then we have to find the probability of occurrence of E. With the information of occurrence of F, we are sure that the cases in which first coin does not result into a head should not be considered while finding the probability of E. This information reduces our sample space from the set S to its subset F for the event E. In other words, the additional information really amounts to telling us that the situation may be considered as being that of a new random experiment for

which the sample space consists of all those outcomes only which are favourable to the occurrence of the event F. Now, the sample point of F which is favourable to event E is HTT.

Thus, Probability of E considering F as the sample space = $\frac{1}{4}$.

Probability of E given that the event F has occurred = $\frac{1}{4}$.

This probability of the event E is called the *conditional probability of E given that F has already occurred*, and is denoted by P (E|F).

3) Consider the random experiment of drawing a card from a deck of 52 cards. Suppose we draw a queen in successive draws. In first draw the probability of drawing a queen is $\frac{4}{52}$. If the card is not replaced before the second draw, then the event of getting queen in the second draw depends on the first event. Now, there will be only 3 queens. The probability of getting a queen in second draw will also be $\frac{3}{52}$.

In the above examples,we observe that the occurance of one event depends on another event which has already happend. Hence,an event E is said to be a dependent event if the occurnce of A depends on another event F which has already happend. It is denoted by P(E|F).

The above examples leads to the definition of conditional probability as:

Let E and F be two events defined on a sample space S.

Let P(F) > 0, then the conditional probability of E given that F has already occurred denoted by P(E|F) and mathematically written as:

$$P(E|F) = \frac{P(E \cap F)}{P(F)}, \quad P(F) > 0$$

Similarly, $P(F|E) = \frac{P(E \cap F)}{P(E)}$, P(E) > 0

Note: 1) P(E|F) and P(F|E) can also be expressed as follows:

$P(E|F) = \frac{n(E \cap F)}{n(F)}$ and $P(F|E) = \frac{n(E \cap F)}{n(E)}$.

2) P(E|F) is read as probability of event E given event F.

Laws of Conditional probability:

a) If E_1 and E_2 be any two events of a sample space S and A be any event of S such that $P(A) \neq 0$, then

$P(E_1 \cup E_2|A) = P(E_1|A) + P(E_2|A) - P(E_1 \cup E_2|A)$

We have, $P(E_1 \ R_2|A)$

$$= \frac{P[(E_1 \cup E_2) \cap A]}{P(A)}$$

$$= \frac{P(E_1 \cap A) \cup P(E_2 \cap A)}{P(A)}$$

$$= \frac{P(E_1 \cap A) + P(E_2 \cap A) - P(E_1 \cap E_2 \cap A)}{P(A)}$$

$$= \frac{P(E_1 \cap A)}{P(A)} + \frac{P(E_2 \cap A)}{P(A)} - \frac{P(E_1 \cap E_2 \cap A)}{P(A)}$$

$= P(E_1|A) + P(E_2|A) - P(E_1 \cap E_2|A)$

b) $P(E'|F) = 1 - P(E|F)$

We have, $P(S|F) = 1$

$\Rightarrow P(E' \cup E|F) = 1$

$\Rightarrow P(E'|F) + P(E|F) = 1$

$\Rightarrow P(E'|F) = 1 = 1 - P(E|F)$

Independent Events: Events are said to be independent if the occurrence of one does not affect the others. In the experiment of tossing a fair coin, the occurrence of the event 'head' in the first toss is independent of the occurrence of the event ' head' in the second toss and third toss.

Example 1: A card is chosen at random from an ordinary deck of 52 cards.

Let E be the event that the cards is an ace and F the event that the card is a club. Then E and F are independent.

Dependent Events: If the occurrence of one event influences the occurrence of the other, then the second event is said to be dependent on the first. In the above example, if we do not replace the first ball drawn, this

will change the composition of balls in the bag while making the second draw and therefore the event of 'drawing a red ball' in the second will depend on event (first ball is red or white) occurring in first draw. Similarly, if a person draw a card from a full pack and does not replace it, the result of the draw made afterwards will be dependent on the first draw.

Example 2: Two fair dice are tossed. Let E be the event that the first die is a 3, F the event that the sum is 6 and G the event that sum is 7. Then E and F are dependent, but E and G are independent.

Note: If the events A and B are independent, that is the probability of occurrence of any one of them P(E|F) = P(E) and P(F|E) = P(F).

7.4 MULTIPLICATION THEOREM ON PROBABILITIES

Now, we discuss multiplication theorem on probabilities for both independent and dependent events.

Multiplication theorem on probabilities for independent events: If two events E and F are independent, the probability that both of them occur is equal to the product of their individual probabilities.

i.e. $P(E \cap F) = P(E).P(F)$.

Note: 1) The theorem can be extended to three or more independent events. If E, F, G, …, …, ... be independent events, then $P(E \cap F \cap G) = P(E)P(F)P(G)$.

2) If E and F are independent then the complements of E and F are also independent. i.e. $P(\overline{E} \cap \overline{F}) = P(\overline{E})P(\overline{F})$.

Multiplication theorem for dependent events: If E and F be two dependent events, i.e. the occurrence of one event is affected by the occurrence of the other event, then the probability that both A and B will occur is $P(E \cap F) = P(E)P(F|E)$.

Note: In the case of three events E, F, G.

$P(E \cap F \cap G) = P(E)P(F|E)P(G|E \cap F)$

i.e., the probability of occurrence of E, F and G is equal to the probability of E times the probability of F given that E has occurred, times the probability of G given that both E and F have occurred.

Ilustrative Examples:

Example 1: Let E and F are two events,such that

$P(E) = 0.8$, $P(F) = 0.6$, $P(E \cap F) = 0.5$.

Find the value of (i) $P(E \cup F)$, (ii) $P(F|E)$, (iii) $P(E|F)$.

Solution: (i) $P(E \cup F) = P(E) + P(F) - P(E \cap F)$

$= 0.8 + 0.6 - 0.5$

$= 0.9$

ii) $P(F|E) = \frac{P(E \cap F)}{P(E)} = \frac{0.5}{0.8} = \frac{5}{8}$

iii) $P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{0.5}{0.6} = \frac{5}{6}$

Example 2: If $P(E) = 0.5$, $P(F) = 0.3$ and $P(E \cap F) = 0.15$. Find $P(E|F)$.

Solution: By definition of conditional probability

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

$$= \frac{0.15}{0.3}$$

$$= 0.5$$

Example 3: If $P(E) = 0.9$; $P(F|E) = 0.8$, find $P(E \cap F)$.

Solution: By definition of conditional probability

$$P(F|E) = \frac{P(F \cap E)}{P(E)}$$

$\Rightarrow P(F \cap E) = P(E).P(F|E)$

$\Rightarrow 0.9 \times 0.8 = 0.72$

Example 4: Find the chance of drawing 2 white balls in succession from a bag containing 5 red and 7 white balls, the balls drawn not being replaced.

Solution: Let E be the event that ball drawn is white in the first draw and F be the event that ball drawn is white in the second draw.

$\therefore P(F \cap E) = P(E).P(F|E)$

Here $P(E) = \frac{7}{12}$, $P(F|E) = \frac{6}{11}$

$$\therefore P(F \cap E) = \frac{7}{12}.\frac{6}{11} = \frac{7}{22}$$

Example 5: In an examination 30% of the students have failed in Mathematics, 20% of the students have failed in chemistry and 10% have failed in both Mathematics and Chemistry. A student is selected at random (i) What is the probability that the students has failed in Mathematics if it is known that he has failed in Chemistry?

Solution: Let E be the event of failing in Mathematics.

$\therefore P(E) = 30\% = 0.3$

Let F be the event of failing in Chemistry

$\therefore P(F) = 20\% = 0.2$

Let $E \cap F$ be the event of failing in both subjects.

$\therefore P(E \cap F) = 10\% = 0.1$

The probability that the student has failed in Mathematics given that he has failed in Chemistry is:

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{0.1}{0.2} = 0.5$$

Example 6: A bag contains 5 white and 3 black balls.Two balls are drawn at random one after the other without replacement.Find the probability that both balls are black.

Solution: Probability of drawing a black ball in the first attempt is:

$$P(E) = \frac{3}{5+3} = \frac{3}{8}$$

Probability of drawing the second black ball given that the first ball drawn is black:

$$P(F|E) = \frac{2}{5+2} = \frac{2}{7}$$

$\therefore$ The probability that both balls drawn are black is given by

$$P(E \cap F) = P(A) \times P(F|E) = \frac{3}{8} \times \frac{2}{7} = \frac{3}{28}$$



CHECK YOUR PROGRESS

Q.1: Two fair coins are tossed. Given that at least one of them is heads, what is the probability that both of them are heads?

Q.2: Suppose that a box contains 8 red balls and 4 white balls. We randomly draw 3 balls from the box without replacement. Find the probability that all 3 are red.

Q.3: A die is rolled, If the outcome is an odd number. What is the probability that it is prime?

Q.4: A family has two children. Find the probability that both children are girls given that at least one of them is a girl.

7.5 TOTAL PROBABILITY

Theorem of total probability: If E_1 and E_2 are mutually exclusive and exhaustive events and $P(E_1) \neq 0$, $P(E_2) \neq 0$, then for any event A of S,

$$P(A) = P(E_1)P(A|E_1) + P(E_2)P(A|E_2)$$

Proof: Since E_1 and E_2 are mutually exclusive and exhaustive events, we have $E_1 \cup E_2 = S$ and $E_1 \cup E_2 = \varnothing$.

$\therefore A = A \cap S$

$= A \cap (E_1 \cup E_2)$

$= (A \cap E_1) \cup (A \cap E_2)$ ... ... (1)

Now, $(A \cap E_1) \cup (A \cap E_2) = A \cap (E_1 \cup E_2)$

$= A \cap \varnothing = \varnothing$

$\therefore A \cap E_1$ and $A \cap E_2$ are mutually exclusive and exhaustive events.

$\therefore$ From (1), we have

$P(A) = P(A \cap E_1) + P(A \cap E_2)$

$= P(E_1)P(A|E_1) + P(E_2)P(A|E_2)$

[multiplication theorem of probability]

If E_1, E_2, E_3 are mutually exclusive and exhaustive events and $P(E_1) \neq 0$, $P(E_2) \neq 0$, $P(E_3) \neq 0$, then for any event A of S,

$$P(A) = P(E_1)P(A|E_1) + P(E_2)P(A|E_2) + P(E_3)P(A|E_3)$$

7.6 BAYE'S THEOREM

The concept of conditional probability discussed earlier takes into account information about the occurrence of one event to predict the probability of another event. This concept can be extended to revise probabilities based on new information and to determine the probability that a particular effect was due to specific cause. The procedure for revising these probabilities is known as Bayes theorem. The Principle was given by Thomas Bayes in 1763. By this principle, assuming certain prior probabilities, the posteriori probabilities are obtained.

Bayes' Theorem or Rule:

Statement: Let $E_1, E_2, E_3, ..., ..., ..., E_n$ be a set of n mutually exclusive and collectively exhaustive events and $P(E_1), P(E_2), P(E_3), ,,,, ..., ..., P(E_n)$, are their corresponding probabilities. If A is another event such that P(A) is not zero and the priori probabilities $P(A|E_1)$, i = 1, 2, ..., ..., ..., n are also known. Then

$$P(E_i|A) = \frac{P(E_i)P(A \mid E_i)}{P(E_1)P(A \mid E_1) + P(E_2)P(A \mid E_2) + ... + P(E_n)P(A \mid E_n)}$$

$$\text{i.e., } P(E_i|A) = \frac{P(E_iP(A \mid E_i)}{\sum_{i=1}^{n} P(E_i)P(A \mid E_i)}$$

Proof: Let S be the sample space of the random experiment.

The events $P(E_1), P(E_2), P(E_3), ..., ..., ..., P(E_n)$ being exhaustive.

$\therefore S = E_1 \cup E_2 \cup ... \cup E_n$

$A = A \cap S \quad [A \subset S]$

$= A \cap (E_1 \cap E_2 \cap ... \cap E_n)$

$= (A \cap E_1) \cup (A \cap E_2) \cup ... (A \cap E_n)$ [Distributive law]

$\Rightarrow P(A) = P(A \cap E_1) + P(A \cap E_2) + ... + ... + ... + P(A \cap E_n)$

$\Rightarrow P(E_1)P(A|E_1) + P(E_2)P(A|E_2) + ... + ... + ... + P(E_n)P(A|E_n)$

$$= \sum_{i=i}^{n} P(E_1)P(A \mid E_i) \qquad \text{... ... ... (1)}$$

Now, $P(A \cap E_i) = P(A)P(E_i|A)$

$$\Rightarrow P(E_i|A) = \frac{P(A \cap E_i)}{P(A)}$$

$$= \frac{P(E_i)P(A \cap E_i)}{\sum_{i=1}^{n} P(E_i)P(A \mid E_i)} \quad \text{[using (1)]}$$

Note: 1) The probabilities $P(E_1)$, $P(E_2)$, $P(E_3)$, ,,,, ..., ..., $P(E_n)$ are called as the 'a prori probabilites'.

2) The probabilities $P(E_1|A)$, $P(E_2|A)$, ..., ..., ..., $P(E_n|A)$ are called'posteriori probabilities'.

Example 1: A purse contains 6 silver coins and 3 gold coins.Another purse contains 4 silver coins and 5 gold coins.A purse is selected at random and a coin is drawn from it.What is the probability that it is a silver coin.

Solution: Let Let E = Event of selecting first purse

F = Event of selecting the second purse

A = Event of getting a silver coin

$\therefore P(E) = \frac{1}{2}$ and $P(F) = \frac{1}{2}$

Also, $P(E|A) = \frac{6}{9} = \frac{2}{3}$ and $P(F|A) = \frac{4}{9}$

$\therefore$ Required probability $P(A) = P(E)P(A|E) + P(F)P(A|F)$

$$= \frac{1}{2} \times \frac{6}{9} + \frac{1}{2} \times \frac{4}{9}$$

$$= \frac{10}{18}$$

$$= \frac{5}{9}$$

Example 2: A box contains two coins. One coin is heads on both sides and the other is heads on one side and tails on the other. One coin is selected from the box at random and the face of one side is observed. If the face is heads, what is the probability that the other side is heads ?

Solution: Let E = Event of selecting two headed coin

F = Event of selecting the other coin

(one head and one tail)

A = Event of getting the second side also head.

Now, we have to find P(E|A)

Here, $P(E) = P(F) = \frac{1}{2}$

$P(A|E) = 1$, $P(A|F) = \frac{1}{2}$

By Baye's theorem, we have

$$P(E|A) = \frac{P(E)P(A|E)}{P(E)P(A|E)+P(F)P(A|F)}$$

$$= \frac{\frac{1}{2}\times 1}{\frac{1}{2}\times 1+\frac{1}{2}\times\frac{1}{2}}$$

$$= \frac{2}{3}$$

Example 3: Assume that a factory has two machines. Past records show that machine I produces 20% of the items and machine II produces 80% of the items. Further, it shows that 6% of the items produced by machine I are defective and only 1% of the items produced by machine II were defective. If a defective item is drawn at random,what is the probability that it was produced by machine I?

Solution: Let E_1, E_2 be the events of drawing an item produced by machines I and II repectively. Let A denote the event of drawing a defective item.

We have $P(E_1) = \frac{20}{100} = \frac{1}{5}$, $P(E_2) = \frac{80}{100} = \frac{4}{5}$

Also, we have $P(A|E_1) = \frac{6}{100}$, $P(A|E_2) = \frac{1}{100}$

From Bayes' theorem, the required probability is

$$P(E_1|A) = \frac{P(E_1).P(A|E_1)}{P(E_1)P(A|E_1)+P(E_2)P(A|E_2)}$$

$$= \frac{\frac{1}{5}.\frac{6}{100}}{\frac{1}{5}.\frac{6}{100}+\frac{4}{5}.\frac{1}{100}} = \frac{6}{6+4} = 0.6$$

Example4: A bolt is manufactured by 3 machines A, B and C. A turns out twice as many items as B and machines B and C produce equal number of items. 2% of bolts produced by A and B are defective and 4% of bolts produced by C are defective. All bolts are put into one stock pile and 1 is chosen from this pile.What is the probability that it is defective?

Solution: Let E_1, E_2, E_3 be the events of bolts manufactured by machine A, B, C repectively.

Here we have to apply total theorem on probability.

B and C produce equal no. of bolts and A produces two times that of B.

$$\therefore\ P(E_1) = 50\% = \frac{50}{100} = 0.50$$

$$P(E_2) = 25\% = \frac{25}{100} = 0.25$$

$$P(E_3) = 25\% = \frac{25}{100} = 0.25$$

Let D be the event of getting defective bolt.

$$P(D|E_1) = \frac{2}{100} = 0.02$$

$$P(D|E_2) = \frac{2}{100} = 0.02$$

$$P(D|E_3) = \frac{4}{100} = 0.04$$

By total theorem on probability.

$$P(D) = P(E_1)P(D|E_1) + P(E_2)P(D|E_2) + P(E_3)P(D|E_3)$$
$$= 0.5 \times 0.02 + 0.25 \times 0.02 + 0.25 \times 0.04$$
$$= 0.0250$$

Example 5: A bag A contains 2 white and 3 red balls and bag B contains 4 white and 5 red balls.One ball is drawn at random at random from one of the bags and is found to be red,find the probability that it is drawn from bag B.

Solution: Here given that the ball drawn is red and we have to find the probability of the ball coming from bag B.

Let E_1 be the event of selecting bag A.

$$\therefore P(E_1) = \frac{1}{2}$$

Let E_2 be the event of selecting bag B.

$$\therefore P(E_2) = \frac{1}{2}$$

Let D be the event of selecting red ball.

Probability of drawing red ball from bag A is:

$$P(D|E_1) = \frac{3}{5}$$

Similarly, probability of drawing red ball from bag B is:

$$P(D|E_2) = \frac{5}{9}$$

We have to find the red ball was drawn from bag B. i.e., $P(E_2|D)$.

By Baye's Theorem:

$$P(E_2|D) = \frac{P(E_2)P(D \mid E_2)}{P(E_1)P(D \mid E_1) + P(E_2)P(D \mid E_2)}$$

$$= \frac{\frac{1}{2}.\frac{5}{9}}{\frac{1}{2}.\frac{3}{5} + \frac{1}{2}.\frac{5}{9}}$$

$$= \frac{\frac{5}{18}}{\frac{3}{10} + \frac{5}{18}}$$

$$= \frac{\frac{5}{18}}{\frac{27+25}{90}}$$

$$= \frac{\frac{5}{18}}{\frac{52}{90}}$$

$$= \frac{25}{52}$$

Example 6: In a bolt factory machines A, B, C manufacture repectively 25%, 35%, 40% of the total. Of their output 5%, 4%, 2% are defective bolts. A bolt is drawn at random and found to be defective. What is the probability that it was manufactured by machine B ?

Solution: Let E_1, E_2, E_3 be the events of bolts manufactured by machine A, B, C repectively.

$$P(E_1) = 25\% = \frac{25}{100} = 0.25$$

$$P(E_2) = 35\% = \frac{35}{100} = 0.35$$

$$P(E_3) = 40\% = \frac{40}{100} = 0.40$$

Let D be the event of drawing a defective bolt.

$$\therefore\ P(D|E_1) = \frac{5}{100} = 0.05 \text{ [defective from 'A']}$$

$$P(D|E_2) = \frac{4}{100} = 0.04 \text{ [defective from 'B']}$$

$$P(D|E_3) = \frac{2}{100} = 0.02 \text{ [defective from 'C']}$$

Now, we have to find the probability of defective bolt manufactured by machine B. i.e., $P(E_2|D)$.

By Baye's Theorem:

$$P(E_2|D) = \frac{P(E_2)P(D\,|\,E_2)}{P(E_1)P(D\,|\,E_1)+P(E_2)P(D\,|\,E_2)+P(E_3)P(D\,|\,E_3)}$$

$$= \frac{0.35\times 0.04}{0.25\times 0.05+0.35\times 0.04+0.4\times 0.02}$$



CHECK YOUR PROGRESS

Q.5: Assume that a factory has two machines. Past records show that machine I produces 20% of the items and machine II produces 80% of the items. Further, it shows that 6% of the items produced by machine I are defective and 1% of the items produced by machine II are

defective. If a defective item is drawn at random, what is the probability that it was produced by machine I?

Q.6: Three machines A, B, C produce repectively 60%, 30%,10% of the total number of items of a factory. The percentage of respective outputs of these machines are repectively 2%, 3% and 4%. An item is selected at random and is found to be defective.Find the probability that the item was produced by machine C.

Q.7: A bag X contains 2 white and 3 red balls and a bag Y contains 4 white and 5 red balls. One ball is drawn at random from one of the bags and is found to be red. Find the probability that it was drawn from bag Y.

Q.8: Urn A contains 3 red and 4 white balls and Urn B has 4 red and 5 white balls. One urn is selected at random and one ball is taken out of it. Thrown ball is red. What is the probability that it was taken from urn B?

Q.9: A factory has three production lines I, II and III contributing 20%, 30% and 50% respectively, to its total output. The percentages of substandard items produced by lines I, II and III are, respectively, 15, 10 and 2. If an item chosen at random from the total output is found to be substandard, what is the probability that the item is from line I?



7.7 LET US SUM UP

- For two events E and F, the probability of **an** event **E** under the condition that F has already occumh is the **conditional probability P(E|F)** of **E** and is given by $P(E|F) = \frac{P(E \cap F)}{P(F)}$.
- Two events for which the occurrence of one has no influence on the occurrence or non-occurrence of the other are **independent** events, otherwise, they are **dependent.**

- If the events E and F are independent and P(E) > 0, P(F) > 0, then $P(E \cap F) = P(E).P(F)$ and $P(E|F) = P(E)$, $P(F|E) = P(F)$.



7.8 FURTHER READING

1) Rao. G. Shanker, *Probability and Statistics*. Universitis Press.
2) R. S. N. Pillai and Bagavathi. *Statistics*. New Delhi: S. Chand and Company Ltd.



7.9 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: Let E be the event that both of them are heads and F the event that at least one of them is heads.

Then, $P(E) = \frac{3}{4}$, $P(E \cap F) = \frac{1}{4}$, $P(F) = \frac{1}{4}$

$$\therefore P(E|F) = \frac{P(E \cap F)}{P(F)}$$

$$= \frac{\frac{1}{4}}{\frac{3}{4}}$$

$$= \frac{1}{3}$$

By using the definition of conditional probability, we can easily get that– $P(E \cap F) = P(F).P(E|F) = \frac{1}{4} \times \frac{1}{3} = \frac{1}{12}$

Ans. to Q. No. 2: For i = 1; 2; 3, let E_i be the event that the i-th ball is red.

Then, $P(E_1 \cap E_2 \cap E_3) = P(E_1).P(E_2|E_1).P(E_3|E_1E_2)$

$$= \frac{8}{12}.\frac{7}{11}.\frac{6}{10} = \frac{14}{55}$$

Ans. to Q. No. 3: When a die is rolled, the sample is S = {1, 2, 3, 4, 5, 6}.

Let E = Event of getting an odd number = {1, 3, 5}

F = Event of getting a prime number = {2, 3, 5}

Then, $E \cap F = \{3, 5\}$

$\therefore P(E) = \frac{3}{6} = \frac{1}{2}$, $P(F) = \frac{3}{6} = \frac{1}{2}$ and $P(E \ F) = \frac{1}{3}$

P(getting a prime already when an odd number)

= P(getting a prime number | getting an odd number)

= P(E|F)

= P(E|F)

$= \frac{P(E \cap F)}{P(F)}$

$= \frac{\frac{1}{3}}{\frac{1}{2}}$

$= \frac{2}{3}$

Ans. to Q. No. 4: The sample space S is given by

S = {(b, b), (b, g), (g, b), (g, g)

Let E : Event that both children are girls

F : Event that at least one of the child is a girl.

Then E = {(g, g) and F = {b, g), (g, b), (g, g)}

$E \cap F = \{(g, g)\}$

$\therefore P(E) = \frac{3}{4}$, $P(E \cap F) = \frac{1}{4}$

$\therefore$ Required probability is $P(E|F) = \frac{P(E \cap F)}{P(F)}$

$= \frac{\frac{1}{4}}{\frac{3}{4}}$

$= \frac{1}{3}$

Ans. to Q. No. 5: Let E_1, E_2 be the events of drawing an item produced by machines I and II repectively. Let A denote the event of drawing a defective item.

We have $P(E_1) = \frac{20}{100} = \frac{1}{5}$, $P(E_2) = \frac{80}{100} = \frac{4}{5}$

Also, we have $P(A|E_1) = \frac{6}{100}$, $P(A|E_2) = \frac{1}{100}$

From Bayes's Theorem, the required probability is

$$P(E_1|A) = \frac{P(E_1)P(A \mid E_1)}{P(E_1)P(A \mid E_1) + P(E_2)P(A \mid E_2)}$$

$$= \frac{\frac{1}{5}.\frac{6}{100}}{\frac{1}{5}.\frac{6}{100} + \frac{4}{5}.\frac{1}{100}}$$

$$= \frac{6}{6+4}$$

$$= 0.6$$

Ans. to Q. No 6: Let E, F, G denote the events in which the items are selected by machines A, B, C repectively.

Then, we have $P(E) = \frac{60}{100} = 0.6$, $P(F) = \frac{30}{100} = 0.3$,

$$P(G) = \frac{10}{100} = 0.1$$

Let D denote the event in which the item selected at random is a defective item. Then

$P(D|E) = \frac{2}{100} = 0.02$, $P(D|F) = \frac{3}{100} = 0.03$, $P(D|G) = \frac{4}{100} = 0.04$

Applying Bayes's theorem, we obtain

$$P(G|D) = \frac{P(G).P(D \mid G)}{P(E)P(D \mid G) + P(F)P(D \mid F) + P(G)P(D \mid G)}$$

$$= \frac{(0.1)(0.04)}{(0.6)(0.02) + (0.3)(0.03) + (0.1)(0.04)}$$

$$= \frac{0.004}{0.012 + 0.009 + 0.004}$$

$$= \frac{0.004}{0.025}$$

$$= 0.16$$

Ans. to Q. No. 7: Let E_1 : the ball is drawn from bag X

E_2 : the ball is drawn from bag Y

and A; the ball is red.

Here $P(E_1) = P(E_2) = \frac{1}{2}$

Also, $P(A|E_1)$ = P (a red ball is drawn from bag X) = $\frac{3}{5}$

$P(A|E_2)$ = P (a red ball is drawn from bag Y) = $\frac{5}{9}$

By Baye's Theorem, we have

$$P(E_2|A) = \frac{P(E_1)P(A \mid E_1)}{P(E_1)P(A \mid E_1) + P(E_2)P(A \mid E_2)}$$

$$= \frac{\frac{1}{2} \times \frac{5}{9}}{\frac{1}{2} \times \frac{3}{5} + \frac{1}{2} \times \frac{5}{9}}$$

$$= \frac{25}{52}$$

Ans. to Q. No. 8: Required probability = $\frac{28}{45}$ [By Bayes's Theorem]

Ans. to Q. No. 9: By Bayes formula

$$P(E_1|A) = \frac{P(E_1)P(A \mid E_1)}{P(E_1)P(A \mid E_1) + P(E_2)P(A \mid E_2)}$$

$$= \frac{0.2 \times 0.15}{0.2 \times 0.15 + 0.3 \times 0.10 + 0.5 \times 0.02}$$

$$= 0.069$$



7.10 MODEL QUESTIONS

Q.1: E and F are two events for which $P(E) = \frac{1}{2}$, $P(F) = \frac{1}{3}$, and $P(E \cap F) = \frac{1}{4}$. Find (i) P(E|F), (ii) $P(E \cup F)$ and (iii) $P(E'|F')$.

Q.2: Three persons A, B, C supply raw mterials to a factory in the proportion $\frac{1}{2}, \frac{1}{3}, \frac{1}{6}$ and 5%, 6%, 8% of the materials supplied by them repectively are found to be defective. A material supplied is selected at random from the total supply. What are the probabilities that material was supplied by A, or by B or by C?

Q.3: One of three identical coins is unbiased while the remaining two are biased with probabilities of getting head in a toss are $\frac{1}{3}$ and $\frac{2}{3}$. One of the coins selected randomly shows head when tossed. Find the probability of the coin to be unbiased.

Q.4: Boys from the districts of Kamrup, Nagaon and Jorhat come to a college in Assam in the proportions 50%, 30% and 20% repectively. 80%, 60% and 40% of the boys comingrepectively from the districts wear Dhooti. A boy selected randomly from all the boys of the college is found to wear Dhooti. What is the probability that he comes kamrup?

Q.5: Three trains X, Y, Z carry repectively 25%, 35%, 40% of the total passengers. Of the passengers carried by the tains 5%, 4% and 3% respectively are ladies. If a pssengers chosen at random out of all the passengers of the trains found to be a lady. Prove that the probability that she was carried by X, or Y or Z are repectively $\frac{25}{69}$, $\frac{28}{69}, \frac{16}{69}$.

Q.6: Three are three groups of children. The first group consists of 3 girls and 1 boy, the second consists of 2 girls and 2 boys and the third consists of 1 girl and 3 boys. One child is picked up at random from each group. Find the probability of selecting 1 girl and 2 boys.

Q.7: A bag contains 5 white and 7 red balls. If a ball is drawn twice at random successively, find the probability that one ball is white and the other is red.

Q.8: In a running competition, the probabilities of the two competitiors of winning the race is $\frac{1}{3}$ and $\frac{1}{5}$. What is the probability that both of them will not win ?

Q.9: One card is drawn from a well shuffled pack of 52 cards.Deduce the probability of drawing.

i) a red card or an ace or both

ii) a diamond or an ace or both.

iii) king of diamond or king of heart or both.

Q.10: 4 men play with a pack of cards containing 52 cards. What is the probability that the 4 kings are held by only one of the players?

*** ***** ***

UNIT 8: RANDOM VARIABLES AND ITS PROBABILITY DISTRIBUTION

UNIT STRUCTURE

8.1 Learning Objective

8.2 Introduction

8.3 Random Variable

- 8.3.1 Definition or Random Variable
- 8.3.2 Discrete Random Variable
- 8.3.3 Continuous Random Variable

8.4 Probability Distribution

- 8.4.1 Probability Mass Function
- 8.4.2 Probability Density Function

8.5 Mathematical Expectation and Varience

- 8.5.1 Mathematical Expectation
- 8.5.2 Properties of Mathematical Expectation
- 8.5.3 Variance of Random Variable
- 8.5.4 Properties of Variances
- 8.5.5 Illustrated Examples

8.6 Let Us Sum Up

8.7 Further Reading

8.8 Answers to Check Your Progress

8.9 Model Questions

8.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- distinguish between two types of experiments viz. deterministic experiments and random experiments
- define what is a random variable and discuss when we get a discrete random variable or a continuous random variable
- have knowledge about probability distributions
- discuss the concept ot probability mass function and probability density function along with their properties

- know about mathematical expectations and variance of a random variable which are vital in the study of any distribution, along with their properties and applications
- realise how different measures of location and dispersion as discussed in earlier chapters can be expressed with the help of moments.

8.2 INTRODUCTION

The experiments that are performed can be classified into two broad categories. In one type of experiment the outcomes of the experiments remain same whenever and wherever the experiment is performed. In the second type of experiment, the outcomes are found to vary every time the experiment is performed.

The first types of experiments are known as deterministic experiments, whereas the second types are known as random experiments. As we have already discussed, in this type of experiment, the experimenter knows the set of all possible outcomes, but cannot say with certainity which outcome will turn out at a particular time. If we want to denote thses various outcomes of a random experiment with the help of a variable, then it becomes a random variable.

In this unit we shall learn about random variables, the different types of random variables and their mathematical expectation and some related theorems. We shall also study about moments and try to understand the difference between raw moments, i.e. the moment about any arbitary point and central moment, the moment about mean.

8.3 RANDOM VARIABLE

Most of the random experiments we encounter have outcomes that can be interpreted in terms of real numbers, such as marks obtained by a student in an examination, number of accidents in a day etc. On the other hand, sometimes the outcomes of an experiment can take any value in an interval of real numbers. For example, the heights of students in a class

can take any value within a specified range. Sometimes, the outcomes of a random experiment may be qualitative, e.g. tossing of a coin may result in head or in tail. We may denote head by '1' and tail by '0' and in this way the outcomes of a random experiment, quantitaive or qualitative can be expressed by a real number. these numerical values of the outcomes that changes from experiment to experiment are called random variables. So we can formally define a random variable as follows:

8.3.1 Definition of Random Variable

The real numbers which are associated with the outcomes of a random experiment are called random variables.

It can also be defined as a variable which takes a definite set of values with a probability associated with each of its values is called a random variable. That is why it is also called a chance variable or a stochartic variable.

Random variables may be of two types:

i) Discrete random variables

ii) Continuous random variables

Note: If X and Y are two random variables, then X + Y, X – Y are also random variables. Again, if, a and b are constants, then ax + y, x + by, ax + by are also random variables.

8.3.2 Discrete Random Variable

If a random variable x assumes only a finite number or countably infinite number of values, then it is called a discrete random variable. So a random variable x can take finite number of values x_1, x_2, ..., ..., ..., x_n or it can take countably infinite number of values x_1, x_2, ..., ..., ... The number of letters received by a post office during a particular time period, the number of vehicles arriving at a toll bridge, the number of machines breaking down on a given day etc are some examples of discrete random variables.

8.3.3 Continuous Random Variable

If a random variable is such that it takes any value within a given interval, then it is called a continuous random variable. in this case we say that x takes any value in an interval, say. (a, b), i.e. $a \leq x \leq b$; i.e. x is a continuous random variable in the interval (a, b). Amount of rainfall in a rainy season, heights of individuals, the time between arrival of customers at a servic system etc. some examples of continuous random varible.

8.4 PROBABILITY DISTRIBUTION

The distribution obtained by taking all possible values of a random variable along with their respective probabilities is called a probability distribution. When the random variable is discrete we get discrete probability distribution and when the variable under consideration is continuous, we get a continuous probability distribution. The discrete probability distribution is called probability mass function and the continuous probability distribution is called probability density function.

8.4.1 Probability Mass Function

Let x denote a discrete random variable which takes the values $x_1, x_2, ..., ..., ..., x_n$ with respective probabilities $p_1, p_2, ..., ..., ..., p_n$. Let, $p_i = P(x = x_i) = p(x_i)$, i = 1, 2, ..., ..., ..., n be the probability that x takes the value xi. Then p(xi) is called the Probability Mass Function (p.m.f.) of x if it satisfies the following conditions.

i) $p(x_i) \geq 0$ for all i = 1, 2, ..., ..., ..., n

ii) $\sum_{i=1}^{n} p(x_i) = 1$

In case, the discrete random variable x is countably infinite, then, $p(x_i) = P(x = x_i)$, i = 1, 2, ..., ..., ..., n is said to be the p.m.f. of x if

i) $p(x_i) \geq 0$ for all i = 1, 2, ..., ..., ...

ii) $\sum_{i=1}^{\infty} p(x_i) = 1$

Illustration: Let x denote the uppermost face while throwing a dice. Then x can take the values 1, 2, 3, 4, 5 or 6, each with probability $\frac{1}{6}$. This can be written as,

$$p(x_i) = P(x = x_i) = \frac{1}{6},\ i = 1, 2, ..., ..., ..., 6$$

Here, $p(x_i) > 0$

and $\sum_{i=1}^{6} p(x_i) = 1$

Thus $p(x_i)$ satisfies both the conditions for a p.m.f.

8.4.2 Probability Density Function

The probability distribution of a continuous random variable x is defined by f(x) is called the probability density function (p.d.f.). Let, x be a continuous random variable in the interval (a,b), then the function f(x) = P(X = x) is said to be the p.d.f. of x, if it satisfies the following conditions.

i) $f(x) \geq 0 \ \forall$ x in (a, b)

ii) $\int_a^b f(x)dx = 1$

Illustration: Consider the function, $f(x) = \begin{cases} a, 0 \leq x \leq 5 \\ 0, \text{otherwise} \end{cases}$

For f(x) to be a p.d.f., the condition $\int_a^b f(x)dx = 1$ must be satisfied, which is true if,

$$\int_0^5 a dx = 1$$

$$\Rightarrow a = \frac{1}{5}$$

Since a > 0, the function $f(x) \geq 0$. Thus f(x) satisfies both the conditions for a p.d.f.

8.5 MATHEMATICAL EXPECTATION AND VARIANCE

8.5.1 Mathematical Expectation

The mathematical expectation of a discrete random variable is the sum of the product of the values taken by the random variable and their respective probabilities. Let us consider the discrete random variable x, which assumes the values $x_1, x_2, ..., ..., ..., x_n$ with respective probabilities $p_1, p_2, ..., ..., ..., p_n$, then the mathematical expectation of the random variable x is denoted by E(x) and is given by,

$$E(x) = \sum_{i=1}^{n} x_i p_i$$

Provided the series is convergent and $\sum_{i=1}^{n} p_i = 1$

If x is a continuous random variable with p.d.f. f(x), $-\infty < x < \infty$, then the mathematical expectation of the random variable x is given by,

$$E(x) = \int_{-\infty}^{\infty} xf(x)dx \text{ , provided } \int_{-\infty}^{\infty} f(x)dx = 1$$

Note: The expectation of a random variable x represents the mean of the probability distribution of x.

Mathematical Expectation of a function of a random variable:

If g(x) is a function of a random variable x, then the expectation of g(x) is given by,

$$E[g(x)] = \sum_{i=1}^{n} g(x_i)p(x_i) = \sum_{i=1}^{n} g(x_i)p_i$$

where, x is a discrete random variable taking values $x_1, x_2, ..., ..., ..., x_n$ with respective probabilities $p_1, p_2, ..., ..., ..., p_n$.

Similarly, when x is a continuous random variable in the interval $(-\infty, \infty)$, with p.d.f. f(x), $-\infty < x < \infty$, then,

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

8.5.2 Properties of Mathematical Expectation

The following are the important properties of an expected value of a random variable:

1) The expected value of a constant is the constant itself.
   i.e. E(c) = c, for every constant c.
2) The expected value of the product of a constant a and a random variable x is equal to the times the expected value of the random variable, i.e. E(ax) = a.E(x).
3) The expected value of a linear function of a random variable is same as the linear function of its expectation,
   i.e. E(a + bx) = a + b.E(x).
4) The expected value of the sum of the two random variables is equal to the sum of their individual expected values.
   i.e. E(x + y) = E(x) + E(y).
5) The expected value of the product of two independent random variables is equal to the product of their individual expected values. i.e. E(xy) = E(x).E(y).

8.5.3 Variance of a Random Variable

Variance is an important characteristic of a random variable. It provides the measure of dispersion of the random variable. The variance of a random variable x is denoted by V(x) or σ^2 and is given by,

$$\sigma^2 = v(x) = E\{(x) - E(x)\}^2$$
$$= E(x^2) - \{E(x)\}^2$$

Note: The positive square root of variance is the standard deviation.

8.5.4 Properties of Variance

1) The variance of a constant is zero. i.e. V(c) = 0.
2) The variance of the product of a constant and a random variable is equal to the variance multiplied by the square of the constant.
   i.e. $V(cx) = c^2.V(x)$.

3) The variance of the sum or difference of two independent random variables. i.e. $V(x \pm y) = V(x) + V(y)$.

8.5.5 Illustrated Examples

1) A random variable x has the following probability distribution

x:	–2	–1	0	1	2	3
p(x):	0.1	0.1	0.2	0.2	0.3	0.1

Calculate E(x) and V(x)

Solution:

$E(x) = \sum_x xp(x)$

$= -2 \times 0.1 + (-1) \times 0.1 + 0 \times 0.2 + 1 \times 0.2 + 2 \times 0.3 + 3 \times 0.1$

$= 0.8$

$V(x) = E(x^2) - \{E(x)\}^2$

Now, $E(x^2) = \sum_x x^2 p(x)$

$= (-2)^2 \times 0.1 + (-1)^2 \times 0.1 + 0 + 1^2 \times 0.2 + 2^2 \times 0.3 + 3^2 \times 0.1$

$= 2.8$

$\therefore V(x) = 2.8 - (-0.8)^2$

$= 2.8 - 0.64$

$= 2.16$

2) In an office there are 10 HCL and 20 HP computers. One computer is selected at random. Find the probability distribution of selecting an HP computer. What is the expected value of selecting a HP computer.

Solution: Let x be a random variable such that,

x = 1, if a HP computer is selcted

= 0, if a HP computer is not selected.

$\therefore$ the probability distribution of x is,

x:	0	1
P(X = x):	$\frac{10}{30}$	$\frac{20}{30}$
	$= \frac{1}{3}$	$= \frac{2}{3}$

$\therefore$ E (an HP computer is selected) $= 0 \times \frac{1}{3} + 1 \times \frac{2}{3}$

$= \frac{2}{3}$

3) Eighty percent patients of a hospital are suffering from hypertension. One patient is randomly selected. Find the probability distribution of patients who are suffering from hypertension and find mean of the distribution.

Solution: Let x be a random variable such that,

x = 1, if the patient suffers from hypertension

= 0, otherwise

Hence the probability distribution of x is

x:	0	1
p(X = x):	0.2	0.8

and $E(x) = 0 \times 0.2 + 1 \times 0.8$

$= 0.8$

4) The monthly demand for transistors have the following probability distribution

Demand (n):	1	2	3	4	5	6
Probability (p):	0.10	0.15	0.20	0.25	0.18	0.12

Determine the expected demand for transistors. If the cost of producing n transistors is given by C = 10,000 + 500n, determine the expected cost.

Solution: Expected demand for transistor = E(n)

$= \sum np$

$= 1 \times 0.10 + 2 \times 0.15 + 3 \times 0.20 + 4 \times 0.25 + 5 \times 0.18 + 6 \times 0.12$

$= 3.62$

E(C) = E(10,000 + 500n)

= 10,000 + 500 E(n)

= 10,000 + 500 × 3.62

= 11,810

5) Proprieter of a food stall has invented a new item of food delicacy. He has calculated that the cost per piece will be Rs. 1, but he

wants to sell it for Rs. 3 per piece. It is, however, perishable and the goods unsold at the end of the day is a total loss. If demand is a variable with the following frequency distribution, calculate his net profit if he makes 15 pieces.

No. of pieces demanded:	10	11	12	13	14	15
Probability:	0.07	0.10	0.23	0.38	0.12	0.10

Solution: Units produced = 15

Total cost price = 15 x 1 = Rs. 15

sell price peer unit = Rs. 3

Unit demanded	Profit
10	Rs. (10×3) – 15 = Rs. 15
11	Rs. (11×3) – 15 = Rs. 18
12	Rs. (12×3) – 15 = Rs. 21
13	Rs. (13×3) – 15 = Rs. 24
14	Rs. (14×3) – 15 = Rs. 27
15	Rs. (15×3) – 15 = Rs. 30

$\therefore$ Expected net profit = Σ profit x probability

= 15 × 0.07 + 18 × 0.10 + 21 × 0.23 + 24 × 0.38 + 27 × 0.12 + 30 × 0.10

= Rs. 3.45



CHECK YOUR PROGRESS

Q.1: Check true or false:

i) The probability distribution of a random variable x is as follows

X = x:	1	2	3	
P(X = x):	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	(T/F)

ii) If c is a constant then, (a) E(c) = c (T/F), (b) V(c) = c (T/F)

iii) If x and y are two independent random variables, then,

a) V(x + y) = V(x) + V(y) (T/F)

b) V(x – y) = V(x) – V(y) (T/F)

iii) If x and y are two independent random variables, then, E(xy) = E(x).E(y)

iv) A continuous random variable can take any value within a given range.

Q.2: Define random variable. What are the important properties of a random variable.



8.6 LET US SUM UP

- The experiments can be classified into two broad categories - deterministic experiment and random experiment.
- The real numbers which are associated with the outcomes of a random experiment are called random variables.
- Random variables may be of two types - discrete random variable and continuous random variable.
- The distribution obtained by taking all possible values of a random variable along with their respective probabilities is called a probability distribution.
- The probability distribution for discrete random variable is called probability mass function (p.m.f.) and the same for continuous random variable is known as probability density function (p.d.f.).
- The mathematical expectation of a discrete random variable is the sum of the product of the values taken by the random variable and their respective probabilities.
- The expectation of the random variable x represents the mean of the probability distribution of x.
- The variance of a random variable provides a measure of dispersion of the random variable.
- The expectation of a constant quantity is the constant itself, whereas variance of a constant is zero.
- The expectation of the sum of two or more random variables is the sum of the individual random variables.

- The expectation of the product of two independent random variables is the product of their expectations.
- The variance of the addition or subtraction iof two independent random variables is equal to the sum of their variances.



8.7 FURTHER READING

1) Feller, W. (1968). *An Introduction to Probability Theory and its Application*. Vol. I. Wiley Eastern.
2) Goon, A.M., Gupta, M. K. and Dasgupts, B. (1997). *An Ouline of Statistical Theory*. Vol. I. World Press.
3) Gupta, S. C., Kapoor, V. K. (2009). *Fundamentals of Mathematical Statistics*. Sultan Chand & Sons.
4) Mukhopadhyay, P. (1996). *Mathematical Statistics*. New Central Book Agency (P) Ltd.
5) Mukhopadhyay, P. (1991). *Theory of Probability*. New Central Book Agency (P) Ltd.



8.8 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: i) False (Total probability exceeds 1); ii) (a) True, (b) False; iii) (a) True, (b) False; iv) True; v) True

Ans. to Q. No. 2: The real numbers whcih are associated with the outcomes of a random experiment are called random variables.

The properties of random variable are as follows:

i) If X is a random variable there, $\frac{1}{x}$, x^2 are also random variables.

ii) If x is a random variable and a and b are constants then ax+b is also a random variable.

iii) If x and y are two random variables, then x + y and x – y are also random variables.



8.9 MODEL QUESTIONS

Q.1: Fill in the blanks:

i) Heights or weights of individuals can be represented by a ..... random variable.

ii) If x and y are two random variables then ax+by is a ....., where a and b are constants.

iii) If x and y are independent, then E(xy) = ...

iv) $V(x \pm y) = V(x) + V(y)$, when x and y are ..... variables.

v) V(ax + b) = ..., where a and b are constants.

vi) E(a) = ... and V(a) = ..., where a is a constant.

Q.2: Can you find the mathematical expectation of a random variable x, if x takes the values 1, 2, 3 with corresponding probabilities $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{3}{4}$.

Q.3: Discuss discrete and continous random variables with examples.

Q.4: What do you mean by mathematical expectation of a function of a random variable. State addition and multiplication law of mathematical expectation.

Q.5: A random variable x has the following probability distribution,

x:	–2	–1	0	1	2	3
f(x):	0.1	2k	0.2	2k	0.3	2k

Find the value of k.

Q.6: Write some important properties of mathematical expectation.

*** ***** ***

UNIT 9: THEORETICAL PROBABILITY DISTRIBUTIONS (DISCRETE VARIABLE I)

UNIT STRUCTURE

9.1 Learning Objectives

9.2 Introduction

9.3 Moment Generating Function

- 9.3.1 Definition
- 9.3.2 Generation of Moments
- 9.3.3 Properties of Moment Generating Function
- 9.3.4 Illustrative Examples

9.4 Binomial Distribution

- 9.4.1 Derivation of Binomial Distribution
- 9.4.2 Definition
- 9.4.3 Moments of Binomial Distribution
- 9.4.4 Moment Generating Function of Binomial Distribution
- 9.4.5 Fitting of Binomial Distribution
- 9.4.6 Properties of Binomial Distribution
- 9.4.7 Importance of Binomial Distribution
- 9.4.8 Illustrative Examples

9.5 Let Us Sum Up

9.6 Further Reading

9.7 Answers To Check Your Progress

9.8 Model Questions

9.1 LEARNING OBJECTIVES

After going through this unit, you will able to:

- obtain moment generating function for any given distribution
- generate moments from probability distribution
- learn some important properties of moment generating function
- derive binomial distribution from a coin tossing experiment
- compute moments of binomial distribution
- fit binomial distribution to real life date

- understand the application of the concept of these probability distributions to real life data.

9.2 INTRODUCTION

So far we have brought out the nature of discrete and continuous probability distributions in general and also have defined their mean and variances. In this present unit we shall discuss about how the moments of a particular distribution as discussed in chapter 6, can be generated by a special kind of function called the moment generating function.

Again we know that the rule generating the probability distribution is the same whether we want to find the probability of the number of heads when three coins are lossed simultaueously or for the number of red balls in case of an urn containing 3 red and 2 black balls, out of which, say, two balls are drawn at random with replacement. Given this basic characteristic of a discrete distribution, our main concern is to examine the outcoms generated by an experiment and choose a distribution that best describes those outcomes.

Under this consideration we shall discuss some discrete probability distributions applicable to many real life situations in the subsequent sections of this unit. The distributions we are going to discuss here are binomial and poisson distributions.

9.3 MOMENT GENERATING FUNCTION

The moments of a probability distribution play an important role in theoretical and in applied statistics. With the help of moments we can measure the central tendency of a set of observations, their variability, their asymmetry and the height of the peak their curves would make. That is the knowledge of moments enable us to determine the probability distribution completely. So it would be very useful if a function could be found which will help in generating moments. One such function is moment generating function. So we can define a moment generating function as follows:

9.3.1 Definition

A function that is used to generate moments of a random variable X is called the moment generating function, denoted by Mx(t) and is given as:

$$M_x(t) = E(e^{tx})$$

When X is a discrete random variable with p.m.f. p(x)

$$M_x(t) = E(e^{tx}) = \sum_x e^{tx} p(x)$$

When x is a continous random variable in the interval (a,b) with p.d.f. f(x),

$$M_x(t) = E(e^{tx}) = \int_a^b e^{tx} f(x)dx$$

9.3.2 Generation of Moments

We have,

$M_x(t) = E(e^{tx})$

$$= E\left(1 + tx + \frac{t^2x^2}{2!} + \frac{t^3x^3}{3!} + ... + \frac{t^rx^r}{r!} + ...\right)$$

$$= E(1) + tE(x) + \frac{t^2}{2!}E(x^2) + \frac{t^3}{3!}E(x^3) + ... + \frac{t^r}{r!}E(x^r) + ...$$

$\therefore \mu_1'$ = co-efficient of t in $M_x(t)$

$\therefore \mu_2'$ = co-efficient of $\frac{t^2}{2!}$ in $M_x(t)$

$\therefore \mu_3'$ = co-efficient of $\frac{t^3}{3!}$ in $M_x(t)$

Thus in general we see that,

$\therefore \mu_r'$ = co-efficient of $\frac{t^r}{r!}$ in $M_x(t)$

Another convenient way of obtaining moments is to differentiate $M_x(t)$ sucussively and then put t = 0.

So if we differentiate $M_x(t)$, one time, with respect to t and then put t = 0, we get μ_1',

i.e. $\mu_1' = \left[\frac{d}{dt}M_x(t)\right]_{t=0}$

Differenting $M_x(t)$ times with respect to t and then put t = 0 we get μ_α'.

$$\mu_2' = \left[\frac{d^2}{dt^2}M_\lambda(t)\right]_{t=0}$$

So in general, we get the raw moment of order r, μ_r' as:

$$\mu_r' = \left[\frac{d^r}{dt^r}M_r(t)\right]_{t=0}$$

9.3.3 Properties of Moment Generating Function

i) If X is a random variable and C is a constant, there,

$M_{C_x}(t) = M_\lambda(ct)$

Proof: L.H.S. = $M_{C_x}(t) = E\left(e^{tcx}\right)$

$= E\left(e^{ctx}\right)$

$= M_x(ct)$

= R.H.S.

ii) Additive property of moment generating function.

The moment generating function of sum of independent random variables is equal to the product of their moment generating functions. In symbol, if $X_1, X_2, ..., ..., ..., X_n$ are independent random variables, then,

$Mx_1 + x_2 + ... + x_n(t) = Mx_1(t).Mx_2(t). ... .Mx_n(t)$

Proof: L.H.S. = $Mx_1 + x_2 + ... + x_n(t)$

$= E\left[e^{t(x_1+x_2+...+x_n)}\right]$

$= E\left[e^{tx_1+tx_2+...+tx_n}\right]$

$= E\left(e^{tx_1}\right)E\left(e^{tx_2}\right)....E\left(e^{tx_n}\right)$

Θ $X, X_2, ..., ..., ..., X_n$ are independent

$= Mx_1(t).Mx_2(t). ... .Mx_n(t)$

= R.H.S.

iii) Effect of change of origion and scale on m.g.f.

$$M_{\frac{x-a}{h}}(t) = e^{\frac{-at}{h}} M_x\left(\frac{t}{h}\right)$$

Proof: $M_{\frac{x-a}{h}}(t) = E\left[e^{t\left(\frac{x-a}{h}\right)}\right]$

$$= E\left[e^{\frac{tx}{h}} e^{\frac{-at}{h}}\right]$$

$$= e^{\frac{-at}{h}} E\left[e^{\frac{t}{h}(x)}\right]$$

$$= e^{\frac{-at}{h}} M_x\left(\frac{t}{h}\right)$$

= R.H.S.

Note: i) The moment generating function doesnot always exist.

ii) In some cares the moment generating function exists, but it cannot generate moments.

9.3.4 Illustrative Example

1) Let X be a random variable with p.m.f. $p(x) = q^x p$, $x = 0, 1, 2, ...$ find the moment generating function and have find its mean and variance.

Solution: $M_x(t) = E(e^{tx})$

$$= \sum_{x=0}^{\alpha} e^{tx}.q^x p$$

$$= p\sum_{x=0}^{\alpha}\left(qe^t\right)^x$$

$$= p[1 + qe^t + (qe^t)^2 + ...]$$

$$= p(1 - qe^t)^{-1}$$

$$\text{Mean} = \mu_1' = \left[\frac{d}{dt} M_x(t)\right]_{t=0}$$

$$= [(-1)p(1-qe^t)^{-2}(-qe^t)]_{t=0}$$

$= pq(1-q)^{-2}$

$= \frac{pq}{p^2}$

$= \frac{p}{p}$

$\mu_\alpha' = \left[\frac{d^2}{dt^2} M_x(t)\right]_{t=0}$

$= \left[\frac{d}{dt}\left\{pqe^t\left(1-ge^t\right)^{-2}\right\}\right]_{t=0}$

$= pq[et(1-ge^t)^{-2} + e^t(-2)(1-ge^t)^{-3}(-ge^t)]_{t=0}$

$= pq[(1-q)^{-2} + 2q(1-q)^{-3}]$

$= \frac{pq}{p^2} + \frac{2pq^2}{p^3}$

$= \frac{q}{p} + \frac{2q^2}{p^2}$

The expansion used here is called binomial expansion with negative index, which is $(1-x)^{-1} = 1 + x + x^2 + x^3 + ... + ... + ...$

$\therefore$ variance $= \mu_\alpha = \mu_\alpha' - \mu_\alpha'^2 = \frac{q}{p} + \frac{2q^2}{p^2} - \frac{q^2}{p^2}$

$= \frac{q}{p} + \frac{q^2}{p^2} = \frac{q(p+q)}{p^2} = \frac{q}{p^2}$

2) Final moment generating function of a continuous random variable x having the p.d.f. $f(x) = \theta e^{-\theta}$, $x \geq 0$

Hence find its mean and variance.

Solution: $M_x(t) = E(e^{tx})$

$= \int^{\alpha} e^{tx}\theta e^{-\theta x} dx$

$= \theta^0 \int^{\alpha} e^{(t-\theta)x} dx$

$= \theta^0 \left[\frac{e(t-\theta)x}{t-\theta}\right]\alpha 0$

$= \theta\left[\frac{1}{t-\theta}(0-1)\right]$

$$= \frac{-\theta}{t-\theta}$$

$$= \frac{\theta}{\theta - t}$$

$$= \left(1 - \frac{t}{\theta}\right)^{-1}$$

$$= 1 + \frac{t}{\theta} + \left(\frac{t}{\theta}\right)^2 + \left(\frac{t}{\theta}\right)^3 + \ldots + \ldots + \ldots$$

$\therefore$ μ_1' = mean co-efficient of t in $M_x(t) = \frac{1}{\theta}$

μ_2' = mean co-efficient of t in $M_x(t) = \frac{2}{\theta^2}$

$\therefore$ variance = $\mu_2 = \mu_2' - \mu_1'^2$

$$= \frac{2}{\theta^2} - \left(\frac{1}{\theta}\right)^2$$

$$= \frac{1}{\theta^2}$$



CHECK YOUR PROGRESS

Q.1: State true or false:

i) Since $M_x(t)$ generates moments, it is called moment generating function. (T/F)

ii) If X an Y are independent r. V S then $M_{x+y}(t) = M_x(t) + M_y(t)$ (T/F)

iii) Moment generating function always exits. (T/F)

iv) A moment generating function, if exists, determences the distribution function. (T/F).

9.4 BINOMIAL DISTRIBUTION

The binomial distribution, as derived by James Bernoulli in the year 1700, deals with population whose elements or outcomes can be devided into two catigories with reference to presence or absence of a particular attribute or characteristic. For example, the answer to a question may be

correct or incorrect, results in an examination may be pass or fail, the factory workers may be educated or uneducated.

That is the outcome in each of these situations may be viewed in terms of presence or absence of an attribute. That is we consider only the two outcomes one possessing and the other not possissing the attribute. A convenient method is to denote them as 'success' and 'failure' according as the presence or absence of the attribute.

The mathematical background for binomial distribution describes descrete date resulting from an experiment called. Bernoulli trials, which are repeted finite number of times. Let us now device binomial distribution from a finite number of Bernoulli treals.

Note: A random experiment whoss outcomes has been classified into two categories, called 'success' and 'failure' is called Bernoulli trial. For example, tossing of a coin may result in a head or in a tail etc.

9.4.1 Derivation of Binomial Distribution

Consider a coin tossing experiment which is repeated, say, n number of times. When we get a head we call it a success and when we get a tail, we call it a failure, success is denoted by 's' and failure by 'f'. Probability of success is p and that of failure is q.

$$P(s) = p,\ P(f) = q,\ q = 1 - p.$$

Suppose we are interested in getting x success out of n trials. Now the probability of x successes and consequently (n–x) failures in n independent trials, in a specified order, say, first x trials succes and the remaining (n–x) trials facture is:

$$P(ss \ldots s\ f\ f \ldots f) = P(s)\ P(s) \ldots P(s)\ P(f) \ldots P(f)$$

$$\because \text{ the trials are independent} = \underbrace{p.p.p.}_{x \text{ factor}}\ \underbrace{q.q.q}_{(n-x) \text{ factor}}$$

$$= p^x.q^{n-x}$$

But this is simply one way of getting the probabilety of x successes out of n trials. As we know, x successes in n trials can occur in ncx ways, and the probability for each of these cases is p^x q^{n-x}, Hence by the addition theorem of probability the probability of

getting x success in n trials is given by,

$p^x q^{n-x} + p^x q^{n-x} + ... + ... + ... + p^x q^{n-x}$, ncx times = $ncx\ p^2\ q^{n-x}$

Which is the probability mass function of binomial distribution. Again, x, the number of successes out of n trials can take values 0, 1, 2, ..., ..., ..., n. Therefore, we may define a binomial distribution as follows.

9.4.2 Definition

A random variable X is said to follow binomial distribution if it assumes non-w values and if its p.m.f. is given by,

$p(x) = p(X = x)$ {$n_{c_x}\ p^x\ q^{n-x}$, x = 0, 1, 2, ..., ..., ..., n

0, otherwise.

The two constante n an p are the parameters of this dist^n. It is also abbreviated as B(n, p).

Note: The probability of getting 0 success. 1 success, ..., ..., ..., n successes in this case, i.e. $nc_o\ p^o q^n$, $nc_1 p\ q^{n-1}$, ..., ..., ..., $ncn\ p^0 q^x$ ... are the successive terms in the expansion of $(q + p)^n$, a binomial expansion, that is why this distribution is also termed as 'binomial distribution'.

9.4.3 Moments of Binomial Distribution

We are interested mainly in the first two moments of benomial distribution which are obtained as follows.

μ_1' = mean = E(x)

$$= \sum_x xp(x)$$

$$= \sum_{x=0}^{n} x n_{c_x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x \frac{n^{n-1}}{x} c_{x-1} p^x q^{n-x} \qquad \because\ n_{c_x} = \frac{n}{x}, n-1_{c_x}$$

$$= np \sum_{x=1}^{n} n-1_{c_x} p^{x-1} q^{n-x}$$

$$= np\left[{}^{n-1}C_0 p^0 q^{n-1} + n-1_{c_1} pq^{n-2} + ... + {}^{n-1}C_1 p^{n-1} q^0\right]$$

$$= np(q + p)^{n-1}$$

$$= np \qquad \because p + q = 1$$

Hence the mean of binomial distribution is n p.

Again, $\mu_1' = E(\lambda^2) = E\{X(x-1)\} + E(x)$

$$E\{x(x-1)\} = \sum_{x=0}^{n} x(x-1) n_{c_x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x(x-1) \frac{n(n-1)^{n-2}}{x(x-1)} c_{x-2} p^x q^{n-x}$$

$$= n(n-1)p^2 \sum_{x=0}^{n} n-2 c_{x-2} p^x q^{n-x}$$

$$= n(n-1)p^2(q+p)^{n-2}$$

$$= n(n-1)p^2$$

$\therefore E(x^2) = n(n-1)p^2 + np$

$\therefore \mu_2 = \text{variance} = \mu_2' - \mu_1'^2 = n(n-1)p^2 + np - n^2p^2$

$$= n^2p^2 - np^2 + np - n^2p^2$$

$$= np(1 - p)$$

$$= npq.$$

So the variance of binomial distribution is npq.

Remark: For binomial distribution mean is always greater than variance. This is an important characteristic of binomial distribution.

9.4.4 Moment Generating Function of Binomial Distribution

If X ~ B (n, p), then

$$Mx(t) = E(e^{tx}) = \sum_x e^{tx} p(x)$$

$$= \sum_{x=0}^{n} e^{tx\,n} c_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n} {}^n c_x \left(e^t p\right)^2 q^{n-x}$$

$$= {}^n c_0 \left(e^t p\right)^0 q^n + {}^n c_1 \left(e^t p\right) q + ... + {}^n c_n \left(e^t p\right)^n q^r$$

$= (q + e^t p)^n$

Which is the moment generating function of binomial dist^n.

9.4.5 Fitting of Binomial Distribution

The recurrence relation for probabilities of binomial distribution is,

$$p(x+1) = \frac{n-x}{x+1}\frac{p}{q}p(x) \text{ . } x = 0, 1, 2, ..., ..., ..., n-1$$

Putting x = 0, 1, 2, ..., ..., ..., (n – 1) we get p(1), p(2), ..., ..., ..., p(n). Here, we need to calculate only p(0), which is:

$$p(0) = {}^nc_0 p^0 q^n = q^n$$

For felling a binomial distribution, we need to find the values of n and p. n is generally known, and p = $\frac{np}{n}$.

Where np is the mean of the given set of data.

After calculating the value of p, we calculate the probabilities p(0), p(1), ..., ..., ..., p(n) as discussed above. Now multiplying these probabilities by N, the total frequency, we get the expected frequencies for x = 0, 1, 2, ..., ..., ..., n. Now we shall illustrate the fitting of biomial distribution with the help of an example.

Note: Calculation of expected frequency is called fitting.

Example: 3 coins are tossed 64 times and the following distribution of number of heads were obtained.

Number of heads:	0	1	2	3
Frequency:	10	18	20	16

Fit a binomial distribution under the hypothesis that the coins are unbiased.

Solution: Since the coins are unbiased

$p = q - \frac{1}{2}$ [in case coins are biased p is to be calculated from mean]

Here, N = 64, n = 3

Now, $p(0) = q^n = \left(\frac{1}{2}\right)^3 = \frac{1}{2^3} = \frac{1}{8}$

Now using recurrence relation, $p(x+1) = \frac{n-x}{x+1}\frac{p}{q}p(x)$

We calculate the probabilities as,

$$p(1) = \frac{3-0}{0+1}\frac{p}{q}p(0) = 3.\frac{1}{2^3} = \frac{3}{8}$$

$$p(2) = \frac{3-1}{1+1}\frac{p}{q}p(1) = \frac{3}{2}.\frac{3}{8} = \frac{3}{8}$$

$$p(3) = \frac{3-2}{2+1}\frac{p}{q}p(2) = \frac{1}{3}.\frac{3}{8} = \frac{1}{8}$$

Hence the expected frequenties are,

x	Expected frequency
0	$n \times p(0) = 64 \times \frac{1}{8} = 8$
1	$n \times p(1) = 64 \times \frac{3}{8} = 24$
2	$n \times p(2) = 64 \times \frac{3}{8} = 24$
3	$n \times p(3) = 64 \times \frac{1}{8} = 8$

9.4.6 Properties of Binomial Distribution

i) The binomial distribution is a discrate distribution where the random variable X takes the values 0, 1, 2, ..., ..., ..., n.

ii) The binomial distribution has two parameters, n and p.

iii) The mean is equal to np and variance is given by npq.

iv) The mean of binomial distribution is always greater than variance.

v) If X_1 and X_2 follows binomial distribution with parameters (n_1, p) and (n_2, p) respectively, these $x_1 + x_2$ follows binomial distribution with parameter ($n_1 + n_2$, p). THis is known as the additive property of binomial distribution.

vi) Binomial distribution with parameters (n.p) tends to poisson distribution if n is very largs, p is very small and np is finite.

vii) Binomial distributions also tends to Normal distribution if n is very large and neither non q is very small.

9.4.7 Importance of Binomial Distribution

Theoretical distribution like Binomial distribution provide data on the basis of which observed (empirical or experimental) results can be assessed. In fact, where theoretical distributions are available one need not have to bother for observed distributions. Theoretical distributions provide the decision maker with a sound basis for taking rational and dependable decisions. However in order to apply Binomial distribution, the assumptions of the distribution must hold true which are: two mutually exclusive outcomes, a fixed probability of outcome in any trial and independent trials.

9.4.8 Illustrative Examples

1) A r. v. X follows binomial distribution with mean $\frac{5}{3}$ and P(x = 2) = P (x = 1). Find variance, P(x = at leart 1), P(x = at most 1).

Solution: If X follows binomial distribution with parameters n an p, the p.m.f. of X is,

$p(x) = {}^{n}c_{x}p^{x}q^{n-x}$, x = 0, 1, 2, ..., ..., ..., n; q = 1 – p

Given, mean = np = $\frac{5}{3}$

and P(x = 2) = P(x = 1)

$\Rightarrow {}^{n}c_{2}p^{2}q^{n-2} = {}^{n}c_{1}pq^{n-1}$

$\Rightarrow \frac{n(n-1)}{2}\frac{p}{q} = n$

$\Rightarrow (n-1)p = 2q$

$\Rightarrow np - p = 2q$

$\Rightarrow \frac{5}{3} - p = 2q$

$\Rightarrow 5 - 3p = 6q$

$\Rightarrow 6q + 3p = 5$

$\Rightarrow 6(1-p) + 3p = 5$

$\Rightarrow 6 - 6q + 3p = 5$

$\Rightarrow -3p = -1$

$\Rightarrow p = \frac{1}{3}$

$\therefore q = 1 - p = \frac{2}{3}$

Again, $np = \frac{5}{3}$ $\quad\quad$ $\therefore n = \frac{\frac{5}{3}}{\frac{1}{3}} = 5$

Hence, variance = $npq = 5.\frac{1}{3}.\frac{2}{3} = \frac{10}{9}$

P(x = at least 1) = $p(x \geq 1)$

$= 1 - p(x = 0)$ $\quad \because p(x = 0) + (x \geq 1) = 1$

$= 1 - {}^5C_0\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^5$

$= 1 - \left(\frac{2}{3}\right)^5$

$= \frac{211}{243}$

P(x = at most 1) = P(x = 0) + p(x = 0)

$= {}^5C_0\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^5 + {}^5C_1\left(\frac{1}{3}\right)^1\left(\frac{2}{3}\right)^4$

$= \frac{112}{243}$

2) Probability of follows in statistics examination is 15%, 20 batchs of 5 students each appear for the examination. Obtain in how many batches 3 or more students would pass.

Solution: Given q = probability of failure = 15% = 0.15

$\therefore$ q = probability of success = 0.85

n = 5.

If X denote the number of students who would pass, then

$p(x) = {}^nc_xp^xq^{n-x}$, x = 0, 1, ..., ..., ..., 5.

We need $p(x \geq 3) = p(x = 3) + p(x = 4) + p(x = 5)$

$= {}^5C_3p^3q^{5-3} + {}^5C_4p^4q^{5-4} + {}^5C_5p^5q^0$

$= 10 \times (.85)^3 (.15)^2 + 5(0.5)^4 (0.15) + (0.15)^5$

$= 0.1381 + 0.3915 + 0.4437$

$= 0.9733.$

$\therefore$ Required number of batches in which 30x more students would pass = 20 x 0.9733

$= 19.468 \approx 19.$



CHECK YOUR PROGRESS

Q. 2: If $X \sim B\left(12, \frac{1}{3}\right)$ what is E(x)?

Q. 3: Examine the following statement and add your comment, "The mean of a binomial distribution is 5 and s.d. is 3".

Q. 4: What are the parameters of binomial distribution?

Q. 5: Find the varience of the binomial distribution

$$p(x) = {}^{10}C_x \left(\frac{2}{5}\right)^x \left(\frac{3}{5}\right)^{10-x}, \quad x = 0, 1, 2, ..., ..., ..., 10.$$

Q. 6: Define Bernoulli treats. Find its distributions and also its mean and variance.

Q. 7: Derive Binomial distribution.

or

Obtain the probability of exactly x successes in n independent Berioulli treats.



9.5 LET US SUM UP

- A function, which is used to generate moments of a probability distribution is known as moment generating function.
- The importance of moment generating function in the fact that, a probability distribution can be completely determined if its moments are known.

- The moment generating function of a random variable X is given by, $Mx(t) = E(e^{tx})$, $t > 0$.
- A random experiment whose outcoms are classified into two categories, called 'success' and 'failure' is called a Bernoulli treats.
- The p.m.f. of binomial distribution with parameters n and p is, $p(x) = {}^{n}C_{x}p^{x}q^{n-x}$, x = 0, 1, 2, ..., ..., ..., n.
- The mean of binomial distribution is np and varience is npq.
- For a binomial distribution, mean is always greater than variance.
- Sum of independent binomial variates is also a binomial variate.



9.6 FURTHER READING

1) Das, K. K, Bhattacharjee, D. (2008). *A Treatere on Statistical Inference and Distributions*. Asian Books Pvt. Ltd.
2) Gupta, S. C. and Kapoor, V. K. (2002). *Fundementals of Mathematical Statistics*. Sultan Chand & Sons.
3) Goon, A. M., Gupta. M. P. and Dasgupta, B. (2002). *Fundementals of Statistics* (Vol. I). Seventh Revised Edition. Kolkata: The World Press Pvt. Ltd.
4) Saxena, H. C. (1994). *Probability and its Application*. Sixth Edition. New Delhi: S. Chand and Company.
5) Singh, R. (2012). *An Introduction to Probability and Probability Distributions*. Books and Ellied (P) Ltd.



9.7 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: i) True, ii) False, iii) False, iv) True.

Ans. to Q. No. 2: We know that if X ~ B (n, p) then E(x) = np.

$$\text{Here, } n = 12, = \frac{1}{3}$$

$$\therefore E(x) = np = 12.\frac{1}{3} = 4$$

Ans. to Q. No. 3: Here, given, mean = 5, s.d. = 3

$\therefore$ variance = 9

But we know that for a benomial distribution, mean is always greater than variance. Hences, the above statement is false.

Ans. to Q. No. 4: The parameters of a binomial distribution are n and p.

Ans. to Q. No. 5: Given, $p(x) = {}^{10}C_x\left(\frac{2}{5}\right)^2\left(\frac{3}{5}\right)^{10-x}$, x = 0, 1, 2, ..., ..., ..., 10.

Here, n = 10, $p = \frac{2}{5}$, $q = \frac{3}{5}$

Hence, variance = npq = $10.\frac{2}{5}.\frac{3}{5} = \frac{12}{5}$

Ans. to Q. No. 6: A particular trial having two outcomes only, viz success and failure occurring with probability p and (1–p) respectively is called a Bernoulli trial. e.g. tossing of a coin.

A random variable X is said to follow Beruoulli distribution if its p.m.f. is given by, $p(x) = p^x(1–p)^{1–x}$, x = 0, 1.

P is the only parameter of this distribution

$\therefore$ Mean = E(x) = 0.(1–p) + 1.p = p.

$V(x) = E(x^2) – \{E(x)^2\}$.

$E(x^2) = 0^2.(x–p) + 1^2.p = p$

$\therefore$ $V(x) = p – p^2 = p(1–p)$.

Ans. to Q. No. 7: Consider n independent Bernoulli trials each having two and two outcomes only called success and failure, Probability of success is p and the probability of failure is (1 – p) = q.

The probability of x successes and (n–x) failures in n independent trials in a specific order, say, s s s... f.f.f is given by,

P (s. s. . . . s. f . f. . . . f) = p(s) p(s) s . . . p(s) p(f) . . . p(f)

$\because$ trials are independent = p.p . . . p q.q . . . q

x times (n-x) times

$= p^xq^{n-x}$

But, x successes in n trails can occur in n_{CX} .... which are mutually exclusive and the probability for each of these is $p^x q^{n-x}$. Hence by addition theorem of probability,

$p(x = x) = n_{CX}\, p^x\, q^{n-x}$, x = 0, 1, 2, . . ., n.



9.8 MODEL QUESTIONS

Q.1: Define moment guerating function and state its properties.

Q.2: Write a note on binomial distribution.

Q.3: Define a binomial distribution distribution and also obtain its mean and variance.

Q.4: Obtain m.g.f. of binomial distribution.

Q.5: Compute moments of binomial distribution.

*** ***** ***

UNIT 10: THEORETICAL PROBABILITY DISTRIBUTIONS (DISCRETE VARIABLE II)

UNIT STRUCTURE

10.1 Learning Objectives

10.2 Introduction

10.3 Poisson Distribution as a Limiting Case of Binomial Distribution

10.4 Moments of Poisson Distribution

10.5 Moment Generating Function

10.6 Fitting of Poisson Distribution

10.7 Properties of Poisson Distribution

10.8 Application of Poisson Distribution

10.9 Illustrative Examples

10.10 Let Us Sum Up

10.11 Further Reading

10.12 Answers To Check Your Progress

10.13 Model Questions

10.1 LEARNING OBJECTIVES

After going through this unit, you will able to:

- understand the concept of poisson distribution
- derive poisson distribution as a limiting form of binomial distribution
- compute moments of poisson distribution
- fit poisson distribution to real life date
- understand the application of the concept of poisson probability distribution to real life data.

10.2 INTRODUCTION

In the earlier unit we have discussed about the binomial distribution.In this unit we will discuss the poisson distribution and its application distributions.The Poisson process was first development by French

mathematician and physicist serieon Devis Poisson, who published if in the year 1837. In the class of univaxiate distributions, the poisson distribution is perhaps the most useful one, because of its wids applicability in the description of natural phenomena, for example, demand for service in a big departmental store, major accident rate per day in a certain highway etc. A typical application of poisson distribution is for analysing queing (or waiting live) problems in which arriving customers during an interval of time arrive independently and the number of arrivals depends on the length of the time interval.

10.3 POISSON DISTRIBUTION AS A LIMITING CASE OF BINOMIAL DISRTIBUTION

The binomial probability law of rare events gives rise to poisson probability law. That is poisson distribution can be derived as a limiting case of binomial distribution under the following conditions.

i) n, the number of independent Bernoulli trials is very large, i.e. $n \to \alpha$

ii) p, the probability of success in each trial is very small, i.e. $p \to \alpha$

iii) $np = \lambda$ (say), is finits, where λ is a positive real number. under the above conditions, we can derive the p.m.f. of poisson distribution as follows:

Let us consider the p.m.f. of poisson distribution as,

$$p(x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, ..., ..., ..., n; \; q = 1 - p.$$

$$\text{Now, } \lim_{n \to \alpha} p(x) = \lim_{n \to \alpha} \binom{n}{x} p^x q^{n-x}$$

$$= \lim_{n \to \alpha} \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \quad \because np = \lambda$$

$$= \frac{\lambda^x}{n!} \lim_{n \to \alpha} \frac{n!}{(n-x)!n^x} \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{\lambda^x}{n!} \lim_{n \to \alpha} \frac{n!}{(n-x)!n^x} \lim_{n \to \alpha} \left(1-\frac{\lambda}{n}\right)^n \lim_{n \to \alpha} \left(1-\frac{\lambda}{n}\right)^{-x}$$

$$= \frac{\lambda^x}{n!} \lim_{n \to \alpha} \frac{n(n-1)(n-2)...(n-\overline{x-1})(n-x)!}{(n-x)!n^x} \quad \therefore e^{-\lambda}$$

$$= \frac{\lambda^x}{n!} \lim_{n \to \alpha} \frac{n^x\left\{1.\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)...\left(1-\frac{x-1}{n}\right)\right\}}{n^x}, e^{-\lambda}$$

$$= \frac{e^{-\lambda}\lambda^x}{x!} \quad \because \lim_{n \to \alpha} \frac{1}{n} \to 0$$

Which is the p.m.f of poisson distribution. We give below the formal definition of poisson distribution as,

Definition: A discrate random variable X is said to follow poisson distribution, if it assumes non-negative values only and its p.m.f is given by,

$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, x = 0, 1, 2, ..., ..., ..., λ

= 0, else where.

λ is the only parameter of this distribution.

10.4 MOMENTS OF POISSON DISTRIBUTION

The arithmatic mean of poisson distribution is given by,

$$\mu_1' = E(x) = \sum_x xp(x)$$

$$= \sum_{x=0}^{\alpha} x\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= e^{-\lambda}.\lambda \sum_{x=1}^{\alpha} \frac{x\lambda^{x-1}}{x(x-1)!}$$

$$= e^{-\lambda}.\lambda \sum_{x=1}^{\alpha} \frac{\lambda^{x-1}}{(x-1)!}$$

$$= e^{-\lambda}.\lambda\left[1+\lambda+\frac{\lambda^2}{2!}+...\right]$$

$= e^{-\lambda}.\lambda.e^{\lambda}$

$= \lambda$

$\therefore$ the mean of poisson distribution is λ.

$\mu_2' = E(x^2) = E\{x(x-1)\} + E(x)$

Now, $= E\{x(x-1)\}$

$$= \sum_x x(x-1)p(x)$$

$$= \sum_{x=0}^{\alpha} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!}$$

The exponential exponsion,

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots$$

$$= e^{-\lambda}.\lambda^2 \sum_{x=2}^{\alpha} x(x-1)\frac{\lambda^{x-2}}{x.(x-1)(x-2)!}$$

$$= e^{-\lambda}.\lambda^2 \sum_{x=2}^{\alpha} \frac{\lambda^{x-2}}{(x-2)!}$$

$$= e^{-\lambda}.\lambda^2 \left[1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots\right]$$

$= e^{-\lambda}.\lambda^2.e^{\lambda}$

$= \lambda^2$

$\therefore E(\lambda^2) = \lambda^2 + \lambda$

Now, $\nu(x) = \mu_2 = \mu_2' - \mu_1'^2$

$= \lambda^2 + \lambda - \lambda = \lambda$

Hence the variance of poisson distribution is also λ.

Remark: Poisson distribution is the only discrate distribution where mean is equal to variance.

10.5 MOMENT GENERATING FUNCTION

The moment generating function of poisson distribution is given by,

$$mx(t) = E(e^{tx}) = \sum_{x=0}^{\alpha} e^{tx}.\frac{e^{-\lambda}\lambda}{x!}$$

$$= e^{-\lambda} \sum_{x=0}^{\alpha} \frac{\left(e^t\lambda\right)^2}{x!}$$

$$= e^{-\lambda}\left[1 + e^t\lambda + \frac{\left(e^t\lambda\right)^2}{2!} + \frac{\left(e^t\lambda\right)^3}{3!} + \ldots\right]$$

$$= e^{-\lambda} e^{e^t\lambda}$$

$$= e^{-\lambda}(1 - e^t)$$

10.6 FITTING OF POISSON DISTRIBUTION

The recurrence relation for probabilities of poisson distribution is,

$$p(x+1) = \frac{\lambda}{x+1} p(x), \; x = 0, 1, 2, ..., ..., ... \qquad (1)$$

Putting x = 0, 1, 2, ..., ..., ... in (1) we get the probabilities p(1), p(2), ..., ..., ... etc. so we need to calculate p(0) only, which is,

$$p(0) = \frac{e^{-\lambda}\lambda^0}{0^0} = e^{-\lambda}$$

where, λ is the mean of poisson distribution is to be obtained from data to be fitted.

Ultimately, the expected frequencies are calculated by multiplying the corresponding probabilities by N, the total frequency. We shall explain the fitting of poisson distribution to a given data set.

Example: 200 drivers made the following road accidents

Accidents:	0	1	2	3	4
No. of drivers:	122	60	15	2	1

Fit a theoretical distribution to the data.

Solution: The suitable theoretical distribution to be fitted here may be the poission distribution,

Let, x_i = number of road accidents

f_i = no. of drivers.

Total number of drivers = N = 200

Here, λ = mean of poisson distribution

$$= \frac{1}{N}\sum_{i=1}^{n} fixi$$

$$= \frac{100}{200}$$

$$= 0.5$$

$$\therefore p = E^{-0.5} = 0.60653$$

Using recurrence relation we get,

$$p(1) = \frac{\lambda}{0+1} p(0) = \lambda.p(0) = 0.5 \times 0.60653 = 0.30327$$

$$p(2) = \left\{e^{-1} + e^{-1} + \frac{e^{-1}}{2!}\right\} p(1) = \frac{0.5}{2} \times 0.60653 = 0.07582$$

$$p(3) = \frac{\lambda}{3} p(2) = \frac{0.5}{3} \times 0.60653 = 0.01264$$

$$p(4) = \frac{\lambda}{4} p(3) = \frac{0.5}{4} \times 0.60653 = 0.00158$$

$\therefore$ the expected frequencies are,

$f(0) = N \times p(0) = 200 \times 0.60653 = 121.306 \approx 121$

$f(1) = N \times p(1) = 200 \times 0.30327 = 60.654 \approx 61$

$f(2) = N \times p(2) = 200 \times 0.07582 = 15.164 \approx 15$

$f(3) = N \times p(3) = 200 \times 0.01264 = 2.528 \approx 3$

$f(4) = N \times p(4) = 200 \times 0.00158 = 0.316 \approx 0$

Total frequency = 200

10.7 PROPERTIES OF POISSON DISTRIBUTION

i) Poisson distribution is a discreta distribution where the random variable X lakes the values 0, 1, 2, ..., ..., ..., λ.

ii) The poisson distribution has only one parameter λ.

iii) The mean and the variance of poisson distribution are equal and is equal to λ.

iv) The poisson distribution has only one or two modes.

v) Poisson distribution can be obtained as a lemiting case of beniomial distribution.

vi) If X and Y are independent poisson variates with paremeters λ_1 and λ_2 respectively, then x + y is also a poisson variate with parameter $\lambda_1 + \lambda_2$.

10.8 APPLICATION OF POISSON DISTRIBUTION

Physical situations in daily life where poisson distribution is applicable is

i) Number of telephone calls recieved per minute at a particular switch board.

ii) Number of accidents in a city per unit time.

iii) Number of defective parts produced in a factory per unit time.

iv) Number of printing mistakes in each page of a book.

10.9 ILLUSTRATIVE EXAMPLES

Example 1: Find the mean and variance of poisson distributionsuch that, P(x = 2) = P(x = 1)

Solution: It X follows poisson distribution with parameter λ,

then, $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, x = 0, 1, 2, ..., ..., ...

Now, p(x = 2) = p(x = 1)

$$\Rightarrow \frac{e^{-\lambda}\lambda^2}{2!} = \frac{e^{-\lambda}.\lambda}{1!}$$

$$\Rightarrow \frac{\lambda}{2} = 1$$

$$\Rightarrow \lambda = 2$$

$\therefore$ the man and variance of poisson distribution is equal to $\lambda = 2$.

Example 2: The number of teckellers travallers deteched per checking follows poisson distribution with mean 10. Find the probability of getting two ticketters tracullers.

Solution: Let X denote the no. of ticketters traceallers in a checking. so X follows poisson distribution with given mean.

$\lambda = 10$

$\therefore p(x) = \frac{e^{-10}10^x}{x!}$, x = 0, 1, 2, ..., ..., ...

$\therefore$ P(getting two ticketters tracullers during a checking)

= P(x = 2)

$$= \frac{e^{-10}10^2}{2!}$$

$$= 0.0025.$$

Example 3: What probability model is appropriats to describes a situation where 100 misprints are distributed randomly throughout the 100 pages of a book? For this model what is the probability that a page observed at random will contain at least 3 misprients.

Solution: Since 100 misprints are distrebuted randomly throughout the 100 pages of the book, the probability that there being a mesprint is p = $\frac{1}{100}$ = 0.01, which is very small.

n = 100 pages is quite large.

Hence poisson distribution is but reaited in this case.

Here, $\lambda = np = 0.01 \times 100 = 1$

$\therefore$ Probability that there are alteast three misprints in a page is:

$p(x \geq 3) = 1 - p(x < 3)$

$= 1 - \{P(x = 0) + P(x = 1) + P(x = 2)$

$$= 1 - \left\{e^{-1} + e^{-1} + \frac{e^{-1}}{2!}\right\}$$

$= 1 - 2.5e^{-1}$

$= 1 - 2.5 \times 0.3679$

$= 0.0802$



CHECK YOUR PROGRESS

Q. 1: Fill in the blanks:

i) For a poisson distribution mean is ..... to variance (less/equal/greater)

ii) The s.d. of a poisson distribution with mean 4 is ..... (4/2).

iii) For a poisson distribution $\frac{e^{-1.5}(1.5)^x}{x!}$, x = 0, 1, 2, ... the varience is (1.5/2.5).

iv) If x follows poisson distribution is ..... $\frac{e^{-\lambda(1-e^t)}}{e^{-\lambda}(1-2)^t}$.

v) If x_i, i = 1, 2, ..., ..., ..., n are n poisson variates then the distribution of $\sum_{i=1}^{n} x_i$ is also a ..... variate. (Binomail/ Poisson)

Q.2: "Poisson distribution is also known as probability distribution of rare events." –Explain.

Q.3: Give some examples of poisson distribution.



10.10 LET US SUM UP

- Poisson distribution can be derived as a limiting case of benomial distribution when, $n \to \alpha$, $p \to o$ and $np \to \lambda$ (say).
- The p. m. f. of poisson distribution with parameter λ is given by,

  $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, x = 0, 1, 2, ..., ..., ... λ.
- Poisson distribution plays an important role in the discription of random phenomena in a continuous interval of any kind. For example, telephone calls at a telephone exchange within a specified period of time.
- Sum of independent poisson variates is also a poisson variate.



10.11 FURTHER READING

1) Das, K. K, Bhattacharjee, D. (2008). *A Treatere on Statistical Inference and Distributions*. Asian Books Pvt. Ltd.

2) Gupta, S. C. and Kapoor, V. K. (2002). *Fundementals of Mathematical Statistics*. Sultan Chand & Sons.

3) Goon, A. M., Gupta. M. P. and Dasgupta, B. (2002). *Fundementals of Statistics* (Vol. I). Seventh Revised Edition. Kolkata: The World Press

Pvt. Ltd.

4) Saxena, H. C. (1994). *Probability and its Application*. Sixth Edition. New Delhi: S. Chand and Company.

5) Singh, R. (2012). *An Introduction to Probability and Probability Distributions*. Books and Ellied (P) Ltd.



10.12 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: i) equal, ii) 2, iii) 1.5, iv) $e^{-\lambda(1-e^{t})}$, v) poisson.

Ans. to Q. No. 2: Poisson distribution is also known as probability distribution of rare events, because poisson distribution is a positivaly skewed distribution, i.e probabilities tend to zero for large number of occurrences. For these poisson distribution is also known as probability distribution of rare events.

Ans. to Q. No. 3: Some examples of poisson distribution are,

i) number of deaths from suicide reported in a particular city.

ii) number of defictive items in a lot of manifactures items.

iii) number of printing mistakes in each page of a book.



10.13 MODEL QUESTIONS

Q.1: Write a note on poisson distribution.

Q.2: Obtain poisson distribution as a limiting case of binomial distribution.

Q.3: If X follows poisson distribution with $V(\lambda) = 2$, find its p.m.f.

Q.4: Define poisson distribution, state the conditions under which the binomial p.m.f. can be approximated by the poisson p.m.f.

Q.7: Obtain m.g.f. of poisson distribution and hence find its mean.

*** ***** ***

UNIT 11: THEORETICAL DISTRIBUTIONS (CONTINUOUS VARIABLE)

UNIT STRUCTURE

11.1 Learning Objectives

11.2 Introduction

11.3 Continuous Probability Distribution

11.4 Normal Distribution

11.4.1 Properties of Normal Distribution

11.4.2 Standard Normal Variate

11.5 Let Us Sum Up

11.6 Further Reading

11.7 Answers To Check Your Progress

11.8 Model Questions

11.1 LEARNING OBJECTIVES

After going through this unit, you will able to:

- continuous probability distribution and probability deusity function
- normal distribution
- standard normal variate
- properties of normal distribution.

11.2 INTRODUCTION

Normal distribution is a continuous distribution. It is a limiting case of the binomial distribution. It has immense application. It is the most popular and commonly used distribution. It was discovered by Abrahoun de Moivre (1667–1754) in 1733.

11.3 CONTINUOUS PROBABILITY DISTRIBUTION

If a random variable is a continuous variable, its probability distribution is called a continuous probability distribution.

A continuous probability distribution cannot be expresed in tabular form. An equation or formula is used to describe a continuous probability

distribution. The equation used to describe a continuous probability distribution is called a probability. density function (p.d.f.)

11.4 NORMAL DISTRIBUTION

Normal distribution is derived from the binomial distribution under two conditions, viz.

i) n, the number of trials, is sufficiently large, i.e. $n \to \infty$

and ii) neither p nor q is very small.

A random variable X is said to follow normal distribution, if and only if, its probability density function is given by

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

Where x is the real value of X, i.e. $-\infty < X < \infty$

The variable X is said to be distributed normally with mean μ and variance α^2 symbolically we write $X \sim N(\mu, \alpha^2)$. It is read as 'X is distributed as $N(\mu, \alpha^2)$'.

Note:

1) The probability density function given by (1) has two parameters, viz., μ and α.

   μ is any real number, i.e. $-\infty < \mu < \infty$ whereas α is a positive real number, i.e., $\alpha > 0$.

2) We know that probability can never be negative.

   $\therefore f(x) \geq 0, \ \forall x$

3) If $\mu = 0$, $\alpha = 1$ the p.d.f is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

In this case $X \sim N(0, 1)$. In this case X is called the standard normal variate, and the p.d.f given by (2) is called standard normal distribution.

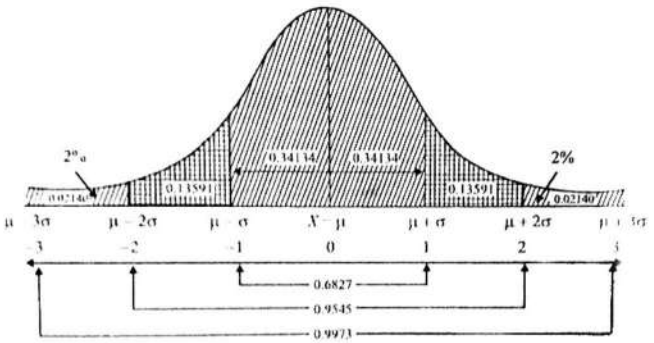
11.4.1 Properties of Normal Distribution

Normal distribution has many properties, some of which are given below.

1) The equation of the normal probability curve is:

$$Y = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

It is of the shape given below. It is bell-shaped.



[*Source:* **Biostatistics** by Dr. P. N. Arora and Dr. P. K. Malhan (Himalya Publishing House, 2010)]

2) The curve is symmetrical about the line x = $\propto$, and x lies between $-\propto$ and $+\propto$.

3) At x = $\propto$, the ordinate is maximum, and it is $Y = \frac{1}{\sigma\sqrt{2\pi}}$.

4) The total area under the normal curve within the limits $-\propto$ to $\propto$ is 1, i.e.,

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1$$

5) The percentage distribution of area under the normal curve given below, and is shown in the figure given above.
   i) About 68% of the area falls between $\mu - \alpha$ and $\mu + \alpha$.
   ii) About 95.5% of the area falls between $\mu - 2\alpha$ and $\mu + 2\alpha$.
   iii) About 99.7% of the area falls between $\mu - 3\alpha$ and $\mu + 3\alpha$.

6) The curve has a single peak, i.e., it is unimodal.

7) The distribution is symmctrical. Mean = median = mode = μ.

8) The curve is asymptotic, i.e., it meets the x-axis at both the ends at infinity. Theeretically the curve never towches the x-axis.

9) The points of inflexion are x = $\mu - \alpha$ and x = $\mu + \alpha$. [Inflexion points are where the function changes concavity.]

10) Coefficient of skewness = 0, and kurtosis = 3.

11) Mean deviation about mean $\simeq \frac{4}{5}\alpha$.

12) Q.D. $\simeq \frac{2}{3}\alpha$

11.4.2 Standard Normal Variate

Let a random variable X be distributed as $N(\mu, \alpha^2)$. The location and shape of the normal curve depends on μ, and α, Where $-\infty < \mu < \infty$, and $\alpha > 0$. Therefore, no master table for the area can be prepared. This hindrance can be removed if we introduce a new variable Z defined by.

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

In this case Z is called the standard normal variate.

Now we prove that the mean of Z is 0, and the standard deviation is 1. Before that we state some related properties.

We know that mean of a random variable X is given by

mean = $E(X) = \Sigma x_i p_i$

We prove that–

i) $E(a) = a$; a is a constant

ii) $E(aX + b) = aE(x) + b$; a, b are constants

iii) $E(X - \bar{X}) = 0$

Proof: i) $E(a) = \Sigma a p_i$

$= a\Sigma p_i$

$= a$, $\because \Sigma p_i = 1$

ii) $(aX + b) = \Sigma(ax_i + b) = \Sigma(ax_i + b)p_i$

$= a\Sigma x_i p_i + b\Sigma p_i$

$= aE(X) + b$, $\because \Sigma p_i = 1$

iii) $E(X - \bar{X}) = E(X) - E(\bar{X})$

$= \bar{X} - \bar{X}$, $\because \bar{X}$ is constant

$= 0$

Also, we note that–

iv) Var (aX) = a^2 Var (X), where a is a constant

v) Var (aX $\pm$ b) = a^2 Var (X), where a, b are constants.

Now we prove that the mean of Z is 0, and the standard deviation is 1.

Mean = E (Z)

$$= E\left(\frac{X-\mu}{\alpha}\right)$$

$$= \frac{1}{\alpha}E(X-\mu), \qquad \text{property (ii)}$$

$$= \frac{1}{\alpha}[E(X)-\mu], \qquad \text{property (iii)}$$

$$= \frac{1}{\alpha}[\mu-\mu]$$

$$= 0$$

Variance = Var (Z)

$$= Var\left(\frac{X-\mu}{\alpha}\right)$$

$$= \frac{1}{\alpha^2}var(X-\mu), \qquad \text{property (v)}$$

$$= \frac{1}{\alpha^2}var\ X, \qquad \text{property (iv)}$$

$$= \frac{1}{\alpha^2}.\alpha^2$$

$$= 1$$

Thus, $Z = \frac{X-\mu}{\alpha}$ is distributed as N (0, 1), i.e., Z ~ N(0, 1).

All the properties for a normal variable hold good for standard normal variable (μ = 0, α = 1).

Let Z be a standard normal variate N(0, 1). The area under the standard normal curve is given by a table which is addended at the end of this chapter.

We know that the total area under the normal curve within the limits $-\infty$ to ∞ is 1. The standard normal curve is symmetrical about Z = 0. Therefore, the area to the left of z = 0 is 0.5, and the

area to the right of z = 0 is 0.5.

Suppose, we are to find the area between z = 0, and z = 1.89, i.e, we are to find out $P(0 \leq Z \leq 1.89)$.

First we find 1.8 at the left of the table and single out that row.

Next we find 0.09 at the top of the table and single out that column.

Then we obtain the entry where the row for 1.8 and the column for 0.09 meet. the entry is 0.47062.

Thus $P(0 \leq Z \leq 1.89)$

= (the are from z = 0 to z = 1.89)

= 0.47062.

Example 1: Find the probability of the standard normal variate z ehen z is greater than 1.09.

Solution: We know that the total area under the normal curve within the limits $-\infty$ to ∞ is 1. Therefore, the area to the right of z = 0 is 0.5, as the curve is symmetrical about z = 0.

$\therefore$ P(Z > 1.09)

= 0.5 – 0.36214 = 0.13786

Example 2: Find the probability of the standard normal variate z when z is less than 1.09.

Solution: We know that the total area under the normal curve within the limits $-\infty$ to $+\infty$ is 1. Therefore, the area to the left of z = 0 is 0.5, and the area to the right or z = 0 is 0.5.

Again, from the table, we get that the area from z = 0 to z = 1.09 is 0.36214.

$\therefore$ P(z < 1.09)

= (area to the left of z = 0) + (area from z = 0 to z = 1.09)

= 0.5 + 0.36214

= 0.86214

Example 3: Find P(0.5 < z < 1.09)

Solution: We know that the area to the right of z = 0 is 0.5. From the table, the area from z = 0 to z = 0.5 is 0.19146. Again, the area from z = 0 to z = 1.09 is 0.36214.

$\therefore$ P(0.5 < z < 1.09)

= (Area from z = 0 to z = 1.09) – (area from z = 0 to x = 0.5)

= 0.36214 – 0.19146

= 0.17068

Example 4: A large number of measurements is normally distributed with mean 65.5 cm and standard deviation 6.2 cm. Find the percentage of measurements that fall between 54.8 cm and 68.8 cm.

Solution: Let X be the normal variate showing the large number of measurements.

We have $\mu = 65.5$, $\alpha = 6.2$

Now, $Z = \frac{X - \mu}{\alpha} = \frac{X - 65.5}{6.2}$

When, X = 54.8 $Z = \frac{54.8 - 65.5}{6.2} = \frac{-10.7}{6.2} = -1.7$

When, X = 68.8, $Z = \frac{68.8 - 65.5}{6.2} = 0.5$

Thus, when X lies between 54.8 and 68.8, Z lies between –1.7 and 0.5

Now P(–1.7 < X < 0.5)

= P(–1.7 < Z < 0) + P(0 < Z < 0.5)

= 0.45543 + 0.19146

= 0.64689 = 64.69%

$\therefore$ the percentage of measurements is 64.69%

Example 5: For the distribution of makrks of candidates in an examination, mean is 14 and standard deviation is 2.5. Assuming the normality of the distribution, find the probability that a randomly selected student will score above 15.

Solution: Let X be the normal variate showing the scores of candidates.

We have $\mu = 14$, $\alpha = 2.5$

Now, $Z = \frac{X - \mu}{\alpha} = \frac{X - 14}{2.5}$

When X, $Z = \frac{15 - 14}{2.5} = \frac{1}{2.5} = 0.4$

$\therefore$ P(x > 15) = P(z > 0.4)

= 0.5 – (Area between z = 0 and Z = 0.4)

= 0.5 – 0.15542

= 0.34458.



CHECK YOUR PROGRESS

Q.1: Find: (i) P (Z > 2.58), (ii) P(Z < 1.96), (iii) P (Z > –3)

Q.2: For a certain normal distribution, mean is 50, and standard deviation is 5. Find P($45 \leq X \leq 55$), X being a normal variate.

Q.3: A normal variate X has mean 100 and standard deviation 4. Find the probability that X lies between 92 and 10.8

Q.4: The distribution of scores of students in an examination is normal. If the mean is 45, and the standard deviation is 5, find the percentage of students who scored between 40 and 50.

Q.5: The marks obtained in a certain examination follow normal distribution with mean 45, and standard deviation 10. If 1000 students appear at the examination, calculate the number of students scoring (i) less than 30 marks, and (ii) more thaw 80 marks.



11.5 LET US SUM UP

- Normal distribution is a continous distribution. It is a limiting case of the binomial distribution.
- Normal distribution is derived from the binomial distribution under two conditions, viz.,

  i) n, the number of trials, is sufficiently large, i.e., $n \to \propto$ and

  ii) neither p nor q is very small.

  A random variable X is said to follow normal distribution, if and only if, its probability density function is given by

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

Where x is the real value of X, i.e., $-\infty < X < \infty$

The variable X is said to be distributed normally with mean μ and variance α^2. Symbolically, we write $X \sim N(\mu, \alpha^2)$. It is read as 'X' is distributed as $N(\mu, \alpha^2)$.

- The normal curve is bell-shaped, and the curve is symmetrical about the line $x = \mu$, and x lies between $-\infty$, and $+\infty$.
- The total area under the normal curve within the limits $-\infty$ to ∞ is 1.
- If a new variable Z is defined by $Z = \frac{X - \mu}{\alpha}$, Z is called the standard normal variate. The mean of Z is 0, and the standard deviation is 1.



11.6 FURTHER READING

1) Das, K. K, Bhattacharjee, D. (2008). *A Treatere on Statistical Inference and Distributions*. Asian Books Pvt. Ltd.
2) Gupta, S. C. and Kapoor, V. K. (2002). *Fundementals of Mathematical Statistics*. Sultan Chand & Sons.
3) Goon, A. M., Gupta. M. P. and Dasgupta, B. (2002). *Fundementals of Statistics* (Vol. I). Seventh Revised Edition. Kolkata: The World Press Pvt. Ltd.
4) Kakoti, S. C. (2003). *Mathematical Statistics: Theory and Applications. Dibrugarh*, Assam: Kasturba Prakashans.
5) Saxena, H. C. (1994). *Probability and its Application*. Sixth Edition. New Delhi: S. Chand and Company.



11.7 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: i) We know that tha total area under the normal curve within the limits $-\infty$ to $+\infty$ is 1. Therefore, the area to the left of z = 0 is 0.5, and the area to the night of z = 0 is 0.5. Again, from the table, we get that the area from z = 0 to z = 2.58 is 0.49506

$\therefore$ P (z > 2.58)

= (area to the right of z = 0) – (area from z = 0 to z = 2.58)

= 0.5 – 0.49506

= 0.00494

ii) We know that the total area under the normal curve within the limits $-\infty$ to $+\infty$ is 1. Therefore, the area to the left of z = 0 is 0.5, and the area to the right of z = 0 is 0.5. Again from the table, we get that the area from z = 0 to z = 1.96 is 0.47500.

$\therefore$ P (z < 1.96)

= (area to the left of z = 0) + (area from z = 0 to z = 1.96)

= 0.5 + 0.47500

= 0.97500

iii) We know that the total area under the normal curve within the limits $-\infty$ to ∞ is 1. Therefore, the area to the left of z = 0 is 0.5, and the area to the right of z = 0 is 0.5.

Now, P(Z > –3)

= (area from z = –3 to z = 0) + (area to the right of z = 0)

= (area from z = 0 to z = 3) + (area to the right of z = 0)

= 0.49865 + 0.5

= 0.99865

Ans. to Q. No. 2: X is a normal variate.

We have $\mu = 50$, $\alpha = 5$

Now, $Z = \frac{X - \mu}{\alpha} = \frac{X - 50}{5}$

When X = 45, $Z = \frac{45 - 50}{5} = -1$

When X = 55, $Z = \frac{55 - 50}{5} = 1$

Thus, when X lies between 45 and 55, Z lies between –1 and 1.

Now $P(-1 \le Z \le 1)$

$= P(-1 \le Z \le 0) + P(0 \le Z \le 1)$

= 0.34134 + 0.34134

= 0.68268

Ans. to Q. No. 3: Do yourself.

Ans. to Q. No. 4: Do yourself.

Ans. to Q. No. 5: Let X be the normal variate showing marks.

We have $\mu = 45$, $\alpha = 10$

Now $Z = \frac{X - \mu}{\alpha} = \frac{X - 45}{10}$

When X = 30, $Z = \frac{30 - 45}{10} = \frac{-15}{10} = -1.5$

When X = 80, $Z = \frac{80 - 45}{10} = \frac{35}{10} = 3.5$

i) Now $P(X \leq 30)$

= (area to the left of z = 0) – (area from z = – 1.5 to z = 0)

(area to the left of z = 0) – (area from z = 0) to z = 1.5)

= 0.5 – 0.43319 = 0.06681

$\therefore$ the required number of students

= 1000 x 0.06681

= 67 (approximately)

ii) $P(Z \geq 80)$

= $P(Z \geq 3.5)$

= (area to the right of z = 0) – (area from z = 0 to z = 3.5)

= 0.5 – 0.49977

= 0.00023

Now 1000 × 0.00023 = 0.23

$\therefore$ there are no students scoring more than 80 marks.



11.8 MODEL QUESTIONS

Q.1: Define standard normal variate and write down the probability density function of this variate.

Q.2: State the conditions under which the normal distribution is obtained as a dimiting case of the binomial distribution.

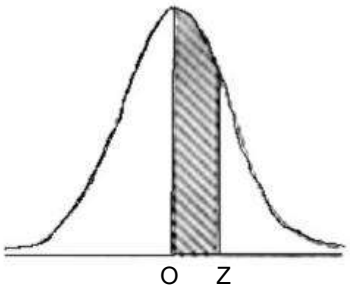
Q.3: Find (i) $P(Z \leq -1.96)$, (ii) $P(Z \leq 1.96)$, (iii) $P(Z > 2.58)$

Q.4: A normal variate X has mean 50 and standard deviation 5. Find

$P(40 \leq X \leq 60)$.

Q.5: The mean of a normal distribution is 50 and the standard deviation is 4. What percentage of total frequency lies between 42.16 and 57.84?

Q.6: Let a continuous variable X follow a normal distribution with mean 12 and standard deviation 2. What is the probability that the value of X selected at random lies between 11 and 14?



Area under the Standard Normal Curve

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.00000	.00399	.00798	.01197	.01595	.01994	.02392	.02790	.03188	.03586
0.1	.03983	.04383	.04776	.05172	.05567	.05962	.06356	.06749	.07142	.07535
0.2	.07926	.08317	.08706	.09095	.09483	.09871	.10257	.10642	.11026	.11409
0.3	.11791	.12172	.12552	.12930	.13307	.13683	.14058	.14431	.14803	.15173
0.4	.15542	.15910-	.16276	.16640	.17003	.17364	.17724	.18082	.18439	.18793
0.5	.19146	.19497	.19847	.20194	.20540	.20884	.21226	.21566	.21904	.22240
0.6	.22575	.22907	.23237	.23565	.23891	.24215	.24537	.24857	.25175	.25490
0.7	.25804	.26115	.26424	.26730	.27035	.27337	.27637	.27935	.28230	.28524
0.8	.28814	.29103	.29389	.29673	.29955	.30234	.30511	.30785	.31057	.31372
0.9	.31594	.31859	.32121	.32381	.32639	.32894	.33147	.33398	.33646	.33891
1.0	.34134	.34375	.34614	.34849	.35083	.35314	.35543	.35769	.35993	.36214
1.1	.36433	.36650	.36864	.37076	.37286	.37493	.37698	:37900	.38100	.38298
1.2	.38493	.38686	.38877	.39065	.39251'	.39435	.39617	.39796	.39973	.40147
1.3	.40320	.40490	.40658	.40824	.40988	.41198	.41309	.41466	.41621	.41774
1.4	.41924	.42073-	.42220	.42364	.42507	.42647	.42785	.42922	.43056	.43189
1.5	.43319	.43448	.43574	.43699	.43822	.43943	.44062	.44179	.44295	.44408
1.6	.44520	.44630	.44738	.44845	.44950	.45053	.45154	.45254	.45352	.45449
1.7	.45543	.45637	.45728	.45818	.45907	.45994	.46080	.46164	.46246	.46327

1.8	.46407	.46485	.46562	.46638	.46712	.46784	.46856	.46926	.46995	.47062
1.9	.47128	.47193	.47257	.47320	.47381	.47441	.47500	.47558	.47615	.47670
2.0	.47725	.47784	.47831	.47882	.47932	.47982	.48030	.48077	.48124	.48169
2.1	.48214	.48257	.48300	.48341	.48382	.48422	.48461	.48500	.48537	.48574
2.2	.48610	.48645	.48679	.48713	.48745	.48778	.48809	.48840	.48870	.48899
2.3	.48928	.48956	.48983	.49010	.49036	.49061	.49086	.49111	.49134	.49158
2.4	.49180	.49202	.49224	.49245	.49266	.49286	.49305	.49324	.49343	.49361
2.5	.49379	.49396	.49413	.49430	.49446	.49461'	.49477	.49492	.49506	.49520
2.6	.49534	.49547	.49560	.49573	.49585	.49598	.49609	.49621	.49632.	.49643
2.7	.49653	.49664	.49674	.49683	.49693	.49702	.49711	.49720	.49728	.49736
2.8	.49744	.49752	.49760	.49767	.49774	.49781	.49788	.49795	.49801	.49807
2.9	.49813	.49819	.49825	.49831	.49836	.49841	.49846	.49851	.49856	.49861
3.0	.49865	.49869	.49874	.49878	.49882	.49886	.49889	.49893	.49896	.49900
3.1	.49903	.49906	.49910	.49913	.49916	.49918	.49921	.49924	.49926	.49929
3.2	.49931	.49934	.49936	.49938	.49940	.49942	.49944	.49946	.49948	.49950
3.3	.49952	.49953	.49955	.49957	.49958	.49960	.49961	.49962	.49964	.49965
3.4	.49966	.49968	.49969	.49970	.49971	.49972	.49973	.49974	.49975	.49976
3.5	.49977	.49978	.49978	.49979	.49980	.49981	.49981	.49982	.49983	.49983
3.6	.49984	.49985	.49985	.49986	.49986	.49987	.49987	.49988	.49988	.49989
3.7	.49989	.49990	.49990	.49990	.49991	.49991	.49992	.49992	.49992	.49992
3.8	.49993	.49993	.49993	.49994	.49994	.49994	.49994	.49995	.49995	.49995
3.9	.49995	.49995	.49996	.49996	.49996	.49996	.49996	.49996	.49997	.49997
4.0	.49997	–	–	–	–	–	–	–	–	–
4.5	.499997	–	–	–	–	–	–	–	–	–
5.0	.4999997	–	–	–	–	–	–	–	–	–

[*Source:* Basic statisties by B. L. Agarwal (New Age International Publishers, 2009)]

*** ***** ***

UNIT 12: INDEX NUMBERS

UNIT STRUCTURE

12.1 Learning Objectives

12.2 Introduction

12.3 Uses of Index Numbers

12.4 Types of Simple Index Numbers

12.5 Properties of Relatives

12.6 Construction of Simple Index Numbers

12.7 Construction of Weighted Index Numbers

12.8 Test of Adequacy of Index Numbers

12.9 Let Us Sum Up

12.10 Further Reading

12.11 Answers to Check Your Progress

12.12 Model Questions

12.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- discuss about the simple and weighted index numbers
- explain the uses of index numbers
- discuss the different types of index numbers
- explain the properties of different index numbers
- discuss the test of accuracy of index numbers.

12.2 INTRODUCTION

Index number is a statistical device (measure) with a purpose of showing average changes in one or more related variables (like price or quantity) between two periods of time (say, between 1991 and 1996) or two places like cities (say, Delhi and Mumbai) or countries (say, India and Japan). For example, we may be interested in knowing which city of India out of Delhi, Mumbai, Kolkata and Chennai is the costliest or the cheapest in terms of price level. A tourist may be interested in knowing about the cost of

living at different tourist places. Consumer price index number or cost of living index number helps in taking such decisions.

Different commodities are measured in different units. For example, rice and wheat are measured in kilograms, cloth in meters and milk in liters etc. Index numbers attempt some averages relating to commodities which are measured in different units. For purpose of comparison, we are interested in knowing relative changes. These relative changes are expressed in percentage terms.

Index numbers are the indicators of the various trends in an economy. Price index numbers indicate the position of prices whether they are rising or falling and at what rate. Similarly, index numbers regarding agricultural production indicates the trend of change whether it is rising or falling at what rate over a period of time. An index number is an economic data figure reflecting price or quantity compared with a standard or base value. The base usually equals 100 and the index number is usually expressed as 100 times the ratio to the base value. For example, if a commodity costs twice as much in 1970 as it did in 1960, its index number would be 200 relative to 1960. Index numbers are used especially to compare business activity, the cost of living, and employment.

An index number is specialized average. Index numbers may be simple or weighted depending on whether we assign equal importance to every commodities or different importance to different commodities according to the percentage of income spent on them or on the basis of some other criteria. In this chapter, we shall discuss both simple and weighted index numbers.

12.3 USES OF INDEX NUMBERS

In business an index is a statistical measure of changes in a representative group of individual data points.

Index numbers are constructed with various aims and purposes. Index numbers constructed are put into several uses. Some of them are described below:

1) The most commonly used types of index numbers are those that relate to changes in level in prices of a commodity or a group of commodities. Such index numbers are price index numbers. Price index numbers are useful in studying price movements with a view to analyzing their causes, as well as to determine their effects on the economy. To determine economic relationships, it is often useful to compare changes in the market price level with changes in other series, such as gold, bank deposits, bank loans, and the physical volume of production.
2) Indices of physical changes over a period of time in production, in marketing, in sales, in imports and exports, and so on have important uses. Such index numbers over a period of time are known as index time series. These index numbers are extremely useful in the study of the movement of the characteristics over time, in the study of seasonal and cyclic behaviour, and in the study of business cycles, etc.
3) Index numbers can be used for forecasting future trend in demand of commodities or in volume of production etc. Index numbers of industrial and agricultural production not only reflect the trend but also are useful in forecasting production. Index numbers of unemployment which reflect the trend of unemployment are useful in determining factors leading to unemployment and in forecasting future trends.
4) Index numbers which reflect the general price level, and in particular, the consumer price index numbers are useful in determining the quantum of additional wages or dearness allowances to compensate for the changes in the cost of living. The Government of India and most of the State Governments have index linked formulae for grant of dearness allowances to their employees.
5) The changes in consumer price index can be used to determine the real income over a period. Suppose that a person's income in 1980 was 150 times than that in 1970, that is, there is an increase of 50% in money income. Suppose that the cost of living index number in

1980 was 175 times than that in 1970, that is, there is an increase of 75%. Then the person's real income in 1980 is $\frac{150}{175}$ = 85.7% of his income in 1970. Thus index numbers are useful in determining the real income.

12.4 TYPES OF SIMPLE INDEX NUMBERS

1) **Price Relatives:** One of the simplest types of index numbers is a price relative. It is the ratio of the price of a single commodity in a given period or point of time to its price in another period or point of time, called the reference period or base period. If the prices for a period, instead of a point of time, are considered, then suitable price average for the period is taken and these prices are expressed in the same units. If p_0 and p_1 denote the price of a commodity during the base period or reference period (0) and the given period (1) then the price relative of the period 1 with respect to the base period 0 is defined by

Price relative in percentage (of period 1 with respect to 0)

$$= \frac{p_1}{p_0} \times 100 \qquad \ldots \ldots \ldots (7.1)$$

We denote price relative in percentage or without percentage by $\frac{p_1}{p_0}$.

Remark: In the formula of price relative we multiply $\frac{p_1}{p_0}$ by 100 only to get a better expression. This expression is therefore called price relative in percentage.

Exercise 1: If the retail price of fine quality of rice in the year 1980 was Rs.3.75 and that for the year 1983 was Rs.4.50, then find the price relative.

Solution: Here the base period is 1980. The price of rice in the base period was Rs. 3.75. Also in the given period 1983, the price of rice was Rs. 4.50. Using the formula (7.1), the required price relative is:

$$p_{\frac{1980}{1983}} = \frac{\text{Rs. } 4.50}{\text{Rs. } 3.75} \times 100 = 120\%$$

Exercise 2: The exchange rate of a US dollar was Rs. 40.00 in July 2008 and was Rs. 50.00 in February 2009. Find the price relative.

Solution: The price relative of a dollar in February 2009 (say period F) with respect to that in July 2008 (say period J) is given by

$$p_{\frac{F}{J}} = \frac{\text{Rs. } 50.00}{\text{Rs. } 40.00} \times 100 = 125\%$$

2) **Quantity Relatives:** Another simple type of index number is a quantity relative. This is useful when we are interested in changes in quantum or volumes of a commodity such as quantities of production or sale or consumption. Here the commodity is used in a more general sense. It may mean the volume of goods (in tonnes) carried by roadways or the volume of export to a country or import from a country. In such cases, we consider of quantity or volume relatives. If quantities or volumes are taken for a period instead of a point of time, a suitable average is to be taken and the quantities or volumes are to be expressed in the same units. If q_0 and q_1 denote the quantity or volume produced, consumed or transacted during the base period (0) and the given period (1) then quantity relative of the period 1 with respect to the base period 0 is defined by

Quantity relative in percentage (of period 1 with respect to 0)

$$= \frac{q_1}{q_0} \times 100 \qquad \ldots \ldots \ldots (7.2)$$

We denote quantity relative in percentage or without percentage by $q_{\frac{0}{1}}$.

Exercise 3: The production of tea in Assam in the month of January 2009 was 3608 tonnes and that in February 2009 was 3700 tonnes. Find the quantity relative.

Solution: The quantity relative of tea in February 2009 (say period F) with respect to that in January 2009 (say period J) is given by

$$q_{\frac{F}{J}} = \frac{3700}{3608} \times 100 = 102.55\%$$

3) Value Relatives: A value relative is another type of simple index number. It is usable when we wish to compare changes in the money value of the transaction, consumption or sale in two different periods or points of time. Multiplying the quantity q by the price p of the commodity produced, transacted or sold gives the total money value pq of the production, transaction or sale. If instead of point of time, period of time is considered, a suitable average is to be taken and is to be expressed in the same units.

Suppose p_0 and q_0 denote the price and quantity of the commodity during the base period (0) and p_1 and q_1 denote the corresponding price and quantity during a given period (1). Then the total value of the commodity during the base period is $v_0 = p_0q_0$ and the corresponding total value during the given period is $v_1 = p_1q_1$. The value relative of the period 1 with respect to the base period 0 is defined by

Value relative in percentage (of period 1 with respect to 0)

$$= \frac{v_1}{v_0} \times 100 = \frac{p_1q_1}{p_0q_0} \times 100 \quad \ldots \ldots \ldots (7.3)$$

We denote value relative in percentage or without percentage by $v_{\frac{0}{1}}$.

Exercise 4: The production of tea in Assam in the month of January 2009 was 3608 tonnes and that in February 2009 was 3700 tonnes. The corresponding prices of tea were Rs. 100.00 and Rs. 120.00 per kilogram, respectively. Find the value relative.

Solution: Here the base period is January 2009 (say period J) and the given period is February 2009 (say period F). Then, q_J = 3608000 kg, p_J = Rs. 100.00 and q_F = 3700000 kg, p_F = Rs. 120.00.

Hence, $v_J = q_J \times p_J$ = Rs. 360800000.00

$v_F = q_F \times p_F$ = Rs. 444000000.00

Using formula (7.3), we find the required value relative,

$$v_{\frac{F}{J}} = \frac{444000000}{360800000} \times 100 = 123\%$$



CHECK YOUR PROGRESS

Q.1: Write True (T) or False (F):

i) Index numbers are specialized averages to measure relative changes. (T/F)

ii) Index numbers measure net change in related variables over time. (T/F)

iii) Simple index numbers require weights in their construction. (T/F)

iv) Index numbers do not indicate the trend of change in the economy. (T/F)

v) Price relative is defined as the ratio of current year's price to some reference year's price. (T/F)

Q.2: Fill in the blanks:

i) Index number is a statistical device to express ..... change in related variable.

ii) An index number is a ..... average.

iii) Index number measures net change in a group of ..... variables.

iv) In the index number represented by $P_{\frac{1990}{1998}}$, the base year is ..... and the current year is .....

v) Simple index numbers have ..... weights.

vi) The price relative means ratio of current price to .....

vii) If the price relative increases from 100 to 300, it shows ..... increase in prices.

12.5 PROPERTIES OF RELATIVES

Let p_a, p_b, p_c, ..., ..., ... denote the prices in the periods a, b, c, ..., ..., ... respectively. Also, let q_a, q_b, q_c, ..., ..., ... denote the quantities and v_a, v_b, v_c, ..., ..., ... denote the values for the periods respectively. The relatives satisfy some properties which are directly obtained from the definitions. We

shall state the results for price relatives and write similar results for the quantity and value relatives.

1) **Identity property:** The price relative of a given period with respect to the same period is 1, that is, $p_{\frac{a}{a}} = 1$.

2) **Time reversal property:** If the base period and the reference period are interchanged, then the product of the corresponding relatives is unity. That is, one is the reciprocal of the other. In other words,

$p_{\frac{a}{b}} = \frac{1}{p_{\frac{b}{a}}}$. We know that $p_{\frac{a}{b}} = \frac{p_b}{p_a} = \frac{1}{\frac{p_a}{p_b}} = \frac{1}{p_{\frac{b}{a}}}$

3) **Circular or Cyclic property:** This property states that,

$p_{\frac{a}{b}} \times p_{\frac{b}{c}} \times p_{\frac{c}{a}} = 1$. We can prove this property as follows:

We know that $p_{\frac{a}{b}} = \frac{p_b}{p_a}$, $p_{\frac{b}{c}} = \frac{p_c}{p_b}$, $p_{\frac{c}{a}} = \frac{p_a}{p_c}$

The required property follows by multiplying the above three terms.

4) **Modified circular or cyclic property:** This property states that,

$p_{\frac{a}{b}} \times p_{\frac{b}{c}} = p_{\frac{a}{c}}$. We prove this property as follows:

We have $p_{\frac{a}{b}} \times p_{\frac{b}{c}} = \frac{p_b}{p_a} \times \frac{p_c}{p_b} = \frac{p_c}{p_a} = p_{\frac{a}{c}}$

12.6 CONSTRUCTION OF SIMPLE INDEX NUMBERS

In general we are interested in general price level rather than on the price level of any particular item. So we have to consider a number of items at the same time. If each commodity is of different qualities then all such qualities have also to be taken into consideration. We have to consider simultaneously a number of items or commodities together for measuring change in price or production.

The sources of data must be reliable and care should be taken in selecting the sources of data. For example, official publications as well as other reliable unofficial reports from producers and merchants should be considered for computing index for production. If retail food prices are being

considered, then prices in stores, not in wholesale markets, are to be taken into consideration.

Now we shall discuss how to construct simple index numbers from a given set of data, like prices of certain commodities, production of certain commodities etc.

There are two methods of constructing simple index numbers. These are:

1) Simple aggregate method and
2) Method of simple average of relatives.

Remember that all items in a simple index number are assigned equal weights. These are unweighted index numbers.

1) Simple Aggregate Method: In this case each item is given equal weights. If we give an equal weight to each item, it means the same thing, whether each item is given a weight or not. It is the simplest method of constructing index numbers.

1.1) Simple aggregate method to find price index numbers: We use the following three steps to find price index number.

Step 1: Find the sum of the current year (that is, given period) prices of all the items included in the list. If 1 denote the current year, then find Σp_1.

Step 2: Find the sum of the base year prices of all the items included in the list. If 0 denote the base year, then find Σp_0.

Step 3: Use the formula $p_{\frac{0}{1}} = \frac{\sum p_1}{\sum p_0} \times 100$

Exercise 5: Find the price index number, using simple aggregate method, for the following information.

Commodities	Units	Prices (in Rupees)	
		2000	2005
Wheat	Quintal	800	1200
Rice	Quintal	1200	1600
Milk	Litres	12	18
Tea	Kilograms	100	120
Clothing	Metres	30	40
Meat	Kilograms	80	120

Solution: Represent prices of 2005 (current year) as p_1 and that of 2000 (base year) as p_0 and reconstruct the table as follows:

Commodities	p_0	p_1
Wheat	800	1200
Rice	1200	1600
Milk	12	18
Tea	100	120
Clothing	30	40
Meat	80	120
	$\Sigma p_0 = 2222$	$\Sigma p_1 = 3098$

Therefore, the price index of 2005 with base year 2000 is:

$$p_{\frac{0}{1}} = \frac{3098}{2222} \times 100 = 139.4$$

Thus, the increase in prices (or price level) is 139.40 – 100 = 39.4%.

Exercise 6: The prices of three commodities A, B and C increased from Rs. 50, Rs. 7 and Rs. 3 in 2005 to Rs. 52, Rs. 8 and Rs. 5 respectively in 2006. Using the simple aggregate method, find by how much on an average the prices have increased?

Solution: We know construct the following table:

Commodities	Price in 2005 (p_0)	Price in 2006 (p_1)
A	50	52
B	7	8
C	3	5
	$\Sigma p_0 = 60$	$\Sigma p_1 = 65$

Therefore, the price index of 2006 with respect to the base period 2005 is:

$$p_{\frac{2005}{2006}} = \frac{\sum p_1}{\sum p_0} \times 100 = \frac{65}{60} \times 100 = 108.33$$

This shows that the price index of current year 2006 with respect to base year 2005 is 108.33%. Hence the price increase in one year is 108.33 – 100 = 8.33%.

1.2) Simple aggregate method to find quantity index numbers:

We use the following three steps to find quantity index number.

Step 1: Find the sum of the current year (that is, given period) quantity or volume produced, consumed or transacted of all the items included in the list. If 1 denote the current year, then find Σq_1.

Step 2: Find the sum of the base year quantity or volume produced, consumed or transacted of all the items included in the list. If 0 denote the base year, then find Σq_0.

Step 3: Use the formula $q_{\frac{0}{1}} = \frac{\sum q_1}{\sum q_0} \times 100$

2) **Method of Simple Average of Relatives:** In this method, average of the relatives is obtained by using any one of the measures of central tendency. In particular, here we shall use arithmetic mean for averaging the relatives. We shall discuss briefly this method only for price relatives. Similar formulas on quantity index number and value index number can be obtained considering quantity and value relatives instead of price relative. There are few questions to find quantity and value index numbers in the section ‘‘Possible Questions”. You should work out those problems with the help of the problems on price index numbers which are solved below.

2.1) Method of Simple Average of Price Relatives: We know that a price relative is nothing but the ratio of current year prices to those in the base year. Taking the arithmetic mean of the price relatives, we get the following formula for price index number:

$$p_{\frac{0}{1}} = \frac{\sum\left(\frac{p_1}{p_0} \times 100\right)}{N}$$

where N stands for number of commodities included in the index numbers.

2.2) Steps to calculate price index numbers by Simple Average of Price Relatives Method:

Step 1: Find percentage price relative for each commodity, that is,

$$\frac{p_1}{p_0}\times 100 .$$

Step 2: Find the sum of these percentage price relatives, that is,

$$\sum\left(\frac{p_1}{p_0}\times 100\right).$$

Step 3: Divide $\sum\left(\frac{p_1}{p_0}\times 100\right)$ by the total number of commodities included in the list.

Exercise 7: Give the solution of Exercise 5 using the Simple Average of Price Relatives.

Solution: We construct the following table:

Sl. No.	Commodities	Units	Prices (in Rupees)		Price Relative $\frac{p_1}{p_0}\times 100$
			2000 p_0	2005 p_1	
1	Wheat	Quintal	800	1200	$\frac{1200}{800}\times 100 = 150.00$
2	Rice	Quintal	1200	1600	$\frac{1600}{1200}\times 100 = 133.33$
3	Milk	Litres	12	18	$\frac{18}{12}\times 100 = 150.00$
4	Tea	Kilograms	100	120	$\frac{120}{100}\times 100 = 120.00$
5	Clothing	Metres	30	40	$\frac{40}{30}\times 100 = 133.33$
6	Meat	Kilograms	80	120	$\frac{120}{80}\times 100 = 150.00$
N = 6					$\sum\left(\frac{p_1}{p_0}\times 100\right) =$

Hence, the price index number of 2005 with respect to base year 2000 is:

$$P_{\frac{0}{1}} = \frac{\sum\left(\frac{p_1}{p_0}\times 100\right)}{N} = \frac{836.66}{6} = 139.44$$

Exercise 8: Give the solution of Exercise 6 using the Simple Average of Price Relatives.

Solution: We first prepare the following table:

Sl. No.	Commodities	Prices (in Rupees)		Price Relative
		2005 p_0	2006 p_1	$\frac{p_1}{p_0}\times 100$
1	A	50	52	$\frac{52}{50}\times 100 = 104.00$
2	B	7	8	$\frac{8}{7}\times 100 = 114.29$
3	C	3	5	$\frac{5}{3}\times 100 = 166.66$
N = 3				$\sum\left(\frac{p_1}{p_0}\times 100\right) = 384.95$

Hence, the price index number of 2006 with respect to base year 2005 is:

$$P_{\frac{0}{1}} = \frac{\sum\left(\frac{p_1}{p_0}\times 100\right)}{N} = \frac{384.95}{3} = 128.32$$

12.7 CONSTRUCTION OF WEIGHTED INDEX NUMBERS

In weighted index number each item is given weight according to the importance it occupies in the list. There are two groups of methods to calculate index number of this category: (a) Weighted aggregate method and (b) Weighted average of price relative methods. In this unit, we will study three types of weighted aggregate method. These methods are

popularly known as: (1) Laspeyre's Method, (2) Paasche's Method and (3) Fisher's Method. All the methods are used to calculate weighted index numbers. The main difference between the Laspeyre's Method and Paasche's Method is that Laspeyre's uses base year quantities of commodities as their relative weights, while Paasche's uses current year quantities of commodities as their relative weights for preparing a price index.

1) Laspeyre's Method:

Laspeyre's Price Index Number: It uses base year quantities as the weights. Accordingly, the formula is:

$$P_L = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

Laspeyre's Quantity Index Number: If we multiply the quantities by the base year price, then we get Laspeyre's quantity index number. Accordingly the formula for Laspeyre's quantity index number is:

$$Q_L = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$$

Steps to Calculate Weighted Index Number by Laspeyre's Method:

Step 1: Multiply current year price (p_1) with the base year quantity (q_0) to get p_1q_0 for each item/commodity.

Step 2: Multiply base year price (p_0) with the base year quantity to get p_0q_0 for each item/ commodity.

Step 3: Add all p_1q_0 and p_0q_0 separately to get Σp_1q_0 and Σp_0q_0.

Step 4: Divide Σp_1q_0 by Σp_0q_0 and multiply by 100 to obtain Laspeyre's price index number.

Exercise 9: Find Laspeyre's price index number from the following data.

Commodity	Price (in Rs.)		Quantity bought in units	
	1995	1996	1995	1996
A	50	52	10	12
B	7	9	3	4
C	3	5	7	6

Solution: Represent prices and quantities in current year by p_1 and q_1 respectively and represent prices and quantities in base year by p_0 and q_0 respectively. Let us construct the following table following the steps 1-3.

Commodities	p_0	q_0	p_1	p_0q_0	p_1q_0
A	50	10	52	500	520
B	7	3	9	21	27
C	3	7	5	21	35
				$\Sigma p_0q_0 = 542$	$\Sigma p_1q_0 = 582$

Therefore, the Laspeyre's price index number is:

$$P_L = \frac{582}{542} \times 100 = 107.38$$

2) Paasche's Method:

Paasche's Price Index Number: It uses current year quantities (q_1) as the weights. Accordingly, the formula is:

$$P_P = \frac{\sum p_1q_1}{\sum p_0q_1} \times 100$$

Paasche's Quantity Index Number: It uses current year prices (p_1) as the weights. Accordingly, the formula is:

$$Q_P = \frac{\sum q_1p_1}{\sum q_0p_1} \times 100$$

Exercise 10: Find Paasche's price index number from the data given in Exercise 9.

Solution: Represent prices and quantities in current year by p_1 and q_1 respectively and represent prices and quantities in base year by p_0 and q_0 respectively. Let us construct the following table.

Commodities	p_0	q_1	p_1	p_1q_1	p_1p_1
A	50	10	52	500	520
B	7	3	9	21	27
C	3	7	5	21	35
				$\Sigma p_0q_1 = 646$	$\Sigma p_1q_1 = 690$

Therefore the Paasche's price index number is:

$$P_P = \frac{690}{646} \times 100 = 106.8$$

Remark: You have noticed that the value of Laspeyre's index number (107.38) and the value of Paasche's index number (106.8) are different for the same data. The difference has arisen because of differences in weights.

3) **Fisher's Method:** The Fisher's price index number is nothing but the geometric mean of the Laspeyre's and Paasche's price index numbers. Hence the formula for Fisher's price index number is:

$$P_F = \sqrt{P_L \times P_p} = \sqrt{\left(\frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100\right) \times \left(\frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100\right)}$$

$$= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 \qquad \text{... ... ... (1)}$$

Similarly, the Fisher's quantity index number is given by:

$$Q_F = \sqrt{Q_L \times Q_P} = \sqrt{\left(\frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100\right) \times \left(\frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100\right)}$$

$$= \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100 \qquad \text{... ... ... (2)}$$

Exercise 11: Find Fisher's price index number from the data given in Exercise 9.

Solution: We have already found that for the given data,

$P_L = 107.38$ and $P_P = 106.8$

Hence the Fisher's price index number is:

$$P_F = \sqrt{107.38 \times 106.8} = 107.09$$



CHECK YOUR PROGRESS

Q.3: Write true (T) or False (F):

i) There are two types of index numbers: simple or unweighted and weighted index numbers. (T/F)

ii) There are two methods of constructing simple index numbers: simple aggregate method and simple average relatives. (T/F)

iii) We get the same index number whether we use simple aggregate method or simple average of price relatives' method. (T/F)

iv) Laspeyre's uses current year quantities as weights. (T/F)

v) Paasche's uses base year quantities as weights. (T/F)

Q.4: Fill up the blanks:

i) In Laspeyre's method of finding price index number we use ..... quantities as weights.

ii) In Paasche's method of finding price index number we use ..... quantities as weights.

iii) $I = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$ is ..... weighted index number.

iv) $I = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$ is ..... weighted index number.

Q.5: Calculate quantity index by (i) Laspeyre's method (ii) Paasche's method (iii) Fisher's method.

Commodity	2004		2006	
	Price (p_0)	Total value (p_0q_0)	Price (p_1)	Total value (p_1q_1)
A	10	100	12	144
B	12	144	14	196
C	14	196	16	256
D	16	256	18	324
E	18	324	20	400

Q.6: From the following date calculate Price index numbers for 2007 as base by (1) Laspeyre's method (2) Paasche's method and (3) Fisher's method.

Commodity	2006		2007	
	Price	Quantity	Price	Quantity
A	20	8	40	6
B	50	10	60	5
C	40	15	50	15
D	20	20	20	25

12.8 TEST OF ADEQUACY OF INDEX NUMBERS

We have seen in 13.6 section that the relatives have certain important properties. What is true for an individual commodity should also true for a group of commodities. The index numbers as an aggregative relative should also satisfy the same set of properties. We shall examine the properties that a good index number should have.

1) **Time Reversal Test or Property:** If the two periods, the base period and the reference period are interchanged, the product of the two index numbers should be unity. In other words, one should be the reciprocal of the other.

Denote the index number of the given year (1) relative to the base year (0) by $I_{\frac{0}{1}}$. Then $I_{\frac{1}{0}}$ is the index number of the year (0) relative to the year (1), that is, with the years interchanged time reversal test requires that,

$$I_{\frac{0}{1}} \times I_{\frac{1}{0}} = 1, \text{ or } I_{\frac{1}{0}} = \frac{1}{I_{\frac{0}{1}}}$$

An index number which satisfies the above property is said to satisfy time reversal test.

Remark: As we have seen in section 13.6, all the relatives satisfy the time reversal property.

Let us now examine whether Laspeyre's index number satisfy this property or not. We know that Laspeyre's index number (without

percentage) is $\frac{\sum p_1 q_0}{\sum p_0 q_0}$. Interchanging the time subscripts, we get

$\frac{\sum p_0 q_1}{\sum p_1 q_1}$ and hence their product is $\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} \neq 1$.

Consequently, Laspeyre's price index number does not satisfy the time reversal property. Similarly, it can be seen that Laspeyre's quantity index number also does not satisfy the time reversal property. Also, Paasche's price and quantity index numbers do not satisfy this property.

Now consider the Fisher's price index number. Interchanging the time subscripts we find another index number. The product of the two is given by

$$\sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}}$$

which is equal to 1. Hence Fisher's price index number satisfies the time reversal property.

2) **Factor Reversal Test or Property:** If the two factors *p* and *q* in a price index formula are interchanged, so that a quantity index number is obtained, then the product of the two index numbers should give the true value ratio, $\frac{\sum p_1 q_1}{\sum p_0 q_0}$

In other words, the price index number multiplied by the corresponding quantity index number should give the true ratio of value in the given year (1) to the value in the base year (0). This property holds for a single commodity, since

$$\frac{p_1}{p_0} \times \frac{q_1}{q_0} = \frac{p_1 q_1}{p_0 q_0}$$

but it does not hold for most of the index numbers. For example,

$$P_L \times Q_L = \frac{\sum p_1 q_1}{\sum p_0 q_0} \times \frac{\sum q_1 p_0}{\sum q_0 p_0} \neq \frac{p_1 q_1}{p_0 q_0}$$

Now consider the Fisher's index number. We have from the formulas (1) and (2),

$$P_F \times Q_F = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} \times \sqrt{\frac{\sum q_1p_0}{\sum q_0p_0} \times \frac{\sum q_1p_1}{\sum q_0p_1}}$$

$$= \frac{\sum p_1q_1}{\sum p_0q_0}$$

This proves that Fisher's index number also satisfy factor reversal test. Fisher's index number is one of the rare index numbers which satisfy both the time reversal and factor reversal tests. So it is called ''Fisher's Ideal Index Number''.

3) **Circular Test:** This test is based on the shifting of the base period. If $I_{\frac{i}{j}}$ denotes the index number for the given period j with respect to the base period i, then this test requires that, $I_{\frac{a}{b}} \times I_{\frac{b}{c}} \times I_{\frac{c}{a}} = 1$

Remark: As we have seen in section 13.6, all the relatives satisfy the circular property.



12.9 LET US SUM UP

- Index number is a statistical measure or device with a purpose of showing average change in one or more related variables over time and space.
- Price index numbers are more commonly used. It measures relative changes in prices over a time period.
- We can have either simple or weighted index number. Simple index is also called unweighted index number or index number with equal weights. In this unit, we have discussed about the simple index numbers as well as weighted index number.
- We have discussed four properties of relatives: Identity property, time reversal property, circular property and modified circular property.
- We have learnt about two methods of constructing simple index

numbers: Simple aggregate method and Method of simple average of relatives. The formula to find index number using simple aggregate method is given by $P_{\frac{0}{1}} = \frac{\sum p_1}{\sum p_0} \times 100$.

Again, the formula to find price index number using the method of simple averages of relatives is given by $P_{\frac{0}{1}} = \frac{\sum p_1}{\sum p_0} \times 100$.

- We have also leant about weighted index numbers. There are two groups of methods to calculate index number of this category: (a) Weighted aggregate method and (b) Weighted average of price relative methods. We briefly discussed about three types of weighted aggregate method. These methods are popularly known as (1) Laspeyre's Method, (2) Paasche's Method and (3) Fisher's Method.
- We have also learnt about the test of adequacy of index numbers. We mentioned three properties that a good index number should have. Fisher's index number satisfies both the time reversal and factor reversal tests and hence it is known as Fisher's Ideal Index Number.



12.10 FURTHER READING

1) Agarwal, D. R. (2006). *Business Statistics*. Delhi: Vrinda Publications.
2) Gupta S. C. (1994). *Fundamentals of Statistics*. New Delhi: Himalayan Publishing House.
3) Rajagopalan, S. P. & Sattanthan, R. (2009). *Business Statistics and Operations Research*. New Delhi: Tata McGraw-Hill
4) Sharma, J. K. (2007). *Business Statistics*. New Delhi: Pearson Education Ltd.
5) Verma, A. P. (2007). *Business Statistics*. Guwahat: Asian Books Private Limited.

12.11 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: i) T, ii) T, iii) T, iv) F, v) F, vi) T

Ans. to Q. No. 2: i) average, ii) specialized, iii) related, iv) 1990, 1998, v) equal, vi) base year price, vii) 200%

Ans. to Q. No. 3: i) T, ii) T, iii) F, iv) F, v) F

Ans. to Q. No. 4: i) base year, ii) current year, iii) Laspeyre's, iv) Paache's

Ans. to Q. No. 5: Construction of various Indicer

p_0	q_0	p_1	q_1	p_0q_0	p_0q_1	p_1q_1	p_1q_0
10	10	12	12	100	120	144	120
12	12	14	14	144	168	196	168
14	14	16	16	196	224	256	224
16	16	18	18	256	288	324	288
18	18	20	20	324	360	400	360
			Total	1020	1160	1320	1160

i) Laspeyr's Price Index is: $\frac{\sum p_1q_0}{\sum p_0q_0} \times 100$

$= \frac{1160}{1020} \times 100$

$= 113.7$

ii) Paasche's Price Index is: $\frac{\sum p_1q_1}{\sum p_0q_1} \times 100$

$= \frac{1320}{1160} \times 100$

$= 113.8$

iii) Fisher's Price Index is: $\sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} \times 100$

$= \sqrt{113.7 \times 113.8}$

$= 113.75$

Ans. to Q. No. 6: Computation of Index Numbers

Commodity	2006		2007					
	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	20	8	40	6	320	160	240	120
B	50	10	60	5	600	500	300	250
C	40	15	50	15	750	600	750	600
D	20	20	20	25	400	400	500	500
					Σp_1q_0 2070	Σp_0q_0 1660	Σp_1q_1 1790	Σp_0q_1 1470

i) Laspayer's Index: $\frac{\sum p_1q_0}{\sum p_0q_0} \times 100$

$= \frac{2070}{1660} \times 100$

$= 124.70$

ii) Paasche's Index: $\frac{\sum p_1q_1}{\sum p_0q_1} \times 100$

$= \frac{1790}{1470} \times 100$

$= 121.77$

iii) Fisher's Index: $\sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} \times 100$

$= \sqrt{\frac{2070}{1660} \times \frac{1790}{1470}} \times 100$

$= \sqrt{1.5184} \times 100$

$= 1.2322 \times 100$

$= 123.22$



12.12 MODEL QUESTIONS

Q.1: Explain the meaning of an index number and its main characteristics.

Q.2: "Index numbers are economic barometers." Why are index numbers so called? Explain.

Q.3: What is price relative? Show with the help of an example how it is calculated?

Q.4: Define quantity and value relatives. What do they measure?

Q.5: State the important properties of relatives.

Q.6: Explain the method of construction of unweighted index numbers by Simple Aggregate Method.

Q.7: Explain the method of construction of unweighted index numbers by Simple Aggregate of Price Relative Method.

Q.8: The exchange rate of a US dollar was Rs. 44.00 in July 2008 and was Rs. 50.00 in February 2009. Find the price relative.

Q.9: The production of rice in Assam in the month of January 2008 was 3608 tonnes and that in February 2009 was 3700 tonnes. Find the quantity relative.

Q.10: The following table shows the prices of the commodities consumed in 1970 and 1978.

		1970	1978
Commodities	Unit	Price (Rs.)	Price (Rs.)
A	Kilogram	4.00	4.60
B	Litre	8.00	8.75
C	Metre	7.00	8.00
D	Kilogram	14.00	16.00
E	Litre	3.00	4.00

Find the price index number using:

a) simple aggregate method

b) method of simple aggregate of price relatives.

Q.11: The production of rice, milk, cloth, wheat and petrol are given in the following table. Taking 1980 as base year, compute quantity index number using:

a) simple aggregate method

b) method of simple aggregate of quantity relatives.

		Quantity Produced	
Commodities	Unit	1980	1990
Rice	Kilogram	400000	700000
Milk	Litre	800000	905000
Cloth	Metre	100000	105000
Wheat	Kilogram	140000	160000
Petrol	Litre	300500	400000

Q.12: What do you mean by tests of index numbers?

Q.13: Explain the important criterion or test that a good index number should satisfy.

Q.14: The prices and quantity of five commodities are given in the following table. Taking 2000 as base year, compute value index number using:

a) simple aggregate method

b) method of simple aggregate of value relatives.

Commodities	2000		2002	
	Price (Rs.)	Quantity (Kg.)	Price (Rs.)	Quantity (Kg.)
A	400	10	700	12
B	800	15	1905	17
C	100	5	125	5
D	140	7	160	7
E	300	6	400	6

Q.15: Find Laspeyre's and Paasche's price and quantity index number from the following data. Also find the Fisher's price and quantity index numbers.

Commodities	Base year		Current year	
	Price (Rs.) p_0	Quantity (Kg.) q_0	Price (Rs.) p_1	Quantity (Kg.) q_1
A	5	25	6	30
B	10	5	15	4
C	3	40	2	50
D	6	30	8	35

Q.16: The following table shows the price quantity of the commodities consumed in 2000 and 2006.

Commodities	2000		2006	
	Price (Rs.)	Quantity (Kg.)	Price (Rs.)	Quantity (Kg.)
I	400	35	460	35
II	800	10	1000	10
III	300	15	50	12
IV	4000	10	5000	8
V	50	20	300	20

Find the following index numbers.

1) Laspeyre's Price and Quantity index numbers.
2) Paasche's Price and Quantity index numbers.
3) Fisher's Price and Quantity index numbers.

*** ***** ***

UNIT 13 : TIME SERIES

UNIT STRUCTURE

13.1 Learning Objectives

13.2 Introduction

13.3 Definition of Time Series

13.4 Importance of Time Series Analysis

13.5 Components of a Time Series

13.6 Methods of Measuring Secular Trend

13.7 Estimation of the Trend by the Method of Moving Average

13.8 Let Us Sum Up

13.9 Further Reading

13.10 Answers To Check Your Progress

13.11 Model Questions

13.1 LEARNING OBJECTIVES

After going through this unit, you will able to:

- understand the concept of time series
- explain the importance of time series analysis
- known about the various factors which affect a time series
- know about the methods for determining secular trend in time series.

13.2 INTRODUCTION

In Economics, Business and Commerce, it is important to make estimate for the future e.g. an economist is interested to know the figures of national income, prices and wages, population etc. for his future planning or a businesman likes to estimate his likely sales in the coming years to adjust his production accordingly. For making such estimates, one has to collect information from the past i.e. one has to deal with statistical data collected and recorded at successive intervals of time (or points of time). Such statistical data relating to time all refered to as Time Series. The analysis of time series has an important role in Economics, Business and

Commerce. Basically, most of the statistical technigues for the analysis of time series data have been developed by economists. However, these techniques can also be applied for study of behaviour of any phenomenon collected chronologically over a period of time in any discipline relating to natural and social sciences, though not directly related to business and economics.

13.3 DEFINITION OF TIME SERIES

A Time Series is a set of observations taken at specified times, usually (but not always) at equal intervals. Thus a set of data depending on time (which may be year, half-year, quarter, month, week, days etc.) is called a time series.

Some examples of time series are:

i) The annual production of steed in India over the last ten year.

ii) The sale of a departmental store in different months of the year.

iii) The daily closing price of a share in the Bombay Stock exchange for a month.

iv) Hourly Temperature recorded by the meteorological office in a city.

v) Quarterly Consumer Price Index numbers of a country for last three years.

vi) Daily rainfall of a particular city in the month of May, June, July and August.

Mathematically, a time series is defined by the functional relationship:

$$Y = f(t)$$

Where y is the value of the phenomenon (or variable) under consideration at time t. Thus, if the values of a phenomenan (variable) at times $t_1, t_2, ..., ..., ..., t_n$ are $y_1, y_2, ..., ..., ..., y_n$ respectively, then the series

t:	t_1	t_2	t_3	...,	...,	...,	t_n
y:	y_1	y_2	y_3	...,	...,	...,	y_n

Constitutes a time series.

13.4 IMPORTANCE OF TIME SERIES ANALYSIS

Analysis of time series is of special importance to Businessman, Economist, Scientist, Sociologist, Geologists and Research workers in various discipline. The following points indicate the importance of time series analysis:

1) **Past Behaviour:** It helps us to understand the past behaviour of a variable. Analysis of past data discloses effect of various factors on the variable under study.
2) **Forecasting:** With the knowledge of past behaviour, it would be possible, within certain limits, to forecast the future behaviour of the variable under study. This helps in making future plans of action. Various five year plans of our country are based on the analysis of past data.
3) **Evaluation:** It halps in evaluation of current achievements. The review and evaluation of progress made on the basis of a plan are done on the basis of time series data. For example, the progress of various plans in our country is judged by the yearly rate of growth in GNP (Gross National Product).
4) **Comparison:** Analysis of time series helps in making comparison between one time period to another.It also facilitate to compare the actual values with the expected values. It provides a scientific basis for making comparisons by studying and isolating the causes of variation of the variable under study.

13.5 COMPONENTS OF A TIME SERIES

Analysis of time series shows that if the values of a phenomenon are observed at different periods of time, the values so obtained are always fluctuating. The reason behing this type fluctuation is the joint effect of various factors (or forces). For example, the sales (y) of a product depends on: (i) advertisement expenditure (ii) the price of the product (iii) the income of the consumer (iv) other competitive product in the market (v) tastes, fashions, habits and customs of the people and so on similarly, the price of a particular

product depends on its (i) demands (ii) various competitive products in the market (iii) availability of raw matirials (iv) transportation cost etc.

To know the proper nature of a time series, the statisticians have broadly classified the various operating factors (forces) affecting the values of a shenomenon in a time series into four parts. These are commonly known as **Components of a time series.** The four components of a time series are:

a) Secular Trend or Trend (T)
b) Seasonal Variation (S)
c) Cyclical Variation (C)
d) Irregular or Random Variation (I)

N.B.: Instead of the word 'variation', the word 'fluctuation' or 'movement' may also be used.

The changes in time series data are the result of the combined effect of these four components. Now, we proceed to a brief discussion on the various components of time series.

a) Secular Trend or Trend (T): The general tendency of the time series data showing continious growth (increase), stagnation or decline during a long period of time is called the **Secular Trend** or simply **Trend.** This phenomenon is usually observed in most of the time series relating to Business and Economics. e.g. an upward tendency is usually observed in time series relating to population, prices, income, production and sales of products, money in circulation etc, while a downward tendency is noticed in time series relating to birth rate, death rate, death rate due to epidemics etc due to the advancement in medical science, improved in medical facilities, better sanitation diet etc. Simpson and Kafka defined secular trend as: **"Trend, also called Secular Trend or Long term trend, is the basic tendency of a time series to grow or decline over a period of time. The concept of trend does not include short-range oscillations, but rather the steady movement over a long time."**

Remarks: It should be understood that the term 'long period' is a relative term and can not be defined exactly. In certain phenomenon,

a period of few hours may be long enough whereas in some others even a period as long as 3-4 years may not be regarded as a long period. For example, if some agricultural or industrial production of a partcular product shows an increase over the past 3 years, this can not be termed as a long term tendency, for which we must have data for 7-8 years. On the other hand, while a patient is having high fever and as per advice of the attending physician, temperature reading are needed to be recorded on every 4 hours, in order to observe whether the effect of a particular medicine shows a decline in temperature or not, then in such a case, taking reading for a period of 48 hrs to 72 hrs would be termed as a long period.

Uses of Trend: The study of the data over a long period of time enables us to have a general idea about the pattern of the behaviour of the phenomenon under consideration. This helps in business forecasting and planning of the future operations. Trend analysis enables us to compase two or more time series over different periods of time and draw important conclusions about them.

b) **Seasonal Variations (S):** The fluctuations or variations in the values of a time series over a period of one year or less are termed as seasonal variation. This variation occurs due to forces which are rhythmic in nature and which repeat themselves periodically over a span of less than a year. Such type of variations are present in a time series if the data are recorded quarterly (after every three months), monthly, weekly or daily basis. **Thus in a time series data where only annual figures are given, there can not be seasonal variation.** Most of the economic time series are influenced by swings e.s. prices, production and consumption of commodities, sales and profits of departmental stores, bank clearings and bank deposits etc. are all affected by seasonal variation. Seasonal variation takes place due to the following two causes–

   i) **Narural Causes:** As the name suggests, the various seasons or weather conditions and climatic changes play an important role in seasonal variation. For example, the sale of umbrella

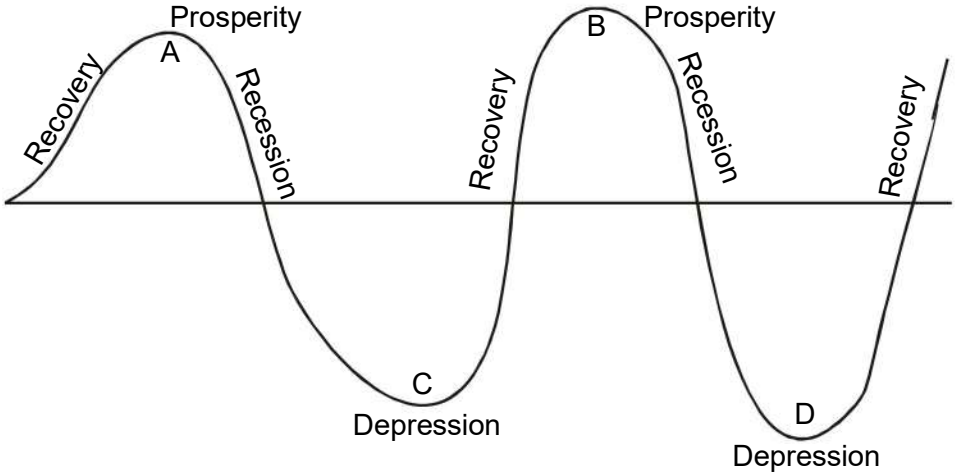
picks up very fast in raing season, the sale of woollen garment is more in winter, during summer demand for cold drinks and ice-cream increases. The prices of agricultural commodities always go down at the time of harvest and then pick up gradually.

ii) **Rituals and Social Customs:** Man made rituals, social customs and teaditions are also responsible for seasonal fluctuation of a time series. For example, sale of jewellary goes up during diwali and marriage season, in the beginning of an academic session sale of books, papers, uniforms etc. go up, on the eve if new year, sale of greetings card increases.

Uses of Seasonal Variation: The main objective of seasonal variations is to isolate them from trend and study their effects. A study of the seasonal pattern is extremely useful for a businesman, producers sales managers etc. in planing future operations and in formulation of policy decisions regarding purchase, productions, inventory control and advertising programmes. In the absence of any knowledge of seasonal variation, a seasonal upswing may be mistaken as indicator of better business conditions while a seasonal slump may be misinterpreted as deteriorating business conditions.

c) **Cyclical Variation (C):** Cyclical variations are the oscillatory mavement in a time series with period of oscillation greater than one year. The cyclical vanations though more or less regular, are not necessairly uniformly periodic i.e. they may or may not follow exactly similar patterns after equal intervals in time one cyclic period normally lasts from 7 to 9 years.

Cyclical variations are found to exist in almost all busimess and economic time series where it is known as **business cycle** or **trade cycle**. The ups and downs (or rises and decline) in business recurring at intervals of time are the effects of cyclical variation. A **business cycle consists of four phasses** namaly **prosperity** (boom), **recession, depression** and **recovery.** Each phase changes gradually into the next phase in the given order until one business cycle is completed:



(Business Cycle)

The time period between two successive booms or depressions (e.g. time period between A and B or C and D) is known as the length of a cycle.

Uses of Cyclical Variations: The study of cyclical variations is of great importance to business executives in the formulation of policies animed at stabilising the level of business activity. A knowledge of cyclical variations enables a businessman to have an idea about the periodicity of booms and depressions and accordingly timely steps may be taken for maintaing stable market for his product.

Note: Though seasonal variations and syclical variations are both periodic in nature, there is a significant difference between the two types of variations. These are mentioned below:

	Seasonal Variations		**Cyclical Variations**
i)	Seasonal variation takes place within one yeas.	i)	Cyclical variation takes place in a period more than one year. It voually takes place in 3 to 10 years time period.
ii)	In seasonal variation the periodic time remains the same.	ii)	In cyclical variation, the periodic time does not remain the same. It takes place only with some rough regularity.

iii)	It is mainly attributed to climatic changes and man made rituals and social customs.	iii)	Economic factors like price, production, sales, demand etc. are responsible for cyclical variations.

d) Irregular Variations (I): Irregular or Random Variations are such variations inherent in a time series which are caused by factors irregular (or erratic) in nature. These are purely random and unpredictable. These include all movements not already covered in trend, seasonal variation and cyclical variation. Irregular variations are caused by unforeseen factors like flood, famines, wars, earth. quakes, fire, epidemics, strike of trade union, terrorision etc. As there are unpredictable, they can not be controlled by human hand. For almost all the time series, irregular variations are inevitable. Normally, these are short term fluctuation but sometimes their effect is so intense that they may give rise to new cyclical or other movements.

Because of their absolutely random character, it is not possible to study them exclusively, nor we can forecast or estimate these precisely. The best that can be done about such variations is to obtain their rough estimates (from the past experience) and accordingly make provisions for such abnormalities during normal time in business.

13.6 METHODS OF MEASURING SECULAR TREND

There are four methods which are generally used for the study and measurement of secular trend in a time series. There are:

i) Graphical Method
ii) Semi-Average Method
iii) Moving Average Method
iv) Method of Leaot Squars.

Note: Here we will restrict our discussion to the **Method of moving average** only.

13.7 ESTIMATION OF TREND BY THE METHOD OF MOVING AVERAGE

The method of moving average **is a very simple and flexible method of measuring trend. This method consists in taking averages (arithmetic mean) of the values for a certain time span and placing it at the centre of the time span.**

The moving average is characterioed by a constant known as **period** of the moving average. Thus the moving average of period 'm' is a series of successive averages (Arithmetic means) of 'm' overlapping values at a time starting with 1st, 2nd, 3rd value and so on. Thus, for the time series values y_1, y_2, y_3, y_4, y_5, ..., ..., ..., the moving average (M.A.) values of period 'm' are given by:

$$\text{1st M.A.} = \frac{1}{m}(y_1 + y_2 + ... + y_m)$$

$$\text{2nd M.A.} = \frac{1}{m}(y_2 + y_3 + ... + y_{m+1})$$

$$\text{3rd M.A.} = \frac{1}{m}(y_3 + y_4 + ... + y_{m+2})$$

and so on we shall discuss two cases:

Case (i) **When period is odd:** It the period 'm' of the moving average is odd, then the successive values of the moving averages are placed against the middle most observation of the corresponding time period. For example, if m = 5, the first moving average value is placed against the 3rd value of the series, the second moving average value is placed against the 4th value of the time series and so on.

Case (ii) **When period is even:** It the period 'm' of the moving average is even, then there are two middle periods and the moving average values are placed in between the two middle periods of the time interval it covers. Obviously, in this case the moving average values will not coincide with a period of the given time series and an attempt is made to synchronise them with the original data by taking a two-

period average of the moving averages and placing them in between the corresponding time periods. This technique is called **centering** and the corresponding moving average values are called **centred moving average.** For example, if m = 4, the first moving average value is placed against the middle of 2nd and 3rd time intervals, the second moving average value is placed in between 3rd and 4th time period and so on. These values are given by–

$$\overline{y_1} = \frac{1}{4}(y_1 + y_2 + y_3 + y_4)$$

$$\overline{y_2} = \frac{1}{4}(y_2 + y_3 + y_4 + y_5)$$

$$\overline{y_3} = \frac{1}{4}(y_3 + y_4 + y_5 + y_6) \text{ and}$$

The centered moving averages are obtained on taking 2 period moving average of $\overline{y_1}$, $\overline{y_2}$, $\overline{y_3}$ and so on.

Thus, First centered moving average = $\frac{1}{2}\left(\overline{y_1} + \overline{y_2}\right)$

Second centered moving average = $\frac{1}{2}\left(\overline{y_2} + \overline{y_3}\right)$ and so on.

These centered moving averages are placed against 3rd, 4th, 5th period and so on.

MERITS AND DEMERITS OF MOVING AVERAGE METHOD:

Merits:

i) This method does not involve any mathematical complexities and is quite simple to understand and use.

ii) There is some inherent flexibility in this method. Addition of few more observations to the existing data does nt affect the trend values already obtained. This will simple result in some more values at the end.

iii) If the period of moving averages coincides with the period of cyclic fluctuations in the data, then this method completely eliminalis the effect of cyclical fluctuations.

iv) This method is also used for measurement of seasonal, cyclical and irregular variations.

Demerits:

i) One of the prime limitations of moving average method is that we can not obtain the trend values for all the given observations. Some trend values in the beginning and at the end have to forego depending on the period of moving average.

ii) This method is not suitable for the purpose of forecasting.

iii) The selection of period of moving average is a difficult task. Therefore, great care has to be taken in selecting the period, particularly, when there is no business cycle is the time series.

iv) If the time series reveals lineartrend only then this method is applicable.

Example 1: Calculate 3 yearly moving average from the data given below:

Year:	1980	1981	1982	1983	1984
Production (’oootons):	112	117	120	123	129
Year :	1985	1986	1987	1988	1989
Production (’oootons):	134	140	146	149	152

Solution: Table for calculation of 3 yearly moving averages.

Year	Prodiction (’ooo tons)	3 Yearly Moving Total	3 Yearly Moving Average
(1)	(2)	(3)	(4)
1980	112		
1981	117	349	116.3
1982	120	360	120
1983	123	372	124
1984	129	386	128.7
1985	134	403	134.3
1986	140	420	140
1987	146	435	14.5
1988	149	447	149
1989	152		

Explanation: In column (3) total of the first three entries namely 112, 117 and 120 is written against the middle value i.e. 117 of this set. Then total of next. three entries i.e. 117, 120 and 123 is written

against 120 and so on. Figures of column (4) are obtained by dividing the columns of figures (3) by 3.

Note: These 3 yearly moving average values are the trend values for the given time series.

Example 2: Calculate trend values by the method of 4 yearly moving average from the following data.

Year:	1991	1992	1993	1994	1995	1996
Value:	44	64	59	52	57	71
Year:	1997	1998	1999	2000	2001	2002
Value:	66	64	71	77	82	80

Solution: Table for calculation of 3 yearly mouing averages

Year	Value moving	4 yearly moving Total	4 yearly moving average	2-period average column	Centred (trend value)
(1)	(2)	(3)	(4)	(5)	(6) = (5) ÷ 2
1991	44				
1992	64 219	54.75			
1993	59 232	58		112.75	56.375
1994	52 239	59.75		117.75	58.875
1995	57 246	61.5		121.25	60.625
1996	71 272	68		126	63
1997	66 272	69.5		132.5	68.75
1998	64 278	69.5		137.5	68.75
1999	71 294	73.5		143	71.5
2000	77 310	77.5		151	75.5
2001	82				
2002	80				

Explanation: In column (3) total of the first four entries namely 44, 64, 59 and 52 is written against the middle of this set i.e. in between 64 and 59. Then the total of next four entries i.e. 64, 59, 52 and 57 is written in between 59 and 52 and so on.

Figures of column (4) are obtained by dividing the figures of column (3) by 4.

In column (5) the total of the first two entries of column (4) i.e. 54.76 and 58 is put in between them i.e. corresponding to the year 1993. Then the total of second and third entries of column (4) i.e. 58 and 59.75 is put corresponding to the year 1994 and so on.

The figures of the column (6) are obtained by dividing the entries of column (5) by 2.

Note: The 4 yearly centred moving averages are the trend values for the given time series.

Example 3: From the data given below, determine trend values by 5 yearly moving average method.

Solution:

Year:	1970	1971	1972	1973	1974	1975	1976	1977	1978
Sales (Rs. lakhs)	130	127	124	135	140	132	129	127	145

Table for calculation trend values by 5 yearly moving average method

Year	Sales (Rs. Lakhs)	5 Yearly Moving Total	5 Yearly Moving Average (Trend value)
(1)	(2)	(3)	(4)
1970	130		
1971	127		
1972	124	656	131.2
1973	135	658	131.6
1974	140	660	132.0
1975	132	663	132.6
1976	129	673	
1977	127		
1978	145		

Note: The first moving total 656 of column (3) is the sum of first five values 14, 130, 127, 124, 135 and 140. The second moving total is 127 + 124 + 135 + 140 + 132 = 658 which can also be obtained by adding (132 – 130) = 2 with first moving total 656. Similarly, 3rd moving Total is 658 + (129 – 127) = 660 and so on.



CHECK YOUR PROGRESS

Q.1: Fill in the blanks:

i) A time series consists of data arranged .....

ii) An overall rise or fall in the time series is called .....

iii) ..... can be made with the help of time series.

iv) Cyclical variations are caused by .....

v) ..... is the overall tendency of the time series data ..... or ..... over a ..... period of time.

vi) Short term variations are classified as (a) ..... (b) .....

vii) The four phases of business cycle (in order) are ....., ..., ...

viii) For the annual data ..... component is absent.

ix) The most important factors causing seasonal variations are .....

Q.2: With which component of time series would you associate each of the following.

i) An era of prosperity.

ii) An after puja sale in a departmental store.

iii) A strike in a factory delaying production by 4 weeks.

iv) The sale of umberlla in rainy season.

v) Secrease in death rate due to advancement of medical science.

vi) A need for increased rice production due to constant increase in population.

Q.3: Select the correct alternatives out of the given one–

a) A time series is a set of data recorded:

i) at time intervals ii) at successive point of time

iii) periodically iv) All the above

b) The decline in birth rate is attached to:

i) Secular Trend ii) Irregular Variation

iii) Seasonal Variation iv) Cyclical Variation

c) The sale of cold drink in summer season is attached to:

i) Secular Trend ii) Seasonal Variation

iii) Cyclical Variation iv) Irregular Variation

d) Seasonal variation repeats during a period of:

i) 5 years ii) 4 years

iii) 7 years iv) None of the above

e) Salient factors responsible for seasonal variation are:

i) Weather ii) Festival

iii) Social Custom iv) All the alove

f) Increase in death rate due to an epidemic is attached to:

i) Secular Trend ii) Cyclical Variation

iii) Seasonal Variation iv) None of the alove



13.8 LET US SUM UP

In this unit we have learnt about the following:

- Time series is a set of observation recorded at successive periods of time.
- The factors which affect the components of time series.
- The component of time series responsible for long term movement in a time series is called secular trend or simply trend.
- The component of time series that shows a regular and periodic pattern of movement over a time span of less than 1 year is called seasonal component.
- Cyclical variation is an oscillatory movement in a time series with period

of oscillation greater than 1 year.

- Irregular movement are such variations inherent in a time series.
- Analysis of time series helps us to understand past behaviour, forecasting, evaluation of current achievement and making comparison between two time period.
- There are four methods for studying and measuring secular trend.
- One of the methods of measuring trend is 'Moving average method' which involve smoothing a time series by averaging succesive groups of data for a certain time span.



13.9 FURTHER READING

1) Agarwal D. R. *Business Statistics*. Delhi: Vrinda Publication.

2) Arora, P. N.; Arora, Sumeet; Arora, S. *Comprehensive Statistical Methods*. S. Chand & Company.

3) Ghosh, R. K. and Saha, S. *Business Mathematics and Statistics.* New Central Book Agency (P) Ltd.

4) Gupta, S. C. *Fundamentals of Statistics*. New Delhi: Himalaya Publishing House.

5) Gupta S. C and Gupta Indira. *Business Statistics.* Himalaya Publishing House.

6) Sharma J. K. *Business Statistics.* New Delhi: Pearsor Education Ltd.



13.10 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: i) Chronologically, ii) Trend / Secular Trend, iii) Forecasts, iv) Business / Trade Cycles, v) Secular Trend, increase, decrease, long, vi) (a) Seasonal (b) Cyclical (vii) Boom, Recession, Depression, Recovery, viii) Seasonal, ix) Weather and Social Customs.

Ans. to Q. No. 2: i) Cyclical, ii) Seasonal, iii) Irregular, iv) Seasonal, v) Secular, vi) Secular.

Ans. to Q. No. 3: a) (iv), b) (i), c) (ii), d) (iv), e) (iv), f) (iv)



13.11 MODEL QUESTIONS

Q.1: Define time series.

Q.2: Explain various components of time series.

Q.3: What is meant by Time Series Analysis? Discuss its importance in business.

Q.4: What do yoy meand by Secular Trend? Explain by giving examples.

Q.5: Explain the meaning of seasonal variations, with illustrations.

Q.6: What are cyclical variation? How are they caused?

Q.7: How do cyclical variations differ from seasonal variation.

Q.8: Which component of time series is associated with the following:

i) Recession

ii) A fire in a factory delaying production for three weeks.

iii) An era of depression.

iv) Sale of ice-cream in summer.

v) Increase in the production of rice in Assam due to application of modern technique.

Q.9: Obtain Trend Value by 5 yearly moving average method.

Year:	1990	1991	1992	1993	1994	1995	1996	1997	1998
Value:	242	250	252	249	253	255	251	257	260
Year:	1999	2000							
Value:	265	262							

Q.10: Calucate Trend Values by 3 yearly moving average method.

Year:	1981	1982	1983	1984	1985	1986	1987	1988
Sugar Production (lakh tons):	37.4	31.2	38.7	39.6	47.6	42.6	48.5	56.2
Year:	1989	1990						
Sugar Production (lakh tons):	60.6	58.2						

Q.11: Calculate trend values by 4 yearly moving average method.

Year:	1985	1986	1987	1988	1989	1990	1991	1992	1993
Value:	39	59	53	46	51	65	59	58	65
Year:	1994	1995	1996						
Value:	71	76	74						

Q.12: Determine the period of the moving average for the data given below and calculate moving average value for that period.

Year:	1	2	3	4	5	6	7	8	9
Value:	130	127	124	135	140	132	129	127	145
Year:	10	11	12	13	14	15			
Value:	158	153	146	145	164	170.			

[**Hint :** Since the peaks of the given data occur at the years 1, 5, 10 and 15, the data clearly exhibits a regular cyclical movement with period 5, hence the period of moving average for determining trend valus is also 5 i.e. period of cyclical variations]

*** ***** ***

UNIT 14: MEASUREMENT OF ECONOMIC INEQUALITY

UNIT STRUCTURE

14.1 Learning Objectives

14.2 Introduction

14.3 Pareto's Law of Income Distribution

14.4 Log normal Distribution (Concept only)

14.5 Lorenz Curve

14.6 Coefficient of Inequality : Gini Coefficient

14.7 Let Us Sum Up

14.8 Further Reading

14.9 Answers to Check Your Progress

14.10 Model Questions

14.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- discuss Pareto's law of income distribution
- describe the concept of log normal Distribution
- write the concept of Lorenz curve and Gini coefficient.

14.2 INTRODUCTION

The measurement of economic inequality has remained a major subject of discussion in Economics. Different concepts and theories have been put forward to deliberate on this issue. A few important concepts in this area has been taken up in this unit. Among them are the Pareto's law of distribution, Lorenze curve and the Gini Coefficient. Apart from these concepts, the concept of log normal distribution has also been discussed.

14.3 PARETO'S LAW OF INCOME DISTRIBUTION

The concept of Pareto Distribution was put forward by Vilfredo Pareto. This is also referred to as the Pareto Principle or the 80-20 Rule. The

Pareto distribution is used in describing social, scientific, and geophysical phenomena in a society. Pareto created a mathematical formula in the early 20th century that described the inequalities in wealth distribution that existed in his native country of Italy.

Pareto observed that 80 percent of the country's wealth was concentrated in the hands of only 20 percent of the population. The theory is now applied in many disciplines such as incomes, productivity, populations, and other variables.

Empirical validity of the Theory: In 1906, Vilfredo Pareto introduced the concept of the Pareto Distribution when he observed that 20 percent of the pea pods were responsible for 80 percent of the peas planted in his garden. He related this phenomenon to the nature of wealth distribution in Italy, and he found that 80 percent of the country's wealth was owned by about 20 percent of its population. In terms of land ownership as well, he further observed that 80 percent of the land was owned by a handful of wealthy citizens, who comprised about 20 percent of the population.

The definition of the Pareto Distribution was later expanded in the 1940s by Dr. Joseph M. Juran, a prominent product quality guru. Juran applied the Pareto principle to quality control for business production to show that 20 percent of the production process defects are responsible for 80 percent of the problems in most products.

Pareto Distribution Formula: The formula for calculating the Pareto Distribution is as follows:

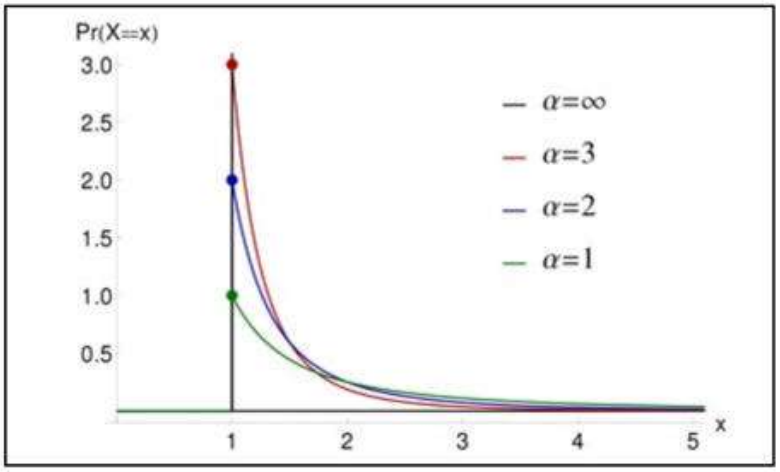
$$F(x) = 1 – (k/x)\alpha$$

Where, x – random variable, k – lower bound on data, α – shape parameter.

Graphical Representation:

On a chart, the Pareto distribution is represented by a slowly declining tail, as shown in Figure 14.1:

Figure 14.1: Pareto Distribution Chart



Source: Wikipedia Commons

The chart is defined by the variables α and x. It provides two main applications. One of the applications is to model the distribution of wealth among individuals in a country. The chart shows the extent to which a large portion of wealth in any country is owned by a small percentage of the people living in that country.

The second application is to model the distribution of city populations, where a large percentage of the population is concentrated in the urban centers and a lower amount in the rural areas. The population in urban centers continues to increase while the rural population continues to decline as younger members of the population migrate to urban centers.

Some Practical Applications of the Pareto Distribution

- **Business Management:** One of the applications of the Pareto concept is in business management. A business may observe that 20 percent of the effort dedicated to a specific business activity generates 80 percent of the business results. A business can use this ratio to identify the most important segments that it can focus on and thereby increase its efficiency.

  For example, if marketing contributed to increased business results, the business can allocate more time and resources into marketing activities to increase the company's revenues and profits.

- **Company Revenues:** The 80-20 Pareto rule may also apply in evaluating the source of the company revenues. For example, when

the company observes that 80 percent of reported annual revenues come from 20 percent of its current customers, it can focus its attention on increasing the customer satisfaction of influential customers.

From this observation, the company can also deduce that 80 percent of customer complaints come from 20 percent of customers who form the bulk of its transactions. Also, focusing on solving the complaints of 20 percent of its customers can increase the overall customer satisfaction of the company. The company should focus on retaining 20 percent of its influential customers and on acquiring new customers.

- **Employee Evaluation:** A company can also use the 80-20 rule to evaluate the performance of its employees. The company may observe that 80 percent of its overall output is the direct result of about 20 percent of its employees. Using the ratio, the company can focus on rewarding the 20 percent most productive employees as a way of motivating them and encouraging the lower cluster of employees to work harder. The productivity ratio could also show the company that 80 percent of human resource problems are caused by 20 percent of the company's employees.

Limitations of the Pareto Distribution: Some of the limitations of this principle are:

- While the 80-20 Pareto distribution rule applies to many disciplines, it does not necessarily mean that the input and output must be equal to 100 percent. For example, 20 percent of the company's customers could contribute 70 percent of the company's revenues. The ratio brings a total of 90 percent.
- Pareto concept is merely an observation that suggests that the company should focus on certain inputs more than others.

14.4 LOG NORMAL DISTRIBUTION (CONCEPT ONLY)

A log-normal distribution is a statistical distribution of logarithmic values from a related normal distribution. A log-normal distribution can be translated to a normal distribution and vice versa using associated logarithmic calculations.

A normal distributions is a probability distribution of outcomes that is symmetrical or forms a bell curve. In a normal distribution 68% of the results fall within one standard deviation and 95% fall within two standard deviations.

While most people are familiar with a normal distribution, they may not be as familiar with log-normal distribution. A normal distribution can be converted to a log-normal distribution using logarithmic mathematics. That is primarily the basis as log-normal distributions can only come from a normally distributed set of random variables.

There can be a few reasons for using log-normal distributions in conjunction with normal distributions. In general most log-normal distributions are the result of taking the natural log where the base is equal to e=2.718. However, the log-normal distribution can be scaled using a different base which affects the shape of the lognormal distribution.

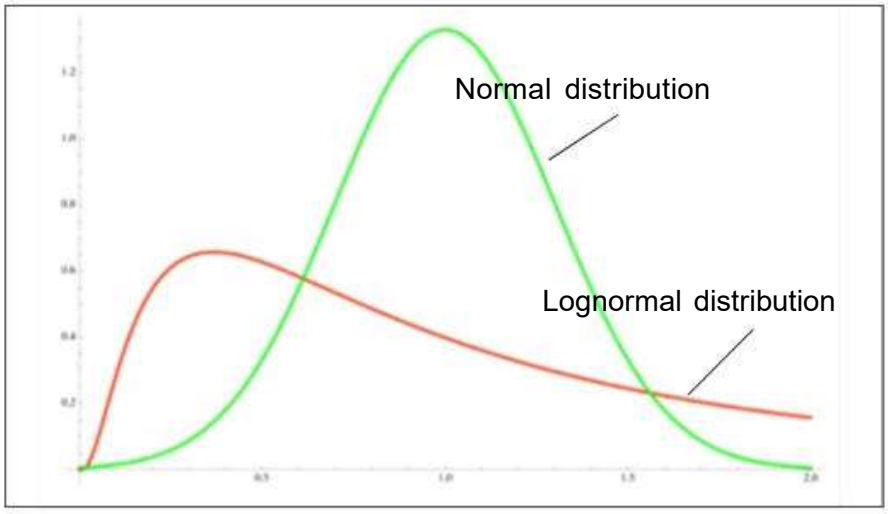
Overall the log-normal distribution plots the log of random variables from a normal distribution curve. In general, the log is known as the exponent to which a base number must be raised in order to produce the random variable (x) that is found along a normally distributed curve.

Applications and Uses of Log-Normal Distribution: Normal distributions may present a few problems that log-normal distributions can solve. Mainly, normal distributions can allow for negative random variables while log-normal distributions include all positive variables.

One of the most common applications where log-normal distributions are used in finance is in the analysis of stock prices. The potential returns of a stock can be graphed in a normal distribution. The prices of the stock however can be graphed in a log-normal distribution. The log-normal distribution curve can therefore be used to help better identify the compound return that the stock can expect to achieve over a period of time.

Note that log-normal distributions are positively skewed with long right tails due to low mean values and high variances in the random variables. This may be viewed from Figure 14.2.

Figure 14.2: Normal vs Lognormal Distribution: Graphical Shapes



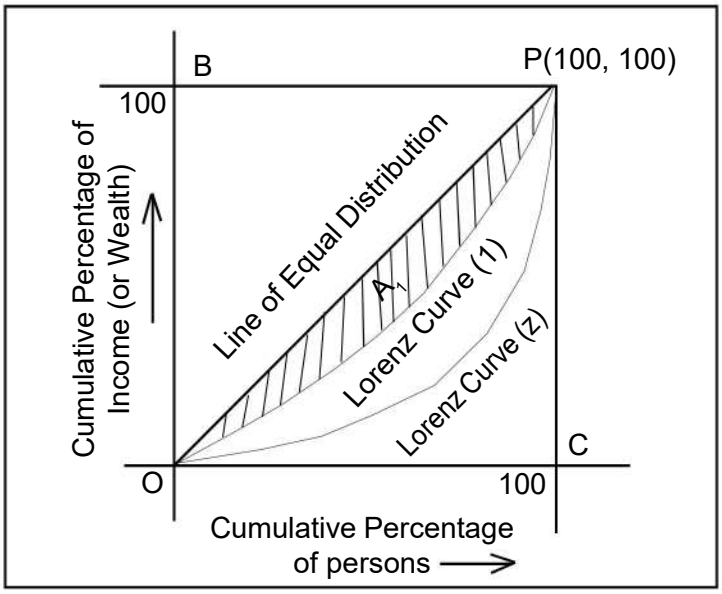
14.5 LORENZ CURVE

The graphical method of measurement of dispersion was developed by a famous economic statisticion Dr. Max O. Lorenz. He used this method to measure the inequalities of income or wealth of a society. In this method, cumulative percentage of income or wealth, i.e., the percentage of income (or wealth) less than a given value is plotted against the cumulative percentage of persons. The curve thus obtained, is known as Lorenz Curve. Another name of Lorenz Curve is 'Cumulative percentage curve'.

In figure, the cumulative percentage of persons is shown on X-axis and cumulative percentage of other characteristics (income or wealth) is shown on Y-axis)

In Figure 3.3 a point P with coordinates (100, 100) is plotted on the figure and perpendiculars PB and PC are dropped on the Y-axis and X-axis respectively from this point. The line joining the point P and O represents the line of equal distribution. Any departure from this line denotes the extent of inequality. For example, the extent of inequalities, in the above figure, are represented by the area A_1 between the Lorenz Curve and the line of equal distribution OP.

Figure 14.3: Lorenz Curve



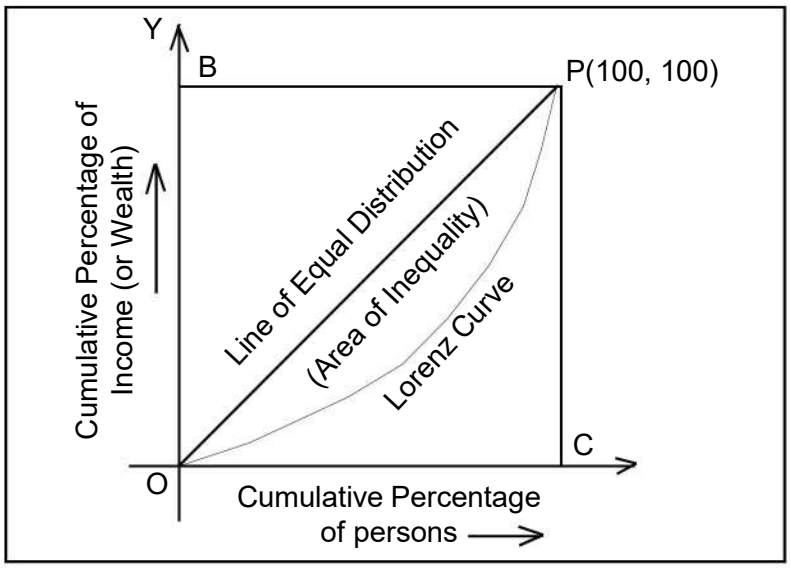
14.6 COEFFICIENT OF INEQUALITY : GINI COEFFICIENT

Gini coefficient is based on the Lorenz Curve and is defined as the ratio of the area between the diagonal and the Lorenz curve to the total area of the half-square in which the curve lies.

Let the area of the triangle POC be denoted by A. The coefficient of inequality i.e. the Gini coefficient is defined by the ratio $\frac{A_1}{A}$, where A_1 is the area between the Lorenz curve (1) and the line of equal distribution OP. Higher the degree of inequality higher is the ratio and accordingly hogher is the value of the Gini coefficient.

The value of the Gini coefficient can theoretically be between the two extreme values of O (indicating perfect equality) to 1 (indicating perfect inequality). This has been shown in Figure 14.4.

Figure 14.4: Lorenz Curve and Gini Coefficient



Example : Use Lorenz Curve to compare the extent of inequalities of income distribution in the two groups.

Monthly Income (Rs.)	No. of Persons in	
	Group A	Group B
1200–1400	800	4800
1400–1600	960	6400
1600–1800	1040	9600
1800–2000	600	3600
2000–2200	480	8000
2200–2400	120	4000

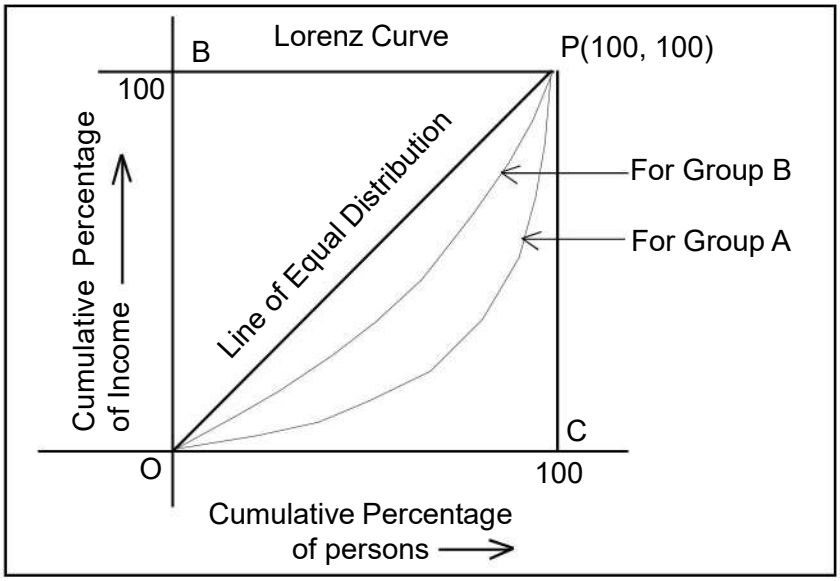
Solution : Let us prepare the following table :

Monthly Income (Rs.)	Mid Values (Rs.)	Cum. Income	Cum. %	Group A			Group B		
				No. of persons	Cum. Total	Cum. %	No of persons	Cum. Total	Cum. %
1200–1400	1300	1300	12	800	800	20	4800	4800	13
1400–1600	1500	2800	26	960	1760	44	6400	11200	31
1600–1800	1700	4500	42	1040	2800	70	9600	20800	57
1800–2000	1900	6400	59	600	3400	85	3600	24400	67
2000–2200	2100	8500	79	480	3880	97	8000	32400	89
2200–2400	2300	10800	100	120	4000	100	4000	36400	100

Note : The percentages are approximated to the nearest whole number.

This may be represented with the help of Figure 3.5.

Figure 14.5: Income inequalities between Group A & B



From Figure 3.5 it is obvious from the above figure that inequalities in the distribution of income are more in group A than in group B.



CHECK YOUR PROGRESS

Q 1: How did Pareto formulated the law of income distribution? (Answer in about 60 words).

.....

.....

.....

.....

.....

.....

Q 2: What does the Lorenz curve reflect upon? (Answer in about 40 words).

.....

.....

.....

.....

Q 3: Define the Gini Coefficient. (Answer in about 30 words).

.....

.....

.....



14.7 LET US SUM UP

- The concept of Pareto Distribution was put forward by Vilfredo Pareto. This is also referred to as the Pareto Principle or the 80-20 Rule. The Pareto distribution is used in describing social, scientific, and geophysical phenomena in a society. Pareto created a mathematical formula in the early 20th century that described the inequalities in wealth distribution that existed in his native country of Italy.
- The definition of the Pareto Distribution was later expanded in the 1940s by Dr. Joseph M. Juran, a prominent product quality guru. Juran applied the Pareto principle to quality control for business production to show that 20 percent of the production process defects are responsible for 80 percent of the problems in most products.
- A log-normal distribution is a statistical distribution of logarithmic values from a related normal distribution. A log-normal distribution can be translated to a normal distribution and vice versa using associated logarithmic calculations.
- The graphical method of measurement of dispersion was developed by a famous economic statisticion Dr. Max O. Lorenz. He used this method to measure the inequalities of income or wealth of a society. In this method, cumulative percentage of income or wealth, i.e., the percentage of income (or wealth) less than a given value is plotted against the cumulative percentage of persons. The curve thus obtained, is known as Lorenz Curve.
- Gini coefficient is based on the Lorenz Curve and is defined as the ratio of the area between the diagonal and the Lorenz curve to the total area of the half-square in which the curve lies.



14.8 FURTHER READING

1) A.M. Gun, M.K. Gupta, B. Dasgupta, "Fundamentals Of Statistics", Volume two, Seventh Edition.

2) Richard I. Levin, David S. Rubin,"Statistics For Management", Seventh Edition.



14.9 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: In 1906, Vilfredo Pareto observed that 20 percent of the pea pods were responsible for 80 percent of the peas planted in his garden. Again, he also found that 80 percent of his country's wealth was owned by about 20 percent of its population. Further, he also found that 80 percent of the land was owned by a handful of wealthy citizens, who comprised about 20 percent of the population. On the basis of such observations, Paretor formulated his law of income distribution.

Ans to Q No 2: Lorenz curves tries to measure the inequalities of income or wealth of a society. In this method, cumulative percentage of income or wealth, i.e., the percentage of income (or wealth) less than a given value is plotted against the cumulative percentage of persons.

Ans to Q No 3: Gini coefficient is based on the Lorenz Curve and is defined as the ratio of the area between the diagonal and the Lorenz curve to the total area of the half-square in which the curve lies.



14.10 MODEL QUESTIONS

Answer the following question in around 150 words

Q 1: Describe Pareto's law of income distribution and its empirical validity.

Q 2: Describe the concept of Lorenz curve.

Q 3: Discuss the concept of log-normal distribution. Mention its vaious uses.

*** ***** ***