PGEC-S2 03
Exam. Code: SM

Statistical Methods

SEMESTER - II

ECONOMICS

BLOCK 1

KRISHNA KANTA HANDIQUI STATE OPEN UNIVERSITY

Subject Experts

Professor Madhurjya Prasad Bezbaruah, Dept. of Economics, Gauhati University

Professor Nissar Ahmed Barua, Dept. of Economics, Gauhati University

Dr. Gautam Mazumdar, Dept. of Economics, Cotton University, Guwahati

Course Coordinator: Dr. Bhaskar Sarmah, Economics, KKHSOU

SLM Preparation Team

Units	Contributors
1, 2, 4 & 5	**Dr. Pranjal Sarma**, Department of Statistics L. C. B. College, Guwahati
3	**Dr. Pranjal Sarma**, Department of Statistics L. C. B. College, Guwahati, & **Dr. Bhaskar Sarmah,** KKHSOU
6, 7	**Dr. Rijusmita Sarma**, HoD, Department of Statistics L. C. B. College, Guwahati

Editorial Team

Content : **Dr. Parag Dutta,** Associate Professor in Economics, KKHSOU

Structure, Format & Graphics : **Dr. Bhaskar Sarmah,** Assistant Professor in Economics, KKHSOU
Dr. Harekrishna Deka, Assistant Professor in Mathematics, KKHSOU

March, 2020.

ISBN: 978-93-89955-67-5

This Self Learning Material (SLM) of the Krishna KantaHandiqui State Open University is made available under a Creative Commons Attribution-NonCommercial-ShareAlike4.0 License (international): http://creativecommons.org/licenses/by-nc-sa/4.0/

Printed and published by Registrar on behalf of the Krishna Kanta Handiqui State Open University.

The university acknowledges with thanks the financial support provided by the **Distance Education Bureau, UGC, New Delhi**, for the preparation of this study material.

Headquarters: Patgaon, Rani Gate, Guwahati 781017; Web : www.kkhsou.in/web/

City Office: Housefed Complex, Dispur, Guwahati-781 006.

COURSE INTRODUCTION

This course entitled *'Statistical Methods'* is common to both MA in Economics programme offered by this University. This is the third course of the 2nd Semester curriculum. As statistical tools are widely used in various economic theories and illustrations; this course has been designed to offer us the basic concept of the subject Statistics and their economic applications. This course comprises 14 units and has been divided into two blocks. The first block consists of seven units while the second block consists also of seven units.

BLOCK INTRODUCTION

This is the first block of the course. This block consists of seven units. Unit 1 gives us a broad overview of measures of central tendency. In unit 2 we will discuss the measures of dispersion. In Unit 3, we have disucssed the concept of economic inquality, in which we have basically discussed the concepts like Pareto's law of income distribution, Lorenz curve as as well as Gini coefficient. Unit 4 discusses about correlation, scatter diagram, types of correlation, correlation coefficient, methods of measuring correlation etc. Unit 5 is about regression. Here the concept of linear regression, lines of regression, regression equation etc. have been discussed. In unit 6 the concept of probability has been discussed. Here we will discuss different theorems on probability. In the last unit of this block expectations and moments have been discussed.

While going through a unit, you will notice some along-sode boxes, which have been included to help you know some of the difficult, unseen terms. Some 'ACTIVITY' (s) has been included to help you apply your own thoughts. Again, we have included some relevant concepts in 'LET US KNOW' along with the text. At the end of sections, you will get 'CHECK YOUR PROGRESS' questions. These have been designed to self-check your progress of study. It will be better if you solve the problems put in these boxes immediately after you go through the sections of the units and then match your answers with 'ANSWERS TO CHECK YOUR PROGRESS' given at the end of each unit.

CONTENTS

		Pages :
UNIT 1:	**Measures of Central Tendency** Arithmatic Mean; Median; Mode; Geometric Mean; Harmonic Mean; Weighted Average	**05-30**
UNIT 2:	**Measures of Dispersion** Range; Quartile Deviation; Mean Deviation; Standard Deviation, Standard Deviation; Coefficient of Variation	**31-52**
UNIT 3:	**Measurement of Economic Inequality** Pareto's Law of Income Distribution; Log normal Distribution (Concept only); Lorenz Curve; Coefficient of Inequality : Gini Coefficient	**53-63**
UNIT 4:	**Correlation** Correlation Analysis; Types of Correlation; Methods of Measuring Correlation; Rank Correlation Coefficient; Tied or Repeated Ranks	**64-86**
UNIT 5:	**Regression Analysis** Regression; Two-Variable Regression Model (Least Square Method); Properties of Regression, Coeffieicents; Correlation vs. Regression	**87-109**
UNIT 6:	**Probability : Basic Ideas** Definition of the Basic Terms; Classical or Mathematical or Apriori Definition of Probability; Addition Theorems or Theorems on Total Probability; The Theorem of Compound Probability or the Multiplicative Law of Probability; Baye's Theorem (Concept only)	**110-127**
UNIT 7:	**Expectations and Moments** Random Variable: Definition or Random Variable, Discrete Random Variable, Continuous Random Variable; Probability Distribution: Probability Mass Function, Probability Density Function; Mathematical Expectation and Variance: Mathematical Expectation, Properties of Mathematical Expectation, Variance of Random Variable, Properties of Variances; Moments; Skewness: Karl Pearson's Co-efficient of Skewness, Boweley's Co-efficient of Skewness; Kurtosis: Measures of Kurtosis	**128-161**

UNIT 1: MEASURES OF CENTRAL TENDENCY

UNIT STRUCTURE

1.1 Learning Objectives
1.2 Introduction
1.3 Arithmatic Mean
1.4 Median
1.5 Mode
1.6 Geometric Mean (GM)
1.7 Harmonic Mean (HM)
1.8 Weighted Average
1.9 Exercises
1.10 Let Us Sum Up
1.11 Further Reading
1.12 Answers to Check Your Progress
1.13 Model Questions

1.1 LEARNING OBJECTIVES

After going through this unit, you will able to:

- give the different measures of central tendency
- write the computational tichniques of different measures of central tendency
- state the properties and respective limitations of these measures.

1.2 INTRODUCTION

Summarisation of the data is a necessary function for any statistical analysis and interpretation. Hence, as a first step, the huge mass of unorganised data are summarised in the form of tables and frequency distributions. Further, to bring out different characteristics of the data into sharp focus, these tables and frequency distributions are to be summarised again using different statistical tools. A measure of central tendency or an avrage is a statistical tool that is used for summarisation of data for statistical analysis.

Definition of Central Tendency

In many a case, like the distributions of height, weight, marks, profit, wage and so on, it has been noted that starting with rather low frequency, the class frequency gradually increases till it reaches its maximum somewhere near the central part of the distribution and after which the class frequency steadily falls to its minimum value towards the end. Thus, central tendency may be difined as the tendency of a given set of observations to cluster around a single central or middle value and the single value that represents the given set of observations is described as a measure of central tendency or average. Hence, it is possible to condense a vast mass of data by a single representative value. Furthermore, the central tendency also facilitates us in providing a basis for comparison between different distributions. Following are the different measures of central tendency :

(i) Arithmetic Mean (AM)

(ii) Median (Me)

(iii) Mode (Mo)

(iv) Geometric Mean (GM)

(v) Harmonic Mean (HM)

Criteria for an Ideal Measure of Central Tendency

Following are the criteria for an ideal measure of central tendency:

(i) It should be properly and unambignously defined.

(ii) It should be easy to compute.

(iii) It should be easy to understand.

(iv) It should be based on all the observations.

(v) It should be capable of further mathematical treatment.

(vi) It should be least affected by the presence of extreme observations.

(vii) It should not be much affected by the fluctuation of sampling.

1.3 ARITHMATIC MEAN

For a given set of observations, the Aritnmatic Mean (AM) may be defined as the sum of all the observations to be divided by the number of observations. Thus, if a variable X assumes n values $x_1, x_2, ..., x_n$, then the AM of X, to be denoted by $\overline{X}$, is given by,

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i \text{ or } \frac{1}{n}\Sigma x$$

In case of a simple frequency distribution relating to an attribute, we have

$$\overline{X} = \frac{f_1x_1 + f_2x_2 + \ldots + f_nx_n}{f_1 + f_2 + \ldots + f_n}$$

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

$$= \frac{1}{N}\sum_{i=1}^{n} f_i x_i \text{ or } \frac{1}{N}\Sigma fx$$

Assuming the observation x_i occurs f_i times, i = 1, 2, . . ., n and $N = \sum_{i=1}^{n} f_i = \Sigma f$ = total frequency.

In case of grouped frequency distribution also we may use the above formula with x_i as the mid value of the i^{th} class interval, on the assumption that all the values belonging to the i^{th} class interval are equal to x_i.

Assumed Mean Method :

This method is used when the magnitude of individual observations is large. The method is helpful in the simplification of calculation work.

Let A be any constant or assumed mean. Now, let us define a new variable $d_i = x_i - A$, i = 1, 2, . . ., n. The formula for mean of X, i.e. $\overline{X}$, in terms of d_i for an individual series of X is given as,

$$\overline{X} = A + \frac{\sum_{i=1}^{n} d_i}{n} = A + \frac{\Sigma d}{N}$$

Where, n = number of observations

A = assumed mean

Again, for a grouped frequency distribution the formula is given as,

$$\bar{X} = A + \frac{\sum_{i=1}^{n} f_i d_i}{N} = A + \frac{\Sigma fd}{N}$$

Where, N = total frequency = $\sum_{i=1}^{n} f_i$

A = assumed mean

Assuming the observation x_i occurs f_i times, i = 1, 2, . . ., n and $d_i = x_i - A$, i = 1, 2, . . ., n

Step Deviation Method or Coding Method

In a grouped frequency distribution, if all the classes are of equal width, say h, the successive mid-values of various classes will differ from each other by this width. This fact can be utilised for reducing the work of computations.

Let us define $d_i' = \frac{x_i - A}{h}$, i = 1, 2, . . ., n

Where, A = assumed mean

h = class width

The formula for mean of X, i.e. $\bar{X}$, in terms of d_i' is given as,

$$\bar{X} = A + \frac{\sum_{i=1}^{n} f_i d_i'}{N} \times h = A + \frac{\Sigma fd'}{N} \times h$$

Example 1: Following are the daily wages in rupees of a sample of 9 workers: 58, 62, 48, 53, 70, 52, 60, 84, 75. Compute the mean wage.

Solution : Let x denote the daily wage in rupees.

Then as given, x_1 = 58, x_2 = 62, x_3 = 48, x_4 = 53, x_5 = 70, x_6 = 52, x_7 = 60, x_8 = 84 and x_9 = 75.

Thus, the mean wage, i.e. $\bar{X}$, is given by,

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

$$= \frac{1}{9} \times (58+62+48+53+70+52+60+84+75)$$

$$= \frac{1}{9} \times 562$$

= Rs. 62.44

Example 2: Compute the mean weight of a group of students of a college from the following data :

Weight in kgs. :	44-48	49-53	54-58	59-63	64-68	69-73
No. of students :	3	4	5	7	9	8

Solution : Let us prepare the following table :

(1) Weight in kgs	(2) No. of students (f_i)	(3) Mid-Value (x_i)	(4) = (2) x (3) $f_i x_i$
44-48	3	46	138
49-53	4	51	204
54-58	5	56	280
59-63	7	61	427
64-68	9	66	594
69-73	8	71	568
Total = N = 36			$\Sigma fi\, xi = 2211$

$\therefore$ Mean weight of 36 students is given as,

$$\bar{X} = \frac{1}{n}\Sigma f_i x_i$$

$$= \frac{1}{36} \times 2211 \text{ kgs.}$$

= 61.42 kgs.

Example 3: The following is the frequency distribution of age of 670 students of a school. Compute the arithmetic mean of the data.

Age (in years) :	5	6	7	8	9	10	11	12	13	14
No. of students :	25	45	90	165	112	96	81	26	18	12

Solution : To apply Assumed Mean formula, let us prepare the following table taking A = 8 as assumed mean in the formula.

x (Age)	f (No. of student)	d = x = A = x – 8	fd
5	25	–3	–75
6	45	–2	–90
7	90	–1	–90

8	165	0	0
9	112	1	112
10	96	2	192
11	81	3	243
12	26	4	104
13	18	5	90
14	12	6	72
	N = 670		$\Sigma fd = 558$

$\therefore$ AM of age of the 670 students is given as,

$$\bar{x} = A + \frac{\Sigma fd}{N} = 8 + \frac{558}{670}$$

= 8 + 0.83 = 8.83 years

Example 4: Find the AM for the following distribution :

Class Interval :	350-369	370-389	390-409	410-429
Frequency :	23	38	58	82
Class Interval :	430-449	450-469	470-489	
Frequency :	65	31	11	

Solution : To find the AM of the distribution, let us use the Step-Deviation Method to simplify our calculations. Now, we prepare the following table:

Class Interval	Frequency (f)	Mid-value (x)	$d' = \frac{x - A}{h} = \frac{x - 419.50}{20}$	fd'
350-369	23	359.50	–3	–69
370-389	38	379.50	–2	–76
390-409	58	399.50	–1	–58
410-429	82	419.50 (A)	0	0
430-449	65	439.50	1	65
450-469	31	459.50	2	62
470-489	11	479.50	3	33
	N = 308			$\Sigma fd' = -43$

$\therefore$ The required AM is given by,

$$\bar{X} = A + \frac{\Sigma fd'}{N} \times h$$

$= 419.50 + \frac{(-43)}{308} = 419.50 - 2.79 = 416.71$

Example 5: The following distribution gives the pattern of overtime work done by 100 employees of a company. Calculate the average overtime work done per employee.

Overtime (in hrs) :	10-15	15-20	20-25	25-30	30-35	35-40
No. of employees :	11	20	35	20	8	6

Solution: Let us prepare the following table taking A=22.5 as assumed mean in the formula.

Overtime (in hrs)	No. of employee	Mid-value (x)	$d' = \frac{x-A}{h}$ $= \frac{x-225}{5}$	$f.d'$
10-15	11	12.5	-2	-22
15-20	20	17.5	-1	-20
20-25	35	22.5 ← A	0	0
25-30	20	27.5	1	20
30-35	8	32.5	2	16
35-40	6	37.5	3	18
	N=100			$\sum fd' = 12$

$\therefore$ The required AM is given by,

$$\overline{X} = A + \frac{\sum fd'}{N} \times h$$

$$= 22.5 + \frac{12}{100} \times 5 = 23.1 \text{ hrs.}$$

Properties of Arithmetic Mean

(i) If all the observations assumed by a variable are constants, say k, then the AM is also k.

For example, if the height of every student in a group of 10 students is 170 cm, then the mean height is, of course, 170 cm.

(ii) The algebric sum of deviations of a set of observations from their AM is zero. i.e. for individual series, $\Sigma(x_i - \bar{x}) = 0$ and for frequency distribution, $\Sigma f_i(x_i - \bar{x}) = 0$

For example, if a variable X assumes five observations, say, 58, 63,

37, 45, 29, then $\bar{x}$ = 46.4. Hence, the deviations of the observations from the AM i.e. $(x_i - \bar{x})$ are (58 – 46.4) = 11.60, (63–46.4) = 16.60, (37–46.4) = –9.40, (45–46.4) = –1.40, (29–46.4) = –17.40. Now, $\Sigma(x_i - \bar{x})$ = 11.60+16.60+(–9.40)+(–1.40)+(–17.40) = 0.

(iii) AM is affected due to change of origin and/or scale which implies that if the original variable x is changed to another variable y so that, $y = \frac{x \pm a}{b}$ then $\bar{y} = \frac{\bar{x} \pm a}{b}$.

For example, suppose the AM of the values of a variable x is 25. Now, if each value of x is increased by 5, we will get a new variable defined as,

$y = x + 5$

$\therefore$ the AM of the new variable would be,

$\bar{y} = \bar{x} + 5$ = 25 + 5 = 30

i.e. after increament in each value of x, the new AM is 30.

(iv) If there are two groups containing n_1 and n_2 observations and $\bar{x}_1$ and $\bar{x}_2$ are the respective arithmetic means, then the combined AM is given by, $\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$

This property could be extended to k groups and we may write

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i \bullet \bar{x}_i}{\sum_{i=1}^{n} n_i}$$

Example 6: The mean salary for a group of 40 female workers is Rs. 5200 per month and that for a group of 60 male workers is Rs. 6800 per month. What is the combined salary?

Solution : As given, n_1 = 40, n_2 = 60, $\bar{x}_1$ = Rs. 5200 and $\bar{x}_2$ = Rs. 6800 Hence, the combined mean salary per month is,

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

$$= \frac{40 \times Rs.\ 5200 + 60 \times Rs.\ 6800}{40 + 60} = \text{Rs. } 6160.$$



CHECK YOUR PROGRESS

Q 1: State true or false

(i) Wtih nominal data, the mean should be used as a measure of central tendency. (T/F)

(ii) When the date are interval or ratio, we can use the mean as a measure of central tendency. (T/F)

(iii) The mean of a data set remains unaffected if an observation equal to mean is included in it. (T/F)

(iv) The sum of deviations from mean is zero. (T/F)

1.4 MEDIAN

Median, for a given set of observations, may be defined as the middle-most value when the observations are arranged either in an ascending order or a descending order of magnitude. In other words, median is termed as a positional average which means that the value of the median is dependent upon the position of the given set of observations for which the median is wanted.

In order to determine the median of an individual series, first of all we have to arrange the observations either in ascending or in descending order of magnitude. Now, suppose we assume that there are n observations in the series. If n is odd i.e. there are odd number of observations in the series, then the $\left(\frac{n+1}{2}\right)^{th}$ value from the beginning will be the median. If n is even, i.e. there are even number of observations in the series, 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.

Again, to determine the median of an ungrouped frequency distribution, first of all we have to construct a cumulative frequency table. If the total frequency (i.e. the total no. of observations) N is odd then the $\left(\frac{N+1}{2}\right)^{th}$ term will be the median. Otherwise, if N is even then the AM of the

$\frac{N}{2}^{th}$ and the $\left(\frac{N}{2}+1\right)^{th}$ terms would be the median.

To determine median from a grouped frequency distribution, first of all we have to construct a cumulative frequency table. Then we determine the median class from the table. The median class is that which contains the median value. If the total frequency N is odd then the median class will be that class which will contain the $\left(\frac{N+1}{2}\right)^{th}$ observation. Again, if N is even then the median class will be that class which will contain the $\frac{N}{2}^{th}$ observation. After determining the median class the particular median value is calculated by using the following formula :

$$\text{Median (Me)} = L + \frac{\frac{N}{2} - fc}{f} \times h$$

Where, L = lower limit of the median class

f = frequency of the median class

fc = cumulative frequency just before the median class

h = width of the median class

N = total frequency

Example 7: Find the medians of the following two series :

(i) 20, 15, 25, 28, 18, 16, 30

(ii) 245, 230, 265, 236, 220, 250

Solution : (i) Arranging the observations in ascending order,

we get, 15, 16, 18, 20, 25, 28, 30.

Since, no. of observations n = 7 i.e. odd, the median would be the $\left(\frac{7+1}{2}\right)^{th}$ term, i.e. the 4th observation.

Hence, median of the series = 20

(ii) Arranging the observations in ascending order,

we get, 220, 230, 236, 245, 250, 265

Here, no. of observations n = 6, i.e. even.

$\therefore$ Median will be AM of the $\left(\frac{n}{2}\right)^{th}$ and the $\left(\frac{n}{2}+1\right)^{th}$ terms, i.e. AM of the 3rd and 4th term.

Hence, median of the series $= \frac{236+245}{2} = 240.5$

Example 8: Find the median of the following distribution :

Variable (x) :	10	11	12	13	14	15	16
Frequency (f) :	8	15	25	20	12	10	5

Solution : Let as prepare the cumulative frequency table :

x	f	c.f.
10	8	8
11	15	23
12	25	48
13	20	68
14	12	80
15	10	90
16	5	95
	N = 95	

Here, N = 95, which is odd.

Hence, $\left(\frac{N}{2}+1\right)^{th}$ term would be the median,

i.e. $\left(\frac{95+1}{2}\right)^{th}$ term would be the median,

i.e. 48th term would be the median,

i.e. X = 12 is the median.

Example 9: Find the median of 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 : Let us prepare the following cumulative frequency table :

Wages (Rs.)	No. of worker (frequency)	Cumulative frequency
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, N = 124, which is even.

Hence, median = AM of the $\left(\frac{N}{2}\right)^{th}$ and the $\left(\frac{N}{2}+1\right)^{th}$ terms

= AM of the $\left(\frac{124}{2}\right)^{th}$ and the $\left(\frac{124}{2}+1\right)^{th}$ terms

= AM of the 62nd and 63rd terms

= $\frac{24+24}{2}$ ($\because$ both the terms are living between c.f. 45 and 65)

= Rs. 24

Example 10: Determin median for the following distribution :

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

Solution : Let us prepare the following cumulative frequency table :

Daily Wages	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

$\therefore \left(\frac{N+1}{2}\right)^{th}$ term = $\left(\frac{111+1}{2}\right)^{th}$ term = 56th term

From the cumulative frequency table we find that the 56th term lies in the class 65-70.

$\therefore$ 65-70 is the median class.

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

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

$\therefore$ 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

Example 11: Calculate median of the following data :

Weight in inches : 3-4 4-5 5-6 6-7 7-8 8-9 9-10 10-11

No. of saplings : 3 7 12 16 22 20 13 7

Solution : Let us prepare the following table :

Class Intervals	Frequency (f)	Cumulative frequency (fc)
3-5	3	3
4-5	7	10
5-6	12	22
6-7	16	38
7-8	22	60

8-9	20	80
9-10	13	93
10-11	7	100

N =100

Here, N = 100 which is even.

$\therefore \frac{N}{2}^{th}$ term = $\frac{100}{2}^{th}$ term = 50th term.

From the cumulative frequency table we find that the 50th term lies in the class 7-8.

$\therefore$ 7-8 is the median class.

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

Here, L =7, f = 22, fc = 38, N = 100, h = 1.

$\therefore$ Median $= 7 + \frac{\frac{100}{2} - 38}{22} \times 1$

$= 7 + \frac{50 - 38}{22} \times 1$

$= 7 + \frac{12}{22}$

= 7 + 0.55

= 7.55 inches



CHECK YOUR PROGRESS

Q 2: Fill in the blanks :

(i) Sum of obsolute deviations from median is _____.

(ii) The median is less affected than mean by _____ values of observations in a distribution.

(iii) The average which divides a distribution into two equal parts is _____.

1.5 MODE

For a given set of observations, mode may be defined as the value that occurs maximum number of times. Thus, mode is that value which has the maximum concentration of the observations around it. This can also be deseribed as the most common value with which, even a layman may be familiar with.

To determine mode from an individual series, we identify the observation which appears maximum number of times. If more than one observation appears maximum number of times, all those observations are termed as mode. But, if all the observations of the series appears equal number of times, we say that, there is no mode of the given series.

To obtain mode for a grouped frequency distribution, first of all we have to identify the modal class. The class against which the frequency is maximum, is termed as modal class. Further, to find exact value of mode we apply the following formula :

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

Where, L = lower limit of the modal class

f_1 = frequency of the modal class

f_0 = frequency of the class preceding the modal class

f_2 = frequency of the class succeeding the modal class.

h = width of the modal class.

Example 12: Determine mode/modes of the following series, if any :

(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 appears maximum number of times. Hence, 4 is the mode of the given series.

(ii) The numbers 7 and 9 appears maximum number of times i.e. 2 times. Hence, 7 and 9 both are modes of the given series.

(iii) Since, all the numbers appear equal number of times i.e. 2 times, hence the given series has no mode.

Example 13: Compute mode for the following distribution :

Class Interval :	350-369	370-389	390-409	410-429
Frequency :	23	38	58	82
Class Interval :	430-449	450-469	470-489	
Frequency :	65	31	11	

Solution : Since the frequency of the class 410-429 is the maximum, hence this class is the modal class. The continuous form of this modal class would be 409.5 – 429.5.

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

Here, L = 409.5, f_1 = 82, f_0 = 58, f_2 = 65, h = 20

$\therefore$ Mode $= 409.5 + \frac{82 - 58}{2 \times 82 - 58 - 65} \times 20$

= 421.21

Example 14: The following table gives the incomplete income distribution of 300 workers of a firm, where the frequencies of the classes 3000-4000 and 5000-6000 are missing. If the mode of the distribution is Rs. 4428.57 find the missing frequencies.

Monthly Income (Rs.)	No. of Workers
1000-2000	30
2000-3000	35
3000-4000	?
4000-5000	75
5000-6000	?
6000-7000	30
7000-8000	15

Solution : Let the frequency of the class 3000-4000 be fi. Then the frequency of the class 5000-6000 will be equal to 300-30-35-fi-75-30-15=115-fi. It is given that, mode=4428.57, therefore, modal class is 4000-5000.

Thus, L=4000, h=1000

$$= L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2}$$

$$\Rightarrow 4428.57 = 4000 + \frac{75 - fi}{2 \times 75 - fi - (115 - fi)}$$

$\Rightarrow fi = 60$ (after simplification)

$\therefore$ frequency of the class 3000-4000 is 60 and the frequency of the class 5000-6000 is (115-fi=115-60)=55.



CHECK YOUR PROGRESS

Q 3: Fill in the blanks :

(i) The mode is always found at the _____ point of a graph of frequency distribution.

(ii) Most appropriate average to determine the size of oranges on a tree is _____

(iii) Mode _____ based on all the observations.

(iv) Like mean or median, mode is not affected by _____ observations.

Relationship among Mean, Median and Mode

Prof. Karl Pearson has established a relationship empirically i.e. experimentally among mean, median and mode. Please note, for any symmetcial distribution, Mean=Median=Mode. However, for asymmetrical distribution, the following relationship among mean, median and mode approximately holds for:

Mean – Mode = 3 (Mean – Median)

1.6 GEOMETRIC MEAN

The geometric mean (GM) is a type of mean or average, which indicates the central tendency or typical value of a set of numbers by using the product of their values (as opposed to the arithmetic mean which uses their sum). If a variable X assumes n values $x_1, x_2, x_3, ..., x_n$, all the values being positive, then the Geometric Mean (GM) of X, denoted by G, is given by

$$G = (x_1 \times x_2 \times x_3 \times \ldots \times x_n)^{\frac{1}{n}}$$

Again, if the frequencies of the values $x_1, x_2, x_3, ..., x_n$ of X are $f_1, f_2, f_3, ..., f_n$ respectively then the geometric mean G of X is given by

$$G = \left(x_1^{f_1} \times x_2^{f_2} \times x_3^{f_3} \times \ldots \times x_n^{f_n}\right)^{\frac{1}{N}}$$

Where $N = \Sigma f$ = total frequency

Example 15: Find the GM of 2, 6 and 12

Solution : As given $x_1 = 2$, $x_2 = 6$, $x_3 = 12$ and n = 3

$$\therefore GM = (x_1 \times x_2 \times x_3)^{\frac{1}{n}}$$

$$= (2 \times 6 \times 12)^{\frac{1}{3}}$$

$$= (6^3)^{\frac{1}{3}}$$

= 6

Example 16: Find the GM for the following distribution :

x :	2	4	8	16
y :	2	3	3	2

Solution : Here, N = total frequency $= \Sigma f = 10$

$$\therefore GM = \left({x_1}^{f^1} \times {x_2}^{f2} \times {x_3}^{f3} \times {x_4}^{f^n}\right)^{\frac{1}{N}}$$

$$= \left(2^2 \times 4^3 \times 8^3 \times 16^2\right)^{\frac{1}{10}}$$

$$= (2)^{2.50}$$

$$= 4\sqrt{2}$$

= 5.66

1.7 HARMONIC MEAN

The reciprocal of the arithmetic mean of the reciprocals of the values of a variable is called the harmonic mean (HM). So, if a variable x assumes n non-zero values $x_1, x_2, x_3, ..., x_n$ then the HM of x denoted by H is given by

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots + \frac{1}{x_n}} = \frac{n}{\Sigma \frac{1}{x}}$$

Again, if the frequencies of x are $f_1, f_2, f_3, ..., f_n$ then the harmonic mean of x is given by

$$H = \frac{f_1 + f_2 + f_3 + \ldots + f_n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \frac{f_3}{x_3} + \ldots + \frac{f_n}{x_n}} = \frac{N}{\Sigma \frac{f}{x}}$$, where $N = \Sigma f$

Example 16: Find the HM of 4, 6 and 10

Solution : Here, x_1 = 4, x_2 = 6, x_3 = 10 and n = 3

$$\therefore HM = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}} = \frac{3}{\frac{1}{4}+\frac{1}{6}+\frac{1}{10}}$$

$$= \frac{3}{0.25+0.17+0.10} = 6.77$$

Example 17: Find the HM for the following data

x :	2	4	8	16
f :	2	3	3	2

Solution : Here, N = total frequency = (Σf) = 10

$$\therefore HM = \frac{N}{\frac{f_1}{x_1}+\frac{f_2}{x_2}+\frac{f_3}{x_3}+\frac{f_3}{x_n}}$$

$$= \frac{10}{\frac{2}{2}+\frac{3}{4}+\frac{3}{8}+\frac{2}{16}}$$

= 4.44

Example 18: An investor buys Rs. 20,000 worth of shares of a company each month. During the first 3 months he bought the shares at a price of Rs. 120, Rs. 160 and Rs. 210. After 3 months what is the average price paid by him for the shares?

Solution : Since the value of shares are changing after every one month, therefore the required average price per share is the harmonic mean of the prices paid in first three months.

$$\therefore H.M = \frac{3}{\frac{1}{120}+\frac{1}{160}+\frac{1}{210}} = \frac{3}{0.008+0.006+0.004}$$

$$= \frac{3}{0.018} = \text{Rs. } 166.66$$



CHECK YOUR PROGRESS

Q 4: Fill in the blanks :

(i) If an deservation in the data set is zero, then its geometric mean is_____.

(ii) _____ is most suitable for averaging ratios and exponential

rates of changes.

(iii) _____ is the reciprocal of the A.M. of reciprocal of observations.

Relation among AM, GM and HM

(a) For any finite number of positive values of a variable, its $AM \geq GM \geq HM$

(b) For any two positive numbers, AM x HM = $(GM)^2$

1.8 WEIGHTED AVERAGE

The numbers which represent the relative importance of the quantities of a series are called the weights of these quantities. Weights may be actual or assumed. The average obtained giving due importance to the quantities of a series or distribution is called the weighted average of the series.

Let the weights attached to the quantities $x_1, x_2, x_3, ..., x_n$ be $w_1, w_2, w_3, ..., w_n$ respectively. The weighted arithmetic mean of these quantities, denoted by $\bar{x}_w$, is given by

Withted $AM = \bar{x}_w = \dfrac{w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n}{w_1 + w_2 + w_3 + \ldots + w_n}$

$$= \frac{\Sigma wx}{\Sigma w}$$

Similarly, the geometric mean of the quantities is given by

Weighted $GM = Anti\log\left(\dfrac{\Sigma w \log x}{\Sigma w}\right)$

Again, the harmonic mean of the same quantities is given by

Weighted $HM = \dfrac{\Sigma w}{\Sigma\left(\dfrac{w}{x}\right)}$

Example 19: Calculate the simple and the weighted arithmetic mean prices per quintal of coal purchased by a factory during the first six months of a year from the following data. When will the two averages be equal?

Month :	January	February	March	April	May	June
Price per quintal (Rs.) :	212.50	256.25	250	260	221.25	270
Quantity purchased (in quintal) :	25	30	40	50	10	45

Solution : Simple AM = Rs. $\left(\frac{212.50 + 256.25 + \ldots + 270}{6}\right)$

= Rs. $\frac{1470}{6}$

= Rs. 245.00

i.e the simple AM of price per quintal of coal = Rs. 245.00

In calculating weighted mean price, the quantities of coal purchased are to be taken as the weights of various prices.

$\therefore$ Weighted AM = Rs. $\left[\frac{(25 \times 212.50) + (30 \times 256.25) + \ldots + (45 \times 270)}{25 + 30 + \ldots + 45}\right]$

= Rs. $\frac{50362.50}{200}$

= Rs. 251.81

1.9 EXERCISES

Q 1: What is measures of central tendency? Explain the different methods of measuring central tendency.

Q 2: What are the characteristics of an ideal average? How far these are satisfied by the different methods?

Q 3: Define weighted average. Distinguish between a simple average and a weighted average.

Q 4: Complute arithmetic mean of the following data :

Marks :	0-10	10-20	20-30	30-40	40-50	50-60
No. of students :	12	18	27	20	17	6

Q 5: Calculate mean and mode from the following data :

Class Interval :	350-369	370-389	390-409	410-429
Frequency :	15	27	31	19

Class Interval :	430-449	450-469
Frequency :	13	6

(Ans. 400.58 and 394.50)

Q 6: An aeroplance flies from A to B at the rate of 500 km/hour and comes back from B to A at the rate of 700 km/hour. Find the average speed of the aeroplane.

(Ans. 583.33 km/hour)

Q 7: Following is an incomplete distribution having modal mark as 44.

Marks :	0–20	20–40	40–60	60–80	80–100
No. of students :	5	18	?	12	5

Find the missing value. (Ans. 48)

Q 8: Multiple choice questions :

(i) If the AM and GM for two numbers are 6.50 and 6 respectively then the two numbers are

(a) 6 and 7 (b) 9 and 4

(c) 10 and 3 (d) 8 and 5

[Ans. (b)]

(ii) If the AM and HM for two numbers are 5 and 3.2 respectively then the GM will be

(a) 16.00 (b) 4.10

(c) 4.05 (d) 4.00

[Ans. (d)]

(iii) For open-end classification, which of the following is the best measure of central tendency?

(a) AM (b) GM

(c) Median (d) Mode

[Ans. (c)]

(iv) When a firm registers both profits and losses, which of the following measure of central tendency cannot be considered?

(a) AM (b) GM

(c) Median (d) Mode

[Ans. (b)]

(v) If a constant 5 is added to each value of a set, the AM of the new

set is

(a) increased by 5 (b) decreased by 5

(c) not affected (d) ten times the original AM

[Ans. (a)]

(vi) If each observation of a set is divided by 2, then the mean of the new values

(a) is same as the original mean (b) is decreased by 2

(b) is half of the original mean (d) remaining the same

[Ans. (c)]

(vii) The geometric mean of $\frac{1}{32}$ and $\frac{8}{25}$ is

(a) 5 (b) 15

(c) 10 (d) 100

[Ans. (c)]

(viii) Which one of the following is not uniquely defined?

(a) Mean (b) Median

(c) Mode (d) All of these measures

[Ans. (c)]

(ix) What is the median for the following observations?

5, 8, 6, 9, 11, 4

(a) 6 (b) 7

(c) 8 (d) None of these

[Ans. (b)]

(x) If the AM and GM for 10 observations are both 15, then the value of HM is

(a) Less than 15 (b) More than 15

(c) 15 (d) cannot be datermined

[Ans. (c)]



1.10 LET US SUM UP

- Central tendency may be difined as the tendency of a given set of observations to cluster around a single central or middle value and the

single value that represents the given set of observations is described as a measure of central tendency or average.

- Following are the different measures of central tendency :
  (i) Arithmetic Mean (AM) or Mean
  (ii) Median (Me)
  (iii) Mode (Mo)
  (iv) Geometric Mean (GM)
  (v) Harmonic Mean (HM)
- For a given set of observations, the Arithmatic Mean (AM) may be defined as the sum of all the observations to be divided by the number of observations.
- Median, for a given set of observations, may be defined as the middle-most value when the observations are arranged either in an ascending order or a descending order of magnitude.
- Mode is that value which has the maximum concentration of the observations around it.
- Prof. Karl Pearson has established a relationship empirically i.e. experimentally among mean, median and mode. For any distribution the following relationship among mean, median and mode approximately holds:

  Mean – Mode = 3 (Mean – Median)
- The geometric mean is a type of mean or average, which indicates the central tendency or typical value of a set of numbers by using the product of their values (as opposed to the arithmetic mean which uses their sum).
- The reciprocal of the arithmetic mean of the reciprocals of the values of a variable is called the harmonic mean.
- Relation among AM, GM and HM
  (a) For any finite number of positive values of a variable, its $AM \geq GM \geq HM$
  (b) For any two positive numbers, AM x HM = $(GM)^2$



1.11 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.



1.12 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: (i) F (ii) T (iii) T (iv) T

Ans. to Q. No. 2: (i) minimum (ii) Extreme (iii) median

Ans. to Q. No. 3: (i) highest (ii) mode (iii) isn't (iv) extreme

Ans. to Q. No. 4: (i) zero (ii) geometric mean

(iii) Harmonic Mean



1.13 MODEL QUESTIONS

A. SHORT TYPE QUESTIONS

Q 1: Write down a few properties of Arithmatic Mean.

Q 2: Show the relationship between Mean, Median and Mode.

Q 3: State the relationship between Arithmetic Mean, Geometric Mean and Harmonic Mean.

B. LONG QUESTIONS

Q 1: Discuss the step deviation method of calculating Arithmetic Mean with suitable example.

Q 2: 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 3: What is the value of mean and median for the following data :

Marks :	5-14	15-24	25-34	35-44	45-54	55-64
No. of Students :	10	18	32	26	14	10

*** ***** ***

UNIT 2 : MEASURES OF DISPERSION

UNIT STRUCTURE

2.1 Learning Objectives

2.2 Introduction

2.3 Range

2.4 Quartile Deviation

2.5 Mean Deviation

2.6 Standard Deviation

2.7 Coefficient of Variation

2.8 Multiple Choice Questions

2.9 Exercises

2.10 Let Us Sum Up

2.11 Further Reading

2.12 Answers to Check Your Progress

2.13 Model Questions

2.1 LEARNING OBJECTIVES

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

- discuss the concept of dispersion and discuss its various measures
- use the computational techniques of different measures of dispersion
- write the concept of Lorenz curve and Gini coefficient.

2.2 INTRODUCTION

A measure of central tendency summarises the distribution of a variable into a single figure which can be regarded as its representative. This measure alone, however, is not sufficient to describe a distribution because there may be a situation where two or more different distributions have the same central value. Conversely, it is possible that the pattern of distribution in two or more situations is same but the values of their central tendency are different. Hence, it is necessary to define some additional summary measures to adequately represent the characteristics of a

distribution. One such measure is known as the measure of dispersion or the measure of variation.

The concept of dispersion is related to the extent of scatter or variability in observations. In short, dispersion for a given set of observations may be defined as the amount of deviation of the observations, usually, from an appropriate measure of central tendency. Measures of dispersion may be broadly classified into

1. Absolute measures of dispersion
2. Relative measures of dispersion

Absolute measures of dispersion are classified into

(i) Range (ii) Quartile Deviation (iii) Mean Deviation

(iv) Standard Deviation.

Likewise, we have the following relative measures of dispersion :

(i) Coefficient of Range

(ii) Coefficient of Quartile Deviation

(iii) Coefficient of Mean Deviation

(iv) Coefficient of Variation.

Characteristics of a Good Measure of Dispersion

A good measure of dispersion should posses the following characteristies :

(1) It should be easy to calculate.

(2) It should be easy to understand.

(3) It should be rigidly defined.

(4) It should be based on all the observations.

(5) It should be capable of further mathematical treatment.

(6) It should not be unduly affected by extreme observations.

(7) It should not be much affected by the fluctuations of sampling.

2.3 RANGE

For a given set of observations, range may be defined as a difference between the largest and smallest observation. Thus if L and S denote the largest and smallest observations respectively then we have

Range = L – S

The corresponding relative measure of dispersion, known as coefficient of range, is given by

$$\text{Coefficient of range} = \frac{L-S}{L+S} \times 100$$

Example 1: Find range and coefficient of range for each of the following data :

(a) Weekly wages of 10 workers of a factory are :

310, 350, 420, 105, 115, 290, 245, 450, 300, 375

(b) The distribution of marks obtained by 100 students,

Marks :	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of Students :	6	14	21	20	18	10	5
Marks :	70-80	80-90	90-100				
No. of Students :	3	2	1				

(c) The age distribution of 60 school going children :

Age (in years) :	5-7	8-10	11-13	14-16	17-19
Frequency :	20	18	10	8	4

Solution :

(a) The largest and smallest wages are L = 450 and S = 105.

$\therefore$ Range = = L – S = 450 – 105 = 345

and coefficient of range $= \frac{L-S}{L+S} = \frac{450-105}{450+105} = 0.62$

(b) Here, the lowest class limit, that is, 5 = 0

and the highest class limit, that in, L = 100

$\therefore$ Range = L – S = 100 – 0 = 100

and coefficient of range $= \frac{L-S}{L+S} = \frac{100-0}{100+0} = 1$

(c) Here, the lowest class boundary, that is, S = 4.5

and the highest class boundary, that is, L = 19.5

$\therefore$ Range = L – S = 19.5 – 4.5 15

and coefficient of range $= \frac{L-S}{L+S} = \frac{19.5-4.5}{19.5+4.5} = 0.625$

2.4 QUARTILE DEVIATION

Half of the deviation between the third quartile (Q_3) and the first quartile

(Q_1) of a distribution is called the quartile deviation (Q. D.) of the distribution. Thus,

Quantile Deviation (Q. D.) = $\frac{Q_3 - Q_1}{2}$

and coefficient of quartile deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100$

Example 2: For a distribution, the values of the first quartile and the third quartile are 39.25 and 80.25. Find the quartile deviation and coefficient of quartile deviation.

Solution: We are given, Q_1 = 39.25 and Q_3 = 80.25

$\therefore$ Quantile Deviation (Q. D.) = $\frac{Q_3 - Q_1}{2}$

$$= \frac{80.25 - 39.25}{2} = \frac{41}{2} = 20.50$$

Example 3: Compute the quantite deviation and co-efficient of quartile deviation from the following distribution :

Weekly wages (in Rs.)	No. of Workers
0-100	12
100-200	18
200-300	35
300-400	42
400-500	50
500-600	45
600-700	20
700-800	8

Solution : Let us prepare the following table :

C.I.	Frequency	Comulative frequency
0-100	12	12
100-200	18	30
200-300	35	65
300-400	42	107
400-500	50	157
500-600	45	202
600-700	20	222
700-800	8	230
	N=230	

Calculation of Q_1

Since $\frac{N}{4} = \frac{230}{4} = 57.5$, therefore Q_1 in the class interval 200-300.

$\therefore L_1 =$ lower limit of the Q_1 class = 200

$f_1 =$ frequency of the Q_1 class = 35

$h_1 =$ width of the Q_1 class = 100

$f_{c_1} =$ cumulative frequency just before Q_1 class=30

Hence, $Q_1 = L_1 + \frac{\frac{N}{4} - f_{c_1}}{f_1} \times h_1$

$$= 200 + \frac{57.5 - 30}{35} \times 100 = \text{Rs. } 278.57$$

Calculation of Q_3

Since $\frac{3N}{4} = \frac{3 \times 230}{4} = 172.5$, therefore Q_3 lies in the class interval 500-600.

$\therefore L_3 =$ lower limit of the Q_3 class = 500

$f_3 =$ frequency of the Q_3 class = 45

$h_3 =$ width of the Q_3 class = 100

$f_{c_3} =$ cumulative frequency just before Q_3 class=157

Hence, $Q_3 = L_3 + \frac{\frac{3N}{4} f_{c_3}}{f_3} \times h_3$

$$= 500 + \frac{172.5 - 157}{45} \times 100$$

= Rs. 534.44

$$\therefore Q.D. = \frac{Q_3 - Q_1}{2} = \frac{534.44 - 278.57}{2} = \text{Rs } 127.94$$

and coefficient $Q.D. = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{534.44 - 278.57}{534.44 + 278.57} = 0.315$



CHECK YOUR PROGRESS

Q 1: Fill in the blanks :

(i) Range is a measure of dispersion most affected by _____ observations.

(ii) An appropriate measure of dispersion with open end class interval is_____.

(iii) Quartile deviation takes into account_____ 50% of the observations.

2.5 MEAN DEVIATION

For a given set of observation, mean deviation is defined as the arithmetic mean of the absolute deviation of the observations from an appropriate measure of central tendency. Hence, if a variable X assumes n values $x_1, x_2, ..., x_n$, then the mean deviation of X about an average A is given by,

$$MD_A = \frac{1}{n}\sum_{i=1}^{n}|x_i - A| = \frac{1}{n}\sum|x - A|$$

For a grouped frequency diftribution, mean deviation about an average A is given by

$$MD_A = \frac{1}{N}\sum_{i=1}^{n} f_i|x_i - A|$$

$$\frac{1}{N}\sum f|x - A|$$

Where x_i and f_i denote the mid value and frequency of the ith-class interval and

$N = \sum f_i$ = total frequency

A relative measure of dispersion applying mean deviation is given by

Coefficient of mean deviation = $\frac{Mean\ deviation\ about\ A}{A} \times 100$

Example 4: What is the mean deviation about mean for the following numbers?

5, 8, 10, 10, 12, 9

Solution : The mean is given by

$$\bar{x} = \frac{5+8+10+10+12+9}{6}$$

= 9

Now we prepare the following table :

x	$\lvert x-\bar{x}\rvert$
5	4
8	1
10	1
10	1
12	3
9	0
	Total = $\Sigma \lvert x-\bar{x}\rvert = 10$

Thus mean deviation about mean is given by

$$M.D. = \frac{1}{n}\Sigma \lvert x-\bar{x}\rvert = \frac{10}{6} = 1.67$$

Example 5: Compute the mean deviation about the arithmetic mean for the following date :

X :	1	3	5	7	9
f :	5	8	9	2	1

Also find the coefficient of the mean deviation about the AM.

Solution : Let us prepare the following table :

x	f	fx	$\lvert x-\bar{x}\rvert = \lvert x-3.88\rvert$	$f \lvert x-\bar{x}\rvert$
1	5	5	2.88	14.40
3	8	24	0.88	7.04
5	9	45	1.12	10.08
7	2	14	3.12	6.24
9	1	9	5.12	5.12
	$N = \Sigma f = 25$	$\Sigma fx = 97$		$\Sigma f \lvert x-\bar{x}\rvert = 42.88$

Here, $AM = \bar{x} = \frac{1}{N}\Sigma fx$

$$= \frac{1}{25} \times 97 = 3.88$$

$\therefore$ Mean deviation about AM = $= \frac{1}{N}\sum f\left\lvert x-\bar{x}\right\rvert = \frac{1}{25} \times 42.88$

Also the coefficient of MD about its AM is $= 1.72$

$$\frac{MD \text{ about } AM}{AM} \times 100$$

$$= \frac{1.72}{3.88} \times 100$$

= 44.33

Example 6: Compute the coefficient of mean deviation about median for the following distribution :

Weight in kgs :	40-50	50-60	60-70	70-80
No. of persons :	8	12	20	10

Solution : We need to compute the median weight in the first stage.

Let us prepare the following table :

Wt. (in kgs)	No. of persons (f)	Cumulative trequency (fi)	Mid value (x)	$\lvert x - Me \rvert = \lvert x - 62.5 \rvert$	$f \lvert x - Me \rvert$
40-50	8	8	45	17.5	140
50-60	12	29	55	7.5	90
60-70	20	40	65	2.5	50
70-80	10	50	75	12.5	125
	N = 50				$\Sigma f \lvert x - Me \rvert = 405$

Here, $\frac{N}{2}^{th}$ term = $\frac{50}{2}^{th}$ term = 25th term. From the table we find that this term is included in the class 60-70. Hence this class is the median class.

$$\therefore \text{ Median, } Me = L + \frac{\frac{N}{2} - fc}{f} \times h$$

$$= 60 + \frac{25 - 20}{20} \times 10$$

= 62.50 kg.

Thus, mean deviation about median

$$= \frac{1}{N} \Sigma f \lvert x - Me \rvert = \frac{1}{50} \times 405\, kg$$

= 8.10 kg.

Hence, coefficient of mean deviation about median

$$= \frac{M.D.\ about\ median}{Median} \times 100$$

$$= \frac{8.10}{62.50} \times 100$$

= 12.96

2.6 STANDARD DEVIATION

Although mean deviation is an improvement over range so far as a measure of dispersion is concerned, mean deviation is difficult to compute is further more, it cannot be treated mathematically. The best measure of dispersion is, usually, standard deviation which does not possess the demerits of range and mean deviation.

Standard Deviation for a given set of observations is defined as the root mean square deviation when the deviations are taken from the AM of the observations. If a variable x assumes n values x_1, x_2, x_3, ..., x_n then its standard diviation, generally denoted by σ, is given by

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n}\Sigma(x-\bar{x})^2}$$

$$= \sqrt{\frac{1}{n}\Sigma x^2 - \bar{x}^2} \qquad \text{(simplified form)}$$

For a grouped frequency distribution, the standard deviation is given by

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

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

$$= \sqrt{\frac{1}{N}\Sigma fx^2 - \bar{x}^2} \qquad \text{(simplified form)}$$

Assumed Mean Method

In case of individual series we have the following formula :

$$\sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}, \qquad \text{d = x–A where A = assumed mean}$$

In case of frequency distribution,

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

Step Deviation Method

If the deviations from assumed mean A 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 $d' = \frac{x - A}{h}$ where h is the H.C.F. of the deviations d = x = A. In this method, we have the following formula :

For individual series,

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

For frequency distributions,

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

where $d' = \frac{d}{h} = \frac{x - A}{h}$, A = assumed mean

Empirisal 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.

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

Or 6 Q. D. = 5 M. D. = 4 S. D.

Coefficient of Variation

The standard deviation is an absolute measure of dispersion and is expressed in the same units as the units of variable X. A relative measure of dispersion, based on standard deviation is known as coefficient of standard deviation or coefficient of variation and is given by

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

This measure introduced by karl pearson, is used to compare the

variability or homogeneity or stability or uniformity or consistency of two or more sets of data having different units of measurement. The data having a higher value of the coefficient of variation is said to be more dispersed or less uniform, etc.

Combined or Composite Standard Deviation

Let a distribution having n_1 observations has mean $\bar{x}_1$ and standard divition σ_1 and let another distribution having n_2 observations has mean $\bar{x}_2$ and standard deviation σ_2. Then the standard deviation σ of the distribution obtained by combining the two distributions having altogether $(n_1 + n_2)$ observations is given by

$$\sigma = \sqrt{\frac{n_1\sigma_1^{\,2} + n_2\sigma_2^{\,2} + n_1d_1^{\,2} + n_2d_2^{\,2}}{n_1 + n_2}}$$

Where $d_1 = \bar{x}_1 - \bar{x}, \quad d_2 = \bar{x}_2 = \bar{x}$

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

Variance : The square of standard deviation is called variance. Conversely, the positive square root of variance is standard deviation, since standard deviation is by definition non-negative. If standard deviation of a variable X in given as σ_x or σ, then the variance of X is given as

Var (x) or V (x) = $\sigma_x^{\,2}$ or σ^2

Properties of Standard Deviation

1. If all the observations assumed by a variable are constant, i.e. equal, then the S. D. is zero. This means that if all the values taken by a variable x is k, say, then S. D. of x i.e. $\sigma_x = 0$. This result applies to range as well as mean deviation.
2. Standard deviation is independent of change of origin but not of change of scale. This property implies that if a constant is added (or subtracted) to all the observations, the value of standard deviation remains unaffected. On the other hand, if all the observations are multiplied or divided by a constant, the standard deviation also gets multiplied (or divided) by this constant.
3. Standard deviation of a given set of observations is not less than mean

deviation from mean, i.e., Standard Deviation $\geq$ Mean Deviation from mean.

4. Standard deviation of a given set of observations is not greater than any other root mean square deviation, i.e. $\sqrt{\frac{1}{n}\Sigma(x-\bar{x})^2} \leq \sqrt{\frac{1}{n}\Sigma(x-A)^2}$

Example 7: Find the standard deviation of the following distribution :

Weights (kgs) :	50-52	52-54	54-56	56-58	58-60
No. of students :	17	35	28	15	5

Solution : Let us prepare the following table :

Weight (kgs)	No. of students (f)	Mid Value (x)	$d' = \frac{x-A}{h} = \frac{x-55}{2}$	fd'	fd'^2
50-52	17	51	–2	–34	68
52-54	35	53	–1	–35	35
54-56	28	55	0	0	0
56-58	15	57	1	15	15
58-60	5	59	2	10	20
	$N = \Sigma f = 100$			$\Sigma fd' = -44$	$\Sigma fd'^2 = 138$

Applying step deviation method we have,

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

$$= \sqrt{\frac{138}{100} - \left(\frac{-44}{100}\right)^2} \times 2 \text{ kgs.}$$

$$= \sqrt{1.38 - 0.1936} \times 2 \text{ kgs.}$$

= 2.18 kgs.

Example 8: 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 students :	3	61	132	153	140	51	2

Solution : Let us prepare the following table :

Age	No. of students (f)	Mid Value (x)	$d' = \frac{x-A}{h} = \frac{x-55}{10}$	fd'	fd'^2
20-30	3	25	–3	–9	27
30-40	61	35	–2	–122	244
40-50	132	45	–1	–132	132
50-60	153	55	0	0	0
60-70	140	65	1	140	140
70-80	51	75	2	102	204
80-90	2	85	3	6	18
	$N = \Sigma f = 542$			$\Sigma fd' = -15$	$\Sigma fd'^2 = 765$

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

$$= \sqrt{\frac{765}{542} - \left(\frac{-15}{542}\right)^2} \times 10$$

$$= \sqrt{1.4114 - 0.0009} \times 10$$

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

= 11.876 years

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

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

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

= 55 – 0.277 = 54.723 years

$$\therefore C.V. = \frac{11.876}{54.723} \times 100 = 21.4$$

Example 9: For a group of 60 boy students, the mean and SD of statistics marks are 45 and 2 respectively. The same figures for a group of 40 girl students are 55 and 3 respectively. What is the mean and SD of marks if the two groups are pooled together?

Solution : As given, $x_1 = 60,\ \bar{x}_1 = 45,\ \sigma_1 = 2$

$n_2 = 40,\ \bar{x}_2 = 55,\ \sigma_2 = 3$

Thus the combined mean is given by

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{60\times 45 + 40\times 55}{60 + 40}$$

= 49

Thus, $d_1 = \bar{x}_1 - \bar{x} = 45 - 49 = -4$

$d_2 = \bar{x}_2 - \bar{x} = 55 - 49 = 6$

$\therefore$ combined SD is given as

$$\sigma = \sqrt{\frac{n_1 s_1^{\,2} + n_2 s_2^{\,2} + n_1 d_1^{\,2} + n_2 d_2^{\,2}}{n_1 + n_2}}$$

$$= \sqrt{\frac{60\times 2^2 + 40\times 3^2 + 60\times(-4)^2 + 40\times 6^2}{60 + 40}}$$

$= \sqrt{30}$ = 5.48

Example 10: Two teams of football players A and B gave goals as shown below. Determine with the help of C. V., the team that is more consistent.

No. of Goals	No. of Games	
	A	B
0	27	17
1	9	9
2	8	6
3	5	5
4	4	3

Solution : Let us prepare the following table :

No. of goals(x)	d= x -A = x –2	d²	A			B		
			f_1	f_1d	f_1d^2	f_2	f_2d	f_2d^2
0	–2	4	27	–54	108	17	–34	68
1	–1	1	9	–9	9	9	–9	9
2	0	0	8	8	0	6	0	0
3	1	1	5	5	5	5	5	5
4	2	4	4	4	16	3	6	12
			$N_2 = 53$	$\Sigma f_1 d = 50$	$\Sigma f_1 d^2 = 138$	$N_2 = 40$	$\Sigma f_2 d = -32$	$\Sigma f_2 d^2 = 94$

Now,

$$\bar{x}_A = A + \frac{\Sigma f_1 d}{N_1} \qquad \bar{x}_B = A + \frac{\Sigma f_2 d}{N_2}$$

$$= 2 + \frac{-50}{53} = 1.1 \qquad = 2 + \frac{-32}{42} = 1.2$$

$$\sigma_A = \sqrt{\frac{\Sigma f_1 d^2}{N_1} - \left(\frac{\Sigma f_1 d}{N_1}\right)^2} \qquad \sigma_B = \sqrt{\frac{\Sigma f_2 d^2}{N_2} - \left(\frac{\Sigma f_2 d}{N_2}\right)^2}$$

$$= \sqrt{\frac{138}{58} - \left(\frac{-50}{53}\right)^2} \qquad = \sqrt{\frac{94}{40} - \left(\frac{-32}{40}\right)^2}$$

$$= \sqrt{2.60 - 0.81} \qquad = \sqrt{2.35 - 0.64}$$

$$= \sqrt{1.79} \qquad = \sqrt{1.71}$$

= 1.34 = 1.31

$$C.V.(A) = \frac{\sigma_A}{\bar{x}_A} \times 100 \qquad C.V.(B) = \frac{\sigma_B}{\bar{x}_B} \times 100$$

$$= \frac{1.34}{1.1} \times 100 \qquad = \frac{1.31}{1.2} \times 100$$

= 122 = 109

We have seen that C.V. (B) < C.V. (A). This implies that team A is more consistent than team B.

Example 11: Prove that for the first n natural numbers, standard deviation (SD) is $\sqrt{\frac{n^2 - 1}{12}}$

Solution : For the first n natural numbers, AM is given by,

$$\bar{x} = \frac{1 + 2 + 3 + ... + n}{n}$$

$$= \frac{1}{n}\left[\frac{n(n+1)}{2}\right] = \frac{n+1}{2}$$

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

$$= \sqrt{\frac{1}{n}\sum x^2 - \bar{x}^2}$$

$$= \sqrt{\frac{1^2+2^2+3^2+....+n^2}{n} - \left(\frac{n+1}{2}\right)^2}$$

$$= \sqrt{\frac{n(n+1)(2n+1)}{6n} - \frac{(n+1)^2}{4}}$$

$$= \sqrt{\frac{(n+1)(4n+2-3n-3)}{12}}$$

$$= \sqrt{\frac{(n+1)(n-1)}{12}}$$

$$= \sqrt{\frac{n^2-1}{12}}$$

Example 12: If AM and coefficient of variation of x are 10 and 40 respectively, what is the variance of (15-2x)?

Solution : Let, y=15-2x

$\therefore \sigma_y = 2\sigma_x$, $\because$ SD is independent of change of origin.

Again, we are given,

C.V.=40 and $\bar{x} = 10$

Thus, C.V. $= \frac{\sigma_x}{x} \times 100$

$$\Rightarrow 40 = \frac{\sigma_x}{10} \times 100$$

$$\Rightarrow \sigma_x = 4$$

$\therefore \sigma_y = 2\sigma_x = 2 \times 4 = 8$

$\therefore$ Variance of $(15-2x)$ $= \{SD \ \ of(15-2x)\}^2$

$= \{SD \ of \ \ y\}^2$

$= \{\sigma_y\}^2$

$= 8^2$

= 64



CHECK YOUR PROGRESS

Q 2: Fill in the blanks :

(i) Standard deviation is always a non-negative number.

(ii) Variance indicates the average distance of any observation in the data set from the mean.

(iii) Standard deviation can be calculated by taking deviation from any measure of central tendency.

(iv) Mean deviation is menimum when deviations are taken from median.

2.7 COEFFICIENT OF VARIATION

Let the mean and the standard deviation of a distribution be $\bar{x}$ and σ respectively. The coefficient of variation (CV) is defined by

$$CV = \frac{\sigma}{\bar{x}} \times 100, \quad \bar{x} \neq 0$$

This measure was introduced by statistician karl pearson.

Note : 1. CV is a unitless measure of dispersion.

2. CV is the best measure to compare the variability of two distributions.

3. A distribution with less CV is considered more consistent (less variable).

4. A distribution with more CV is considered less consistent (more variable).

Example 13: The mean and standard deviation of heights of adults (age 30 years), and children (age 10 months) are given below -

	Adults	Children
Mean	160 cm	60 cm
SD	10 cm	5 cm

Show that there is a greater variation in the heights of children than that of adults.

Solution :

$$\text{CV (adults)} = \frac{\sigma}{\bar{x}} \times 100 = \frac{10}{160} \times 100 = 6.25$$

$$\text{CV (children)} \ \frac{\sigma}{\bar{x}} \times 100 = \frac{5}{60} \times 100 = 8.33$$

Now CV (children) > CV (adults)

$\therefore$ there is greater variation in the heights of children than that of adults.



CHECK YOUR PROGRESS

Q 3: The mean blood pressure of 25 boys is 120 mm Hg, and SD is 10 mm Hg. The mean height and SD of the same set of boys is 160 cm and 5 cm respectively. SHow that blood pressure is more variable than height.

2.8 MULTIPLE CHOICE QUESTIONS

Q 1: Which of the following statements is correct?

(a) Two distributions may have identical measures of central tendency and dispersion.

(b) Two distributions may have the identical measures of central tendency but different measures of dispersion.

(c) Two distributions may have the different measures of central tendency but identical measures of dispersion.

(d) All the statements (a), (b) and (c)

[Ans. (d)]

Q 2: If the mean diviation of a distribution is 6.4, the standard deviation is:

(a) 6.8 (b) 8.0

(c) 6.4 (d) None of the above

[Ans. (b)]

Q 3: The appropriate measure of dispersions for open end classifisation is

(a) Standard Deviation (b) Mean Deviation

(c) Quartile Deviation (d) All these measures

[Ans. (c)]

Q 4: A shift of origin has no impact on

(a) Range (b) Mean Deviation

(c) Standard Deviation (d) All these and quartile deviation

[Ans. (d)]

Q 5: The standard deviation of 10, 16, 10, 16, 10, 10, 16, 16 is

(a) 4 (b) 6 (c) 3 (d) 0

(Ans. 3)

Q 6: For any two numbers standard deviation is always

(a) Taric the range (b) Half of the range

(c) Square of the range (d) None of these

[Ans. (b)]

Q 7: If each observation of a series is divided by 5, the S. D. of the new observations is

(a) $\frac{1}{10}^{th}$ of S. D. of original observations

(b) 5 times of S. D. of original observations

(c) not changed

(d) $\frac{1}{5}^{th}$ of S. D. of original observations

[Ans. (d)]

Q 8: Which of the following is a unitless measure of dispersion?

(a) Standard Deviation (b) Mean Deviation

(c) Coefficient of Variation (d) Range

[Ans. (c)]

2.9 EXERCISES

Q 1: The mean and standard deviation of 10 observations are 35 and 2 respectively. Find the changed mean and standard deviation if,

(a) each observation is increased by 4 and

(b) each observation is multiplied by 5

[Ans. (a) 39, 2, (b) 175, 10]

Q 2: Determine mean deviation from median and standard deviation from the following distribution :

X :	10	11	12	13	14
f :	3	12	11	12	3

(Ans. 0.75, 1)

Q 3: Find the coefficient of variation for the following distribution of wages.

Daily wages (Rs.) :	30-40	40-50	50-60	60-70	70-80	80-90
No. of workers :	17	28	21	15	13	6

(Ans. 26.93)

Q 4: Determine quantile deviation and its coefficient from the following distribution.

Marks :	6	7	8	9	10	11	12
No. of students :	3	6	9	13	8	5	4

(Ans. 1, 0.11)

Q 5: Which of the following companies A and B is more consistent so far as the payment of dividend are concerned ?

Dividend paid by A :	5	9	6	12	15	10	8	10
Dividend paid by B :	4	8	7	15	18	9	6	6

(Ans. Company A)



2.10 LET US SUM UP

- A measure of central tendency summarises the distribution of a variable into a single figure which can be regarded as its representative.
- The concept of dispersion is related to the extent of scatter or variability in observations. In short, dispersion for a given set of observations may be defined as the amount of deviation of the observations, usually, from an appropriate measure of central tendency.
- Measures of dispersion may be broadly classified into
  1. Absolute measures of dispersion
  2. Relative measures of dispersion
- Absolute measures of dispersion are classified into
  (i) Range (ii) Quartile Deviation (iii) Mean Deviation
  (iv) Standard Deviation.
- Various relative measures of dispersion are Coefficient of Range, Coefficient of Quartile Deviation, Coefficient of Mean Deviation, Coefficient of Variation.
- For a given set of observations, range may be defined as a difference between the largest and smallest observation.
- Half of the deviation between the third quartile (Q_3) and the first quartile (Q_1) of a distribution is called the quartile deviation (Q. D.) of the distribution.
- For a given set of observation, mean deviation is defined as the arithmetic mean of the absolute deviation of the observations from an appropriate

measure of central tendency.

- Standard Deviation for a given set of observations is defined as the root mean square deviation when the deviations are taken from the AM of the observations.
- The square of standard deviation is called variance.
- 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.
- 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.



2.11 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.



2.12 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: (i) extreme (ii) Q.D. (iii) central

Ans to Q No 2: (i) True (ii) True (iii) False (iv) True

Ans to Q No 3: CV (blood pressure) $= \frac{\sigma}{\bar{x}} \times 100 = \frac{10}{120} \times 100 = 8.3$ (approx)

CV (height) $= \frac{\sigma}{\bar{x}} \times 100 = 3.125$

$= \frac{5}{100} \times 100 = 3.125$

now CV (blood pressure) > CV (height)

$\therefore$ blood pressure is more variable than height.



2.15 MODEL QUESTIONS

A. SHORT TYPE QUESTIONS

Q 1: Which measures of dispersion is not affected by the presence of extreme observations ?

(a) Range (b) Mean Deviation

(c) Standard Deviation (d) Quartile Deviation

Q 2: If the mean of a series is 10 and its coefficient of variation is 40 percent, the variance of the series is :

(a) 4 (b) 8

(c) 12 (d) 16

Q 3: Define coefficient of variation (C. V.). Write a short note on this measure.

B. LONG QUESTIONS

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

Q 2: The following distribution is the distribution of ages of 100 students selected at random from a college consisting of 3500 students. Find the standard distribution for this distribution.

Age (in yrs.) :	16-17	17-18	18-19	19-20	20-21	21-22	22-33
No. of students. :	4	14	18	28	20	12	4

(Ans. 1.47 yrs.)

Q 3: Determine the mean deviation from the mean for the following distribution :

Class Interval :	12-15	15-18	18-21	21-24	24-27	27-30
Frequency :	2	4	5	6	3	1

(Ans. 3.33)

Q 4: Coefficients of variation of two sets of observations are 58 and 69. Their standard deviations are 21.2 and 15.6 respectively. FInd their arithmetic means.

*** ***** ***

UNIT 3 : MEASUREMENT OF ECONOMIC INEQUALITY

UNIT STRUCTURE

3.1 Learning Objectives

3.2 Introduction

3.3 Pareto's Law of Income Distribution

3.4 Log normal Distribution (Concept only)

3.5 Lorenz Curve

3.6 Coefficient of Inequality : Gini Coefficient

3.7 Let Us Sum Up

3.8 Further Reading

3.9 Answers to Check Your Progress

3.11 Model Questions

3.1 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.

3.2 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.

3.3 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:

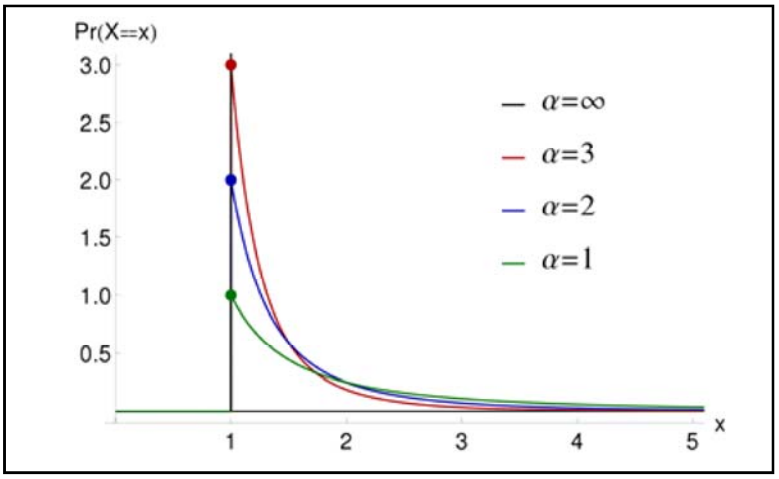
$$F(x) = 1 – (k/x)\alpha$$

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 3.1:

Figure 3.1: 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.

3.4 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.

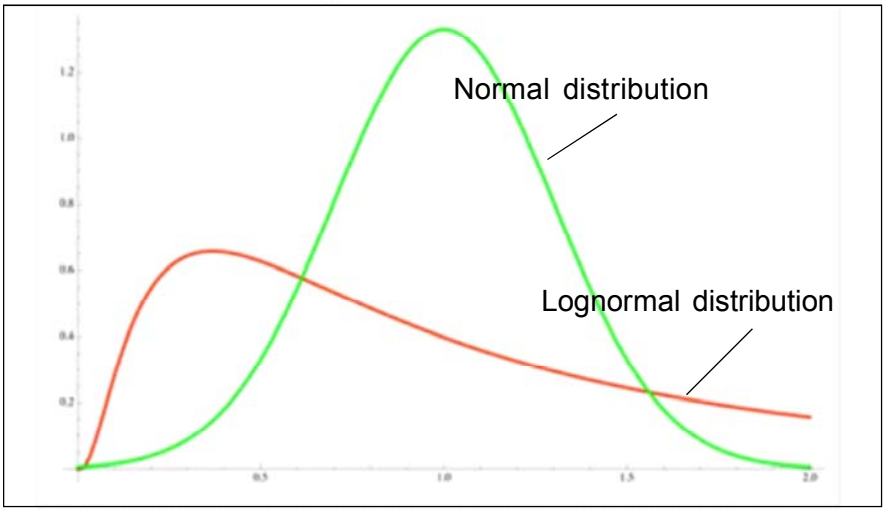
Overall the log-normal distribution plots the log of random variables from a normal distribution curve. In general, the log is known as the exponent to which a base number must be raised in order to produce the random variable (x) that is found along a normally distributed curve.

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 3.2.

Figure 3.2: Normal vs Lognormal Distribution: Graphical Shapes



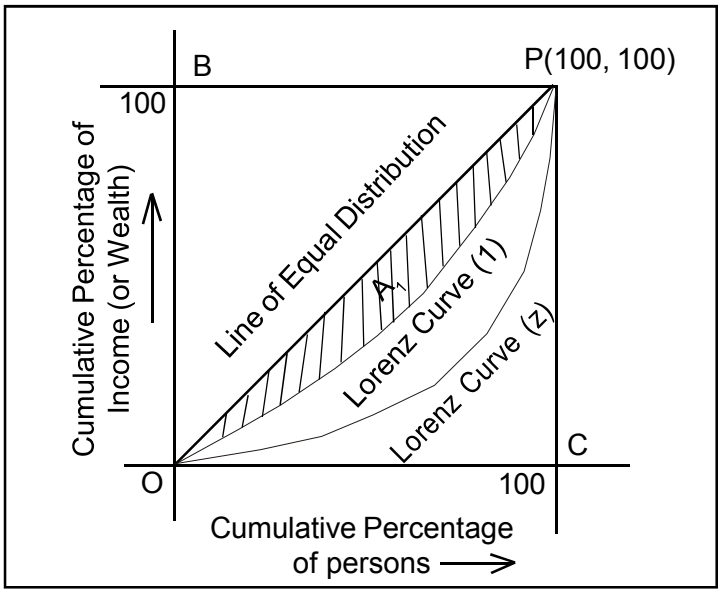
3.5 LORENZ CURVE

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 3.3: Lorenz Curve



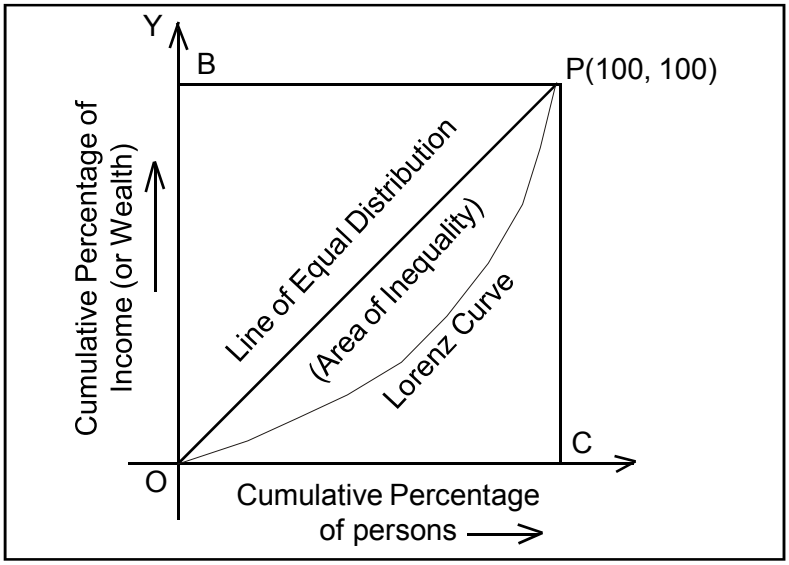
3.6 COEFFICIENT OF INEQUALITY : GINI COEFFICIENT

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 3.4.

Figure 3.4: 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

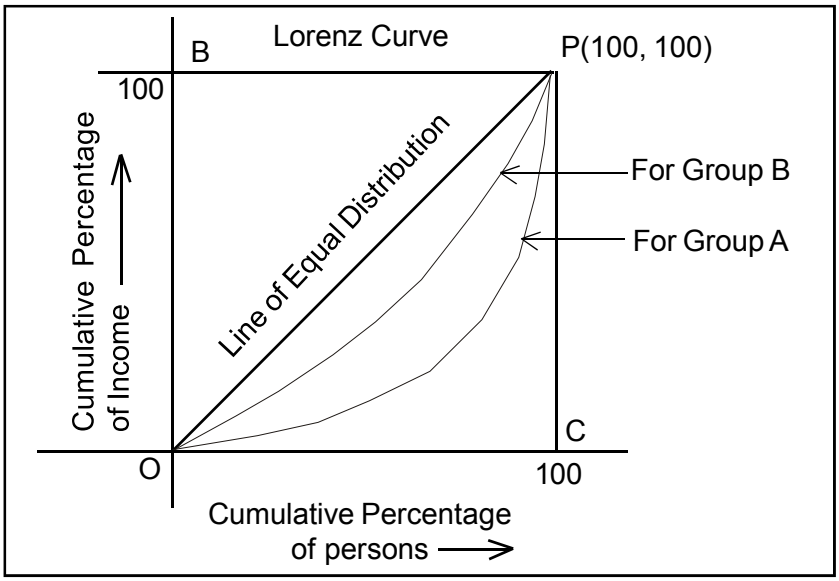
Solution : Let us prepare the following table :

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 3.5: 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).

.....

.....

.....



3.7 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.



3.8 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.



3.9 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.



3.10 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.

*** ***** ***

UNIT 4 : CORRELATION

UNIT STRUCTURE

4.1 Learning Objectives

4.2 Introduction

4.3 Correlation Analysis

4.4 Types of Correlation

4.5 Methods of Measuring Correlation

4.6 Properties of Correlation Coefficient

4.7 Rank Correlation Coefficient

4.8 Tied or Repeated Ranks

4.9 Multiple Choice Questions

4.10 Exercises

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-

- put forward the meanings of bivariate data
- state the concept of correlation between two variables and quantitative measurement of correlation including the interpretation of positive, negative and zero correlation
- use the technique of finding correlation between two quatitative characteristics and its interpretation.

4.2 INTRODUCTION

In the previous chapters, we discussed many statistical measures relating to univariate distribution, i.e. distribution of one variable like height, weight, mark, profit, wages and so on. However, there are situations that demand study of more than one variable simultaneously. A distribution of

paired values of two variables is called a bivariate distribution and a distribution of the values of more than two variables is called a multivariate distribution. If x and y denote marks in Mathematics and Statistics for a group of 30 students, then the corresponding bivariate data would be (x_i, y_i) for i = 1, 2, . . ., 30 where (x_i, y_i) denotes the marks in Mathematics and Statistics for the student with serial number i and so on.

If two variables are such that when one changes the other also changes then the two variables are said to be correlated. In the practical field we have to study the type of relationship that might exist between the ages of husbands and wives, the heights of fathers and sons, the volume of production of a particular crop, the price of a commodity and the demand for it, and so on. When we study the relationship between two variables only then it is called simple correlation or correlation.

4.3 CORRELATION ANALYSIS

Which studying two variables at the same time, if it is found that the change in one variable is reciprocated by a corresponding change in the other variable either directly or inversely, then the two variables are known to be associated or correlated. Otherwise, the two variables are known to be dissociated or uncorrelated or independent. For example, if the price of a commodity falls, the demand for it rises; if price rises, demand of that commodity falls. Thus there is correlation between the two variables price and demand.

Causation and Correlation

Causation or causality, is the capacity of one variable to influence another. The first variable may bring the second into existence or may cause the incidence of the second variable to fluctuate.

Causation is often confused with correlation, which indicates the extent to which two variables tend to increase or decrease in parallel. However, correlation by itself does not imply causation. There may be a third factor, for example, that is responsible for the fluctuations in both variables.

A statistically significant correlation has been reported, for example, between yellow cars and a lower incidence of accidents. That does not

indicate that yellow cars are safer, but just that fewer yellow cars are involved in accidents. A third factor, such as the personality type of the purchaser, is more likely to be responsible.

4.4 TYPES OF CORRELATION

Correlation for two variables may be classified into the following three types :

(i) Positive correlation,

(ii) Negative correlation and

(iii) Zero correlation

(i) Positive correlation : If two related variables move in the same derection i.e. an increase (or decrease) on the part of one variable results an increase (or decrease) on the part of the other variable, then the two variables are said to be positively correlated. As for example, height and weight yield, profit and investment etc. are positively correlated.

When the changes in two related variables are exactly proportional and are in the same direction then we say that there is perfect positive correlation between them. For example, when temparature of a gas increases, the volume of the gas also increases at the same rate. In this case temparature and volume are perfectly positively correlated.

(ii) Negative correlation : If two related variables move in the opposite directions i.e. an increase (or a decrease) on the part of one variable results a decrease (or an increase) on the part of the other variable, then the two variable are known to have a negative carrelation. For example if price of a certain commodity increases(decreases) then in normal situation, demand for that commodity decreases(increases). Therefore, price and demand of a commodity is negatively correlated.

(iii) Zero correlation : Two variables are said to have zero correlation or no correlation between them if they tend to change with no connection to each other. In this case the variables are said to be uncorrelated. As for example, shoe size and intelligence are said to have zero correlation.

4.5 METHODS OF MEASURING CORRELATION

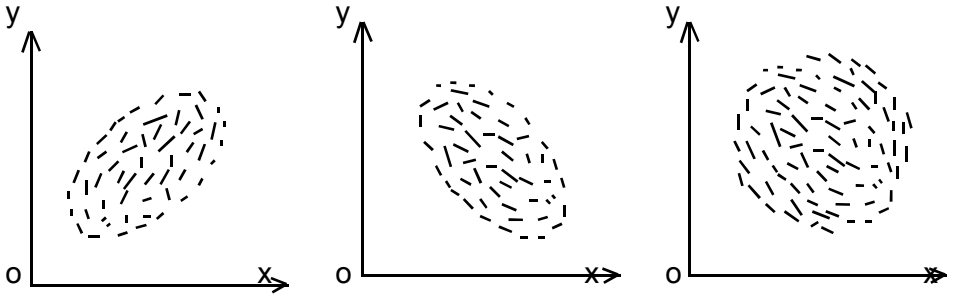
There are basically two methods of measuring correlation between two variables, viz

(i) Diagramatic method and

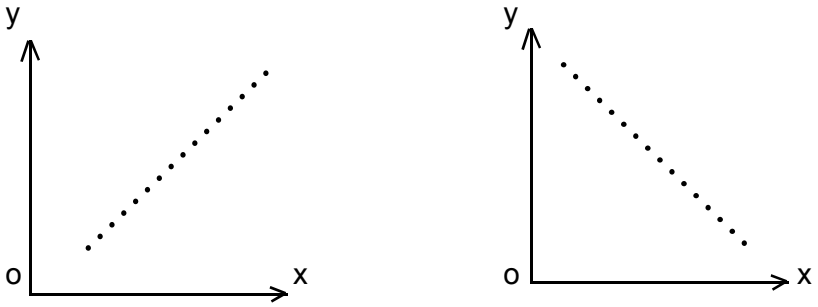
(ii) Mathematical method

(i) Diagramatic method : The diagramatic method is also known as the scatter diagram method. In this method, we can have an idea about the extent of association between two variables graphically.

Yet the bivariate data be denoted by (x_i, y_i), where i = 1, 2, . . ., n. If these paired values are plotted on a graph paper by taking the values of x on the x -axis and the values of y on the y -axis, then the diagram obtained on the graph paper is called a scatter diagram.



(a) Positive correlation (b) Negative correlation (c) Zero correlation



(d) Perfect positive correlation (e) Perfect Negative correlation

If the plotted points lie from lower left corner to upper right corner (as shown in fig. (a)) then it indicates that there is positive correlation between the two variables.

If the plotted point concentrate from upper left to lower right (as shown in fig. (b)) then it reveals that there is negative correlation between the two variables.

If the plotted points fall haphazardly without depicting any particular pattern (as shown in fig. (c)) then it indicates that there is no correlation or zero correlation between the two variables.

If all the points in the scatter diagram fall on a straight line with positive slope (as shown in fig. (d)) then it implies that there is perfect positive correlation between the two variables.

Again, if all the points fall on a straight line with negative slope (as shown in fig. (e)) then it indicates that there is perfect negative correlation between the two variables.

(ii) Mathematical Method : The following are mathematical methods of measuring correlation between two variables :

(a) Karl Pearson's coefficient of correlation

(b) Spearman's rank correlation coefficient

(a) Karl Pearson's Coefficient of Correlation : Karl Pearson's coefficient of correlation is a statistical tool of measuring degree linear relationship between two variables numerically. The coefficient may be defined as the ratio of covariance between the two variables to the product of the standard deviations of the two variables. If the two variables are denoted by x and y and if the corresponding bivariate data are (x_i, y_i), for i = 1, 2, 3, . . ., n, then the coefficient of correlation between x and y, due to Karl Pearson, is given by :

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

Where $Cov(x, y)$ = $\frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y})$

$= \frac{1}{n}\Sigma xy - \bar{x}\,\bar{y}$ (after simplification)

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

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

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

$$= \sqrt{\frac{1}{n}\Sigma y^2 - \bar{y}^2} \qquad \text{(after simplification)}$$

Thus, the simplified form of the coefficient of correlation is given as:

$$r_{xy} = \frac{\frac{1}{n}\Sigma xy - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\Sigma x^2 - \bar{x}^2}\sqrt{\frac{1}{n}\Sigma y^2 - \bar{y}^2}}$$

$$= \frac{\frac{1}{n}\Sigma xy - \frac{\Sigma x}{n}.\frac{\Sigma y}{n}}{\sqrt{\frac{1}{n}\Sigma x^2 - \left(\frac{\Sigma x}{n}\right)^2}\sqrt{\frac{1}{n}\Sigma y^2 - \left(\frac{\Sigma y}{n}\right)^2}}$$

$$= \frac{\frac{1}{n}\left[\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}\right]}{\sqrt{\frac{1}{n}\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right].\frac{1}{n}\left[\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right]}}$$

$$= \frac{\frac{1}{n}\left[\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}\right]}{\sqrt{\left(\frac{1}{n}\right)^2\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right]\left[\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right]}}$$

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

Assumed Mean Method

When deviations are taken from assumed means, i.e. if we assume $d_x = x - A, d_y = y - B$, A and B are assumed means of x and y respectively, then the following formula is used.

$$r = \frac{\Sigma d_x d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{n}}{\sqrt{\left[\Sigma {d_x}^2 - \frac{(\Sigma d_x)^2}{n}\right]\left[\Sigma {d_y}^2 - \frac{(\Sigma d_y)^2}{n}\right]}}$$

4.6 PROPERTIES OF CORRELATION COEFFICIENT

1. The coefficient of correlation is a unit-free measure.

   This means that if x denotes height of a group of students expressed in cm and y denotes their weight expressed in kg, then the correlation coefficient between height and weight would be free from any unit.

2. The coefficient of correlation remains invariant under a change of origin and / or scale of the variables under consideration.

   This property states that if the original pair of variables x and y is changed to a new pair of variables u and v by effecting a change of origin and scale for both x and y i.e.

   $u=\frac{x-a}{b}$ and $v=\frac{y-c}{d}$

   Where a and c are the origins of x and y and b and d are the respective scales, then we have $r_{xy}=\frac{bd}{/b//d/}r_{uv}$

3. The coefficient of correlation always lies between -1 and 1, including both the limiting values i.e. $-1\le r\le 1$.

Example 1 : Compute the correlation coefficient between x and y from the following data :

$$n=10,\ \Sigma xy=220,\ \Sigma x^2-200,\ \Sigma y^2-262,\ \Sigma x=40,\ \Sigma y=50$$

Solution : We know that,

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

$$=\frac{220-\frac{40\times 50}{10}}{\sqrt{\left[200-\frac{40^2}{10}\right]\left[262-\frac{50^2}{10}\right]}}$$

$$=\frac{220-200}{\sqrt{(200-160)(262-250)}}$$

$$= \frac{20}{\sqrt{40 \times 12}}$$

= 0.91 (approx)

Example 2: Find Karl Pearsons correlation coefficient from the following information :

X :	2	3	5	5	6	8
Y :	9	8	8	6	5	3

Solution : Let us prepare the following table :

x	y	xy	x^2	y^2
2	9	18	4	81
3	8	24	9	64
5	8	40	25	64
5	6	30	25	36
6	5	30	36	25
8	3	24	64	9
$\Sigma x = 29$	$\Sigma y = 39$	$\Sigma xy = 166$	$\Sigma x^2 = 163$	$\Sigma y^2 = 279$

Now,

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

$$= \frac{166 - \frac{29 \times 39}{6}}{\sqrt{\left[163 - \frac{29^2}{6}\right]}\sqrt{\left[279 - \frac{39^2}{6}\right]}}$$

= 0.93 (approx) (after simplification)

Example 3 : Find out the coefficient of correlation between the per capita income and the price level from the following data :

Year :	1980	1981	1982	1983	1984	1985
Per capita income :	360	420	500	556	600	590
Price index :	100	104	115	160	280	290

Solution : Let, x denotes per capita income and y denotes the price level or price index.

Now, we prepare the following table :

x	y	d_x = x-A = x-500	d_y = y-B = y-160	dx^2	dy^2	dx.dy
360	100	–140	–60	19600	3600	8400
420	104	–80	–56	6400	3136	4480
500	115	0	–45	0	2025	0
556	160	56	0	3136	0	0
600	280	100	120	10000	14400	12000
590	290	90	130	8100	16900	11700
		$\Sigma dx = 26$	$\Sigma dy = 89$	$\Sigma dx^2 = 47236$		$\Sigma d_x d_y = 36580$

Now using formula for assumed mean method we have,

$$r_{xy} = \frac{\Sigma d_x d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{n}}{\sqrt{\left\{\Sigma d^2{}_x - \frac{(\Sigma d_x)^2}{n}\right\}}\sqrt{\left\{\Sigma d^2{}_y - \frac{(\Sigma d_y)^2}{n}\right\}}}$$

$$= \frac{36580 - \frac{26 \times 89}{6}}{\sqrt{\left\{47236 - \frac{26^2}{6}\right\}}\sqrt{\left\{40061 - \frac{89^2}{6}\right\}}}$$

$$= \frac{36580 - 385.67}{\sqrt{47123.33 \times 38740.83}}$$

$$= \frac{36194.33}{42727.87}$$

= 0.847 (approx)

Example 4 : Determine the value of the correlation coefficient r from the following data :

$$\sum_{i=1}^{8}(x_i - \bar{x})^2 = 32, \qquad \sum_{i=1}^{8}(y_i - \bar{y})^2 = 72$$

$$\sum_{i=1}^{8}(x_i - \bar{x})(y_i - \bar{y}) = 43$$

Solution : Now, $Cov(x,y) = \frac{1}{n}\sum_{i=1}^{8}(x_i - \bar{x})(y_i - \bar{y})$

$$= \frac{1}{8}\sum_{i=1}^{8}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{8} \times 43 \quad = 5.375$$

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

$$= \sqrt{\frac{1}{8}\sum_{i=1}^{8}(x_i - \bar{x})^2}$$

$$= \sqrt{\frac{1}{8} \times 32}$$

$$= \sqrt{4} \qquad = 2$$

$$\sigma y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= \sqrt{\frac{1}{8}\sum_{i=1}^{8}(y_i - \bar{y})^2}$$

$$= \sqrt{\frac{1}{8} \times 72}$$

$$= \sqrt{9} \qquad = 3$$

$$\therefore r = \frac{Cov(x,y)}{\sigma x . \sigma y}$$

$$= \frac{5.375}{2 \times 3}$$

$$= \frac{5.375}{6}$$

= 0.8958

Exmple 5: Calculate the Karl Pearson's Coefficient of correlation from the following data :

(i) Sum of deviations of x values = 5

(ii) Sum of deviations of y values = 4

(iii) Sum of squares of deviations of x values = 40

(iv) Sum of squares of deviations of y values = 50

(v) Sum of the product of deviations of x and y values = 32

(vi) No. of pairs of observations = 10

Solution : Let, d_x=X–A and d_y=Y–b be the deviations of X and Y values.

We are given, $\sum d_x = 5,\ \sum d_y = 4$

$$\sum {d_x}^2 = 40,\ \sum {d_y}^2 = 50$$

$$\sum d_x \cdot d_y = 32 \text{ and } n = 10$$

Now, using formula for assumed mean method, we have,

$$r_{xy} = \frac{\sum d_x d_y - \frac{\left(\sum d_x\right)\left(\sum d_y\right)}{n}}{\sqrt{\left\{\sum {d_x}^2 - \frac{\left(\sum d_x\right)^2}{n}\right\}\left\{\sum {d_y}^2 - \frac{\left(\sum d_y\right)^2}{n}\right\}}}$$

$$= \frac{32 - \frac{5 \times 4}{10}}{\sqrt{\left\{40 - \frac{5^2}{10}\right\}\left\{50 - \frac{4^2}{10}\right\}}}$$

$= 0.704$ (after simplification)

Example 6: From the following data compute coefficient of correlation between X and Y.

	X Series	Y Series
Arithmetic Mean (AM)	25	18
Sum of Squares of deviations from AM	136	138

Summation of products of deviations of X and Y series from their respective means=122

Number of pairs of values = 15

Solution : We are given,

$$\bar{x} = 25,\ \bar{y} = 18$$

$$\sum \left(x - \bar{x}\right)^2 = 136,\ \sum \left(y - \bar{y}\right)^2 = 138$$

$$\sum \left(x - \bar{x}\right)\left(y - \bar{y}\right) = 122,\ n = 15$$

$$\therefore r_{xy} = \frac{Cov(x, y)}{\sigma_x \cdot \sigma_y}$$

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

$$= \frac{122}{\sqrt{136 \times 138}} = \frac{122}{137} = 0.89$$



CHECK YOUR PROGRESS

Q.1: Fill in the blanks :

(i) Correlation coefficient is _____ of the units of measurement.

(ii) Coefficient of correlation is a measure of the strength of the _____relationship between two variables.

(iii) Coefficient of correlation lies between _____ and _____

(iv) If two variables are independent they are_____.

4.7 RANK CORRELATION COEFFICIENT

Rank correlation coefficient is a method of usually measuring the relationship between two qualitative variable or attributes such as honesty, beauty, morality, etc. which cannot be measured numerically. In this method, various items are assigned ranks according to the two characteristics and a correlation is computed between these ranks. Charles Edward Spearman, a British psychologist developed the rank correlation coefficient formula and hence rank correlation coefficient is also known as spearman's rank correlation coefficient.

Spearman's rank correlation coefficient is denoted by R or Y and is given as $R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$

Where d is the difference between the pair of ranks of the same individual in the two characteristics and n is the number of pairs.

Note : The value of R always lees between – 1 and 1, i.e. $-1 \le R \le 1$.

Example 7 : Compute rank correlation from the following data relating to ranks given by two judges in a contest.

Serial No. of candidate :	1	2	3	4	5	6	7	8	9	10
Rank by Judge A :	10	5	6	1	2	3	4	7	8	9
Rank by Judge B :	5	6	9	2	8	7	3	4	10	1

Solution : Let R_1 denotes the ranks given by Judge A and R_2 denotes the ranks given by Judge B.

Now, let us prepare the following table :

Serial No.	R_1	R_2	$d = R_1 - R_2$	d^2
1	10	5	5	25
2	5	6	–1	1
3	6	9	–3	9
4	1	2	–1	1
5	2	8	–6	36
6	3	7	–4	16
7	4	3	1	1
8	7	4	3	9
9	8	10	–2	4
10	9	1	8	64
				$\Sigma d^2 = 166$

$\therefore$ The rank correlation coefficient is given by,

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 166}{10(10^2-1)}$$

= 0.006

Example 8 : Compute the coefficient of rank correlation between sales and advertisement expressed in thousands of rupees from the following data :

Sales :	90	85	68	75	82	80	95	70
Advertisement :	7	6	2	3	4	5	8	1

Solution : Let R_1 denotes ranks for sales and R_2 denotes ranks for advertisement.

Now, let us prepare the following table :

Sales	Advertisement	Ranks for Sales(R_1)	Ranks for Advertisement (R_2)	d = R_1–R_2	d^2
90	7	2	2	0	0
85	6	3	3	0	0
68	2	8	7	1	1
75	3	6	6	0	0
82	4	4	5	–1	1
80	5	5	4	1	1
95	8	1	1	0	0
70	1	7	8	–1	1
					$\Sigma d^2 = 4$

$\therefore$ Rank correlation coefficient,

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 4}{8(8^2-1)}$$

= 1 – 0.0476

= 0.95

Example 9: The following are the marks obtained by a group of students in two papers. Calculate the rank coefficient of correlation.

Economics :	78	36	98	25	75	82	92
Statistics :	84	51	91	69	68	62	86
Economics :	62	65	39				
Statistics :	58	35	49				

Solution : Let R_1 denotes the ranks of marks obtained in Economics and R_2 denotes the ranks of marks obtained in Statistics.

Now, let us prepare the following table :

Marks in Economics	Marks in Statistics	R_1	R_2	d=R_1–R_2	d^2
78	84	4	3	1	1
36	51	9	8	1	1
98	91	1	1	0	0

25	69	10	4	6	36
75	68	5	5	0	0
82	62	3	6	–3	9
92	86	2	2	0	0
62	58	7	7	0	0
65	35	6	10	–4	16
39	49	8	9	–1	1
					$\Sigma d^2 = 64$

$\therefore$ Rank correlation coefficient,

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 64}{10(10^2-1)}$$

= 1 – 0.388

= 0.612 (approx)

4.8 TIED OR REPEATED RANKS

In case of a tie, i.e., when two or more individuals have the same rank, each individual is assigned a rank equal to the mean of the ranks.

In the formula for tied rank we are to add the factor $\frac{m(m^2-1)}{12}$ to Σd^2 where m is the number of times an item is repeated. The correction factor is to be added for each repeated value in both the series. Thus spearmain's rank correlation coefficient formula for tied ranks becomes,

$$R = 1 - \frac{6\left[\Sigma d^2 + \Sigma \frac{m(m^2-1)}{12}\right]}{n(n^2-1)}$$

Example 10: Calculate the spearman's ranks correlation coefficient between the series A and B given below :

Series A :	57	59	62	63	64	65	55	58	57
Series B :	113	117	126	126	130	129	111	116	112

Solution : Let, R_1 denotes ranks of the series A and

R_2 denotes ranks of the series B

Now, let us prepare the following table :

Series A	Series B	Rank of Series A(R_1)	Rank of Series B(R_2)	d = R_1-R_2	d^2
57	113	2.5	3	–0.5	0.25
59	117	5	5	0	0
62	126	6	6.5	–0.5	0.25
63	126	7	6.5	0.5	0.25
64	130	8	9	–1	1.00
65	129	9	8	1	1.00
55	111	1	1	0	0
58	116	4	4	0	0
57	112	2.5	2	0.5	0.25
					$\Sigma d^2 = 3.00$

The observation 57 in series A is repeated twice and the observation 126 in series B is also repeated twice.

$$\therefore \Sigma\frac{m(m^2-1)}{12} = \frac{m_1(m_1^2-1)}{12} + \frac{m_2(m_2^2-1)}{12} \quad \text{(where } m_1 = 2,\ m_2 = 2\text{)}$$

$$= \frac{2(2^2-1)}{12} + \frac{2(2^2-1)}{12}$$

$$= \frac{4-1}{6} + \frac{4-1}{6}$$

$$= \frac{3}{6} + \frac{3}{6}$$

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

$\therefore$ Rank correlation coefficient

$$R = 1 - \frac{6\left[\Sigma d^2 + \Sigma\frac{m(m^2-1)}{12}\right]}{n(n^2-1)}$$

$$= 1 - \frac{6\left[3+1\right]}{9(9^2-1)}$$

$$= 1 - \frac{6 \times 4}{9 \times 80}$$

= 0.967



CHECK YOUR PROGRESS

Q.2: Fill in the blanks :

(i) Rank correlation coefficient lies between _____/

(ii) In rank correlation coefficient the association need not be_____.

(iii) The sum of the difference of rank is_____.

(iv) Great advantage of_____ is that it can be used to rank attributes which can not be expressed by way of numerical value.

4.9 MULTIPLE CHOICE QUESTIONS

Q 1: If r_{xy} = 0, the variables x and y are

(a) linearly related (b) independent

(c) not linearly related (d) none of the above

Ans : (c)

Q 2: If $r_{xy} = \pm 1$, the two lines of regression are

(a) coincident (b) paralled

(c) perpendicular to each other (d) none of the above

Ans : (a)

Q 3: For finding correlation between two attributes, we consider

(a) Pearson's correlation coefficient

(b) Scatter diagram

(c) Spearman's rank correlation coefficient

(d) All these

Ans : (c)

Q 4: If the relationship between two variables x and y is given by $2x + 3y + 4 = 0$, then the value of the correlation coefficient between x and y is

(a) 0 (b) 1

(c) –1 (d) negative

Ans : (c)

Q 5: The coefficient of correlation between two variables

(a) can have any unit

(b) is expressed as the product of units of the two variables

(c) is a unitfree measure

(d) none of these

Ans : (c)

Q 6: Scatter diagram is considered for measuring

(a) Linear relationship between two variables

(b) Curvitinear relationship between two variables

(c) Neither (a) nor (b)

(d) Both (a) and (b)

Ans : (d)

Q 7: The correlation between the speed of an automobile and the distance travelled by it after applying the brakes is

(a) Negative (b) Zero

(c) Positive (d) None of these

Ans : (a)

Q 8: Pearson's correlation coefficient is used for finding

(a) Correlation for any type of relation

(b) Correlation for linear relation only

(c) Correlation for curvilinear relation only

(d) Both (b) and (c)

Ans : (b)

Q 9: If the plotted points in a scatter diagram are evenly distributed, then the correlation is

(a) Zero (b) Negative

(c) Positive (d) (a) or (b)

Ans : (a)

Q 10: In rank correlation coefficient the association need not be linear

(a) true (b) false

(c) both (d) none

Ans : (a)

4.10 EXERCISES

Q 1: What is the value of correlation coefficient due to Pearson on the basis of the following data :

x :	–5	–4	–3	–2	–1	0	1	2	3	4	5
y :	27	18	11	6	3	2	3	6	11	18	27

Ans : Zero

Q 2: If for two variables x and y, the covariance, variance of x and variance of y are 40, 16 and 256 respectively, what is the value of the correlation coefficient ?

Ans : 0.625

Q 3: What is the coefficient of correlation from the following data ?

x :	1	2	3	4	5
y :	8	6	7	5	5

Ans : –0.85

Q 4: Given that for twenty pairs of observations,

$\Sigma xy = 525,\ \Sigma x = 129, \Sigma y = 97, \Sigma x^2 = 687$

$\Sigma y^2 = 427$ and u = 10 – 3 y . Find the coefficient of correlation between x and u .

Ans : –0.74

Q 5: Eight contestants in a musical contist were ranked by two judges A and B in the following manner :

Serial number of the contestants :	1	2	3	4	5	6	7	8
Rank by Judge A :	7	6	2	4	5	3	1	8
Rank by Judge B :	5	4	6	3	8	2	1	7

Find the rank correlation coefficient.

Ans : 0.57

Q 6: Find the value of Rank correlation coefficient between the following marks in Economics and Statistics.

Roll No. :	1	2	3	4	5	6
Marks in Economics :	25	30	46	30	55	80
Marks in Statistics :	30	25	50	40	50	78

Ans : 0.857

Q 7: Discuss the different types of correlation.

Q 8: Write short notes on :

(a) Rank correlation

(b) Scatter diagram

(c) Positive and Negative correlation.

Q 9: From the followings data, calculate Karl Pearson's coefficient of correlation :

(i) $\sum_{i=1}^{9}(x_i-\bar{x})^2=120$, $\sum_{i=1}^{9}(y_i-\bar{y})^2=346$, $\sum_{i=1}^{9}(x_i-\bar{x})(y_i-\bar{y})=193$

(ii) $\Sigma x=125,\ \Sigma y=80,\ n=10,\ \Sigma x^2=1586,\ \Sigma y^2=650,\ \Sigma xy=1007$

(iii) $n=10,\ \Sigma x=140,\ \Sigma y=150,\ \Sigma(x-14)^2=180$,

$\Sigma(y-15)^2=215,\ \Sigma(x-14)(y-15)=60$

Ans : (i) 0.95, (ii) 0.47, (iii) 0.92



4.11 LET US SUM UP

- If two variables are such that when one changes the other also changes then the two variables are said to be correlated.
- Correlation for two variables may be classified into the following three types (i) Positive correlation,
  (ii) Negative correlation and
  (iii) Zero correlation
- If two related variables move in the same derection i.e. an increase (or decrease) on the part of one variable results an increase (or decrease) on the part of the other variable, then the two variables are said to be positively correlated.
- If two related variables move in the opposite directions i.e. an increase (or a decrease) on the part of one variable results a decrease (or an

increase) on the part of the other variable, then the two variable are known to have a negative carrelation.

- Two variables are said to have zero correlation or no correlation between them if they tend to change with no connection to each other.
- There are basically two methods of measuring correlation between two variables, viz (i) Diagramatic method and
  (ii) Mathematical method
- The following are mathematical methods of measuring correlation between two variables :
  (a) Karl Pearson's coefficient of correlation
  (b) Spearman's rank correlation coefficient
- Karl Pearson's coefficient of correlation is a statistical tool of measuring degree linear relationship between two variables numerically. The coefficient may be defined as the ratio of covariance between the two variables to the product of the standard deviations of the two variables.
- Rank correlation coefficient is a method of usually measuring the relationship between two qualitative variable or attributes such as honesty, beauty, morality, etc. which cannot be measured numerically.
- Spearman's rank correlation coefficient is denoted by R or Y and is given as $R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$

  Where d is the difference between the pair of ranks of the same individual in the two characteristics and n is the number of pairs.
- In case of a tie, i.e., when two or more individuals have the same rank, each individual is assigned a rank equal to the mean of the ranks.



4.12 FURTHER READING

1) Gujrati, D (1995); Basic Econometrics (third edition), McGraw-Hill, Inc; Singapore
2) Gun, A M; Gupta, M K & Dasgupta, B (2002); Fundamentals of Statistics, Volume One; World Press Private Ltd. Kolkata

3) Gupta, S C (1994); Fundamentals of Statistics; Himalayan Publishing House, New Delhi.

4.13 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q. No. 1 : (i) independent (ii) linear
(iii) -1, 1 (iv) uncorrelated

Ans to Q. No. 2 : (i) -1 to +1 (ii) linear
(iii) zero (iv) rank correlation

4.14 MODEL QUESTIONS

A. SHORT TYPE QUESTIONS

Q 1: Give the concept of correlation.

Q 2: Distinguish between causation and correlation.

Q 3: What is meant by zero correlation?

Q 4: Write down the properties of correlation coefficcient.

B. LONG QUESTIONS

Q 1: Find the coefficient of correlation from the given data in each of the following cases.

(i) x :	6	8	15	18	20	24	28	31		
y :	10	12	15	18	25	22	26	28		
(ii) x :	39	65	62	90	82	75	25	98	36	78
y :	47	53	58	86	62	68	60	91	51	84
(iii) x :	9	8	7	6	5	4	3	2	1	
y :	15	16	14	12	10	13	11	8	9	

Ans : (i) 0.96, (ii) 0.78, (iii) 0.88

Q 2: Calculate rank correlation coefficient from the following data :

x :	75	88	95	70	60	80	81	50
y :	120	134	150	115	110	140	142	100

Ans : 0.92

Q 3: Discuss various methods of measuring correlation.

Q 4: The following table provides the distribution of items according to size groups and also the number of defectives :

Size group :	9-11	11-13	13-15	15-17	17-19
No. of items :	250	350	400	300	150
No. of defective items :	25	70	60	45	20

Find the correlation coefficient between size and defectives.

Ans : 0.07

*** ***** ***

UNIT 5 : REGRESSION ANALYSIS

UNIT STRUCTURE

5.1 Learning Objectives

5.2 Introduction

5.3 Regression

5.4 Two-variable Regression Model (Least Square Method)

5.5 Properties of Regression Coefficients

5.6 Correlation vs. Regression

5.7 Multiple Choice Questions

5.8 Exercises

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:

- put foward the concept of regression, regression equations and regression lines
- state the different kinds of regression and their applications
- utilise the technique of estimation of a variable from known set of data.

5.2 INTRODUCTION

If the coefficient of correlation calculated for bivariate data (X_i, Y_i), i = 1, 2, . . ., n, is reasonably high and a cause and effect type of relation is also believed to be existing between them, the next logical step is to obtain a functional relation between these variables. This functional relation is known as regression equation in Statistics.

The regression equations are useful for predicting the value of dependent variable for given value of the independent variable. The nature

of a regression equation is different from the nature of a mathematical equation, e.g., if Y = 10 + 2x is a mathematical equation then it implies that Y is exactly equal to 20 when X = 5. Howerver, if Y = 10 + 2x is a regression equation, then Y = 20 is an average or most likely or most probable value of Y when X = 5. The variable corresponding to whose value the average value of the other variable is estimated is called the independent or explanatory or exogenous variable. The variable whose average value is estimated is called the dependent or exphained or endogenous variable.

5.3 REGRESSION

Regression is the measure of the average relationship between two or more variables in terms of the original units of the data. In other words, regression is a technique of estimating the most probable or most likely or average values of one variable for specified values of the other variable on the basis of the given values of the two variables. This is done by estimating a relationship between the two variables.

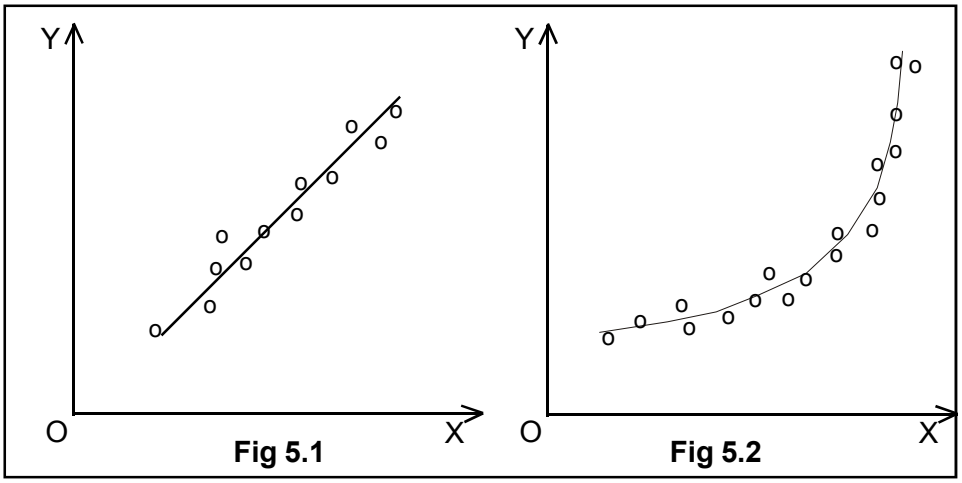
Regression Equations

The two equations which are estimated on the basis of the given values of the two variables to estimate the most likely values of one variable for specified values of the other are called regression equations. The geometrical representations of regression equations are called regression lines. The terms regression equations and regression lines are used interchangeably.

Linear and nonlinear or curvilinear regression

If the points of the scatter diagram of two correlated variables clusters around a straight line then we say that there is linear regreassion between the two variables. If the scatter diagram clusters around a curve then we say that these is nonlinear or curvilinear regression between the two variables. For example let x denotes the height of n persons and y denotes their weight. Yet the values of x be $x_1, x_2, ..., x_n$ and the corresponding values of y be $y_1, y_2, ..., y_n$. Now from these bivariate data we get the ordered pairs $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, each of which represents a point in the xy - plane. If these point cluster around a line as shown in figure 5.1 then it

indicates that there is a linear regression between height (x) and weight (y) and if the point cluster around a curve as shown in figure 5.2 then it indicates that there is non-linear or curvilinear regression between x and y .



5.4 TWO-VARIABLE REGRESSION MODEL (LEAST SQUARE MEHTOD)

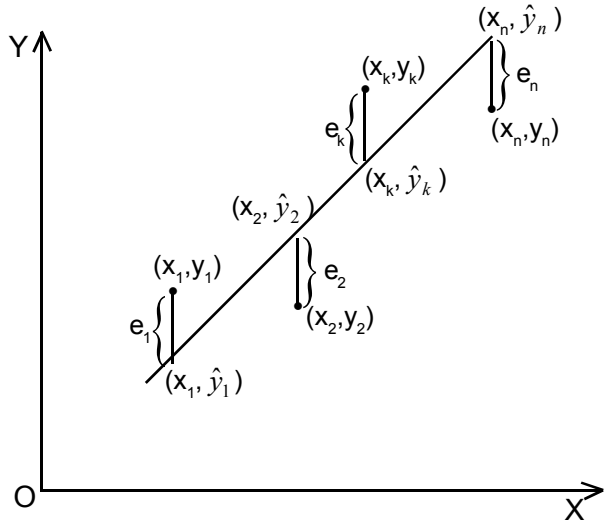
When only two variables are associated with regression analysis then we call it single regression. When we say regression it means simple regression. In case of a simple regression model if depends on x , then the regression model is termed as regression lines of y on x and is given by,

Y=a+bx -(1)

Here a and b are two constants and they are also known as regression parameters. Furthermore b is also known as the regression coefficient of y on x and is also denoted by b_{yx}. We may define the regression line of y on x as the line of best fit obtained by the method of least squares and used for estimating the value of the dependent variable y for a known value of the independent variable x .

Estimation of Regression Parameters

Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ be n ordered pains of corresponding values of the two correlated variables x and y . Let the line to be estimated on the basis of these given n pairs of values of x and y by applying the method of least squares is of the form :



Let the estimated values of y corresponding to the values $x_1, x_2, ..., x_n$ of X be $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ respectively. Again, the actual values of y corresponding to the values $x_1, x_2, ..., x_n$ are $y_1, y_2, ..., y_n$ respectively. Corresponding to $x = x_1$, the value of y as estimated by the line (1) is $\hat{y}_1 = a + bx_1$.

The estimated value of y corresponding to $x = x_2$ is $\hat{y}_2 = a + bx_2$. Similarly, the estimated value of y corresponding to $x = x_n$ is $\hat{y}_n = a + bx_n$. If we denote the errors of estimates by $e_1, e_2, ..., e_n$ respectively then we have,

$$e_1 = y_1 - \hat{y}_1 = y_1 - (a + bx_1) = y_1 - a - bx_1$$

$$e_2 = y_2 - \hat{y}_2 = y_2 - (a + bx_2) = y_2 - a - bx_2$$

Similarly, $e_n = y_n - \hat{y}_n = y_n - (a + bx_n) = y_n - a - bx_n$

For a point above the line, the error of estimate will positive and for a point below the line, the error of estimate will be negative. Now the values of the two parameters a and b of the line (1) are to be determined in such a way that the sum of the squares of the errors is the least.

i.e. $\sum_{i-1}^{n} e^2{}_i = \sum_{i-1}^{n} (y_i - a - bx_i)^2$ is the least.

or $\Sigma e^2 = \Sigma (y - a - bx)^2$ is the least.

By applying the principle of least squares of Differential Calculus we find that if we put the values of a and b obtained by solving the following two equations in $y = a + bx$ then Σe^2 will the least with respect to this line.

$$\Sigma y = na + b\Sigma x \qquad \text{– (2)}$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \qquad \text{– (3)}$$

Equations (2) and (3) are called normal equations. From the given values of x and y we get Σx, Σy, Σx^2 and Σxy. Putting these values in (2) and (3) and solving these two equations simerltaneously for a and b we get the values of a and b. Putting these values of a and b in (1) we get the regression equation of y on x.

Dividing both sides of (2) by n we get,

$\bar{y} = a + b\bar{x}$ – (4)

Now, subtracting (4) from (1) we get,

$y - \bar{y} = b(x - \bar{x})$ – (5)

Again, multiplying (2) by Σx and (3) by n we get

$(\Sigma x)(\Sigma y) = na(\Sigma x) + b(\Sigma x)^2$

and $n(\Sigma xy) = na(\Sigma x) + bn(\Sigma x^2)$

Subtracting, $(\Sigma x)(\Sigma y) - n(\Sigma xy) = b\left[(\Sigma x)^2 - n\Sigma x^2\right]$

or, $b = \dfrac{(\Sigma x)(\Sigma y) - n(\Sigma xy)}{(\Sigma x)^2 - n\Sigma x^2}$

or, $b = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$ –(6)

In equation (5), b is called the regression coefficient of y on x and it is denoted by b_{yx}. Thus, if we write b_{yx} for b in the equation (5).

We get, $y - \bar{y} = b_{yx}\left(x - \bar{x}\right)$ –(7)

Where $b_{yx} = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$ –(8)

Equation (7) is called the regression equation of y on x where b_{yx} is given by (8). b_{yx} can also be expressed as

$$b_{yx} = \frac{\Sigma xy - \dfrac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}} = \frac{\Sigma d_x d_y - \dfrac{(\Sigma d_x)(\Sigma d_y)}{n}}{\Sigma {d_x}^2 - \dfrac{(\Sigma d_x)^2}{n}}$$ (Assumed mean formula)

where $d_x = x - A$, A = assumed mean x

$d_y = y - B$, B = assumed mean y

There may be cases when the variable x depends on y and we may take the regression line of x on y as

$$x = a' + b'y \quad -(9)$$

Unlike the minimization of the sum of the squares of the errors for obtaining the estimates of a' and b' we get the following normal equations in a' and b' :

$$\Sigma x = na' + b'\Sigma y$$

$$\Sigma xy = a'\Sigma y + b'\Sigma y^2$$

By solving the above normal equations simultaniously for a' and b' and putting these values in (9) we get the regression equation of x on y as below :

$$x - \bar{x} = b'(y - \bar{y})$$

or $x - \bar{x} = b_{xy}(y - \bar{y})$ (we write b_{xy} for b')

Where $b_{xy} = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2}$

$$= \frac{\Sigma xy - \dfrac{(\Sigma x)(\Sigma y)}{n}}{\Sigma y^2 - \dfrac{(\Sigma y)^2}{n}} = \frac{\Sigma d_x \Sigma d_y - \dfrac{(\Sigma d_x)(\Sigma d_y)}{n}}{\Sigma {d_y}^2 - \dfrac{(\Sigma d_y)^2}{n}}$$ (Assumed mean formula)

Where $d_x = x - A$, A = assumed mean of x

$d_y = y - B$, B = assumed mean of y .

Note : (i) The two lines of regression must intersect at the point $(\bar{x}, \bar{y})$. That is why this point satifies both the regression lines.

(ii) The regression equation of x on y is used to obtain the average value of y corresponding to a given value of x. Similarly, the regression equation of y on x is used to obtain the average value of x corresponding to a given of y .

5.5 PROPERTIES OF REGRESSION COEFFICIENTS

Property 1 : The product of the two regression coefficients is equal to the square of the correlation coefficient.

or

The correlation coefficient is the geometric mean of the regression coefficients.

Proof : We know that,

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

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

$$= \frac{Cov(x,y)}{\sigma x^2} = \frac{Cov(x,y)}{\sigma x.\sigma y}.\frac{\sigma_y}{\sigma_x} = r = \frac{\sigma_y}{\sigma_x}$$

Similarly, we can have,

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

Now, $b_{yx}.b_{xy} = r\frac{\sigma_y}{\sigma_x}.r\frac{\sigma_x}{\sigma_y}$

$$\Rightarrow b_{yx}.b_{xy} = r^2$$

This shows that the product of the two regression coefficients is equal to the square of the correlation coefficient.

Again, from $r^2 = b_{yx}.b_{xy}$ we have,

$$r = \pm\sqrt{b_{yx}.b_{xy}}$$

Which shows that correlation coefficient r is the geometric mean of the two regression coefficients.

Property 2 : Both the regression coefficients must have the same sign.

Proof : We know that,

$$b_{yx} = \frac{Cov(x,\ y)}{\sigma x^2}$$

and $b_{xy} = \frac{Cov(x,\ y)}{\sigma y^2}$

In these two expressions, we must get $\sigma x^2 > 0$ and $\sigma y^2 > 0$ as squared quantities can't be negative. Moreover, both expressions have the same numerator $Cov(x,y)$.

Therefore, the sign of the two regression coefficients depends on the same term $Cov(x,y)$. Thus, if $Cov(x,y)$ is positive, both the regression

coefficients would be positive. Otherwise if $Cov(x, y)$ is negative, both the regression coefficients would be negative.

Property 3 : The two regression coefficients are independent of the change of origin but are dependent on the change of scale.

Proof : Let u and υ be the new variables obtained by changing the origin and scale of the original variable x and y as follows :

$$u = \frac{x-a}{h},\ \upsilon = \frac{y-b}{k} \qquad -(1)$$

Where, a, b, h and k (>0) are constants. Since correlation coefficient is independent of the change of origin and scale, we have,

$$r_{xy} = r_{uv} \qquad -(2)$$

As per transformation we have,

$$\sigma_x = h\sigma_u \text{ and } \sigma_y = h\sigma_\upsilon$$

Now, $$b_{yx} = r_{xy}\frac{\sigma_y}{\sigma_x} = r_{u\upsilon}\frac{k\sigma_\upsilon}{h\sigma_u} = \frac{k}{h}\left(r_{u\upsilon}\frac{k_\upsilon}{h_u}\right) \qquad -(3)$$

$$b_{xy} = r_{xy}\frac{\sigma_x}{\sigma_y} = r_{u\upsilon}\frac{k\sigma_u}{h\sigma_\upsilon} = \frac{h}{k}\left(r_{u\upsilon}\frac{k_u}{h_\upsilon}\right) \qquad -(4)$$

From (3) and (4) it is clear that the two regression coefficients are independent of the change of origin but are dependent on the change of scale.



CHECK YOUR PROGRESS

Q.1: Fill in the blanks :

(i) The two regressions coefficients are of _____ sign.

(ii) Regression analysis is used to study_____ between the variables.

(iii) If a regression coefficient is negative then the correlation between the variables would also be _____.

(iv) If one of regression coefficient is greater than unity, the other one must be _____ than unity.

5.6 CORRELATION VS. REGRESSION

- With the help of correlation one measures the covariation between two variables. In correlation neither variable may be termed as dependent or independent variable. Since correlation does not establish a relationship between the two variables as such one can't estimate the value of one variable corresponding to a given value of the other variable.

  Regression establishes a functional relationship between two variables and hence one can estimate the value of one variable corresponding to a given value of the other variable.
- With the help of correlation one can't study which variable is the cause and which variable is the effect.

  Regression establishes the cause and effect relationship between two variables. The variable that indicates cause is the independent variable and the variable that indicates effect is the dependent variable.
- Correlatuon coefficient between two variables x and y is always symmetric, i.e. $r_{xy} = r_{yx}$. But regression coefficient is not symmetric in general, i.e. $b_{xy} \neq b_{yx}$ in general.
- Correlation coefficient is a relative measure of the linear relationship between the two variables x and y and hence it is a pure number, i.e. a unitless number whatever may be the unit of measurement of x and y. This pure number always lies between –1 and 1. On the other hand, regression coefficient is an absolute measure of change in one variable corresponding to one unit change in the other variable. Moreover, if one of the regression coefficient is greater than unity, the other regression coefficient must be less than unity.
- While correlation coefficient is independent of change of origin and scale, regression coefficients are independent of change of origin only but not of scale.

Examples

1. The following results were worked out from scores in Statisties and Mathematics in a certain examination :

	Scores in Statistics (X)	Scores in Mathematics (Y)
Mean	39.5	47.5
Standard Deviation	10.8	17.8

Karl Pearson's correlation coefficient between X and Y is 0.42.

Find both the regression lines. Use these lines to estimate the value of Y when X = 50 and the value of X when Y = 30.

Solution :

We are given,

$\overline{X} = 39.5,\ \overline{Y} = 47.5$

$\sigma_x = 10.8,\ \sigma_y = 17.8$

and r_{xy} = 0.42

Now, $b_{YX} = r_{xy}\frac{\sigma_y}{\sigma_x} = 0.42\frac{17.8}{10.8}$ =0.69

$$b_{XY} = r_{xy}\frac{\sigma_y}{\sigma_x} = 0.42\frac{10.8}{17.8} = 0.25$$

$\therefore$ Regression line of Y on X is

$$Y - \overline{Y} = b_{YX}\left(X - \overline{X}\right)$$

$$\Rightarrow Y - 47.5 = 0.69(X - 39.5)$$

$$\Rightarrow Y = 0.69X - 0.69 \times 39.5 + 47.5$$

$$\Rightarrow Y = 0.69X + 20.24 \quad -(1)$$

$\therefore$ Regression line of X on Y is

$$X - \overline{X} = b_{XY}\left(Y - \overline{Y}\right)$$

$$\Rightarrow X - 39.5 = 0.25(Y - 47.5)$$

$$\Rightarrow X = 0.25Y - 0.25 \times 47.5 + 39.5$$

$$\Rightarrow X = 0.25Y + 27.62 \quad -(2)$$

Again, the predicted value of Y when X = 50 is obtained from the regression line of Y on X and is given by,

Y = 0.69 x 50 + 20.24 = 54.74

and the predicted value of X when Y = 30 is obtained from the regression line of X on Y and is given by,

X = 0.25 x 30 + 27.62 = 35.12

Example 2 : Given the bivariate data

x :	2	4	5	6	8	11
y :	18	12	10	8	7	5

(i) Fit the regression line of Y on X and predict when X = 10.

(ii) Fit the regression line of X on Y and predict when Y = 8.5.

(iii) Calculate Karl Pearson's coefficient of correlation.

Solution : Let us prepare the following table :

x	$x-\bar{x}$ $=x-6$	$(x-\bar{x})^2$	y	$y-\bar{y}$ $=y-10$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
2	–4	16	18	8	64	–32
4	–2	4	12	2	4	–4
5	–1	1	10	0	0	0
6	0	0	8	–2	4	-
8	2	4	7	–3	9	–6
11	5	25	5	–5	25	–25
$\Sigma X = 36$		$\Sigma(X-\bar{X})^2 = 50$	$\Sigma Y = 60$		$\Sigma(Y-\bar{Y})^2 = 106$	$\Sigma(X-\bar{X})(Y-\bar{Y}) = -67$

Now, $\bar{X} = \dfrac{\Sigma X}{n} = \dfrac{36}{6} = 6$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{60}{6} = 10$$

$$\therefore\ b_{YX} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2}$$

$$= \frac{-67}{50} = -1.34$$

$$b_{XY} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(Y-\bar{Y})^2}$$

$$= \frac{-67}{106} = -0.53$$

(i) The regression equation of Y on X is,

$$Y-\bar{Y} = b_{YX}(X-\bar{X})$$

$$\Rightarrow Y-10 = -1.34(X-6)$$

$$\Rightarrow Y = 1.34X + 1.34 \times 6 + 10$$

$$\Rightarrow Y = -1.34X + 18.04$$

When X = 10, the predicted value of y would be

y = –1.34 x 10 + 18.04

= 4.64

(ii) The regression equation of x on y is

$$X - \bar{X} = b_{XY}\left(Y - \bar{Y}\right)$$

$$\Rightarrow X - 6 = -0.63(Y - 10)$$

$$\Rightarrow X = -0.63Y + 0.63 \times 10 + 6$$

$$\Rightarrow X = -0.63Y + 12.3$$

When, y = 8.5 the predicted value of x would be,

x = –0.63 x 8.5 + 12.3

= 6.945

(iii) We know that,

$$r_{xy} = \pm\sqrt{b_{yx} . b_{xy}}$$

$$= -\sqrt{(-1.34) \times (-0.63)}$$

$$= -\sqrt{0.8442}$$

= –0.92 (approx)

Example 3 : Given below are the data for heights of some fathers and sons. Find out the two lines of regression. Calculate the expected height of the son when the height of the father is 67.5 inches.

Heights of father : (in inches)	65	66	67	67	68	69	71	73
Heights of son : (in inches)	67	68	64	58	72	70	69	70

Solution : Let, X denots the height of father and Y denotes the height of son.

Let us prepare the following table :

x	d_x=x-A =x-68	y	d_y=y-B =y-67	d^2x	d^2y	d_xd_y
65	–3	67	0	9	0	0
66	–2	68	1	4	1	–2
67	–1	64	–3	1	9	3
67	–1	58	–9	1	81	9
68	0	72	5	0	25	0
69	1	70	3	1	9	3
71	3	69	2	9	4	6
73	5	70	3	25	9	15
$\Sigma x = 546$	$\Sigma d_x = 2$	$\Sigma y = 538$	$\Sigma d_y = 2$	$\Sigma d^2{}_y = 50$	$\Sigma d^2{}_y = 138$	$\Sigma d_y d_y = 34$

Here, $\bar{x} = \frac{\Sigma x}{n} = \frac{546}{8} = 68.25$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{538}{8} = 67.25$$

Now, $b_{yx} = \frac{\Sigma d_x d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{n}}{\Sigma d_x{}^2 - \frac{(\Sigma d_x)^2}{n}}$

$$= \frac{34 - \frac{2 \times 2}{8}}{50 - \frac{2^2}{8}}$$

$$= \frac{34 - 0.5}{50 - 0.5} = \frac{33.5}{49.5} = 0.68$$

$$b_{xy} = \frac{\Sigma d_x d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{n}}{\Sigma d_y{}^2 - \frac{(\Sigma d_y)^2}{n}}$$

$$= \frac{33.5}{138 - \frac{2^2}{8}} = \frac{33.5}{137.5} = 0.24$$

$\therefore$ The regression line of y on x is,

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\Rightarrow y - 67.25 = -0.68(x - 68.25)$$

$$\Rightarrow y = -0.68x - 0.68 \times 68.25 + 67.25$$

$$\Rightarrow y = -0.68x + 20.84$$

Again, the regression equation of x on y is,

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$\Rightarrow x - 68.25 = -0.24(y - 67.25$$

$$\Rightarrow x = 0.24y - 0.24 \times 67.25 + 68.25$$

$$\Rightarrow x = 0.24y + 52.11$$

The expected height of son when the height of father is 67.5 inches is to be obtained from the regression equation of y on x. Thus when x = 67.5 inches, the expected height of son, i.e. y = (0.68 x 67.5 + 20.84) inches = 66.74 inches.

Example 4 : For the variable x and y, the regression equations are given as 7x – 3y – 18 = 0 and 4x – y – 11 = 0

(i) Find the arithmetic means of x and y.

(ii) Identify the regression equation of y on x.

(iii) Compute the correlation coefficient between x and y.

(iv) Given the variances of x is 9, find the S.D. of y.

Solution :

(i) Since the two lines of regression intusect at the point $(\bar{x}, \bar{y})$, replacing x and y by $\bar{y}$ respectivelt in the given regression equations, we get,

$$7\bar{x} - 3\bar{y} - 18 = 0$$

and $4\bar{x} - \bar{y} - 11 = 0$

Solving these two equation, we get $\bar{x} = 3$ and $\bar{y} = 1$. Thus the arithmetic means of x and y is given by 3 and 1 respectively.

(ii) Let us assume that 7x – 3y – 18 = 0 represents the regression line of y on x and 4x – y – 11 = 0 represents the regression line of x on y.

Now, 7x – 3y – 8 = 0

$$\Rightarrow y = \frac{7}{3}x - 6$$

$$\therefore b_{yx} = \frac{7}{3}$$

Again 4x – y – 11 = 0

$$\Rightarrow x = \frac{1}{4}y + \frac{11}{4}$$

$$\therefore\ b_{xy} = \frac{1}{4}$$

Thus $r^2 = b_{yx} \times b_{xy} = \frac{7}{3} \times \frac{1}{4}$

$$= \frac{7}{12} < 1$$

Since $|r| \leq 1 \Rightarrow r^2 \leq 1$, our assumptions are correct. Thus, 7x – 3y – 18 = 0 truly represents the regression line of y on x.

(iii) Since $r^2 = \frac{7}{12}$

$$\therefore\ r = \sqrt{\frac{7}{12}} = 0.7638$$

(iv) We know that,

$$b_{yx} = r_{xy}\frac{\sigma_y}{\sigma_x}$$

$$\Rightarrow \frac{7}{3} = 0.7638 \times \frac{\sigma_y}{3} \qquad [\because\ Var(x) = \sigma_x{}^2 = 9]$$

$$\Rightarrow \sigma_y = \frac{7}{0.7638} = 9.1647$$

Example 5 : For a bivariate data, you are given the following information :

$\Sigma(x-58) = 46 \qquad \Sigma(x-58)^2 = 3086$

$\Sigma(y-68) = 9 \quad \Sigma(y-68)^2 = 483$

$\Sigma(x-58)(y-68) = 1095,\ \bar{x} = 58,\ \bar{y} = 68$

Number of pairs of observations = 7. You are required to determains.

(i) the two regression equations and

(ii) the coefficient of correlation between x and y.

Solution : We are given, $\bar{x} = 58,\ \bar{y} = 68$

$\Sigma(x-\bar{x}) = \Sigma(x-58) = 46$

$\Sigma(x-\bar{x})^2 = \Sigma(x-58)^2 = 3086$

$\Sigma(y-\bar{y}) = \Sigma(y-68) = 9$

$\Sigma(y-\bar{y})^2 = \Sigma(y-68)^2 = 483$

$$\Sigma(x-\bar{x})(y-\bar{y}) = \Sigma(x-58)(y-68) = 1095$$

Now, $b_{yx} = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \dfrac{1095}{3086} = 0.35$

$$b_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2} = \frac{1095}{483} = 2.27$$

(ii) $\therefore$ The regression line of y on x is

$$y-\bar{y} = b_{yx}(x-\bar{x})$$

$$\Rightarrow y-68 = 0.35(x-58)$$

$$\Rightarrow y = 0.35x - 0.35\times 58 + 68$$

$$\Rightarrow y = 0.35x + 47.7$$

The regression line of x on y is

$$x-\bar{x} = b_{xy}(y-\bar{y})$$

$$\Rightarrow x-58 = 2.27(y-68)$$

$$\Rightarrow x = 2.27y - 2.27\times 68 + 58$$

$$\Rightarrow x = 2.27y - 96.36$$

(ii) The coefficient of correlation

$$r_{xy} = \pm\sqrt{b_{yx}\,.\,b_{xy}}$$

$$= +\sqrt{0.35\times 2.27} = 0.89$$

Example : Find out the regression coefficients Y on X, X on Y and correlation coefficient between X and Y on the basis of the following data :

$\sum xy = 350,\ \sum x = 50,\ \sum y = 60,\ n = 10,$

Variance of x=4 and Variance of Y=9

Solution : Regression coefficient of Y on X is given by

$$b_{yx} = \frac{Cov(x,y)}{\sigma_x^{\,2}}$$

$$= \frac{\frac{1}{n}\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{Variance\ of\ X}$$

$$= \frac{\frac{1}{n}\sum xy - \bar{x}\bar{y}}{Variance\ of\ X}$$

$$= \frac{\frac{1}{n}\sum xy - \left(\frac{\sum x}{n}\right)\left(\frac{\sum y}{n}\right)}{Variance\ of\ X}$$

$$= \frac{\frac{350}{10} - \frac{50}{10} \times \frac{60}{10}}{4} = 1.25$$

Regression Coefficient of X on Y is given by

$$b_{xy} = \frac{Cov(x, y)}{\sigma_y^{\,2}}$$

$$= \frac{\frac{1}{n}\sum\left(x - \bar{x}\right)\left(y - \bar{y}\right)}{Variance\ of\ Y}$$

$$= \frac{\frac{1}{n}\sum xy - \bar{x}\bar{y}}{Variance\ of\ Y}$$

$$= \frac{\frac{1}{n}\sum xy - \left(\frac{\sum x}{n}\right)\left(\frac{\sum y}{n}\right)}{Variance\ of\ Y}$$

$$= \frac{\frac{350}{10} - \frac{50}{10} \times \frac{60}{10}}{9}$$

$$= \frac{\frac{350}{10} - \frac{50}{10} \times \frac{60}{10}}{9}$$

Hence, Coefficient of correlation between X and Y is given by

$$r_{xy} = \pm\sqrt{b_{yx} \cdot b_{xy}}$$

$$= \pm\sqrt{1.25 \times 0.55}$$

$$= 0.83$$

5.7 MULTIPLE CHOICE QUESTIONS

Q 1: The regression line of y on x is derived by

(a) the minimisation of vertical distances in the scatter diagram.

(b) the minimisation of horizontal distances in the scatter diagram.

(c) both (a) and (b)

(d) (a) or (b)

Ans : (a)

Q 2: The two lines of regression become identical when

(a) r = 1 (b) r = –1

(c) r = 0 (d) (a) or (b)

Ans : (d)

Q 3: What are the limits of the two regression coefficients ?

(a) No limit (b) Must be positive

(c) One positive and the other negative

(d) Product of the regression coefficients must be numerically less than unity.

Ans : (d)

Q 4: The regression coefficients remain unchanged due to

(a) shift of origin (b) shift of scale

(c) both (a) and (b) (d) (a) or (b)

Ans : (a)

Q 5: The two line of regression intersect at the point

(a) (0, 0) (b) (1, 1)

(c) (–1, –1) (d) $(\bar{x}, \bar{y})$

Ans : (d)

Q 6: If x and y are two variables, there can be at most

(a) one regression line

(b) two regression line

(c) three regression line

(d) an infinite number of regression lines

Ans : (b)

Q 7: The errors in case of regression equations are

(a) positive (b) negative

(c) zero (d) All these

Ans : (d)

Q 8: The regression lines are perpendicular to each other if r is equal to

(a) 0 (b) +1

(c) –1 (d) ± 1

Ans : (a)

Q 9: The regression coefficients are zero if r is equal to

(a) 2 (b) –1

(c) 1 (d) 0

Ans : (d)

Q 10: In the regression line Y = a + bx, b is called the

(a) slope of the line

(b) intercept of the line

(c) neither (a) nor (b)

(d) both (a) and (b)

Ans : (a)

5.8 EXERCISES

Q 1: What do you mean by linear regression? Why there are two regression lines ? When do these become identical ?

Q 2: Obtain the equations of the two lines of regression for a bivariate distribution.

Q 3: If the regression coefficient of y on x, the coefficient of correlation between x and y and variance of y are $-3/4$, $-\sqrt{3/4}$ and 4 respectively, what is the variance of x ?

Ans : $16/3$

Q 4: The following data are about sale and advertisement expenditures of a firm :

	Sale	Advertisement expenditure (in crores of Rs.)
Mean	40	6
S. D.	10	1.5

Coefficient of correlation r = 0.9

(a) Estimate the likely sale for a proposed advertisement expenditure of Rs. 10 crores.

(b) What would be the advertisement expenditure if the firm fixes a sale target of 60 crores of rupees ?

Q 5: The following results relate to bivariate data on (x, y) :

$\Sigma x = 142,$ $\Sigma y = 166$

$\Sigma xy = 2434.$ $\Sigma x^2 = 2085$

$\Sigma y^2 = 2897$

Determine the correlation coefficient between x and y. Estimate y at x = 20 using linear regression.

(Ans : r = 0.78, y = 23.56)

Q 6: Given the variance of x = 9. The regression equations are

8x – 10y + 66 = 0 and 40x – 18y – 214 = 0

What were

(a) The means of x and y

(b) The coefficient of correlation between x and y

(c) The standard deviation of y

(Ans : (a) $\bar{x} = 13,\ \bar{y} = 17$

(b) r = 0.6

(c) σ_y =4)

Q 7: Find the two regression lines from the following data :

x :	1	2	3	4	5
y :	2	3	5	4	6

Estimate y if x = 2.5

(Ans : y = 0.9x + 1.3, x = 0.9y – 0.6, y = 3.55)

Q 8: The following data were obtained from two variable x and y :

$\bar{x} = 20,\ \bar{y} = 15,\ \sigma_x = 4,\ \sigma_y = 3,\ r = 0.7$

Obtain the two regression equations and find the most likely value of y when x = 24.

(Ans : y = 0.53x + 4.5, x = 0.93y + 6, y = 17.22)

Q 9: Assume that the two statistical variables x and y follow the regression equations

y = 2x + 2

and x = ay + b

If the correlation coefficient between x and y is 0.93, find the value of a.

[Ans : a = 0.43 (approx)]



5.9 LET US SUM UP

- Regression is the measure of the average relationship between two or more variables in terms of the original units of the data.
- The two equations which are estimated on the basis of the given values

of the two variables to estimate the most likely values of one variable for specified values of the other are called regression equations.

- If the points of the scatter diagram of two correlated variables clusters around a straight line then we say that there is linear regreassion between the two variables.
- When only two variables are associated with regression analysis then we call it single regression.
- The product of the two regression coefficients is equal to the square of the correlation coefficient or The correlation coefficient is the geometric mean of the regression coefficients.
- Both the regression coefficients must have the same sign.
- The two regression coefficients are independent of the change of origin but are dependent on the change of scale.
- With the help of correlation one measures the covariation between two variables.
- In correlation neither variable may be termed as dependent or independent variable.
- Since correlation does not establish a reltionship between the two variables as such one can't estimate the value of one variable corresponding to a given value of the other variable.
- Regression establishes a functional relationship between two variables and hence one can estimate the value of one variable corresponding to a given value of the other variable.
- With the help of correlation one can't studt which variable is the cause and which variable is the effect.
- Regression establishes the cause and effect relationship between two variables.
- The variable that indecates cause is the independent variable and the variable that indicates effect is the dependent variable.
- Correlatuon coefficient between two variables x and y is always symmetric, i.e. $r_{xy} = r_{yx}$. But regression coefficient is not symmetric in general, i.e. $b_{xy} \neq b_{yx}$ in general.

- Correlation coefficient is a relative measure of the linear relationship between the two variables x and y and hence it is a pure number, i.e. a unitless number whatever may be the unit of measurement of x and y. This pure number always lies between –1 and 1.
- Regression coefficient is an absolute measure of change in one variable corresponding to one unit change in the other variable. Moreover, if one of the regression coefficient is greater than unity, the other regression coefficient must be less than unity.
- While correlation coefficient is independent of change of origin and scale, regression coefficients are independent of change of origin only but not of scale.



5.10 FURTHER READING

1) Gujrati, D (1995); Basic Econometrics (third edition), McGraw-Hill, Inc; Singapore
2) Gun, A M; Gupta, M K & Dasgupta, B (2002); Fundamentals of Statistics, Volume One; World Press Private Ltd. Kolkata
3) Gupta, S C (1994); Fundamentals of Statistics; Himalayan Publishing House, New Delhi.



5.11 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q. No. 1: (i) same (ii) dependence
(iii) negative (iv) less



5.12 MODEL QUESTIONS

A. SHORT QUESTIONS

Q 1: What is meant by regression? What are its types?

Q 2: Distinguish between linear and non-linear regression.

Q 3: State the limits of two regression coefficient.

Q 4: What is a two variable regression model?

B. LONG QUESTIONS

Q 1: State the properties of regression coefficients and prove them.

Q 2: Discuss the relationship between correlation and regression.

Q 3: Given below the information about the capital employed and profit earned by a company over the last twenty five years :

	Mean	S.D.
Capital employed ('0000 Rs.)	62	5
Profit earned ('000 Rs.)	25	6

Correlation coefficient between capital and profit = 0.92. Find the sum of the regression coefficients for the above data.

(Ans : 1.871)

*** ***** ***

UNIT 6: PROBABILITY : BASIC IDEAS

UNIT STRUCTURE

6.1 Learning Objectives

6.2 Introduction

6.3 Definition of the Basic Terms

6.4 Classical or Mathematical or Apriori Definition of Probability

6.5 Addition Theorems or Theorems on Total Probability

6.6 The Theorem of Compound Probability or the Multiplicative Law of Probability

6.7 Baye's Theorem (Concept only)

6.8 Exercises

6.9 Let Us Sum Up

6.10 Further Reading

6.11 Answers to Check Your Progress

6.12 Model Questions

6.1 LEARNING OBJECTIVES

After studying this chapter, you will be able to :

- discuss the basic concept of probability
- define the meanings of the terms related to probability
- put foward the different rules and their applications
- explain the different computational techniques of solving practical problems.

6.2 INTRODUCTION

The concept of probability originated from the analysis of the games of chance in the 17th century. In recent time, probability has developed itself into a full fledged subject and become an integral part of Statistics. The theopry of probability is a study of statistical or Random experiments. It is the back bone of Statistical inference and Decision Theory that are essential tools of the analysis of most of the modern business and economic

problems.

The broad divisions of probability are Subjective Probability and Objective Probability. Subjective Probability is basically dependent on personal judgement and experience and, as such, it may be influenced by the personal belief, attitude and bias of the person applying it. However, in the field of uncertainity, this would be quite helpful and it is being applied in the area of decision making management. This Subjective Probability is beyond the scope of our present discussion. We ar going to discuss Objective Probability in the remaining sections.

6.3 DEFINITION OF THE BASIC TERMS

In this section, we shall discuss the basic definitions of a few terms relationg to the concept of probalility.

Random Experiment: If the result or outcome of an experiment conducted again and again under the same conditions may be any one of the possible outcomes then the experiment is called a random experiment. For example, if a coin is tossed again and again under the same conditions then we get two outcomes - Head (H) and Tail (T). It is impossible to say in advance whether a Head or a Tail would turn up when we toss the coin once. Thus, tossing a coin is an example of a random experiment.

Trial: The task of performing a random experiment is called a trial. For example, if a coin is tossed then it would be a trial.

Event: The outcome or result of a trial is called an event. For example, if a coin is tossed and we get a head it would be called an event.

Equally likely events: Two or more events are said to be equally likely if neither of them occurs in preference to the other. For example, if an unbiased coin is tossed, the two possible outcomes or events, a head or a tail are equally likely.

Mutually Exclusive events: Two or more events are said to be mutually exclusive if they are such that when one occurs, other (others) cannot occur. For example, the two possible outcomes or events of toss of a coin are mutually exclusive.

Exhaustive events: The events associated with a random experiment are said to be exhaustive if they include all possible outcomes

of the experiment. Thus, if a set of events are exhaustive then any one of them is sure to occur. For example, the two events Head and Tail, when a coin is tossed once, are exhaustive as no other event except these two can occur.

Favourable cases: The outcomes or events of a random experiment which imply the occurrence of an event are called the favourable cases to the event. For example, the number of favourable cases to the event of getting an even number in throwing a dice is 3, since there are 3 even numbers 2, 4, and 6 in between 1 and 6.

Sample Space: The set of all possible outcomes of a random experiment is called a Sample Space and generally it is denoted by the symbol S or Ω. For example, if a dice is thrown then the Sample Space would be: S = {1,2,3,4,5,6}

6.4 CLASSICAL OR MATHEMATICAL OR APRIORI DEFINITION OF PROBABILITY

If n is the number of equally likely, mutually exclusive and exhaustive outcomes of a random experiment out of which m outcomes are favourable to the occurrence of an event A, then the probability that A occurs, denoted by P(A), is given by:

$$P(A) = \frac{Number\ of\ outcomes\ favourable\ to\ A}{Number\ of\ exhaustive\ outcomes\ or\ total\ number\ of\ probable\ outcomes}$$

$$= \frac{m}{n}$$

The probability of non-occurrence of the event A is denoted by $P(\overline{A})$ or $P(A^c)$ and is given by

$$P(\overline{A}) = \frac{Number\ of\ outcomes\ unfavourable\ to\ A}{Total\ number\ of\ probable\ outcomes}$$

$$= \frac{n-m}{n} = \frac{n}{n} - \frac{m}{n}$$

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

Now, $P(\bar{A}) = 1 - \frac{m}{n}$

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

$\Rightarrow P(A) + P(\bar{A}) = 1$

Note 1: If the number of favourable cases, m = n i.e. all the probable cases are favourable, then probability of the event A becomes 1, i.e. P(A) = 1. In this case A is called a certain or sure event.

If the number of favourable cases, m = 0, i.e. all the probable cases are unfavourable, then probability of the event A becomes 0, i.e. P(A) = 0. In this case A is called an impossible event.

Since the value of m lies between 0 and n, the probability of event A lies between 0 and 1, i.e. $0 \leq P(A) \leq 1$.

Note 2: The classical definition of probability has the following limitations:

(a) It is applicable only when the total number of events is finite.

(b) It can be used only when the events are equally likely or equi - probable. This assumption is made well before the experiment is performed.

(c) This definition has only a limited field of application like coin tossing, dice throwing, drawing cards etc. where the possible events are known well in advance. In the field of uncertainity or where no prior knowledge is provided, this definition is inapplicable.

Note 3: The ratio of number of favourable events is known as odds in favour of the event A and its inverse ratio is known as odds against the event A. i.e.

odds in favour of A = $\frac{m}{n-m}$

and odds against A = $\frac{n-m}{m}$



CHECK YOUR PROGRESS

Q 1: State True or False

(i) The probability of an impossible event is always zero.

(ii) If two events are independent, then they will also be mutually exclusive.

(iii) The probability of occurence of an event cannot be a negative number.

Example 1: A coin is tossed 3 times. What is the probability of getting : i) 2 heads ii) atleast 2 heads

Solution: When a coin is tossed 3 times, we shall get the following probable outcomes:

HHH, HHT, HTH, HTT, THH, THT, TTH, TTT i.e. 8 outcomes

(i) Out of these 8 outcomes, 2 heads occur in three cases namely HHT, HTH and THH. If A denotes the event of getting 2 heads, then above three cases are favourable for the event A.

$$\therefore P(A) = \frac{3}{8} = 0.375$$

(ii) Let B denotes the event of getting atleast 2 heads. The favourable cases for the event B are,

HHT, HTH, THH, HHH i.e. 4 cases

$$\therefore P(B) = \frac{4}{8} = 0.50$$

Example 2: Two dice are thrown simultaneously. Find the probability that the sum of points on the two dice would be 7 or more.

Solution: If two dice are thrown, then the probable outcomes would be

(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)
(2,1), (2,2), (2,3), (2,4), (2,5), (2,6)
(3,1), (3,2), (3,3), (3,4), (3,5), (3,6)
(4,1), (4,2), (4,3), (4,4), (4,5), (4,6)
(5,1), (5,2), (5,3), (5,4), (5,5), (5,6)
(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)
i.e. 36 cases

Now, the cases where the sum of points is 7 or more are -

(1,6), (2,5), (2,6), (3,4), (3,5)
(3,6), (4,3), (4,4), (4,5), (4,6)
(5,2), (5,3), (5,4), (5,5), (5,6)
(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)
i.e. 21 cases

$\therefore$ Probability that the sum of points on the two dice is 7 or more

$$= \frac{21}{36} = \frac{7}{12}$$

6.5 ADDITION THEOREMS OR THEOREMS OF TOTAL PROBABILITY

Theorem 1:

Statement: If $A_1, A_2, \ldots\ldots\ldots\ldots, A_k$ are k mutually exclusive and exhaustive events then the probability that any one of them occurs is equal to the sum of the probabilities of occurrences of these events separately. In symbol,

$P(A_1+A_2+\ldots\ldots\ldots+A_k)$ or $P(A_1 \cup A_2 \cup \ldots\ldots\ldots\ldots\ldots \cup A_k) = P(A_1)+P(A_2)+\ldots\ldots\ldots\ldots+P(A_k)$

Theorem 2:

Statement: If A and B are two events which are not mutually exclusive then the probability that atleast one event will occur i.e. P(A+B) or $P(A \cup B)$ is given by

$P(A+B)$ or $P(A \cup B) = P(A)+P(B) - P(AB)$, where $P(AB) = P(A \cap B)$

Example 3: If two dice are thrown what is the probability that a sum of nine points or eleven points will be obtained?

Solution: Let A: the event that the sum of points is nine

B: the event that the sum of points is eleven

Now, when two dices are thrown, we may get 36 different outcomes. Hence the total number of all possible cases = 36.

Again, the cases favourable for the event A are: (6,3), (5,4), (4,5), (3,6), i.e. 4 cases.

The cases favourable for the event B are: (6,5), (5,6) i.e. 2 cases.

$$\therefore P(A) = \frac{4}{36}$$

and, $$P(B) = \frac{2}{36}$$

Here, the events A and B are mutually exclusive.

$\therefore$ P (that a sum of nine points or eleven points)

$= P(A \cup B)$

$= P(A) + P(B)$

$= \frac{4}{36} + \frac{2}{36}$

$= \frac{6}{36} = \frac{1}{6}$

Example 4: There are 20 marbles in an urn which are numbered from 1 to 20. A marble is drawn at random. Find the probability that the number on the marble is a multiple of either 2 or 5.

Solution:

Let, A: the event that the number on the marble is a multiple of 2

B: the event that the number on the marble is a multiple of 5.

The cases favourable for the event A are: 2,4,6,8,10,12,14,16,18,20; i.e. 10 cases

The cases favourable for the event B are: 5, 10, 15, 20; i.e. 4 cases. Again, the cases favourable for the vent both A and B, i.e. favourable for $A \cap B$ are: 10,20; i.e. 2 cases

Here, the events A and B are not mutually exclusive. Therefore, we have

P (that the number is a multiple of either 2 or 5)

$= P(A \cup B) = P(A) + P(B) - P(AB)$

$= \frac{10}{20} + \frac{4}{20} - \frac{2}{20} = \frac{12}{20} = \frac{3}{5}$

6.6 THE THEOREM OF COMPOUND PROBABILITY OR THE MULTIPLICATIVE LAW OF PROBABILITY

Statement: For any two events A and B, the probability that A and B occur simultaneously is given by the product of the unconditional probability of A and the conditional probability of B given that A has already occured, i.e.

$P(A \cap B)$ or $P(AB) = P(A) \cdot P(B/A)$, provided $P(A) > 0$

Alternatively, we may write

$P(A \cap B)$ or $P(AB) = P(B) \cdot P(A/B)$, provided $P(B) > 0$

Note:

1. If the events A and B independent, then P(B/A) = P(B) and consequently for two independent events A and B the multiplicative law becomes P(AB) = P(A) . P(B). Thus we may also define independent event as

Two events A and B are independent if

$$P(AB) = P(A).P(B)$$

2. From the theorem of conditional probability, we have,

$$P(AB) = P(A).P(B/A)$$

Example 5: An urn contains 6 white and 8 black balls. Two successive draws of one ball are made. Find the probability that the first drawn ball is white and the second drawn ball is black if

(i) the first drawn ball is replaced before the second draw.

(ii) the first drawn ball is not replaced.

Solution:

Let A: the event of drawing a white ball in the first draw

B: the event of drawing a black ball in the second draw.

Case (i): The first drawn ball is replaced

If the first drawn ball is replaced before the second draw, the event A will have no effect on the occurrence of the event B as the total number of balls will remain the same just before the second draw. Hence, the events A and B are independent of each other.

$$\therefore P(AB) = P(A).P(B)$$

$$= \frac{6}{14} \times \frac{8}{14} = \frac{48}{196}$$

$$= \frac{6}{14} \times \frac{8}{14} = \frac{48}{196}$$

Case (ii): The first drawn ball is not replaced

If the first drawn ball is not replaced, the event A will affect the occurrence of the event B. Hence, we have,

$$P(AB) = P(A) . P(B/A)$$

$$= \frac{6}{14} \times \frac{8}{13} = \frac{48}{182}$$

Certain Other Theorems Relating to Probability

Theorem 1: If A and B are independent events then A and $\overline{B}$ are also independent events.

Proof: The event A can be expressed as follows:

$$A = (A \cap \overline{B}) \cup (A \cap B)$$

Here, $A \cap \overline{B}$ and $A \cap B$ are disjoint or mutually exclusive events.

Hence, by Axiomatic definition of probability, we get,

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

$$\Rightarrow P(A \cap \overline{B}) = P(A) - P(A \cap B)$$

$$= P(A) - P(A).P(B) \quad [\because \text{A and B are independent}]$$

$$= P(A)[1 - P(B)]$$

$$= P(A)P(\overline{B})$$

$\therefore A$ and $\overline{B}$ are independent

Theorem 2: If A and B are independent events then $\overline{A}$ and $\overline{B}$ are also independent events.

Proof: We are given that

$$P(A \cap B) = P(A).P(B)$$

Now, $P(\overline{A} \cap \overline{B}) = P(\overline{A \cup B})$

$$= 1 - P(A \cup B)$$

$$= 1 - [P(A) + P(B) - P(A \cap B)]$$

$$= 1 - [P(A) + P(B) - P(A).P(B)] \quad (\because \text{A and B are independent})$$

$$= 1 - P(A) - P(B) + P(A).P(B)$$

$$= [1 - P(B)] - P(A)[1 - P(B)]$$

$$= [1 - P(B)][1 - P(A)]$$

$$= P(\overline{A})P(\overline{B})$$

Hence $\overline{A}$ and $\overline{B}$ are independent.

Example 6: If A and B are two events such that P(A) = $\frac{2}{3}$, $P(\overline{A} \cap B) = \frac{1}{6}$ and $P(A \cap B) = \frac{1}{3}$, find $P(B), P(A \cup B), P(A/B), P(B/A), P(\overline{A} \cup B), P(\overline{A} \cap \overline{B})$ and $P(\overline{B})$. Also examine whether the events A and B are: (a) Equally likely,

(b) Exhaustive, (c) Mutually exclusive and (d) Independent.

Solution:

We are given, P(A) = $\frac{2}{3}$, $P(\overline{A} \cap B) = \frac{1}{6}$, $P(A \cap B) = \frac{1}{3}$

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

$$= \frac{1}{6} + \frac{1}{3} = \frac{1}{2}$$

Now, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$= \frac{2}{3} + \frac{1}{2} - \frac{1}{3} = \frac{5}{6}$$

Again, $P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{1}{3} \times \frac{2}{1} = \frac{2}{3}$

and, $P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{1}{3} \times \frac{3}{2} = \frac{1}{2}$

Now, $P(\overline{A} \cup B) = P(\overline{A}) + P(B) - P(\overline{A} \cap B)$

$$= [1 - P(A)] + P(B) - P(\overline{A} \cap B)$$

$$= 1 - \frac{2}{3} + \frac{1}{2} - \frac{1}{6}$$

$$= \frac{2}{3}$$

$$P(\overline{A} \cap \overline{B}) = 1 - P(A \cup B)$$

$$= 1 - \frac{5}{6} = \frac{1}{6}$$

$P(\overline{B})$ = 1 - P(B) = 1- $\frac{1}{2}$ = $\frac{1}{2}$

(a) Since $P(A) \neq P(B)$, A and B are not equally likely events.

(b) Since $P(A \cup B) \neq 1$, A and B are not exhaustive events.

(c) Since $P(A \cap B) \neq 0$, A and B are not mutually exclusive events.

(d) Since, $P(A).P(B) = P(A \cap B)$, A and B are independent events.

Example 7: A bag contains 5 red and 3 black balls and second bag contains 4 red and 5 black balls.

(a) If one ball is selected at random from each bag, what is the probability that both of them are of same colour?

(b) If a bag is selected at random and two balls are drawn from it, what is the probability that they are of (i) same colour, (ii) different colour?

Solution:

(a) Required probability = [Probability that ball from both bags are red] + [Probability that balls from both bags are black]

$$= \frac{5}{8}\times\frac{4}{9} + \frac{3}{8}\times\frac{5}{9}$$

$$= \frac{35}{72}$$

(b) Let A be the event that first bag is drawn, so that $\overline{A}$ denotes the event that second bag is drawn. Since the two events are equally likely, mutually exclusive and exhaustive, we have

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

(i) Let R be the event that two drawn balls are red and B be the event that they are black.

$\therefore$ The required probability

$$= [P(A \cap R) + P(A \cap B)] + [P(\overline{A} \cap R) + P(\overline{A} \cap B)]$$

$$= P(A)P(R/A) + P(A)P(B/A) + P(\overline{A})P(R/\overline{A}) + P(\overline{A})P(B/\overline{A})$$

$$= P(A)[P(R/A) + P(B/A)] + P(\overline{A})[P(R/\overline{A}) + P(B/\overline{A})]$$

$$= \frac{1}{2}\left[\frac{{}^5c_2}{{}^8c_2} + \frac{{}^3c_2}{{}^8c_2}\right] + \frac{1}{2}\left[\frac{{}^4c_2}{{}^9c_2} + \frac{{}^5c_2}{{}^9c_2}\right]$$

$$= \frac{1}{2}\left[\frac{10+3}{28}\right] + \frac{1}{2}\left[\frac{6+10}{36}\right] = \frac{229}{504}$$

(ii) Let c denotes the event that the drawn balls are of different colours.

$\therefore$ The required probability

$$P(C) = P(A \cap C) + P(\overline{A} \cap C)$$

$$= P(A)P(C/A) + P(\overline{A})P(C/\overline{A})$$

$$= \frac{1}{2}\cdot\frac{5\times 3}{{}^8c_2} + \frac{1}{2}\cdot\frac{4\times 5}{{}^9c_2}$$

$$= \frac{1}{2}\left[\frac{15}{28} + \frac{20}{36}\right] = \frac{275}{504}$$

Example 8: The odds against A solving a certain problem are 4 to 3 (or 4:3) and odds in favour of B solving the same problem are 7 to 5 (or 7:5). What is the probability that the problem is solved if they both try independently?

Solution:

[Note: If a is the number of cases favourable to an event A and $\propto$ is the number of cases favourable to its compliment event $\overline{A}$, then odds in favour of A are $a:\propto$ and odds against A are $\propto: a$. Obviously $P(A)=\frac{a}{a+\propto}$ and $P(\overline{A})=\frac{\propto}{a+\propto}$]

Let, E_1 denotes the event that A solves the problem

E_2 denotes the event that B solves the problem

Clearly E_1 and E_2 ar independent event.

We are given that, $P(E_1) = \frac{3}{4+3} = \frac{3}{7}$

$$\Rightarrow P(\overline{E}_1)=1-\frac{3}{7}=\frac{4}{7}$$

and, $P(E_2)=\frac{7}{7+5}=\frac{7}{12}$

$$\Rightarrow P(\overline{E}_2)=1-\frac{7}{12}=\frac{5}{12}$$

Now, P(atleast one can solve the problem) = 1- P(none can solve the problem)

$$= 1-P\left(\overline{E}1\cap\overline{E}2\right)$$
$$= 1-P\left(\overline{E}_1\right).P\left(\overline{E}_2\right)$$
$$= 1-\frac{4}{7}\times\frac{5}{12}=\frac{16}{21}$$

Example 9: A sum is given to five students A, B, C, D and E. Their respective chances of solving it are $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, $\frac{1}{4}$ and $\frac{1}{5}$. What is the probability that atleast one of the students solves the sum?

Solution:

Let, E1: the event of solving the sum by A

E2: the event of solving the sum by B

E3: the event of solving the sum by C

E4: the event of solving the sum by D

E5: the event of solving the sum by E

We are given,

$$P(E_1)=\frac{1}{2},\ P(E_2)=\frac{1}{3},\ P(E_3)=\frac{1}{4},\ P(E_4)=\frac{1}{4},\ P(E_5)=\frac{1}{5}$$

Now, P(atleast one can solve the sum) = 1- P(None can solve the sum)

$$= 1-P\left(\overline{E}_1\cap\overline{E}_2\cap\overline{E}_3\cap\overline{E}_4\cap\overline{E}_5\right)$$

$$= 1-P\left(\overline{E}_1\right)P\left(\overline{E}_2\right)P\left(\overline{E}_3\right)P\left(\overline{E}_4\right)P\left(\overline{E}_5\right)$$

$$= 1-\left(1-\frac{1}{2}\right)\left(1-\frac{1}{3}\right)\left(1-\frac{1}{4}\right)\left(1-\frac{1}{5}\right)$$

$$= 1-\left(\frac{1}{2}\times\frac{2}{3}\times\frac{3}{4}\times\frac{3}{4}\times\frac{4}{5}\right)$$

$$= 1-\frac{3}{20}=\frac{17}{20}$$

Example 10: There are two urns containing 5 red and 6 white balls and 3 red and 7 white balls respectively. If two balls are drawn from the first urn without replacement and transferred to the second urn and then a draw of another two balls is made from it, what is the probability that both the balls drawn are red?

Solution:

Since two balls are transferred from the first urn containing 5 red and 6 white balls to the second urn containing 3 red and 7 white balls, we are to consider the following cases:

Case A: Both the balls transferred are red. In this case, the second urn contains 5 red and 7 white balls.

Case B: The two balls transferred are of different colours. Then the second urn contains 4 red and 8 white balls.

Case C: Both the balls transferred are white. Now the second urn contains 3 red and 7 white balls.

Let R denotes the event of drawing two red balls from the second urn.

$\therefore$ The required probability = $P\left(R\cap A\right)+\left(R\cap B\right)+P\left(R\cap C\right)$

$$= P(A)P(R/A)+P(B)P(R/B)+P(C)P(R/C)$$

$$= \frac{{}^5c_2}{{}^{11}c_2} \times \frac{{}^5c_2}{{}^{12}c_2} + \frac{{}^5c_1 \times {}^6c_1}{{}^{11}c_2} \times \frac{{}^4c_2}{{}^{12}c_2} + \frac{{}^6c_2}{{}^{11}c_2} \times \frac{{}^3c_2}{{}^{12}c_2}$$

$$= \frac{10}{55} \times \frac{10}{66} + \frac{30}{55} \times \frac{6}{66} + \frac{15}{55} \times \frac{3}{66}$$

$$= \frac{325}{66 \times 55} = \frac{65}{726}$$



CHECK YOUR PROGRESS

Q 2: Fill in the blanks

(i) Two events A and B are statistically independent when P(A)=_____.

(ii) The probability of getting and odd number in tossing a dice is_____.

(iii) If P(A/B)=P(A), then the events A and B are _____.

(iv) If $P(A \cap B) = 0.60$ and $P(A \cap B) = 0.70$ of two events A and B, then $P(A) + P(B) =$ _____.

6.7 BAYE'S THEOREM (CONCEPT ONLY)

This theorem is also known as theorem of posteriori probability. It was developed by Rev. Thomas Bayes.

The Baye's theorem is basically a technique of determining prior probabilities in the light of new information. The rule is stated below:

Let an event A can ocuur if any one of the set of exhaustive and mutually exclusive events B1, B2, ..., Bn occurs. The probabilities P(B1), P(B2), ..., P(Bn) and the conditional probabilities P(A/Bi), i = 1,2,....., n for A to ocuur are known. Then the conditional probability P(Bi/ A) when A has actually occured is given by:

$$P(B_i / A) = \frac{P(B_i).P(A / B_i)}{P(B_1).P(A / B_1) + P(B_2).P(A / B_2) + \ldots\ldots\ldots\ldots + P(B_n).P(A / B_n')}$$

$i = 1,2,\ldots\ldots,n$

6.8 EXERCISES

Multiple Choice Questions

Q 1: If $P(A \cap B) = 0$, then the two events A and B are

(a) Mutually exclusive (b) Exhaustive

(c) Equally likely (d) Independent

Ans: (a)

Q 2: If for two events A and B, $P(A \cup B) = 1$, then A and B are

(a) Mutually Exclusive events (b) Equally likely events

(c) Exhaustive events (d) Dependent events

Ans: (c)

Q 3: If P(A) = P(B), then the two events A and B are

(a) Independent (b) Dependent

(c) Equally Likely (d) Both (a) and (c)

Ans: (c)

Q 4: If for two events A and B, $P(A \cap B) \neq P(A) \times P(B)$, then the two events A and B are

(a) Independent (b) Dependent

(c) Not equally likely (d) Not exhaustive

Ans: (b)

Q 5: If $P(A/B) = P(A)$, then

(a) A is independent of B (b) B is independent of A

(c) B is dependent of A (d) Both (a) and (b)

Ans: (d)

Q 6: If two events A and B are independent, then

(a) A and the complement of B are independent

(b) B and the complement of A are independent

(c) Complements of A and B are independent

(d) All of these (a), (b) and (c)

Ans: (d)

Q 7: The probability of an event can assume any value between

(a) -1 and 1 (b) 0 and 1

(c) -1 and 0 (d) none of these

Ans: (b)

Q 8: If p:q are the odds in favour of an event, then the probability of that event is

(a) p/q (b) $\frac{p}{p+q}$

(c) $\frac{q}{p+q}$ (d) none of these

Ans: (b)

Q 9: For any two events A and B,

(a) $P(A)+P(B) \geq P(A \cap B)$ (b) $P(A)+P(B) < P(A \cap B)$

(c) $P(A)+P(B) > P(A \cap B)$ (d) $P(A) \times P(B) \leq P(A \cap B)$

Ans: (a)

Q 10: If two events A and B are independent, then $P(A \cap B)$ equals to

(a) P(A)+P(B) (b) P(A).P(B)

(c) P(A) . P(B/A) (d) P(B) . P(A/B)

Ans: (b)

Long Questions:

Q 1: There are two urns. The first urn contains 3 red and 5 white balls whereas the second urn contains 4 red and 6 white balls. A ball is taken at random from the first urn and is transferred to the second urn. Now another ball is selected at random from the second urn. Find the probability that the second ball is red.

Ans: 35/88

Q 2: A problem in probability was given to three students A, B and C whose chances of solving it are $1/3$, $1/5$ and $1/2$ respectively. What is the probability that the problem would be solved?

Ans: $11/15$

Q 3: A number is selected at random from the first 1000 natural numbers. What is the probability that the number so selected would be a multiple of 7 or 11?

Ans: 0.22

Q 4: The odds are 9:7 against a person A, who is now 35 years of age, living till he is 65 and 3:2 against a person B, now 45 years of age, living till he

is 75. Find the chance that atleast one of these persons will be alive 30 years hence.

Ans: $\frac{53}{80}$

Q 5: A bag contains 10 red and 8 black balls. Two balls are drawn at random. Find the probability that (a) both of them are red, (b) one is red and the other is black.

Ans: $\frac{5}{17}, \frac{80}{153}$



6.9 LET US SUM UP

- In this unit, we have discussed a few basic concepts of probability. We have defined the terms random experiment, trial, event and its different types, favourable cases, sample space etc.
- To discuss probability a few fundamental theorems exist, e.g., classical or mathematical or apriori definition of probability, addition theorem, multiplicative theorem etc.
- The Baye's theorem is also known as posteriori theorem. It was developed by Rev. Thomas Bayes.



6.10 FURTHER READING

1. Mukhopadhyay, P. (1996), *Mathematical statistics*. New Central Book Agency (P) Ltd.
2. Goon, A.M., Gupta, M.K. and Dasgupts, B. (1997), *An ouline of statistical Theory.* Vol. I, World Press.
3. Mukhopadhyay, P. (1991). *Theory of Probability.* New Central Book Agency (P) Ltd.
4. Gupta, S.C., Kapoor, V.K. (2009). *Fundamentals of Mathematical Statistics*, Sultan Chand & Sons.

5. Feller, W. (1968). *An Introduction to Probability Theory and its Application*, Vol I, Wiley Eastern.



6.11 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No. 1: (i) True (ii) False (iii) True

Ans to Q No. 2: (i) P(A).P(B) (ii) $\frac{1}{2}$

(iii) independent (iv) 0.10



6.12 MODEL QUESTIONS

Q 1: A pair of unbiased dice is thrown. Find thje probility that (i) the sum of spots is either 5 or 10, (ii) either there is a doublet or a sum less than 6.

Q 2: In a group of 1000 persons, there are 650 who can speak Hindi, 400 can speak English and 150 can speak both Hindi and English. If a person is selected at random, what is the probability that he speaks (i) Hindi only, (ii) English only, (iii) only one of the two languages, (iv) atleast one of the two language?

Q 3: For two events A and B, let P(A) = 0.4, P(A+B) = 0.58 and P(B) = 0.3. Check whether A and B are independent or not.

Q 4: A can hit a target thrice in 5 shots, B twice in 5 shots and C thrice in 4 shots. They fire a volley. What is the probability that 2 shots hit?

Q 5: If A, B and C are mutually exclusive and exhaustive events, and P(A) = $\frac{1}{2}$. P(B) and P(B) = $\frac{2}{3}$. P(C), find P(A), P(B) and P(C).

*** ***** ***

UNIT 7: EXPECTATIONS AND MOMENTS

UNIT STRUCTURE

7.1 Learning Objectives

7.2 Introduction

7.3 Random Variable

- 7.3.1 Definition of Random Variable
- 7.3.2 Discrete Random Variable
- 7.3.3 Continuous Random Variable

7.4 Probability Distribution

- 7.4.1 Probability Mass Function
- 7.4.2 Probability Density Function

7.5 Mathematical Expectation and Variance

- 7.5.1 Mathematical Expectation
- 7.5.2 Properties of Mathematical Expectation
- 7.5.3 Variance of Random Variable
- 7.5.4 Properties of Variance
- 7.5.5 Illustrated Examples

7.6 Moments

7.7 Skewness

- 7.7.1 Karl Pearson's Co-efficient of Skewness
- 7.7.2 Boweley's Co-efficient of Skewness

7.8 Kurtosis

- 7.8.1 Measures of Kurtosis

7.9 Let Us Sum up

7.10 Further Reading

7.11 Answers to Check Your Progress

7.12 Model Questions

7.1 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 expectation and variance of a random variable which are vital in the study of any distribution, along with their properties and applications
- distinguish between raw moments and central moments and how central moments can be expressed in terms of raw moments.
- realise how different measures of location and dispersion as discussed in earlier chapters can be expressed with the help of moments.
- discuss the concepts of skewness and kurtosis.

7.2 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 second 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 these 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.

7.3 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:

7.3.1 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. That is why it is also called a chance variable or a stochastic 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.

7.3.2 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 $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ or it can take countably infinite number of values $x_1, x_2, \ldots\ldots\ldots\ldots$.

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.

7.3.3 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 service station etc are some examples of continuous random varible.

7.4 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.

7.4.1 Probability Mass Function

Let x be a discrete random variable which takes the values $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ with respective probabilities $p_1, p_2, \ldots\ldots\ldots, p_n$. Let, $p_i = P(x = x_i) = P(x_i)$, $i = 1, 2, \ldots\ldots\ldots$, n be the probability that x takes the value x_i. Then $p(x_i)$ 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 $i = 1, 2, \ldots\ldots\ldots\ldots$, 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,....., 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

Example 1: 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 1/6. This can be written as,

p(x_i) = P(x= x_i) = 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.

Example 2: Check whether the following probabilety distribution can be regarded as a Probabilety mass function and draw its graph,

x:	0	1	2	3	4
p(x):	0.10	0.20	0.40	0.20	0.10

Also find $P(x \geq 3)$

Solution : Here, $\sum_{x} p(x) = \sum_{x-0}^{4} p(x)$

= 0.10 + 0.20 + 0.40 + 0.20 + 0.10

= 1.

Also each of $p(x) \geq 0 \quad \forall x.$

The above distribution can be regarded as a probability mass functions.

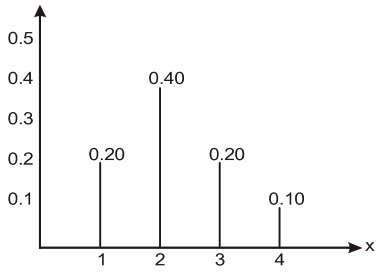


Fig. : probability mass function

Now $P(x \geq 3)$ =P(x=3) + P(x=4)

=p(3) + p(4)

= 0.20 + 0.10 = 0.30.

7.4.2 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

Example 3: Consider the function, $f(x)$ = $\begin{cases} a, 0 \leq x \leq 5 \\ 0, otherwise \end{cases}$

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,

$$\int_0^5 a dx = 1$$

$$\Rightarrow \quad a = \frac{1}{5}$$

Since a>0, the function f(x)$\geq$0. Thus f(x) satisfies both the conditions for a p.d.f.

Example 4: If $f(x) = kx$, $1 \leq x \leq 2$ is to be a probability density functino of the random variable X, find the value of k.

Solution: $\because f(x)$ is the p.d.f. of x, we must have,

$$f(x) \geq 0 \forall x$$

and $$\int_a^b f(x)dx = 1$$

$$\Rightarrow k\int_1^2 x dx = 1$$

$$\Rightarrow k\left[\frac{x^2}{2}\right]_1^2 = 1$$

$$\Rightarrow k\left[2 - \tfrac{1}{2}\right] = 1$$

$$\Rightarrow k = \frac{2}{3}$$

7.5 MATHEMATICAL EXPECTATION AND VARIANCE

7.5.1 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 $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ with respective probabilities $p_1, p_2, \ldots\ldots\ldots, p_n$, 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,

$$E(x) = \int_{-\infty}^{\infty} x f(x) dx \text{, provided } \int_{-\infty}^{\infty} f(x) dx = 1$$

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 $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ with respective probabilities $p_1, p_2, \ldots\ldots\ldots, p_n$.

Similarly, when x is a continuous random variable in the interval $(-\infty, \infty)$, with p.d.f. $f(x)$, $-\infty < x < \infty$, then,

$$E[g(x)] = \int_{-\infty}^{\infty} g(x) f(x) dx$$

7.5.2 Properties of Mathematical Expectation

Following are the important properties of mathematical expectation:

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 a 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).

7.5.3 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.

7.5.4 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. $V(x \pm y) = V(x) + V(y)$.

7.5.5 Illustrated Examples

Example 5: 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 x 0.1 + (-1) x 0.1 + 0 x 0.2 +1 x 0.2 + 2 x 0.3 +3 x 0.1

= 0.8

$V(X) = E(X^2) - \{E(X)\}^2$

Now, $E(X^2) = \sum_x x^2 p(x)$

= $(-2)^2$ x 0.1 + $(-1)^2$ x 0.1 + 0 + 1^2 x 0.2 + 2^2 x 0.3 + 3^2 x 0.1

= 2.8

$\therefore V(x) = 2.8 - (-0.8)^2$

= 2.8 - 0.64

= 2.16

Example 6: The p.d.f. of a r.v. X is

$f(x) = \frac{2}{3}x, \quad 1 \le x \le 2$,

find 9i) E(X), (ii) $P\left(\frac{5}{4} \le x \le \frac{7}{4}\right)$

Solution : We have, $f(x) = \frac{2}{3}x, \quad 1 \le x \le 2$

(i) $\therefore E(x) = \int_a^b f(x)dx = \frac{2}{3}\int_1^2 x.x \quad dx$

$$= \frac{2^1}{3}\int_1^2 x^2 \quad dx$$

$$= \frac{2}{3}\left[\frac{X^3}{3}\right]_1^2$$

$$= \frac{2}{3}\left[\frac{8}{3} - \frac{1}{3}\right]$$

$$= \frac{14}{9}$$

(ii) $$P\left(\frac{5}{4} \le x \le \frac{7}{4}\right)$$

$$= \int_{5/4}^{7/4} f(x)\ dx$$

$$= \frac{2}{3}\int_{5/4}^{7/4} x\ dx$$

$$= \frac{2}{3}\left[\frac{x^2}{2}\right]_{5/4}^{7/4}$$

$$= \frac{2}{3}\left[\frac{49}{32} - \frac{25}{32}\right]$$

$$= \frac{2}{3}.\frac{24}{32}$$

$$= \frac{1}{2}$$

Example 7: x is a random variable with p.d.f

$$f(x) = C(1-x), \qquad 0 < x < 1$$

fine (i) c (ii) E(3x+1)

Solution: Since f(x) is a p.d.f, we must have

$$f(x) \ge 0$$

and $$\int_a^b f(x)dx = 1$$

Now, $$\int_a^b f(x)dx = c\int_0^1 (1-x)dx = 1$$

$$\Rightarrow C\left[x - \frac{x^2}{2}\right]_0^1 = 1$$

$$\Rightarrow C\left[1 - \frac{1}{2}\right] = 1$$

$$\Rightarrow C.\frac{1}{2}=1$$

$$\Rightarrow C=2$$

$$\therefore f(x)=2(1-x),\ 0 \leq x \leq 1$$

Now, E(X) $= \int_a^b x.f(x)dx$

$$= 2\int_0^1 x.(1-x)dx$$

$$= 2\int_0^1 x.(x-x^2)dx$$

$$= 2\left[\frac{x^2}{2}-\frac{23}{3}\right]_0^1$$

$$= 2\left[\frac{1}{2}-\frac{1}{3}\right]$$

$$= 2.\frac{1}{6}$$

$$= \frac{1}{3}$$

$\therefore$ (i) E(3x+1) $= 3E(X)+1$

$$= 3.\frac{1}{3}+1$$

$$= 2.$$

Example 8: 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):	10/30	20/30
	=1/3	= 2/3

$\therefore$ E (an HP computer is selected) = 0x1/3 + 1x2/3

= 2/3

Example 9: 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 x 0.2 + 1 x 0.8

= 0.8

Example 10: 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 x 0.10 + 2 x 0.15 + 3 x 0.20 + 4 x 0.25 + 5 x 0.18 + 6 x 0.12

= 3.62

E (C) = E(10,000 + 500n) = 10,000 + 500 E(n)

= 10,000 + 500 x 3.62

= 11,810

Example 11: 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 per unit = Rs. 3

Unit demanded	Profit
10	Rs. (10x3) - 15 = rs. 15
11	Rs. (11x3) - 15 = Rs. 18
12	Rs. (12x3) - 15 = Rs. 21
13	Rs. (13x3) - 15 = Rs. 24
14	Rs. (14x3) - 15 = Rs. 27
15	Rs. (15x3) - 15 = Rs. 30

$\therefore$ Expected net profit = $\sum$ profit x probability

= 15x0.07 + 18x0.10 + 21x0.23 + 24x0.38 + 27x0.12 + 30x0.10

= Rs. 23.04



CHECK YOUR PROGRESS

Q 1: State whethre the following statemetns are true or false.

(i) The probability distribution of a random variable x is as follows

X = x:	1	2	3	
P(X= x):	1/4	1/2	3/4	(T/F)

(ii) If c is a constant then, (a) E(c) = c (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.

.....

.....

.....

.....

7.6 MOMENTS

Moments are a special type of measure, which can reveal the characteristic of a distribution. They are very much important in the sense that, 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 curve would make. Because of the great convenience in obtaining measures of the various characteristics of a frequency distribution, the calculation of the first four moments may well be made in the first step in the analysis of a frequency distribution. The letter 'μ' is used, generally, to denote the moments. Moments are called 'raw moments' when measured about any arbitary point 'a' and are called 'central moments' when measured about mean. We shall discuss them below.

Raw moments: The r[th] order raw moment about any point 'a' is,

$$\mu_r^{1} = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - a)^r, \qquad N = \sum_{i=1}^{n} f_i$$

Putting r = 0, 1, 2, 3,..... in the above expression, we get

$$\mu_0^{1} = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - a)^0 = \frac{1}{N}\sum_{i=1}^{n} f_i = 1$$

$$\mu_1^{1} = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - a) = \frac{1}{N}\sum_{i=1}^{n} f_i x_i - a\frac{1}{N}\sum_{i=1}^{n} f_i$$

$$= \bar{x} - a = d \text{ (say)}$$

..... and so on

Central Moments: The rth central moment about mean $\bar{x}$ is called the central moment of order r, and is given by,

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x-\bar{x}\right)^r$$

Putting r = 0, 1, 2, 3,..... in the above expression we get,

$$\mu_0 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x-\bar{x}\right)^0 = 1$$

$$\mu_1 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x-\bar{x}\right) = \frac{1}{N}\sum_{i=1}^{n} f_i x_i - \bar{x}\frac{1}{N}\sum_{i=1}^{n} f_i$$

$$= \bar{x} - \bar{x} = 0$$

$$\mu_2 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x-\bar{x}\right)^2 = \text{variance}$$

Now, we can express moments about mean in terms of moments about any point

We have, $\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x-\bar{x}\right)^r$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i\left\{\left(x_i-a\right)-\left(\bar{x}-a\right)\right\}^r$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i\left\{\left(x_i-a\right)-d\right\}^r$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i-a\right)^r - {}^rc_1\frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i-a\right)^{r-1}d +$$

$${}^rc_2\frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i-a\right)^{r-2}d^2 - \ldots\ldots\ldots\ldots + (-1)^r d^r$$

$$\therefore\ \mu_r = \mu_r' - {}^rc_1\mu_{r-1}'\mu_1' + {}^rc_2\mu_{r-2}'\mu_1'^2 + \ldots\ldots\ldots + (-1)^r\mu_1'^r$$

In particular,

$$\mu_2 = \mu_2' - 2\mu_1'\mu_1' + \mu_1'^2 = \mu_2' - \mu_1'^2$$

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

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

We can also express skewness and kurtosis in terms of moments as follows:

Skewness = $\beta_1 = \frac{\mu_3^{\,2}}{\mu_2^{\,3}}$ or $\nu_1 = \sqrt{\beta_1}$

Kurtosis = $\beta_2 = \frac{\mu_4}{\mu_2^{\,2}}$ or $\nu_2 = \beta_2 - 3$

Example 12:: The first three moments about the point x=10 are 5, 35 and 25 respectively. Final $\bar{x}, \mu_2, \mu_3$

Solution: If $\mu_r^{\,1} = \frac{1}{N}\sum_{i=1}^{n} fi(x_1 - a)^r$, we are given,

$$a = 10, \mu_1^{\,1} = 5, \mu_2^{\,1} = 35, \mu_3^{\,1} = 25$$

Now, $\mu_1^{\,1} = \bar{x} - a$

$$\therefore \bar{x} = \mu_1^{\,1} + a = 5 + 10 = 15.$$

$$\mu_2 = \mu_2^{\,1} - \mu_1^{\,1^2} = 35 - 25 = 10$$

$$\mu_3 = \mu_3^{\,1} - 3\mu_2^{\,1}\mu_1^{\,1} + 2\mu_1^{\,1^3}$$

$$= 25 - 3.35.5 + 2.5^3$$

$$= 25 - 525 + 250$$

$$= -250$$

Example 13: The first four moment of a distribution about the value 4 of the variables are -1.5, 17, -30 and 108. Find the moments about mean and β_1 and β_a.

Solution: If usual rotation we have,

$$a = 4, \mu_1^{\,1} = -1.5, \mu_2^{\,1} = 17, \mu_3^{\,1} = -20, \mu_4^{\,1} = 108.$$

Moments about mean,

$$\mu_2 = \mu_2^{\,1} - \mu_1^{\,1^2} = 17 - (1.5)^2 = 17 - 2.25 = 14.75.$$

$$\mu_3 = \mu_3^{\,1} - 3\mu_2^{\,1}\mu_1^{\,1} + 2\mu_1^{\,1^3} = -30 - 3.17.(-1.5) + 2.(-1.5)^3$$

$$= -30 + 76.5 - 6.75$$

$$= 39.75$$

$$\mu_4 = \mu_4^{\,1} - 4\mu_3^{\,1}\mu_1^{\,1} + 6\mu_2^{\,1}\mu_1^{\,1^2} - 3\mu_1^{\,1^4}$$

$$= 108 - 4.(-30).(-1.5) + 6.17.(-1.5)^2 - 3.(-1.5)^4$$

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

$$= 142.3125$$

$$\therefore \beta_1 = \frac{\mu_3^{\,r}}{\mu_2^{\,3}} = \frac{(39.75)^2}{(14.75)^3}$$

$$= 04926$$

$$\therefore \beta_1 = \frac{\mu_4}{\mu_2^{\,2}} = \frac{\mu_4}{\mu_\alpha^{\,2}} = \frac{142.3115}{(14.75)^2}$$

$$= 0.6543$$

Note:

1) For a symmetrical distribution moments of odd order are zero.

2) For Normal distribution β_1 =0 and β_2 =3.

3) First order raw moment about origin, i.e. $\mu_1^{/} = \frac{1}{N}\sum fx$ is called the mean of the distribution.



CHECK YOUR PROGRESS

Q 3: fill in the blanks

(i) Moments are called _____ moments when measured about any arbitary point a.

(ii) Moments are known as _____ moments when measured about mean.

(iii) Second order central moment μ_2 = _____.

(iv) For a symmetrical distribution all moments of odd order are _____.

(v) Skewness is measured in terms of moments as _____.

7.7 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** (Fig 7.2) and if the longer tail lies towards the left, it is said to be negatively skewed (Fig 7.3).

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 (Fig 7.1)

Figure 7.1-7.3: Different Types of Skewness

Figure 7.1: Symmetrical Distribution

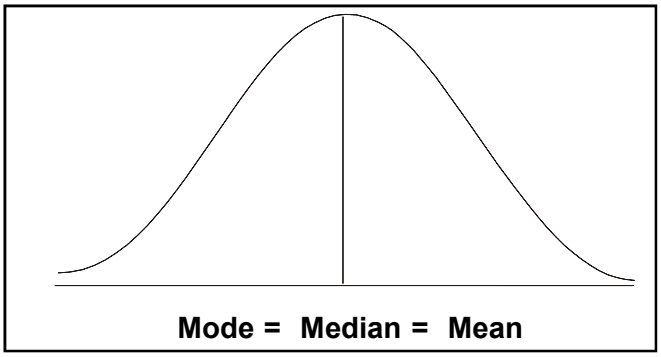


Figure 7.2: Positively Skewed Distribution

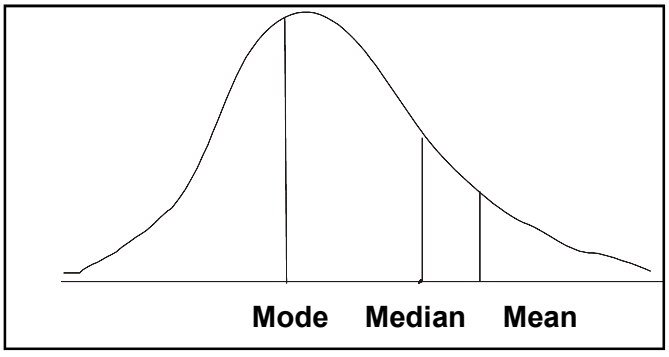
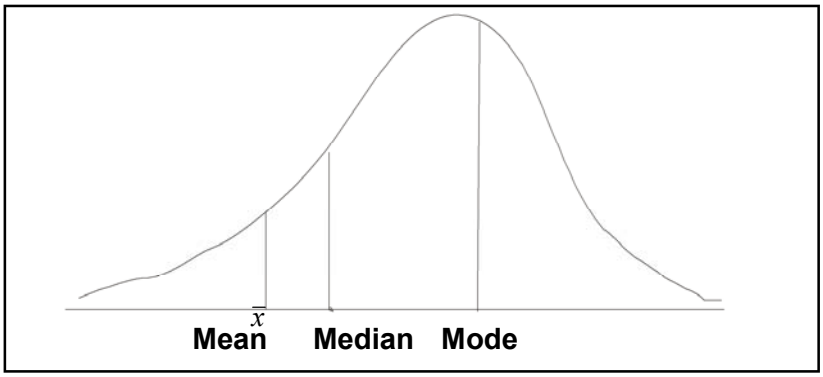


Figure 7.3: Negatively Skewed Distribution



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.

$$Q_3 - \text{Median} \neq \text{Median} - Q_1$$

7.8 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.

7.7.1 Karl Pearson's Co-efficient of Skewness

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

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

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

In that case,

$$Sk_p = \frac{3(Mean - Median)}{S.D.}$$

Remarks :

1. Theoretically, Karl Pearson's Co-efficient of Skewness lies between ± 3. But in practice these limits are 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 $\bar{x}$ > Median > Mode

$\therefore$ For a positively skewed distribution, the value of Mean is the greatest of the three measures and value of Mode is the least.

On the other hand, SK < 0 if Mode > Median > $\bar{x}$ mean

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

7.7.2 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

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

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 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 14 : 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,

$$SK_p = \frac{\bar{x} - Mode}{\sigma}$$

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

$\Rightarrow$ 3.2 = 30 – Mode

$\Rightarrow$ Mode = 30 – 3.2 = 26.8

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

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

$\Rightarrow$ 3.2 = 90 – 3 Median

$\Rightarrow$ 3 Median = 86.8

$\Rightarrow$ Median = 28.93

Example 15 : 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,

$$SK_p = \frac{3(\bar{x} - Median)}{\sigma}$$

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

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

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

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

= 49.84%

Example 16: 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.

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

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 17: 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 $(\bar{x}) = \frac{\Sigma x}{n} = \frac{452}{10_2} = 45.2$

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

$$= \sqrt{\frac{24270}{10} - (45.2)^2} = 19.59$$

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

Example 18: 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 19: 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)	Frequency	d = x –A	$d^/$=d/10	$fd^/$	$fd^{/2}$
(x)	(f)				
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 20: From the data given below, calculate Karl Pearson's co-efficient of Skewness.

Solution : Table for calculating 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

Ht (cm)	(f)	Mid value(x)	d = x–A	$d'=d/5$	fd'	fd'^2	Class Boundarie
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
	N= 100				$\Sigma fd'$ = 4	$\Sigma fd'^2$=208	

Let A = 147 Here i = 5

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

$$= 147 + \frac{4 \times 5}{100} = 147.2\ cm$$

$$\sigma = \sqrt{\frac{\Sigma fd^{12}}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$$
$$= \sqrt{\frac{208}{100} - \left(\frac{4}{100}\right)^2} \times 5$$
$$= \sqrt{2.08 - .0016} \times 5$$

= 7.21 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 cm

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

Example 21 : 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 13 29 46 60 112 94 45

50 and above 21

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
	N = 420	

We know that, Bowley's coefficiant of Skewness

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

Here N = 420

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

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

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

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

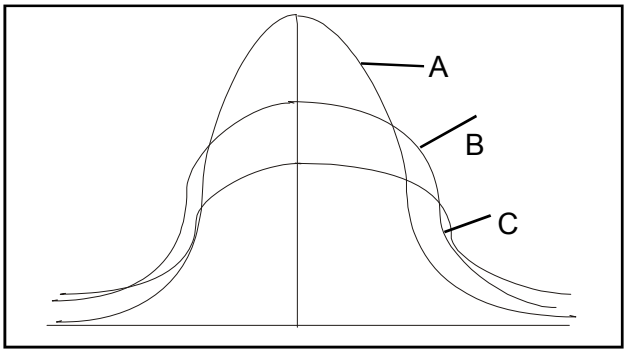
$$= \frac{-1.19}{11.51} = -0.103$$

7.8 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.

Figure 7.4: Lapto kurtic, meso kurtic and platy kurtic curves



A – Lapto Kurtic

B – Meso Kurtic

C – Platy Kurtic

Mean

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 curve and 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**

7.8.1 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:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{6^4}$$

$$\alpha_2 = \beta_2 - 3 = \frac{\mu_4}{6^4} - 3$$

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 22 : 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 $\mu_2 = 16$

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

$= 40 - 3 \times 2 \times 20 + 2\,(2)^3 = -64..7$

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

$= 50 - 4 \times 2 \times 40 + 6 \times 20 \times (2)^2 - 3(2)^4 = 162$

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

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

Example 23 : 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

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

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 24 : 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 4: Explain meaning of skewness using sketches of frequency curves. How does skewness different from dispersion.

Q 5: Distinguish between Karl Pearson's and Bowley's measure of skewness. Which one of these would you prefer and why?

Q 6: How is Kurtosis measured ?

.....

.....

.....

.....



7.9 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.
- Moments are a special type of measure which can measure the characteristic of the distribution - like the central tendency of a set of observations, their variability, symmetry, peakedness etc.
- Moments are of two types - Raw moments and Central momnets.
- Moments are termed as raw moments when measured about any arbitary point, say, a and central moments are those moments when measured about the mean of the distribution.
- Central moments can be expressed in terms of raw moments and vice versa.



7.10 FURTHER READING

1. Mukhopadhyay, P. (1996), *Mathematical statistics*. New Central Book Agency (P) Ltd.
2. Goon, A.M., Gupta, M.K. and Dasgupts, B. (1997), *An ouline of statistical Theory*. Vol. I, World Press.
3. Mukhopadhyay, P. (1991). *Theory of Probability*. New Central Book Agency (P) Ltd.
4. Gupta, S.C., Kapoor, V.K. (2009). *Fundamentals of Mathematical Statistics*, Sultan Chand & Sons.
5. Feller, W. (1968). *An Introduction to Probability Theory and its Application*, Vol I, Wiley Eastern.



7.11 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 which 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 then, 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.

Ans to Q No 3:

(i) raw (ii) central (iii) variance (i v)

zero (v) $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$

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



7.12 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.

(vii) Skewness is given by β_1 = _____.

(viii) μ_1' is the _____ of a distribution.

(ix) μ_2 is the _____ of a distribution.

Q 2: Can you find the mathematical expectation of a random variable x, if x takes the values 1,2,3 with corresponding probabilities 1/4, 1/2 and 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.

Q 7: Let x be a random variable with density function

$f(x) = c(1-x), \quad 0<x<1.$

Q 8: What are moments? Distinguish between raw moments and central moments.

Q 9: 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 10: 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 11: 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.

*** ***** ***

Name of the Paper: .....
Course Code: .....

Centre for Internal Quality Assurance (CIQA)
Krishna Kanta Handiqui State Open University
City Office: Housefed Complex, Guwahati-781006

Learner's Feedback on Course

Dear Learner,

Regarding the course as mentioned above, we would like to know your opinions and comments so as to improve the quality of self learning materials in future. Please respond to the following statements by ticking the number you feel most reflect your opinion. After completion of the additional comments, please detach the page and send/mail the same to us at the address given below.

The Director, Centre for Internal Quality Assurance, KKHandiqui State Open University
Housefed Complex, Dispur, Guwahati-781006
(E-mail id: ciqa@kkhsou.in)

1) Approximately how many hours did you spend for studying the units in the course? ☐☐☐

2) Please give your opinions (by ✓ mark) to the following items based on your reading of the block:

Sl. No.	Statements	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
I)	The SLMs of the course fulfil the learning objectives set out in the beginning of each unit					
II)	The units of the SLM could satisfy my academic needs and expectations					
III)	The Unit writers have excellent knowledge about the course contents					
IV)	Language and contents of the units were presented at a level which I could readily understand					
V)	Ample opportunity for participation in the activities provided in the units					
VI)	Used enough Illustrations (Diagrams, tables etc.) for conceptual clarity					
VII)	Quality of content is engaging, relevant, and up-to-date					
VIII)	The self check questions are very helpful					
IX)	The Possible/model questions and the answers to check my progress have benefited me a lot					

Additional Comments: *(Please feel free to provide your open comments)*

1) Which aspects of the SLM, according to you, worked well?

.....

.....

2) What sort of changes/improvements do you feel KKHSOU could implement to improve the overall quality of the SLM?

.....

.....

Thank you for taking the time to complete this form.