Department of Distance Education

Punjabi University, Patiala

Class : M.A. I (Economics) **Semester : 2**

Paper : III (Basic Quantitative Mathod) **Unit : I**

Medium : English

Lesson No.

1.1 : Geometric Mean, Harmonic Mean and Their Applications

1.2 : Measures of Dispersion

1.3 : Lerenz Curve and its Uses

1.4 : Skewness

1.5 : Probability

1.6 : Correlation Analysis

1.7 : Regression Analysis

Department website : www.pbidde.org

M.A. (ECONOMICS) PART - I **PAPER–III**

SEMESTER- II **(BASIC QUANTITATIVE METHODS)**

LESSON NO. 1.1 **Author : Dr. Vipla Chopra**

GEOMETRIC MEAN, HARMONIC MEAN AND THEIR APPLICATIONS

1.1.1 Introduction
1.1.2 Objectives
1.1.3 Geometric Mean (G.M.)
 1.1.3.1 Meaning
 1.1.3.2 Calculation of G.M.
1.1.4 Mathematical Properties of G.M.
1.1.5 Application of Geometric Mean
1.1.6 Harmonic Mean (H.M.)
 1.1.6.1 Meaning
 1.1.6.2 Calculation of H.M.
1.1.7 Uses of Harmonic Mean
1.1.8 Summary
1.1.9 Key Words
1.1.10 Further Readings
1.1.11 List of Questions
 1.1.11.1 Short Questions
 1.1.11.2 Long Questions

1.1.1 Introduction

In this lesson, we shall study only one feature of frequency distribution i.e. central tendency or averages. When a statistician groups mass of data in a frequency table, he finds a tendency towards centralisation. The items of the distribution show a tendency to cluster or concentrate around a particular value which is called the average or a measure of central tendency. An average is the measure which condenses a huge unwieldy set of numerical data into single numerical value which is representative of the entire distribution. There are various measures of central tendency but we are going to discuss only two such measures namely Geometric Mean and Harmonic mean. The specific uses of both these measures of central tendency have also been given.

1.1.2 Objectives

After studying this lesson you will be able to learn

- the meaning of Geometric Mean and meaning of Harmonic Mean.
- the procedures for calculating G.M. and H.M. for different series.
- the mathematical properties of G.M.
- Uses of both G.M. and H.M.

1.1.3 Geometric Mean (G.M.) :

1.1.3.1 Meaning :

The geometric mean of items a_1, a_2, ...a_n is defined as the n^{th} root of the product of 'n' items. Symbolically it can be expressed as

$$G = (a_1\ a_2\ a_3 \ldots\ldots\ldots\ldots\ldots\ldots a_n)^{1/n}$$

$$\therefore \quad G.M. = \sqrt[n]{(a_1) \times (a_2) \times (a_3) \ldots\ldots\ldots\ldots\ldots (a_n)} \qquad \ldots\ldots\ldots\ldots(i)$$

Where $a_1 a_2 a_3 \ldots\ldots\ldots\ldots\ldots\ldots a_n$ are n items in the series.

When there are two items $a_1 a_2$, geometric mean is the square root of the product of these items.

i.e. $G.M. = \sqrt{a_1 a_2}$

Similarly, when there are three items geometric mean is the cube root of the product of these items.

i.e. $G.M. = \sqrt[3]{(a_1) \times (a_2) \times (a_3)}$

When the number of items is three or more the calculation of geometric mean becomes difficult. In order to facilitate calculation logarithms are used.

Take logarithms of (i), we get

$$\text{Log G.M.} = \text{Log } (a_1 \times a_2 \times a_3 \ldots\ldots\ldots\ldots\ldots \times a_n)^{1/n}$$

$$\text{Log G.M.} = \frac{1}{n} (\log a_1 + \log a_2 + \log a_3 \ldots\ldots\ldots\ldots + \log a_n)$$

$$\text{Log G.M.} = \frac{\sum \log A}{n} \text{ (where } \sum\log A = \log a_1 + \log a_2 + \log a_3 \ldots\ldots\ldots\ldots + \log a_n)$$

Take anti log on both sides, we get

$$G.M. = AL \frac{\sum \log A}{n}$$

Hence, geometric mean is the anti log of the arithmetic average of the logs of the values of a variable.

1.1.3.2 Calculation of Geometric Mean (Individual Series) :

We know that if X_1, X_2, ... X_n are n items of a series then

$$G.M. = AL \frac{\sum \log A}{n}$$

Example 1 :

The monthly income of families in a locality are given below. Calculate Geometric Mean :

85, 15, 500, 45, 40, 70, 75, 08, 250, 36.

Solution:

Family	**X** **Income**	**Log X** **Logarithms**
1	85	1.9294
2	15	1.1761
3	500	2.6990
4	45	1.6532
5	40	1.6021
6	70	1.8451
7	75	1.8751
8	8	0.9031
9	250	2.3979
10	36	1.5563
N = 10		$\sum$log X = 17.6373

$$\text{G.M.} = \sqrt[n]{(x_1)\times(x_2)\times(x_3)\times(x_4)\ldots\ldots\ldots\ldots(x_n)}$$

$$= \sqrt[10]{85\times 15\times 500\times 45\times 40\times 70\times 75\times 8\times 250\times 36}$$

$$\text{Log G.M.} = \frac{1}{10}[\text{Log } 85 + \text{Log } 15 + \text{Log } 500 + \text{Log } 45 + \text{Log } 40$$

$$+ \text{Log } 70 + \text{Log } 75 + \text{Log } 8 + \text{Log } 250 + \text{Log } 36]$$

or G.M. = Anti Log of 17.6373/10 = Anti Log of 1.7637

G.M. = Rs. 58.08

In Discrete Series :

For calculation of G.M. in case of discrete series the following formula has been used

$$\text{G.M.} = \text{Anti Log } \frac{\left[\sum f \log x\right]}{N}$$

Example 2:

From the following data relating to income of 500 families determine geometric mean.

Income (Rs. per Month)	:	100	750	1000	1250	1500
No. of Families	:	100	150	75	90	85

Solution :

Income X **(Rs. per Month)**	**No. of Families** **f**	**Log X**	**f Log X**
100	100	2.0000	200.00
750	150	2.8751	431.265
1000	75	3.0000	225.000
1250	90	3.0969	278.721
1500	85	3.1761	269.9685
	Σf = 500		$\Sigma f \log X$ = 1404.9545

$$\text{G.M.} = \text{Anti Log } \frac{[\sum f \log x]}{N} = \text{Anti Log } \frac{[1404.9545]}{500}$$

G.M. = Anti Log (2.8099) = 645.5

Continuous Series :

$$\text{G.M.} = \text{Anti Log } \frac{[\sum f \log m]}{N}$$

Example 3:

From the following data compute the value of geometric mean:

Marks :	0-10	10-20	20-30	30-40	40-50
No. of Students :	8	12	20	6	4

Solution :

Income X	**No. of Students** **f**	**Mid Points** **m**	**Log m**	**f Log m**
0-10	8	5	0.6990	5.5920
10-20	12	15	1.1761	14.1132
20-30	20	25	1.3978	27.9560
30-40	6	35	1.5441	9.2646
40-50	4	45	1.6523	6.6128
	n = 50			$\Sigma f \log m$ = 63.5386

$$\text{G.M.} = \text{Anti Log } \frac{[\sum f \log m]}{N} = \text{Anti Log } \frac{63.5386}{50} = \text{A.L.}(1.2708) = 18.65$$

1.1.4 Mathematical Properties of G.M. :

Geometric Mean possesses very important properties, which add to its usefulness.

1. If each item is replaced by geometric mean the product of items

remains the same.

Geometric Mean of 5, 10 and $20 = \sqrt[3]{5 \times 10 \times 20} = 10$

If each item is replaced by 10 i.e. G.M. the product of items remains the same.

2. The products of the corresponding ratio on either the same side are always equal. If ratio of the geometric meant to the figures which are equal to or less then the G.M. are multiplied together, this product of the ratios of figures more than the geometric mean.

e.g. G.M. of 4, 5, 20, 25 = $\sqrt[4]{4 \times 5 \times 20 \times 25} = \sqrt[4]{10000} = 10$

Now, $\frac{10}{4} \times \frac{10}{5} = \frac{20}{10} \times \frac{25}{10}$ or $5 = 5$

This property establishes the role of geometric mean as a measure of relative changes.

Example 4:

If the price of one commodity increases from Rs. 4 to Rs. 64 and the price of another commodity decreases from Rs. 16 to Rs. 1, there is no relative change in the price levels. This is because decrease in price of one commodity is compensated by an increase in the price of other commodity.

Commodity	**Original Price**	**New Price**
A	4	64
B	16	1
$\bar{x}$	$\frac{20}{2} = 10$	$\frac{65}{2} = 32.5$
G.M.	$\sqrt{64} = 8$	$\sqrt{64} = 8$

Since there is no change in G.M. there is no change in the price level. Because of this property, the geometric mean is used while averaging ratios or percentages.

3. It is possible to calculate the combined Geometric Mean of two or more series if only their geometric means and the number of items are known.

If G_1, G_2, ...G_k, are the geometric means of K series of sizes n_1, n_2, ... n_k respectively, the Geometric means G of the combined series of sizes $n_1 + n_2 +$ ....$+n_k$ is given by

$$G = A.L. \left(\frac{n_1 LogG_1 + n_2 LogG_2 + ...+ n_k LogG_k}{n_1 + n_2 + ...+ n_k} \right)$$

4. The geometric mean of the ratio of corresponding observation in the two series is equal to the ratio of their geometric means.

X	**Y**	**X/Y**
3	2	1.50
6	4	1.50
8	4	2.00
9	8	1.12
G.M.= 6	4	1.50

5. If any value of series is zero, the value of geometric means is infinity and thus in appropriate.
   e.g. Let there be 3 items 15, 20 and 0. The geometric mean of these items would be $\sqrt[3]{15 \times 20 \times 0} = \sqrt[3]{0} =$ infinity, and hence inappropriate.

6. It cannot be calculated, if the number of negative values is odd. This is because, the product of the values will become negative and we cannot find out the root of a negative product viz.

   $\sqrt[4]{-3 \times -5 \times -4 \times 7} = \sqrt[4]{420}$

1.1.5 Application of Geometric Mean :

1. This is an important method of averaging ratios. Its main use is in the construction of Index number of prices.
2. If a series of figure is in geometric progression, then geometric mean will be the most suitable average.
3. Geometric Mean is recommended for finding average rate of growth or change over a period of time.
4. Geometric mean is useful in calculating the average rate of increase of any sum at compound interest or calculating the average rate of increase of a population. Let P_0 represents the principal at the beginning of a period, P_n the principle at the end of the period, r the rate of interest and n the number of years.

$$P_n = P_0(1+r)^n, \frac{P_n}{P_0} = (1+r)^n, \left(\frac{P_n}{P_0}\right)^{\frac{1}{n}} = 1 + r \text{ } and$$

$$r = n\sqrt{\frac{P_n}{P_0}} - 1$$

Example 5:

A sum of money at compound interest increases from Rs. 1000 to Rs. 1500 during a period of 10 years. Find rate of interest

Now $P_n = P_0\ (1+r)^n$

$P_n = 1500,\ P_0 = 1000,\ n = 10$

$1500 = 1000\ (1+r)^{10}$

Take Log

$\therefore$ $Log\ 1500 - Log\ 1000 = 10 Log\ (1+r)$

$3.1791 – 3.0000 = 10 Log\ (1+r)$

$10 Log\ (1+r) = 0.01761$

$\therefore$ $1+r = A.L.\ 0.1761$

$r = 0.041$ *or* 4.1%

Example 6 :

The price of a commodity increased by 12% from 1983 to 1984. 15% from 1984 to 1985 and 30% from 1985 to 1986. Find the average increase.

Solution :

Since we are required to average percentages the Geometric Mean is an appropriate measure.

Price at the End of the Year

Percentage	Assuming Price at the beginning 100	Log x
12	112	2.0492
15	115	2.0607
30	130	2.1139
		$\sum Log\ x = 6.2238$

$$\text{G.M.} = A.L.\left(\frac{\sum Log\, x}{n}\right) = A.L.\left(\frac{6.2238}{3}\right)$$

G.M. = AL (2.0746)

G.M. = 118.8

Average increase = 118.8 - 100 = 18.8%

The Geometric Mean is very useful and appropriate for averaging ratios, percentages and rates of change. It is also useful in determining the rate of growth and in constructing index numbers.

Example 7:

On 01.01.94 a T.V. set was purchased for Rs. 12000. On 31.12.98, it is valued at Rs. 6000. Find the average rate of depreciation of the T.V. Set.

Solution :

Let P, be the value at the beginning of the period 1997 i.e. Rs. 12000,

A, be the value at the end of the period 1998 i.e. Rs. 6000,

n, be the number of years the T.V. is used i.e. 5 years

and i, the rate of decrease of depreciation per unit i.e. $\frac{r}{100}$

By the formula of gradual decrease, we have

Thus $A = P(1-i)^n$

Substituting the given values in the above, we get,

$$6000 = 12000\left(1-\frac{r}{100}\right)^5$$

$$\Rightarrow \quad \left(1-\frac{r}{100}\right)^5 = \frac{6000}{12000} = \frac{1}{2} = 0.50$$

$$\Rightarrow \quad 1-\frac{r}{100} = \sqrt[5]{0.50}$$

$$= AL\ \frac{1}{5}\ (Log\ 0.50)$$

$$1-\frac{r}{100} = AL\ \frac{1}{5}\ \left(\bar{1}.6990\right)$$

$$= AL\ \frac{1}{5}\ \left(\bar{5}+4.6990\right)$$

$$1-\frac{r}{100} = AL\ \ \bar{1}.9398 = 0.8706$$

$$\Rightarrow \quad \frac{-r}{100} = 0.8706-1, \qquad r \Rightarrow \quad 0.1294 \times 100 = 12.94\%$$

Thus, the average rate of depreciation is 13% p.a.

Merits :

1. It is based on all the items of the series.
2. It is rigidly defined.
3. It is suitable for further algebraic treatment.
4. It measures relative changes.
5. It gives less weight to small items.
6. It also possesses the merit of samplying stability.
7. The use of Geometeric Mean make it easier to shift base year.

Demerits :

1. It is not easy to understand.
2. It's computation is relatively difficult.
3. It may be a value which does not exist in the series.
4. It may be necessary to give small weight to small items and more weight to large items. In such a case Geometric Mean is not suitable.
5. It is not a widely known average.
6. It cannot be used if any value of the variable is zero.
7. It gives imaginary value if any item in the data is negative.

Exercise 1 :

1. Define Geometric Mean.
2. State two properties of Geometric Mean.
3. Discuss merits and demerits of Geometric Mean.

1.1.6 Harmonic Mean (H.M.) :

1.1.6.1 Meaning :

The Harmonic Mean of a series of numbers is the reciprocal of arithmetic mean of the reciprocals of the individual numbers.

Symbolically :

For two items X_1 and X_2

$$H.M. = Reciprocal\ of\ \frac{\frac{1}{X_1}+\frac{1}{X_2}}{2}$$

$$H.M. = \frac{2}{\frac{1}{X_1}+\frac{1}{X_2}}$$

For n items = $X_1, X_2, \ldots\ldots\ldots\ldots\ldots\ldots X_n$

$$H.M. = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + \ldots\ldots\ldots\ldots\ldots + \frac{1}{X_n}}$$

In general, $H.M. = \frac{n}{\sum\left[\frac{1}{x}\right]}$

1.1.6.2 Calculation of Harmonic Mean :

Individual Observation :

If X_1, X_2, ... X_n are n items then

Example 8:

Calculate Harmonic Mean of the following data:

15, 25, 35, 45, 55

x	**1/x**
15	0.067
25	0.040
35	0.028
45	0.022
55	0.108
	$\sum\frac{1}{x} = 0.175$

$$H.M. = \frac{n}{\sum\frac{1}{x}} = \frac{5}{0.175} = 28.57$$

Discrete Series :

$$H.M. = \frac{n}{\sum f \times \frac{1}{x}} = \frac{n}{\sum\frac{f}{x}}$$

Example 9 :

From the following data compute the value of Harmonic Mean :

Marks :	10	20	30	40	50
No. of Students :	20	30	50	15	5

Solution :

Marks (x)	f	f/x
10	20	2.000
20	30	1.500
25	50	2.000
45	15	0.375
50	5	0.100
	N = 20	$\sum f/x$ = 5.975

$$H.M. = \frac{N}{\sum \frac{f}{X}} = \frac{120}{5.975} = 20.08$$

Continuous Series :

Where m = **mid** point of a class

Example 10:

Find the Harmonic Mean for the following distribution:

Class	Frequency
40-50	19
50-60	25
60-70	36
70-80	72
80-90	51
90-100	43

Solution:

Class	Mid Value (m)	f	$\frac{1}{m}$	$f \times \frac{1}{m}$
40-50	45	19	0.0222	0.4218
50-60	55	25	0.0182	0.4550
60-70	65	36	0.0154	0.5544
70-80	75	72	0.0133	0.9572
80-90	85	51	0.0118	0.6018
90-100	95	43	0.0105	0.4515
		N = 246		$\sum \frac{f}{m}$ = 3.4421

$$H.M. = \frac{N}{\sum \frac{f}{m}} = \frac{246}{3.4421} = 71.4$$

1.1.7 Uses of Harmonic Mean:

This average is useful in the cases where time, rate and price are involved. It is useful in calculating average speed at which a journey has been performed or the average price at which an article has been sold. It has been employed in the field of economic statistics in the measurement of price movements.

Example 11 :

An aeroplane flies around a square whose side is 100 miles long taking the first side 100 miles per hour, the second side 200 miles per hour, the third side 300 miles per hour and the fourth side 400 miles per hour. What is the average speed of aeroplane ?

Solution :

$$H.M. = \frac{4}{\frac{1}{100} + \frac{1}{200} + \frac{1}{300} + \frac{1}{400}} = \frac{4}{0.01 + 0.005 + 0.033 + 0.0025}$$

$$= \frac{4}{0.0208} = \frac{1}{0.0052} = 192 m.p.h.$$

Example 12 :

A car travels uphill at a speed of 27 km. per hour and while returning covers the distance at a speed of 4 km. per hour. Find the average speed.

Solution :

Speed (km/h)	**1/x**
27	1/27
45	1/45

$$H.M. = \frac{2}{\frac{1}{27} + \frac{1}{45}} = \frac{2}{0.03704 + 0.2222} = \frac{2}{0.05962}$$

Average speed is 33.75 km. per hour.

Example 13:

A certain store made profits of Rs. 5000, Rs. 10,000, Rs. 80,000 in 1995, 1996 and 1997 respectively. Determine the average rate of growth of this store's profits.

Solution :

Rate of growth of profits from 1995 to 1996 is $= \frac{10000}{5000} \times 100 = 200\%$

Rate of growth of profits from 1996 to 1997 is $= \frac{80000}{10000} \times 100 = 800\%$

The average rate of growth of store's profits from 1995 to 1997 is the geometric mean of 200 and 800 i.e.

Average Rate of Growth = $\sqrt{200 \times 800} = \sqrt{160000} = 400\%$

Merits :

1. Harmonic mean satisfies the test of rigid definition. Its definition is precise and its value is always definite.
2. It is based on all the observations in the given data.
3. It tends itself to algebraic manipulation.
4. It is not much affected by sampling fluctuations.
5. It gives greater importance to small items and as such a single big item cannot push up its value.
6. It measures relative changes and is extremely useful in averaging certain type of ratios and rates.

Demerits :

1. It is difficult to understand and calculate.
2. It is usually a value, which does not exist in a series.
3. It gives large weight to small items.
4. It's value cannot be computed when there are both positive and negative items in a series or when one or more items are zero.
5. Besides these limitations, the Harmonic Mean has very little practical application and is not a good representation of a statistical series, unless the phenomenon is such where small items have to be given a very heavy wieghtage.

Algebraic Properties :

1. H.M. can not be computed from a series if any of its values is zero. This is because the reciprocal of 0 does not exists i.e. $\frac{1}{0} = \infty$.

2. H.M. can be computed from a series with any number of negative values. Thus, the H.M. of the values -5 and -10 will be

$$\frac{2}{\frac{1}{-5}+\frac{1}{-10}} = \frac{-20}{3} = -6.67 .$$

3. For any series in which all the values are not equal nor any value is zero, the value of the H.M. is less than the G.M. and A.M.

Exercise 2:

1. Define Harmonic Mean.
2. What are the merits of Harmonic Mean ?
3. State any two algebraic properties of Harmonic Mean.

1.1.8 Summary

In this lesson we have studied about the two important measures of Central tendency namely geometric mean and harmonic mean. These measures of central tendency give one of the very important characteristics of data. The concepts of Geometric Mean and Harmonic Mean useful for specific type of applications have been understood with illustrations.

1.1.9 Key Words

Geometric Mean : Geometric Mean of N observations is the Nth root of the product of the given values of observations.

Harmonic Mean : Harmonic Mean of N observations is the reciprocal of the arithmetic mean of the reciprocals of the given values of N observations.

1.1.10Further Readings:

S.P. Gupta	:	Statistical Methods
S.C. Gupta	:	Fundamentals of Statistics
Murray R Spiegel	:	Statistics.

1.1.11List of Questions

1.1.11.1 Short Questions

1. What is Geometric Mean.
2. Discuss the mathematical properties of Geometric Mean.
3. Define Harmonic Mean.
4. What are the Specific uses of Harmonic Mean ?

1.1.11.2 Long Questions

1. Find the Geometric Mean for the following distribution :

Marks	:	0-10	10-20	20-30	30-40	40-50
No. of Students	:	5	7	15	25	8

2. A machine was purchased for Rs. 50,000 in 1984. Depreciation on the diminishing balance was charged @ 40% in the first year, 25% in the second year and 15% per annum during the next three years. What is the average depreciation charged during the whole period ?

3. A cyclist pedals from his house to his college at a speed of 10 Km.p.h. and back from the college to his house at 15 Km.p.h. Find the average Speed.

4. Find out Harmonic mean from the following :

   2574, 465, 75, 5, 0.8, 0.08, 0.005, 0.0009

5. In a certain factory, a unit of work in completed by A in 4 minutes, by B in 5 minutes, by C in 6 minutes, by D in 10 minutes and by E in 12 minutes. What is their average rate of working ? What is the average number of units of work completed per minute ?

M.A. (ECONOMICS) PART – I | **PAPER–III**
SEMESTER- II | **(BASIC QUANTITATIVE METHODS)**

LESSON NO. 1.2

MEASURES OF DISPERSION

STRUCTURE:

1.2.1 Introduction
1.2.2 Objectives
1.2.3 Meaning of Dispersion
 1.2.3.1 Absolute Dispersion
 1.2.3.2 Relative Dispersion
1.2.4 Purpose of Measuring Dispersion
1.2.5 Measures of Dispersion
 1.2.5.1 Range
 1.2.5.2 Inter-quartile range
 1.2.5.3 Quartile Deviation or semi-inter-quartile range
 1.2.5.4 Mean Deviation or Average Deviation
 1.2.5.5 Standard Deviation
1.2.6 Summary
1.2.7 Further Readings
1.2.8 List of Questions
 1.2.8.1 Short Questions
 1.2.8.2 Long Questions

1.2.1 Introduction:

Measures of Central Tendency aim at measuring the tendency of mass of data towards centralisation. Thus averages are summary figures. But an average may not be good respresentative figure and it may conceal many characteristics. Some items may be near the average, while others away from the average. In such a case, one must know the extent of the scatter of items around the average. In this lesson the study of measures of dispersion describes this important features of a frequency distribution.

1.2.2 Objectives :

After completion of this lesson you should be able to:

- ♦ Know the meaning of dispersion
- ♦ Learn the purpose of finding the dispersion
- ♦ Describe various types of dispersion

1.**2.3 Meaning of Dispersion :** The extent of the scatter of items around the average, to throw light on the composition of a series, is called dispersion. In a general sense, it refers to lack of uniformity in the size of items in a *series*. If the variation in a series is substantial, the dispersion is said to be considerable. If it is little, dispersion is insignificant. But in a precise sense *dispersion* not only gives an idea about the *variability* of a series, but also a precise measure of this variation. The *deviations* of size of items from the average are then averaged to give a single figure. This figure gives the dispersion of the series. Comparisons can then be made with similar figures representing other series.

Measures of dispersion are called "average of the second order" since they involve the averaging of the deviations of the various items in a series from their average. They are distinguished from the measures of central tendency (mean, mode, median etc.) which are described as "average of the first order."

Dispersion or variation or averages of the second order can be expressed in two ways:

1.2.3.1 Absolute Dispersion

It is called "absolute dispersion" if it is expressed in the unit in which the original data are expressed, say rupees, etc.

1.2.3.2 Relative Dispersion

It is relative dispersion if the measurement is expressed as a ratio or percentage. In this case, dispersion is not expressed in the original data.

1.2.4 Purpose of Dispersion: The average have a great utility in statistical analysis. But they have their own limitations. Even an ideal average can represent a series only "as a single figure can". They do not bring out the entire store of a phenomenon. There may be series :

(1) Whose averages may be identical but which may differ from each other in their respective formations. For example, the daily earnings of two persons A and B are as given below :

A	**B**
8	5
8	12
7	6
8	6
9	11

The average earnings of both persons are Rs. 8, but the scatter of the value of the items in the two series around their averages earnings is different.

(ii) Whose averages may differ in their numerical size, but their formations may be similar. Let us suppose that the two series X and Y are as follows :

X	Y
4	7
6	9
8	11

The averages are respectively 6 and 9, but the scatter of the items in the two series around their respective means in either series is only 2.

Thus the study of dispersion helps to throw more light on the composition of a series.

1.2.5 Measures of Dispersion : The measures of dispersion are obtained in two ways:

(a) By expressing the items as lying between certain limits.

(b) By averaging the differences between the individual measurements and a representative figure, i.e. the average.

Different methods used for computation of dispession are (a) The Range, (b) The Inter-quartile Range, (c) Quartile Deviation, (d) the Mean Deviation, (e) Standard Deviation.

1.2.5.1The Range

The range is the simplest possible measure of dispersion. It is defined as the difference between the largest and the smallest values of a distribution or series. According to W.I. King, "The dispersion of a group may be measured by the difference in size or characteristic of the most extreme items. In other words, the range may be measured by the general deviation of the items from the type." If range is divided by the sum of the extreme items, the resulting figure is called "the ratio of the range" or "the co-efficient of the scatter." If the marks of the students in a particular examination vary between the limits of 81 and 33, the range is 81-33=48. The figure indicates the variability in the marks of students symbolically.

Range : $R = R_1 - R_2$

$$\text{Co-efficient of Range} = \frac{R_1 - R_2}{R_1 + R_2}$$

$$= \frac{81-33}{81+33} = \frac{48}{114} = 0.42$$

Where R_1 = the greatest value of the variable

R_2 = the smallest value of the variable.

Merits of Range : (i) Range is the simplest method of measuring dispersion. In the computation of range it is not necessary to know the value of each item. It can be computed if only the values of extreme items are known. Its chief merit, therefore, is that it can be easily calculated and understood.

(ii) It provides a ready measure of comparison with the help of its coefficient.

Demerits of Range : The range suffers from the following short-comings :

(i) The value of range is entirely dependent on two extreme values, i.e. the lowest and the highest items. The value of the range varies from sample to sample and is never stable. Since the range depends entirely on the size of extreme items, it is not a satisfactory measure of dispersion.

(ii) It does not take into consideration all the items. It is not based on all the observations of a series. It is an unsatisfactory measure of dispersion and should be used with caution.

(iii) It is only a rough measure of dispersion.

In the words of W.I. King, range as a measure of dispersion is too indefinite to be used as a practical measure of dispersion.

Uses of Range : Range as a measure of dispersion is commonly used in some fields.

It is commonly used in those fields where the variations are not much. Its use is popular in quality control of manufactured goods. It is also used for measuring fluctuations in prices or shares or fluctuations in interest rates. But it must be remembered that range is a very rough measure of dispersion and is unsuitable for precise and accurate studies.

1.2.5.2 Inter-Quartile Range : In case of range, the difference of extreme items is found. But if the difference in the values of two quartiles calculated, it is called inter-quartile range. As a measure of dispersion, it is better than range because it is not affected by the values of the extreme items. The inter-quartile range gives a fair measure of variability as 50% of the values of a variable are between the two quartiles.

However, inter-quartile range suffers from the same defects from which range suffers, e.g. it is also affected by fluctuations of sampling. It is also not based on all the value. It ignores the composition of a series, but it is easy to calculate and understand.

In the inter-quartile range, it is generally the difference between the third quartile and the first quartile symbolically.

Inter-quartile range $=Q_3-Q_1$

Example 1. Given below are the marks obtained by a group of 20 students in a test in English and Hindi.

Roll No. of Students

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,
12, 13, 14, 15, 16, 17, 18, 19, 20

Marks in English

53, 54, 52, 32, 30, 60, 47, 46, 35, 28, 25,
42, 33, 48, 72, 51, 45, 33, 65, 29

Marks in Hindi

58, 55, 25, 32, 26, 85, 44, 80, 33, 72, 10,
42, 15, 46, 50, 64, 39, 38, 30, 36.

Find Inter-Quartile range for marks in English and Hindi

Solution : Arrange the data in two series to get the quartiles.

Sr. No.	Marks in English	Marks in Hindi
1.	25	10
2.	28	15
3.	29	25
4.	30	26
5.	32	30
6.	33	32
7.	33	33
8.	35	36
9.	42	38
10.	45	39
11.	46	42
12.	47	44
13.	48	46
14.	51	50
15.	52	55
16.	53	58
17.	54	64
18	60	72
19.	65	80
20.	72	85

(a) English

Q_1 (i.e. the score of the 5th student) is 32.

Q_3 (i.e. the score of the 15th student) is 52.

The Inter-quartile range is : Q_3—Q_1

or 52-32=20

(b) Hindi

As before, Q_1 here is 30

and Q_3 is 55

Inter-quartile range is 55-30 =25

1.2.5.3 Quartile Deviation or Semi-inter-Quartile range: Here the measure of dispersion is based on the lower and upper quartiles divided by two. It is defined as the arithmetic average of the difference of two quartiles, Q_1 and

Q_3

Quartile Deviation or QD $= \frac{Q_3 - Q_1}{2}$

Co-efficient of QD $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$

Example 2 : Calculate the semi-inter-quartile Range and its co-efficient for the marks of 59 students in Economics :

Marks-group	No. of students
0-10	4
10-20	8
20-30	11
30-40	15
40-50	12
50-60	6
60-70	3

Solution : Computation of the semi-inter-Quartile Range :

Marks-group	No. of Students	Cumulative Frequency
0-10	4	4
10-20	8	12
20-30	11	23
30-40	15	38
40-50	12	50
50-60	6	56
60-70	3	59

First Quartile or

Q_1 = the marks of the 59/4, i.e. 14.75th student which lie in the 20-30 marks group.

By interpolation, $Q_1 = 20 + \frac{30-20}{11}(14.75-11) = 22.5$ marks

Third Quartile or

Q_3 =the marks of the $\frac{3(59)}{4}$ i.e. 44.25th student which lie in the 40-50 marks group.

By interpolation, $Q_3 = 40 + \frac{50-40}{12}(44.25\text{-}38) = 45.2$

Semi-inter-Quartile Range $\frac{Q_3 - Q_1}{2} = \frac{45.2 - 22.5}{2} = 11.35$

Co-efficient of the Semi-Inter-Quartile Range = $\dfrac{Q_3 - Q_1}{Q_3 + Q_1} = \dfrac{45.2 - 22.5}{45.2 + 22.5}$ = 0.335

Median or M = $\dfrac{59}{2}$ i.e. marks of the 29.5th student.

lies in 30-40 group

Med. Marks = $30 + \dfrac{(40-30)}{15}(29.5-23)$

= 34.33 marks.

In the above example, the quartile deviation is 11.35 marks. It these marks are added to the lower quartile, the resulting figure would be 33.85, and if they are subtracted from the upper quartile, it will again be 33.85 The actual value of the median is 34.33. It shows that the series is not perfectly symmetrical, though the departure from symmetry is not much. It however, shows that the dispersion of items on the two sides of the median is almost equal.

Merits of Quartile Deviation : The quartile deviation and its co-efficient have to their credit the merits of simplicity. The results are highly satisfactory if the main body of the array only is dealt with and extreme variations are ignored.

Demerits : Quartile Deviation is not based on all the observation of the data. Fluctuations of sampling affect it to great extent. It is not an accurate and precise measure of dispersion.

1.**2.5.4The Mean Deviation or Average Deviation**

The range-methods of dispersion suffer from the fact that they are calculated by taking into account only two values of a series-either the extreme values as in the case of range or values of quartiles as in the case of quartile deviation. They ignore other values of series. These methods are described as “methods of limits”. These methods ignore the composition of the series. This drawback is over-come by the method of averaging deviations in which variations of items are calculated from an average to throw light on the composition of the series and the dispersal of items round a central value.

The Average Deviation or Mean Deviation of a series is the arithmetic average of the deviations of various items from a measure of central tendency-mean, median or mode. It takes all the deviations as positive. It is the numerical sum of the deviations divided by the number of items. The signs of the various deviations are ignored as the sum of the deviations of the items from their mean is always zero.

The mean deviation is sometimes computed in relation to the median or mode. In practice, however, it is calculated from the mean. This is done because it is easier to calculate it from the mean than from the median. Calculation from the mode is uncommon.

Using symbols

(i) Mean Deviation from the mean, $d_a = \frac{\sum|da|}{n}$

Where $\sum da$ = sum of the deviations of the items from their arithmetic mean and n for the number of items.

(ii) Mean Deviation from the median, $d_m = \frac{\sum|dm|}{n}$

Where $\sum dm$ stands for the deviation from the median.

(iii) Mean Deviation from the mode, $d_z = \frac{\sum|dz|}{n}$

Where $\sum|dz|$ stand for the deviation from the mode.

Mean deviation is an absolute measure of dispersion, expressed in the units in which the original data are collected. To convert it into a relative measure, it is divided by the average from which it has been calculated. It is then known as the "Mean Co-efficient of dispersion." Thus the mean Co-efficient of dispersion from mean, median and mode would be respectively :

$$\frac{\delta a}{\text{Mean}}, \frac{\delta m}{\text{Median}}, \frac{\delta z}{\text{Mode}}$$

Methods of Calculation

(a) Ungrouped data

(i) Find the deviations of the items from the mean or median.

(ii) Add all the deviation.

(iii) Divide the sum obtained by n to get the mean deviation.

Example 3 : Calculate mean deviation from median for the following items :

54, 71, 52, 57, 72, 49, 45

Solution :

We arrange the given items and get :

45, 49, 52, 54, 57, 71, 72.

Therefore, median is the size of the fourth item, i.e., median is 54. We now find deviations from 54.

Items	d (signs ignored)
45	9
49	5
52	2
54	0
57	3
71	17
72	18
	$\lvert d_m \rvert = 54$

Mean Deviation from median

$$\text{or } d_m = \frac{\sum|d|}{n} = \frac{54}{7} = 7.71$$

Co-efficient of mean deviation $= \frac{\delta m}{\text{median}} = \frac{7.71}{54} = 0.142$

(b) **Grouped data : (i) Discrete series.**

(i) Obtain the deviation of size of item from the mean or median.

(ii) Multiply the deviations by their corresponding frequencies.

(iii) Add the products.

(iv) Divide the product by the total number of observations.

(i) $\delta m = \frac{\sum f\,|dm|}{n}$ (ii) $\delta a = \frac{\sum f\,|da|}{n}$

Example 4 : Find mean deviation of the distribution given below :

No. of accident	Persons having said number of accidents
0	15
1	16
2	21
3	10
4	17
5	8
6	4
7	2
8	1
9	2
10	2
11	0
12	2
Total	100

Solution : Calculation of mean deviation :

No. of accidents (m)	**Persons having said no. of accidents (f)**	**Deviation from Median (±) signs ignored (|dm|)**	**Total Deviation (f |dm|)**
0	15	2	30
1	16	1	16
2	21	0	0
3	10	1	10
4	17	2	34
5	8	3	24
6	4	4	16
7	2	5	10
8	1	6	6
9	2	7	14
10	2	8	16
11	0	9	0
12	2	10	20
	n = 100		$\sum f\,\lvert dm\rvert = 196$

The value of median = 2

Mean deviation or $= \frac{\sum f|dm|}{n} = \frac{196}{100} = 1.96$ accidents.

(c) Continuous series :

The calculation of mean deviation in case of continuous series is done by the same procedure as is used in discrete series.

Example 5 : Calculate the mean deviation (from median) for the following data :

Class interval	**Frequency**	**Class interval**	**Frequency**
1-3	6	9-11	21
3-5	53	11-13	16
5-7	85	13-15	4
7-9	56	15-17	4

Solution : The median of the above series is 6.5.

Class interval	**Mid points**	**Deviations from actual median 6.5 dm**	**Frequency (f)**	**Deviation Frequency**
1-3	2	4.5	6	27.0
3-5	4	2.5	53	132.5

5-7	6	0.5	85	42.5
7-9	8	1.5	56	84.0
9-11	10	3.5	21	73.5
11-13	12	5.5	16	88.0
13-15	14	7.5	4	30.0
15-17	16	9.5	4	38.0
Total			245	515.5

$$\text{Mean deviation} = \frac{\sum f|dm|}{n} = \frac{515.5}{245} = 2.1$$

Characteristics of Deviation

(1) It takes every item into consideration.

(2) It is easy to calculate and understand.

(3) It is not affected much by the values of extreme items.

(4) It ignores the algebraic signs of the deviations, and as such, it is not capable of further mathematical treatment.

(5) Mean deviation is not a very accurate measure of dispersion, especially when it is computed from the mode, because mode can be unrepresentative. Even when it is computed from median, it cannot be fully relied upon, because, if the degree of variability in a series is high, median is also an unrepresentative average.

(6) It gives weight to deviations according to their size, extreme deviations having more weight than small ones, but not being disproportionately magnified.

Uses : Mean deviation has found favour with businessmen and economists due to simplicity in calculations and also due to the fact that standard deviation gives more importance to extreme values. The co-efficient of mean deviation is good one to use in a large number of economic studies, such as computation of personal distribution of wealth in a community, since the very rich and the very poor both get their due consideration.

1.2.5.5 Standard Deviation : Since the calculation of mean deviation ignores the algebraic signs, it is mathematically unsound. The calculation of standard deviation form of average deviation from the mean. It is based on all the values in a distribution. Here we first find out the sum of the squares of the deviation from the mean and the divide it by the number of observations, and the square root of this number is defined as the standard deviation.

The deviations are calculated from the mean because the sum of the squares of deviation is minimum from mean. It is an improvement on the mean deviation in so far as the plus and minus signs before the deviation becomes positive when squared.

Methods of Calculation :

(a) Ungrouped data :-

(i) Calculate the deviation of the items from the mean.

(ii) Square the deviations.

(iii) Add these squares.

(iv) Divide the sum by the total number of items to get the variance.

(v) Take the square root of the quotient to get the required standard deviation Symbolically.

Standard Deviations or σ (Sigma) = $\sqrt{\frac{\sum d^2}{n}}$

Where σ stands for the S.D. (Standard Deviation), $\sum d^2$ for the sum of the squares of the deviations measured from the arithmetic average, and n for the number of items.

Example 6 : Find the standard deviation of the following set of observation: 58, 59, 60, 53, 66, 66, 75, 52, 69, 52.

Solution :

Sr. No.	**X**	**deviation from the mean 61**	**d^2**
1	58	-3	9
2	59	-2	4
3	60	-1	1
4	53	-8	64
5	66	+5	25
6	66	+5	25
7	75	+14	196
8	52	-9	81
9	69	+8	64
10	52	-9	81
	Total = 610		$\sum d^2$=550

Arithmetic mean, $\overline{X} = \frac{610}{10} = 61$

$$\sigma = \sqrt{\frac{\sum d^2}{n}} = \sqrt{\frac{550}{10}}$$

= 7.416 Answer

Co-efficient of Standard deviation $= \frac{\sigma}{\overline{X}} = \frac{7.416}{61} = 0.121$

(b) Grouped data

(i) Calculate the deviation of the mid-points of the group from the arithmetic mean.

(ii) Multiply the deviations with their corresponding frequencies.

(iii) Multiply the products in (ii) by deviations once again to get the products of the squares of deviations by the frequencies of the corresponding groups (fd^2).

(iv) Sum the products $\left(\sum fd^2\right)$ and divide the sum by the number of items $\left(\sum f\right)$ to get variance.

(v) Find the square root of variance to get the standard deviation.

Symbolically, $(\text{SD or}) = \sqrt{\frac{\sum fd^2}{n}}$

Example : Calculate the standard deviation from the following frequency distribution:

Class interval	Frequency
1-3	40
3-5	30
5-7	20
7-9	10

Solution :

Class interval	**mid Value**	**Deviation from assumed mean 4**	**frequency f**	**fd**	**fd^2**
1-3	2	-2	40	-80	160
3-5	4	0	30	0	0
5-7	6	2	20	40	80
7-9	8	4	10	40	160
			100	0	400

Arithmetic mean $= A + \frac{\sum fd}{n}$

$$= 4 + \frac{0}{100} = 4$$

Standard deviation $=\sqrt{\frac{\sum fd^2}{n}}$

$$=\sqrt{\frac{400}{100}} = 2$$

Co-efficient of standard deviation $=\frac{\sigma}{\overline{X}}=\frac{2}{4}=0.5$

Variance : It is the square of standard deviation. It is the arithmetic mean of the squared deviations about the arithmetic mean of a distribution.

$$\text{Variance} = \sigma^2$$

$$\text{or } \sigma = \sqrt{(\text{var}\,iance)}$$

Characteristics of Standard Deviation

(1) Standard deviation is affected by the value of every item in a distribution.

(2) The standard deviation gives greater weightage to extreme values because of the process of squaring.

(3) The standard deviation has a definite relationship with the area of the normal curve.

(4) The standard deviation is not very much affected by sampling fluctuations and has, therefore, greater stability than any other measure.

(5) It is amenable to algebraic treatment and possesses many mathematical properties.

(6) Standard deviation is not easy to calculate nor it is easily understood.

Use : The standard deviation should always be used as a measure of dispersion unless there is good reason to use another measure.

(f) Co-efficient of Variation : The co-efficient of standard deviation multiplied by 100 gives the co-efficient of variation.

Symbolically,

i.e. $\text{C.V.}=\frac{\text{Standard deviation}}{\text{arithmetic mean}}\times 100 \qquad C.V.=\frac{\sigma}{\overline{X}}\times 100$

This measure or co-efficient is used for comparing the dispersion between two series.

1.2.6 Summary

In this lesson different measures of dispersion have been used to examine the representativeness of an average, range, quartile deviation and mean deviation are mathematically convenient to handle. A comparison of various measures of dispersion indicate that standard deviation is by far, the best measure of dispersion.

1.2.7 Further Readings

1. G.S. Bhalla and Others: Elementary Statistics, Published by NCERT, New Delhi.
2. Samuel Hays: An Outline, Statistics.

1.2.8 List of Questions

1.2.8.1 Short Questions:

(a) Define dispersion.

(b) Explain the difference between absolute and relative measure of dispersion.

(c) What is coefficient of variation.

(d) What are quartiles?

(e) Explain the term standard deviation

1.2.8.2 Long Questions

(a) Explain the term 'dispersion'. What are the various measures of dispersion?

(b) Find quartile deviation for the following data :

Class interval	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	4	15	28	16	6	4

(c) Calculate mean deviation for the following frequency distribution:

Size of item	6	7	8	9	10	11	12
Frequency	3	5	7	12	8	4	2

(d) Find the standard deviation for the following age-distribution of 120 children :

Age (years)	0-2	2-4	4-6	6-8	8-10	10-12	12-14
No. of children	7	16	23	37	22	13	2

(e) Write notes on :

(a) Dispersion

(b) Merits and demerits of mean deviation.

M.A. (ECONOMICS) PART - I **PAPER–III**

SEMESTER- II **(BASIC QUANTITATIVE METHODS)**

LESSON NO. 1.3 **Author : Dr. Vipla Chopra**

LORENZ CURVE AND ITS USES

Structure

1.3.1 Introduction

1.3.2 Objectives

1.3.3 Significance of Measuring Variation

1.3.4 Properties of a good measure of Variation

1.3.5 Absolute and Relative measures of variation

1.3.6 Lorenz Curve

- **1.3.6.1 Meaning**
- **1.3.6.2 Procedure/Steps in drawing Lorenz Curve**
- **1.3.6.3 Illustrations**
- **1.3.6.4 Merits and Demerits**

1.3.7 Summary

1.3.8 Key Words

1.3.9 Suggested Readings.

1.3.10List of Questions.

- **3.10.1Short Questions**
- **3.10.2Long Questions**

1.3.1 Introduction :

In the previous lesson, we discussed some of the measures of central value. These measures used to provide a single representative value of a given set of data. This single value alone cannot adequately describe a set of data. In two or more distributions the central value may be the same but still there can be wide disparities in the formation of distribution. Thus the measures of central tendency must be supported and supplemented by some other measures. One such measure is Dispersion. In the present lesson we study dispersion to have an idea of the homogeneity (compactness) or heterogeneity (scatter) of the distribution.

A measure of dispersion (or variation) describes the spread or scattering of the individual values around the central value. To illustrate the concept of variation, let us consider the data given below:

Firm A **Daily Sales (Rs.)**	**Firm B** **Daily Sales (Rs.)**	**Firm C** **Daily Sales (Rs.)**
2000	**2050**	**2200**
2000	**1950**	**2800**
2000	**2000**	**1000**
$\overline{X}_A = 2000$	$\overline{X}_B = 2000$	$\overline{X}_c = 2000$

Since the average sales for firms A, B and C is the same, we are likely to conclude that the distribution pattern of the sales is similar. But a close examination shall reveal that distribution differ widely from one another.

1) **"Dispersion is the measure of the variation of the items." .....A.L.Bowley.**

2) **"Dispersion is a measure of the extent to which the individual items vary." ........L.R. Connor.**

1.3.2 Objectives :

After going through this lesson, you will be able to learn :

* the concept and significance of measuring variability.
* properties of a good measure of variation.
* the concept of absolute and relative variation.
* the concept of Lorenz Curve.
* the procedure in drawing Lorenz Curve.
* merits and demerits of Lorenz Curve.

1.3.3 Significance of Measuring Variation :

Measuring variation is significant for some of the following purposes :

i. To determine the Reliability of an Average.

ii. To control the variation of the data from the Central Value.

iii. To compare two or more distributions with regard to their variability.

iv. To facilitate the use of other Statistical Measures.

1.3.4 Properties of a good measure of variation :

i. It should be rigidly defined.

ii. It should be easy to calculate and easy to understand.

iii. It should be based on each and every item of the distribution.

iv. It should not be affected much by extreme observations.

v. It should have sampling stability.

1.3.5 Absolute and Relative measures of variation :

Measures of variation may be either absolute or relative. Measures of absolute variation are expressed in terms of the original data. Relative measures of dispersion are obtained as ratios or percentages and are thus pure numbers independent of the unit of measurement. Some of the well known measures of variation which provide a numerical index of the variability of the given data are Range, Mean Deviation, Quartile deviation and Standard Deviation. Lorenz Curve is a graphical method of measuring variation which is our main concern.

Exercise 1

1. What is dispersion ?

Ans ....

....

2. What purpose does a measure of dispersion serve ?

Ans ....

....

3. Distinguish between absolute measure and relative measure of dispersions.

Ans ....

....

1.3.6 Lorenz Curve :

The statistical devices which seek to measure concentration of various types of distributions are called measures of concentration. The most common among these devices is the Lorenz Curve.

Max O. Lorenz, a famous economic statistician of England devised a graphic method of studying dispersion, which was known after his name as Lorenz Curve. It was first used by him for the measurement of economic inequalities such as in the distribution of income and wealth between different countries or between different periods of time.

1.3.6.1 Meaning :

Lorenz Curve is a cummulative percentage curve in which the percentage of items is combined with the percentage of other things as wealth, profit, wages, turnover, production, population etc. It is a graphic method of showing the extent of variation in the size distribution from equal distribution.

1.3.6.2 Steps/Procedure in drawing Lorenz Curve :

A very distinctive feature of the Lorenz Curve consists in dealing with the cumulative value of the variable and cumulative frequencies rather than its absolute values and the given frequencies. The technique of drawing the curve is fairly simple and consists of the following steps.

(1) Cummulate the values of the variable and corresponding frequencies.

(2) Take the total variable as 100 and express the cummulated values of the variable as per centage of the total, i.e.,

$$\frac{\text{Cumulated value} \times 100}{100}$$

(3) As in step (2) express every cumulated frequency as the percentage of the total, i.e.,

$$\frac{\text{Cumulated frequency} \times 100}{\text{Total frequency}}$$

(4) Now take the coordinate axes, X-axis representing the percentages of the cumulated frequencies (X) and (Y)-axis representing the percentages of the cumulated values of the variable (Y). Both X and Y take the values from 0 to 100 as shown in the fig. 1.

(5) Draw the diagnal line Y=X Joining the origin O (0, 0) with the point P (100, 100) as shown in the diagram. The line OP will make an angle of 45° with the X-axis and is called the line of equal distribution/line of equality. In case of equal distribution, 10 per cent of the firms will employ 10 per cent of total persons, 20 per cent of firms will employ 20 per cent of persons etc. i.e. Any point on this diagnal shows that same per cent on X as on Y.

(6) Plot the percentages of the cumulated values of the variable (Y) against the percentages of the corresponding cumulated frequency (X) for the given distribution and join these points with a smooth freehand curve. Obviously, for any given distribution this curve will never cross the line of equal distribution OP. It will always lie below OP unless the distribution is uniform in which case it will coincide with OP. The greater the variability, the greater is the distance of the curve from OP.

Thus when the distribution of items is not proportionately equal, the variability (dispersion) is indicated and the curve is farther from the line of equal distribution OP. The greater the variability, the greater is the distance of the curve from OP. Thus the Lorenz Curve which departs from the line of equal distribution shows the extent of dispersion.

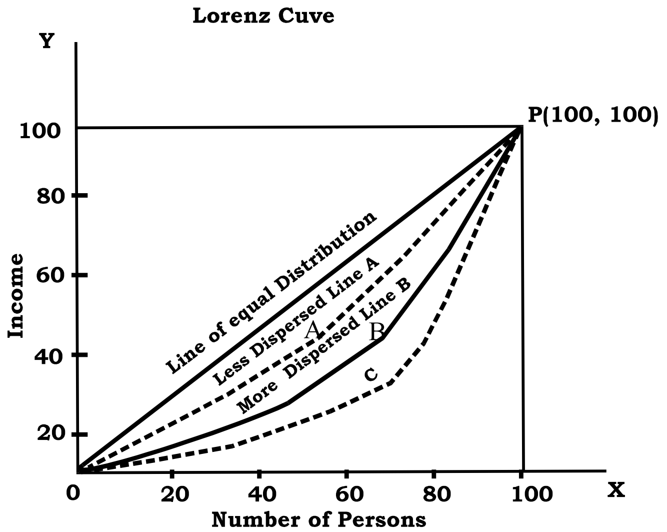


Fig.1

In Fig. 1, OP is the line of equal Distribution of Income. If the plotted cumulative percentages lie on this line, there is no variability in the distribution of income of persons. The points lying on the curve OAP indicate a less degree of variability as compared to the points lying on the curve OBP. Variability is still greater, when the points lie on the curve OCP. Thus a measure of variability is still greater, when the points lie on the curve OCP. Thus a measure of variability of the distribution is provided by the distance of the curve of the cumulated percentages of the given distribution from the line of equal distribution.

Exercise 2

1. What is Lorenz Curve ?

Ans.. ....

....

2. Explain the method of drawing the Lorenz Curve ?

Ans.. ....

....

1.3.6.3 Illustrations :

Example 1: From the following data relating to the purchases made by the customers of the two different localities, study the dispersions by means of the Lorenz Curve.

Purchase in'ooo Rs.	6	25	60	84	105	150	170	400
No. of Customers :								
Locality X :	6	11	13	14	15	17	10	14
Locality Y :	2	38	52	28	38	26	12	4

Computation of the Lorenz Curve

Purchases				Locality X			Locality Y		
	Amou-nt in '000 Rs. X	Cum of X	Cum % of X	No. of Custo-mers *f*	C*f*	C*f* %	No. of Custo-mers F	C*f*	C*f*%
	6	6	0.6	6	6	6	2	2	1
	25	31	3.1	11	17	17	38	40	20
	60	91	9.1	13	30	30	52	92	46
	84	175	17.5	14	44	44	28	120	60
	105	280	28.0	15	59	59	38	158	79
	150	430	43.0	17	76	76	26	184	92
	170	600	60.0	10	86	86	12	196	98
	400	1000	100.0	14	100	100	4	200	100
Total	**1000**		**100**	**100**		**100**	**200**		**100**

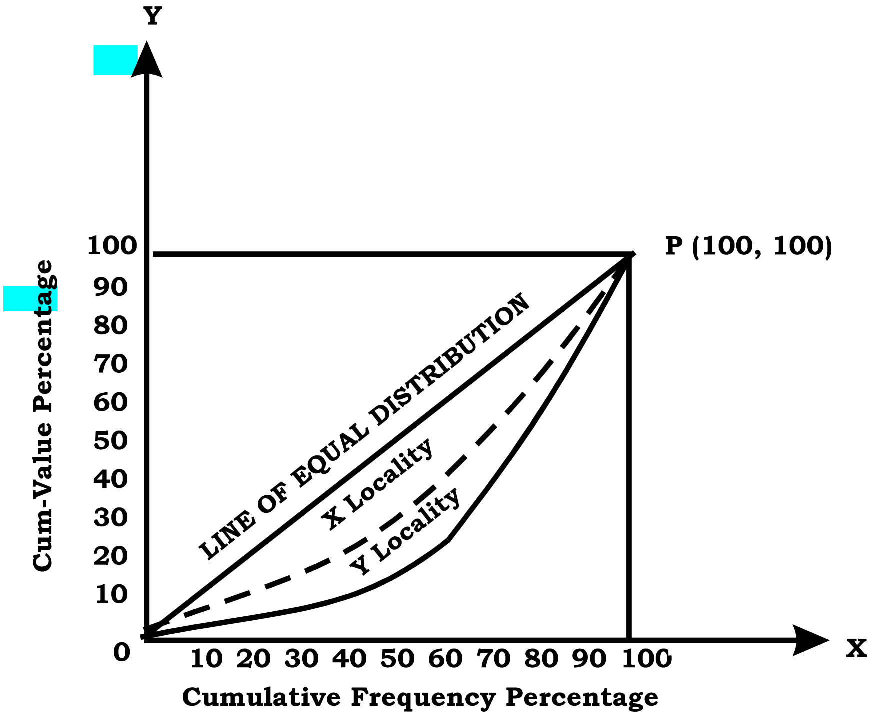


Fig. 2

Since the curve of the locality Y is farther from the line of equal distribution, it represents greater dispersion or incosistency.

Example 2 : From the following table giving data regarding income of workers in a factory, draw a graph (Lorenz Curve) to study the inequality of income:

Income	**No. of Workers in the factory**
Below 500	**6,000**
500-1,000	**4,250**
1,000-2,000	**3,600**
2,000-3,000	**1,500**
3,000-4,000	**650**

Solution : Income (in Rs.)	Mide Value	Cum. income	Per. of cum. income	No. of Workers F	Cum. frequ-ency	Per. of Cum. frequency
(1)	(2)	(3)	(4)	(5)	(6)	(7)
0-500	250	250	2.94	6,000	6,000	37.5
500-1,000	750	1,000	11.76	4,250	10,250	64.1
1,000-2,000	1500	2,500	29.41	3,600	13,850	86.6
2,000-3,000	2500	5,000	58.82	1,500	15,350	95.9
3,000-4,000	3500	8,500	100.00	650	16,000	100.0
Total	8,500			16,000		

The Lorenz Curve (Fig. 3) prominently exhibits the inequality of the distribution of income among the factory workers.

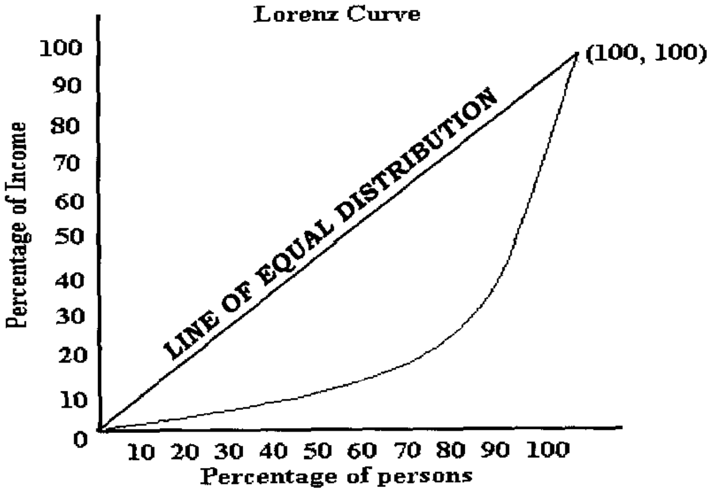


Fig.3

Example 3 : Draw a Lorenz Curve for the data:

Number of Persons	(X)	13	10	5	4	1
Wealth (in thousand rupee)	(Y)	78	100	75	80	25

Solution :

Number of Persons X	Cumulative frequency of X	Percentage Cumulative frequency of X	Wealth Y	Cumulative frequency Y	Percentage Cumulative frequency of Y
1	2	3	4	5	6
13	13	40	78	78	22
10	23	70	100	178	50
5	28	85	75	253	70
4	32	97	80	333	93
1	33	100	25	358	100

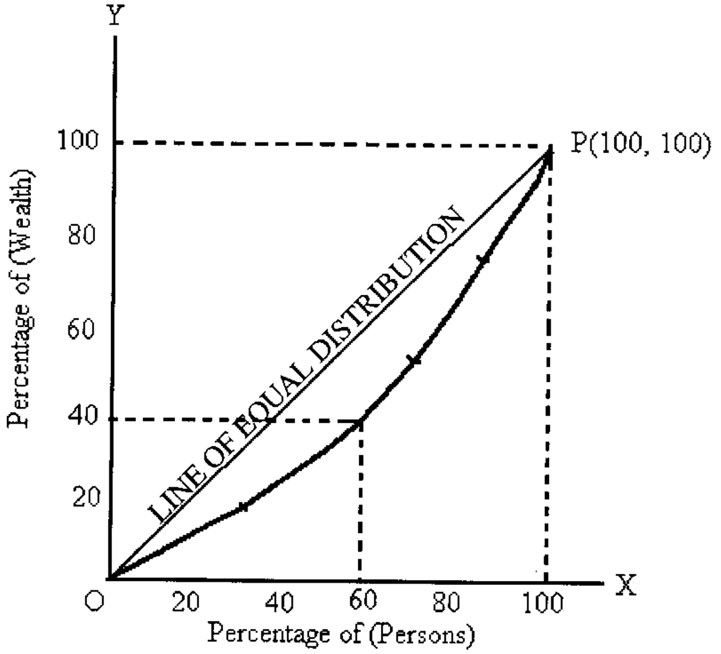


Fig. 4

1.3.6 Merits and Demerits :

Merits

i. It is most attractive and effective. Its effect lasts longer on the mind.

ii. It is simple technique for comparing the variability of two or more services.

iii. The dispersion of the distribution can easily be seen.

iv. It does not involve number in dispersion measure and hence leaves no burden on the mind.

v. It is useful in studying the variability in the distributions particularly relating to income, wealth, wages, profit, lands and capital etc.

Demerits :

i. The dispersion is shown graphically only.

ii. It does not tell us the presence of dispersion in the form of numerical value.

iii. It is difficult to draw as it involves too many calculations like cumulation, percentage etc.

iv. It gives us only a idea of the dispersion as compared with the line of equal distribution.

1.3.7 Summary

In this lesson we have studied the meaning and importance of dispersion. The distinction between absolute and relative measures of dispersion has been explained. Among different measures of variation, Lorenz Curve which is a technique of measuring dispersion of two series graphically has been studied more elaborately. The steps involved in the construction of Lorenz Curve have been explained with illustrations. The Lorenz Curve really illustrates the inequality in the actual size distribution. The various uses of Lorenz Curve have also been highlighted along with its drawbacks.

1.3.8 Key words

Dispersion : Dispersion or spread is the degree of the scatter or variation of the variables about a central value.

Relative Variation : Relative Variation is used to compare two or more distributions by relating the variation of one distribution to the variation of the other.

Line of Equal Distribution : The line of equal distribution is a digonal line on the Lorenz Curve on which any point shows the same per cent on both X and Y axis.

Lorenz Curve : Lorenz Curve is graphic method of showing the extent of variation in the size distribution from equal distribution.

1.3.9 Suggested Readings :

S.P. Gupta	:	Statistical Methods.
S.C. Gupta	:	Fundamentals of Statistics.
Digambar Patri	:	Statistical Methods.
P.G. Enns	:	Business Statistics.

1.3.10List of Questions

1.3.10.1 Short Questions

(a) What is Lorenz Curve ?

(b) Explain the term dispersion.

(c) Give uses of Lorenz Curve.

1.3.10.2. Long Questions

(a) What is Lorenz Curve? How is it constructed? What is its use?

(b) The frequency distribution of marks obtained in (i) Mathematics (M) English (E) are as follows :

Mid value of Range of marks	**No. of Students Scoring in (M)**	**No. of Students, Scoring in (E)**
5	10	1
15	12	2
25	13	26
35	14	50
45	22	59
55	27	40
65	20	10
75	12	5
85	11	3
95	9	1

`Analyse the data by drawing the Lorenz Curve on the same diagram and describe main features you observe.

(c) What do you understand by 'Dispersion'? What purpose does a measure of dispersion serve?

(d) What is a Lorenz Curve? How it is useful in measuring income inequalities between two regions ?

(e) From the following table giving data regarding income of employes in two factories, draw a graph (Lorenz Curve) to show which factory has greater in equalities of income.

Income (Rs.) :	Below 200	200-500	500-1,000	1,000-2,000	2,000-3,000
Factory A :	7,000	1,000	1,200	800	500
Factory B :	800	1,200	1,500	400	200

(f) Write an explanatory note on Lorenz Curve.

(g) The following table gives the population and earnings of residents in towns A and B. Represent the data graphically so as to bring out the inequality of the distribution of the earnings of residents.

Town A	No. of Persons :	50	50	50	50	50	50	50	50	50	50
	Earnings (Rs. daily) :	35	50	75	115	160	180	225	300	425	925
Town B	No. of Persons :	100	140	60	50	200	90	60	40	160	100
	Earnings (Rs. daily) :	160	320	120	280	400	400	280	920	240	960

M.A. (ECONOMICS) PART- I **PAPER–III**
(BASIC QUANTITATIVE METHODS)

LESSON NO. 1.4 **AUTHOR : DR. VIPLA CHOPRA**

SKEWNESS

STRUCTURE

1.4.1 Introduction
1.4.2 Objectives
1.4.3 Meaning of Skewness
1.4.4 Tests of Skewness
1.4.5 Measures of Skewness
    1.4.5.1 Karl Pearson's Measure of Skewness
    1.4.5.2 Bowley's Measure of Skewness
    1.4.5.3 Kelly's Co-efficient of Skewness
    1.4.5.4 Measures of Skewness based on moments.
1.4.6 Summary
1.4.7 Further Readings
1.4.8 List of Questions
    1.4.8.1 Short Questions
    1.4.8.2 Long Questions

1.4.1 Introduction:

In this lesson we will analyse the frequency distribution on the basis of dispersal of items on each side of an average and nature of peak of the frequency curve. This will be done by studying skewness and kurtosis.

1.4.2 Objectives:

After completion of this lesson you will be able to:

- ♦ Know how items on each side of an average are dispersed.
- ♦ Learn about different tests of skewness
- ♦ Measure different measures of skewness
- ♦ Learn about peakedness of distribution

1.4.3 Meaning of Skewness

The term ‘Skewness’ means ‘lack of symmetry’. The distribution may be ‘symmetrical’ or ‘asymmetrical’ (or Skewed).

According to Morris Hamburg : “Skewness refers to the asymmetry or lack of symmetry in the shape of a frequency distribution.”

Croxton and Cowden defined Skewness as follows :

“When a series is not symmetrical, it is said to be asymmetrical or Skewed.”

Symmetrical Distribution

When the distribution is symmetrical, the spread is always the same on both sides of the central point and the values of mean, median and mode are always identical. In such a distribution, the Skewness is zero.

In symbols : $\overline{X}$ = Median = Mode

Graphically, the curve will be bell-shaped as depicted in Part (a).

Asymmetrical or Skewed Distribution : When the distribution is asymmetrical (or Skewed), the spread is not the same on both sides of the central point and the value of mean, median and mode are not identical. In such a distribution there are two possibilities namely :

(a) Positive Skewness

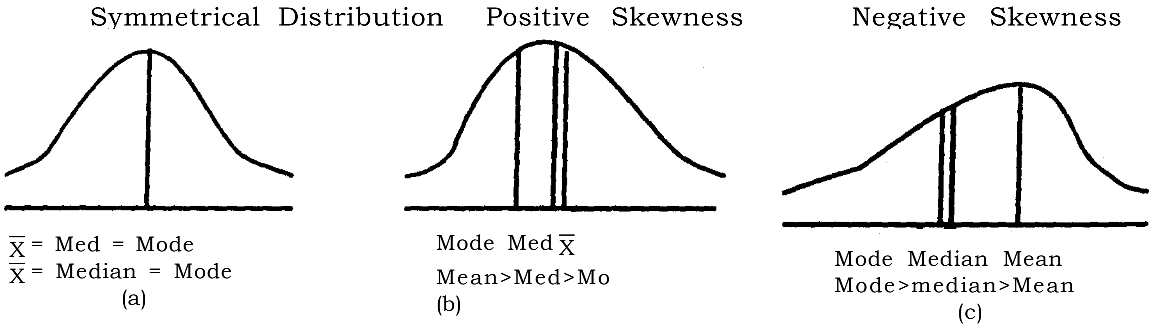
Skewness is said to be positive if the curve is more elongated to the right side (Part b) that is to say, if the mean of the distribution is to the right of or greater than the mode.

In symbols, $\overline{X}$ > Median > Mode

(b) Negative Skewness

Skewness is said to the negative if the curve is more elongated to the left side (Part c), that is to say mode is always greater than median and median is greater than mean.

In symbols, Mode > Median > Mean



Symmetrical Distribution Positive Skewness Negative Skewness

$\overline{X}$ = Med = Mode
$\overline{X}$ = Median = Mode
(a)

Mode Med $\overline{X}$
Mean>Med>Mo
(b)

Mode Median Mean
Mode>median>Mean
(c)

1.4.4 TESTS OF SKEWNESS

The presence or absence of Skewness in a distribution can be judged by applying the following tests :

1. In a Skewed distribution, the values of mean, median and mode are not equal.
2. In a Skewed distribution, the quartiles (Q_1 and Q_3) are not equi-distant from the median.
3. The sum of the positive deviations from the median is not equal to the sum of the negative deviations.
4. The curve of Skewed distribution is not bell-shaped.
5. Frequencies are not equally distributed at points of equal deviations from mode.

1.4.5 Measures of Skewness

Measures of Skewness bring out the extent and direction of a symmetry which exists in a given frequency distribution. These measures can be expressed either in absolute sense or in the relative sense.

Absolute measures of Skewness : The measures which express skewness in the unit in which the values of the series are expressed are called absolute measures. The measures which express skewness as a pure number are called relative measures of skewness.

Relative measures of Skewness

1. Karl Pearson's Coefficient of Skewness.
2. Bowley's Coefficient of Skewness.
3. Kelly's Coefficient of Skewness, and
4. Measures of Skewness based on moments.

1.4.5.1 Karl Pearson's Measures of Skewness

Karl Pearson's Measures of Skewness is based on difference between mean and mode.

Absolute Skewness = Mean - Mode shows both the extent and direction of Skewness.

If Mean > Mode $\Rightarrow$ Skewness is positive and

If Mean < Mode $\Rightarrow$ Skewness is negative.

Further relative measures of Skewness is known as Pearsonian Coefficient of Skewness.

$$\text{Coefficient of Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{S tan dard deviation}}$$

$$Sk_p = \frac{\overline{X} - \text{Mode}}{\sigma}$$

There is no limit to this measure in theory, but in practice its value lies between±1.

When mode is ill-defined, it can be calculated from the following formula:

Mode = 3 Median - 2 Mean

$$\text{Coefficient of Skewness} = \frac{\left[\overline{X} - \left(3\,\text{Med} - 2\overline{X}\right)\right]}{\sigma} = \frac{3\left(\overline{X} - \text{Med}\right)}{\sigma}$$

Theoretically this measure varies between ± 3 ; however, in practice its value lies between ±1.

Example 1 : Find the coefficient of skewness from the following data :

Value	:	6	12	18	24	30	36	42
Frequency	:	4	7	9	18	15	10	5

Solution:

Value x	Frequency f	$d = x–24$	$d^1 = d/6$	$f d^1$	fd'^2
6	4	-18	-3	-12	36
12	7	-12	-2	-14	28
18	9	-6	-1	-9	9
24	18	0	0	0	0
30	15	6	1	15	15
36	10	12	2	20	40
42	5	18	3	15	45
	N = 68			$\sum fd^1 = 15$	$\sum fd'^2 = 173$

$$\overline{x} = A + \frac{\sum fd'}{N} \times c = 24 + \frac{15}{68} \times 6 = 24 + 1.32 = 25.32$$

Mode = 24

$$\text{S.D. or } \sigma = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times \; = \sqrt{\frac{173}{68} - \left(\frac{15}{68}\right)^2} \times 6$$

$$= \sqrt{254 - 0.05} \times 6 = \sqrt{2.49} \times 6 = 1.57779 \times 6 = 9.47$$

$$Sk_p = \frac{x - \text{Mode}}{\sigma} = \frac{25.32 - 24}{9.47} = \frac{1.32}{9.47} = 0.139$$

Example 2 : From the data given below calculate Karl Pearson’s Coefficient of Skewness.

Wages (Rs.)	**No. of Persons f**
70-80	12
80-90	18
90-100	35
100-110	42
110-120	50
120-130	45

130-140	20
140-150	8

Solution :

Wages	**No. of Persons**	**Mid values**	**d' =** $\frac{X-105}{10}$	**fd'**	**fd'^{2}**
X	**f**	**m**			
70-80	12	75	-3	-36	108
80-90	18	85	-2	-36	72
90-100	35	95	-1	-35	35
100-110	42	105	0	0	0
110-120	50	115	1	+50	50
120-130	45	25	2	+90	180
130-140	20	135	3	+60	180
140-150	8	145	4	+32	128
	$\sum$ f = 230			$\sum$ fd' = + 125	$\sum$ fd'^{2} = 753

$$Sk_p = \frac{\overline{x} - \text{Mode}}{\sigma}$$

Now $\overline{x} = A + \frac{\sum fd'}{N} \times i = 105 + \frac{125}{230} \times 10 = 110.435$

By Inspection Mode lies in 110-120 class

$$\therefore \text{ Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i = 110 + \frac{50 - 42}{100 - 42 - 45} \times 10$$

Mode = 116.154

Now

$$\text{S.D.} = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times i \qquad = \sqrt{\frac{753}{230} - \left(\frac{125}{230}\right)^2} \times 10$$

$$= \sqrt{753 \times 230 - (125)^2} \times \frac{10}{230} = \sqrt{173190 - 15625} \times \frac{1}{23} = \sqrt{57565} \times \frac{1}{23}$$

$\therefore \quad \sigma = 10.435$

$$Sk_p = \frac{\overline{x} - \text{Mode}}{\sigma} = \frac{110.435 - 116.154}{10.435} = \frac{-5.719}{10.435}$$

$Sk_p = -0.548$

1.4.5.2 Bowley's Measure of Skewness

Dr. A. L. Bowley propounded another measure of skewness based on the relative positions of the median and the two quartiles. If the distribution is symmetrical, the first and third quartiles are equidistant from the median.

Symbolically $(Q_3 - \text{Med}) = (\text{Med} - Q_1)$

or $Q_3 + Q_1 - 2\text{ Med} = 0$

If the distribution is a symmetrical (or skewed), the quartile will not be equidistant from the median. In such a case skewness can be measured by the following formula.

$$\text{Bowley's Coefficient of Skewness} = \frac{Q_3 + Q_1 - 2\text{Med}}{Q_3 - Q_1}$$

Find the value of Coefficient of Skewness for the following series :

Number of Children (Per Family)	0	1	2	3	4	5	6
Number of families	7	10	16	25	18	11	8

Solution :

Number of Children (per family) (x)	**No. of families (f)**	**C.f.**
0	7	7
1	10	17
2	16	33
3	25	58
4	18	76
5	11	87
6	8	95

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

Now $Q_1 = \frac{(N+1)^{th}}{4}$ item $= \frac{95+1}{4}$th item = 24th item

$\therefore\ Q_1 = 2$

$Q_3 = \frac{3(N+1)^{th}}{4}$ item = 3 × 24th item = 72th item.

$Q_3 = 4$

$$\text{Med} = \frac{(N+1)^{th}}{4} \text{ item} = \frac{95+1}{2}\text{th item} = 48\text{th item}$$

Med = 3

$$Sk_B = \frac{4\times 2 - 2\times 3}{4-2} = \frac{6-6}{2} = \frac{0}{2} = 0$$

Example 4 : Calculate Quartile Measure of Skewness for the following series :

Class :	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
f :	2	3	5	10	25	20	12	10	8	5

Solution

Class (X)	**Frequency f**	**Cumulative (c.f.)**
0-10	2	2
10-20	3	5
20-30	5	10
30-40	10	20
40-50	25	45
50-60	20	65
60-70	12	77
70-80	10	87
80-90	8	95
90-100	5	100

$$\text{Med} = \text{Size of } \frac{N}{2}\text{th item} = \frac{100}{2}\text{th item} = 50\text{th item}$$

Med lies in Class interval 50-60

$$\text{Med} = L + \frac{\frac{N}{2} - c.f.}{f} \times i = 50 + \frac{\frac{100}{2} - 45}{20} \times 10 = 50 + \frac{50-45}{20} \times 10$$

Med = 50+2.5 = 52.5

$$Q_1 = \text{Size of } \frac{N}{4}\text{th item} = \frac{100}{4}\text{or 25 th item}$$

$\therefore$ Q_1 lies in class interval 40-50

$$Q_1 = L + \frac{\frac{N}{4} - c.f.}{f} \times i = 40 + \frac{\frac{100}{4} - 20}{25} \times 10$$

$$Q_1 = 40 + \frac{25-20}{25} \times 10 = 40 + \frac{5}{25} \times 10 = 40 + 2 = 42$$

Q_3 = Size of $\frac{3N}{4}$th or $\frac{3\times 100}{4}$ = 75th item

$\therefore Q_3$ lies in the class interval 60-70

$$Q_3 = L + \frac{3\frac{N}{4} - \text{c.f.}}{f} \times i = 60 + \frac{75-65}{12} \times 10 = 60 + \frac{10}{12} \times 10$$

$$Q_3 = 60 + 8.3 = 68.3$$

$$Sk_B = \frac{Q_3 + Q_1 - 2\,\text{Med}}{Q_3 - Q_1} = \frac{68.3 + 42 - 2\times 52.5}{68.3 - 42} = \frac{110.3 - 105}{26.3} = \frac{5.3}{26.3} = +0.2$$

1.4.5.3 Kelly's Coefficient of Skewness

Bowley's measure of skewness leaves on either extreme of the distribution 25% items hence it is a measure based on the middle 50% items of the distribution. Kelly's measure of skewness is based on Deciles and Percentiles.

$$Sk_K = \frac{P_{90} + P_{10} - 2\,\text{Med}}{P_{90} - P_{10}}$$

Also $$Sk_K = \frac{D_9 + D_1 - 2\,\text{Med}}{D_9 - D_1}$$

This measure is rarely used in practice

Example 5 : Calculate Kelly's co-efficient of skewness of the following distribution:

Variable : X	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency :	2	5	7	13	21	16	8	3

Solution :

Variables (x)	**Frequency f**	**Cumulative Frequency (c.f.)**
0-5	2	2
5-10	5	7
10-15	7	14
15-20	13	27
20-25	21	48
25-30	16	64
30-35	8	72
35-40	3	75

$$Sk_k = \frac{D_9 + D_1 - 2\text{Med}}{D_9 - D_1}$$

Median = Size of $\frac{N}{2}$th item = $\frac{75}{2}$th item = 37.5th item

$\therefore$ Med lies in the class interval 20-25

$$\text{Med} = L + \frac{\frac{N}{2} - \text{c.f.}}{f} \times i = 20 + \frac{37.5 - 27}{21} \times 5 = 20 + 2.5 = 22.5$$

Now D_1 = Size of $\frac{N}{10}$th item = Size of $\frac{75}{10}$th item = 7.5th item

D_1 lies in the class interval 10-15

$$D_1 = D_1 = L + \frac{\frac{N}{10} - \text{c.f.}}{f} \times i = 10 + \frac{75 - 7}{7} \times 5 = 10 + .36 = 10.36$$

D_1 lies in the class interval 30-35

$$\therefore D_9 = L + \frac{\frac{9N}{10} - \text{c.f.}}{f} \times i = 30 + \frac{67.5 - 64}{8} \times 5$$

$$D_9 = 30 + 2.19 = 32.19$$

$$Sk_K = \frac{32.19 + 10.36 - 2(22.5)}{32.19 - 10.36} = \frac{42.55 - 45}{21.83} = \frac{-2.45}{21.83} = -0.112$$

1.4.5.4 Measure of Skewness based on moments

A measure of skewness based on moments is denoted β_1. β_1 is defined as

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

Where

μ_2 = Second moment about mean

μ_3 = Third moment about mean

if β_1 = 0 then the distribution is symmetrical

if β_1 < 0 then the distribution is positively skewed

if β_1 > 0 then the distribution is negatively skewed.

Calculation of Moments

Moment about arithmetic mean	**Moments about an arbitrary origin**
First moment $\mu_1 = \frac{\sum(x-\bar{x})}{N}$ or $\frac{\sum f(x-\bar{x})}{N}$,	$\mu'_1 = \frac{\sum(X-A)}{N}$ or $\frac{\sum f(X-A)}{N}$
Second moment $\mu_2 = \frac{\sum(x-\bar{x})^2}{N}$ or $\frac{\sum f(x-\bar{x})^2}{N}$,	$\mu'_2 = \frac{\sum(X-A)^2}{N}$ or $\frac{\sum f(X-A)^2}{N}$
Third moment $\mu_3 = \frac{\sum(x-\bar{x})^3}{N}$ or $\frac{\sum f(x-\bar{x})^3}{N}$,	$\mu'_3 = \frac{\sum(X-A)^3}{N}$ or $\frac{\sum f(X-A)^3}{N}$
Fourth moment $\mu_4 = \frac{\sum(x-\bar{x})^4}{N}$ or $\frac{\sum f(x-\bar{x})^4}{N}$,	$\mu'_4 = \frac{\sum(X-A)^4}{N}$ or $\frac{\sum f(X-A)^4}{N}$

We can also calculate moments about mean form moments about an arbitrary origin.

$$\mu_1 = \mu'_1 - \mu'_1 = 0$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2$$

$$\mu_3 = \mu'_3 - 3\mu'_2\,\mu'_1 + 3(\mu'_1)^3$$

$$\mu_4 = \mu'_4 - 4\mu'_3\,\mu'_1 + 6\mu'_2\,(\mu'_1)^2 - 3(\mu'_1)^4$$

1.4.6 Summary

In this lesson Skewness analyse the frequency distribution on the basis of dispersal of items about mean and nature of peakedness of the frequency curve. Different measures of Skewness measure the amount and direction of asymmetry.

1.4.7 Further Readings

S.P. Gupta	:	Statistical Methods
S.C. Gupta	:	Fundamentals of Statistics

1.4.8 List of Questions

1.4.8.1 Short Questions

(a) Explain the important of measure of Skewness.

(b) Define the concept of 'Skewness'.

(c) State different measures of Skewness.

(d) State the difference between 'absolute' measures of Skewnes and 'relative' measures of Skewness.

1.4.8.2 Long Questions

(a) Find the co-efficient of skewness from the following:

Value :	6	12	18	24	30	36	42
Frequency:	4	7	9	18	15	10	5

(b) In a certain distribution the following results were obtained:

Mean = 45, Median = 48

Co-efficient of skewness = -0.4

The person who gave you the data failed to give the value of standard deviation and you are required to estimate it with the help of available information.

(c) From the data given below calculate Karl Pearson's coefficient of Skewness and explain its significance.

Wages	No. of Persons	Wages	No. of Persons
70-80	12	110-120	50
80-90	18	120-130	45
90-100	35	130-140	20
100-110	42	140-150	08

(d) What do you mean by "Skewness". What are different tests of Skewness?

Type Setting :
Department of Distance Education, Punjabi University, Patiala.

M.A. (ECONOMICS) PART-I **PAPER : II**
SEMESTER-II **BASIC QUANTATIVE METHODS**

LESSON NO. 1.5 **AUTHOR : DR. S.K. SRIVASTAVA**

PROBABILITY

Random Experiments

In the most varied fields of practical and scientific activity, cases occur where certain experiments or observations may be repeated a large number of times under similar circumstances. On each occasion, our attention is then directed to a result or the observations which is expressed by certain number of characteristic features. In many cases these characteristics directly take quantitative form at each observation, something is counted or measured. In other cases, the characteristics are qualitative, we observe, for example, the colour of a certain object, the occurrence or non-occurrence of some specified event in connection with each experiment etc.

1. If we make a series of throws with an ordinary dice each throw yields as its result one of the number 1, 2.............6.
2. If we measure the length and weight of the body of each number of a group of animals belonging to the same species, every individual gives rise to an observation, the result of which is expressed by two numbers.
3. If we observe at regular time intervals the prices of k different commodities, the result of each observation is expressed by k numbers.
4. If we observe the sex of every child born in a certain district, the result of each observation is either 'boy' or 'girl'.
5. If, in a steel factory we take a sample from daily production and measure its tensile strength, the result of each observation is given by a number.

In some cases we know the phenomenon under investigation sufficiently well to feel justified in making exact predictions with respect to the result of each individual observation. In majority of the cases, however, our knowledge is not precise enough to allow the exact predictions of the result of individual observations. This is the situation, e.g. in all examples 1-5 quoted above. In such a case, we shall say that we are concerned with a sequence of random experiments.

It does not seem possible to give a precise definition of what is meant by the word random. The sense of the word is best conveyed by the following example.

If an ordinary coin is rapidly spun several times, and if we take care to keep the conditions of the experiment as uniform as possible in all respects, we shall find that we are unable to predict whether, in a particular instance, the coin will fall 'heads' or 'tails'. Even if we try to build a machine throwing the coin with perfect regularity, it is not likely that we shall be able to predict the result of individual throws. On the contrary, the result of the experiment will always fluctuate in an uncontrollable way from one instance to another. At first this may seem rather difficult to explain. A moment's reflection will, however, show that practically the conditions of experiment could be exactly repeated only up to certain approximation, even a small change in conditions of the experiment could have a dominating influence on the result. Thus an exact prediction of results will always be practically impossible. Similar remarks apply to the throws of a dice of any other game of chance.

Next let us imagine that we observe a number of men of a given age during a period of say, one year, and not in each case whether the men are alive at the end of the year or not. It will obviously be impossible to make exact predictions with regard to the life or death of one particular person, even if we have detailed information concerning health, occupation, habits etc. Since the causes leading to the ultimate result are too many and too complicated to allow of any precise calculation.

Such examples are representative of large group of random experiments. The fluctuation in the results of a series of experiments may be because of any of the several causes, the essential thing is that, in all cases, an exact prediction of the result of individual experiments becomes impossible.

We have seen, that in sequence of random experiments, it is not possible to predict individual results. However, as soon as we turn our attention from individual experiment to the whole sequence of experiments, an extremely important phenomenon appears : in spite of the irregular behaviour of individual results. The average results of long sequence of random experiments show a similar regularity. While the result of any particular performance of random experiment show a similar regularity. That is the basis for quite predictions.

In order to obtain some idea of this important mode of regularity, we consider a random experiment which is repeated a large number of times. Observe each time whether a possible result A (called an event) takes place or not. If in the first performance of the experiment. 'A' occurs exactly m times, the ratio $\frac{m}{n}$ is called the relative frequency of A in 'n' trails. Now, if we observe the frequency m/n for increasing value of n we shall generally find that it shows a tendency to become more or less constant for large values of n. In mathematical terms, we find that m/n tends to certain limit as n tends to infinity.

This limiting values is called the probability of the event A. Thus we have

$$P(A) = \lim_{n \to \infty} \frac{m}{n}$$

This is referred to as the empirical or statistical definition of probability.

Sample Space and Events

Definition : The sample space, S, of a random experiment, is the set (collection) of all its possible outcomes.

Before discussing the mathematical theory of probability, we must consider the set of all the possible outcomes, of a random experiment, called the sample space. Thus, if the random experiment is the tossing of a coin, its sample space can be described as the set of two points H, T., where H stands for Head and T for tail. Here, we assume that other occurrences such as tossing or breaking the coin or its standing on its edge are impossible. In the experiment in the tossing of two coins, the sample space consists of the 36 ordered pairs (1, 1),(1, 2)........(1, 6)...........(2, 1)...........(6, 6).

Definition : An event is a subset of a sample space.

Consider an experiment of tossing a dice, the sample space will then be the set, 1,2,3,4,5,6. Any sub-set of this set will be an **event**. For example, 3 is an event which in words could be, described as 'No. 3 occurs', or the subset 1,3,5 is an event implying that 'an odd number occurs' etc.

Each of individual outcomes of an experiment is called **simple event** or **elementary event.** In the above examples there are six simple events, namely 1,2,3,4,5 and 6. Thus simple event cannot be decomposed into a combination of other events.

In contrast, consider an event A, which is the occurrences of an odd number, clearly this event can be decomposed into three simple events 1,3 and 5. Such events are called **compound events**.

With any two events A and B, we can associate two new events namely "Both A and B occur" and "Either A or B or both occur". These events are respectively as $A \cap B$ and $(A \cup B)$ and read as "A Intersection B" and "A union B". The definition extends to more than two events in a natural way.

Note : $A \cap B$ is generally written as AB.

Example 1. Consider the experiment of tossing 3 coins. The sample space consists of eight points.

S = (H H H), (H H T), (H T H), (T H H), (H T T), (T H T), (T T H), (T T T)

A, B and C defined below are the events.

A : "Two or more heads occur" i.e.

A = (H H H), (H H T), (H T H), (T H H)

B : "Heads occur on each coin" i.e. (H H H)

C : "Exactly two heads occur" i.e.

(H H T), (H T H), (T H H)

Then

$A \cap B$ or AB = (H H H)

$A \cup B$ = (H H H), (H H T), (H T H), (T H H) = A

$A \cap C$ = (H H T), (H T H), (T H H) = C

$B \cup C = A$

$B \cap C = \phi$

......................Since B & C has no element in common it is called an empty set or impossible event and is denoted by ϕ in this case. We call the events B and C mutually exclusive events.

Definition : Two events are said to be **mutually exclusive**, if the occurrence of one precludes the occurrence of the other.

Example 2 : Let A be the event that a student is 20 years old. Let B be the event that a student is 20 years old and is a smoker. AB is the event that a student is 20 years old and is a smoker.

Axiomatic approach to Probability :

We have earlier described the frequency interpretation of probability, according to which the probability of an event A is a number denoted by P (A). This number can be known to us only by experience as the result of a very long series of observations of independent trails of experiment. But such a definition of probability does not serve the purposes as with many experiments, it may not be possible to repeat them under the same condition. The development of the theory must be accomplished by logical deductions from the basic consumption based on centuries old human experience. In other words it should be built up on the basis of certain axioms like any other well developed mathematical discipline.

Definition : The Axiomatic approach to Probability was introduced by the Russian mathematician A.N. Kolmogorov in the year 1933. When this approach is followed, no precise definition of probability is given rather we give certain axioms on which probability calculations are based. A probability measure P on the events of a finite sample space S is numerical valued function defined on the events of S for which the following three axioms are satisfied.

Axiom 1 : The probability of an event ranges from zero to one. If the event cannot take place, its probability shall be zero and if it is certain, i.e. bound to occur, its probability shall be one.

Axiom 2 : $P(S) = 1$

Axiom 3 : $P(A \cup B) = P(A) + P(B)$ for every pair of mutually exclusive events of A and B of S. The above axioms are sufficient if S contains a finite number of elements.

Classical definition of probability

The classical approach of probability theory is restrictive in two respects. First the total number of elementary events in the sample space is finite, say n. Secondly, it is assumed that the experiment is such that elementary events are equally likely in the sense that when all relevant evidence is taken into account, no one of them can be expected to occur in preference to the others. For example, the two elementary events 'occurrence of heads' and 'occurrence of tails' are equally likely for the random experiment in which a coin is tossed : or the simple events described by a number 1, 2...............or 6 are equally likely for the experiment.

Definition : If an experiment can result in n exhaustive, mutually exclusive and equally likely ways, out of which m are favourable to an event A, the probability of A is then defined by

P (A) = m/n

This definition of probability is called the classical definition of probability and is associated with the names of Laplace and Bernoulli. Such a definition is generally always applicable in most of the games of chance.

For example, if a coin is tossed, there are two exhaustive, mutually exclusive and equally likely simple events, namely, 'head' and 'tail' out of which one is favourable to the event 'head', hence the probability of getting 'head' is 1/2. So 1/2 is the probability of getting 'tail'.

Similarly, if a dice is thrown there are six exhaustive, mutually exclusive and equally likely cases. For the event that an even number occurs, there are 3 favourable cases namely the occurrence of 2, 4 and 6, hence its probability is 3/6 or 1/2.

Example 3 : Suppose the random experiment is tossing of 3 coins. In this case, the sample space consists of 8 points which are exhaustive, mutually exclusive and equally likely.

(H H H), (H H T), (H T H), (T H H), (H T T), (T H T), (T T H), (T T T)

The probability that the number of heads is exactly two = 3/8.

Since the three cases favourable to the event are (HHT), (HTH) and (THH).

The probability that we get atleast two heads is equal to $\frac{4}{8}=\frac{1}{2}$

Example 4 : One card is drawn from a standard pack of 52. What is the probability that it is

(i) a red card

(ii) either a king or queen

(iii) either a spade or a club.

The total number of exhaustive ways which are mutually exclusive and equally likely are 52 in this case.

(i) The number of cases favourable to the event of drawing a red card is clearly 26 and hence its probability $=\frac{26}{52}=\frac{1}{2}$

(ii) There are 4 kings and 4 queens in a pack of 52 cards. Hence a king or a queen can be drawn in 8 ways which are favourable to the event of drawing a king or a queen. Hence its probability = $\frac{8}{52}=\frac{2}{13}$

(iii) There are 13 spades and 13 clubs and the number of cases favourable to the event that either a spade or a club is drawn is clearly 26. Hence its probability

$$= \frac{26}{52} = \frac{1}{2}$$

Example 5 : A bag contains 4 white and 6 black balls. Two balls are drawn from it at random. Find the probability that (i) both are white and (ii) one is white and the other is black.

The bag contains altogether 10 balls, out of which 2 balls can be drawn in

$$10_{c_2} = \frac{10 \times 9}{2 \times 1} = 45 \text{ ways}$$

(i) Two white balls can be drawn in ${}^4C_2 = \frac{4 \times 3}{2 \times 1} = 6$ different ways, hence the probability of drawing two white balls is

$$= \frac{6}{45} = \frac{2}{15}$$

(ii) One white and one black ball can be drawn ${}^4C_1 \times {}^6C_1$ = 4×6 =24 ways, hence the probability of drawing one white and black ball is

$$= \frac{24}{45} = \frac{8}{15}$$

Example 6 : Two dices are rolled. Find the probability of obtaining a total of 8 points.

There are 6×6 = 36 different, exhaustive, mutually exclusive and equally likely cases in which this experiment can result. These are (1,1), (1,2)...............(1,6),(2,1) .... (2,6), (3,1).............(6,6)

A total of 8 points can be obtained in the following ways :

(2, 6), (3,5), (4, 4), (5, 3), (6, 2)

Hence the number of cases favourable to this event are 5 and its probability

$$= \frac{5}{36}$$

Complementary events

Let there be two events A and B. A is called the complementary event of B and vice versa if A and B are mutually exclusive and exhaustive. The complementary event of a given event A is generally denoted by $\overline{A}$. Obviously,

$A \cup \overline{A}$ = S whole sample space

It is easy to verify that

P (A) + P ($\overline{A}$) = 1

Several times it is easy to find the probability of the complementary event $\overline{A}$ than of A, and for finding the probability of A, we use the relation.

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

Example 7 : Five men in a company of 20 are graduates. If 3 men are picked out of 20 at random.

(i) What is probability that all are graduates and

(ii) What is the probability of atleast one graduate amongst them ?

The total number of ways of picking 3 men out of 20 men ${}^{20}C_3$

(i) The number of ways of picking 3 graduates = ${}^{5}C_3$

Hence the probability of 3 graduates

$$\frac{5C_3}{20C_3} = \frac{5 \times 4 \times 3}{20 \times 19 \times 18} = \frac{1}{114}$$

(ii) In this case it is easy to find the probability of the Complementary event that there are no graduates among 3 men selected. The number of ways favourable to this event = ${}^{15}C_3$ and its probability

$$\frac{15C_3}{20C_3} = \frac{15 \times 14 \times 13}{20 \times 19 \times 18} = \frac{91}{228}$$

Hence

P (There is atleast one graduate)

= 1 - P (there is no graduate)

$$= 1 - \frac{91}{228} = \frac{137}{228}$$

Additive Law of Probability

Let two events A and B be mutually exclusive. If the event A can happen in m_1 ways and the event B in m_2 ways, out of a total number of n ways, then by definition, the probability that either A or B occurs is

$$P(A \cap B) = \frac{m_1 + m_2}{n}$$

$P(A \cap B) = \frac{m_1}{n} + \frac{m_2}{n}$ = P(A) + P(B), since P(A) = $\frac{m_1}{n}$ and P (B) = $\frac{m_2}{n}$ by definition

Since the last two expressions are precisely those defining P (A) + P (B). Thus we have proved the following addition formula.

Addition Formula : When A and B are mutually exclusive events

$P(A \cup B) = P(A) + P(B)$

For more than two mutually exclusive events the formula extends in a natural way. For example, let us take the case of three events.

$P(A \cup B \cup C) = P(A) + P(B) + P(C)$

This addition formula is not applicable when the events are not mutually exclusive. In case A and B are non-mutually exclusive, we have

$P(A \cup B) = P(A) + P(B) - P(AB)$

Conditional Probability : Multiplication Law

The notion of conditional probability arises in the following manner. One has at his disposal some information in the form of the happening of event and he wants to know the probability of another event, given the occurrence of this specific event.

Suppose that an experiment can result in n exhaustive, mutually exclusive and equally likely ways. Out of this, let m (A) cases be favourable to an event A and m (AB) cases are favourable to event A and B both.

Then clearly,

$$P(AB) = \frac{m(AB)}{n} \text{ and } P(A) = \frac{m(A)}{n}$$

Now let us suppose that it is known that the event A has occurred and given this, we want to find the probability of B. This is known as the conditional probability of B, given that A has occurred and is denoted by P(B/A). Since we know that A has occurred. We know that the experiment can result in one out of those (A) ways which are favourable to the event. To find the conditional probability of B given A, we must consider ways out of these m (A) ways which are favourable to the event B. This number is clearly m (AB), and we have

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

$$\text{or } P(B / A) = \frac{P(AB)}{P(A)} = \frac{P(A \cap B)}{P(A)}$$

The last formula is taken as the definition of the conditional probability of the event B given that A has occurred.

Definition : The conditional probability of event B given that A has occurred is given by

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

Such probability is defined only if P(A)>0.

When written in product form, this is called the multiplication rule of probability.

Multiplication Law : P (AB) = P (A) P (B/A)

For three events A, B and C, the formula extends to

P (ABC) = P (A) P (B/A) P (C/AB)

Where P (C/AB) is the probability of event C given that both A and B have occurred.

Independent events : If two events A and B are such that the probability that A will occur does not depend upon whether or not B occurs, then B is said to be independent of A, and we can write P (B/A) = P (B).

For this case the multiplication law reduces to :

P (AB) = P (A). P (B)

and this implies that P (A/B) = P (A) and so B is independent of A. We simply say that A and B are independent.

Example 8 : Two cards are drawn from a pack of 52 cards, the first card drawn is replaced before the second card drawn.

(i) What is the probability that both will be spades ?

Let A_1 be the event of getting a spade on the second draw. Since the first card drawn is replaced, the probability of getting a spade on the second draw should not depend upon whether or not a spade was obtained on the first draw, hence A_1 and A_2 are independent. Thus, (both spades).

$P(A_1 A_2) = P(A_1) P(A_2)$

$$= \frac{13}{52} \times \frac{13}{52} = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$$

(ii) What is the probability that the cards will be either two spades or two hearts ? Let A denotes the event of getting two spades and B, the event of getting two hearts. Then since A and B are mutually exclusive, the required probability is

$P(A \cup B) = P(A) + P(B)$

$$= \frac{1}{16} + \frac{1}{16} = \frac{2}{16} = \frac{1}{8}$$

Example 9 : As in example 8, let two cards be drawn, but this time first card drawn is not replaced before the second is drawn. What is the probability that both cards will be spades ?

In this case let A_1, and A_2 be in (i) of example 8. Now A_2 is not independent of A_1 because if a spade is obtained on first draw, the chances of getting a spade on the second draw will be smaller than if a non-spade had been obtained in the first draw. In this case.

$P(A_1 A_2) = P(A_1). P(A_2)$

$$= \frac{13}{52} \times \frac{12}{51} = \frac{1}{17}$$

Example 10 : A bag contain 10 white and 5 black balls. Two balls are drawn one after another without replacement, find the probability that both are white.

Let A and B denote the event of drawing a white ball in the first and second draws respectively. Then the required probability is

$P(AB) = P(A). P(B/A)$

$$= \frac{10}{15} \times \frac{9}{14} = \frac{3}{7}$$

Example 11 : A bag contains 6 white and 4 black balls and second bag contains 4 white and 8 black balls. One of the bags is chosen at random and 2 balls are drawn from it. Find the probability that one is white and the other is black.

Let A denote the event that of two balls drawn, one is white and the other B_1 is black. Also let B_2 be the event that the second bag was selected. Hence the event A can occur in two mutually exclusive ways either with B_1 or B_2 (thus we can write)

$A = (A B_1) \cup (AB_2)$

and since AB_1 and AB_2 are mutually exclusive, By addition formula

$P(A) = P(B_1) P(A/B_1) + P(B_2) P(A/B_2)$

$$= \frac{1}{2} \times \frac{6 \times 4}{10_{C_2}} + \frac{1}{2} \times \frac{4 \times 8}{12_{C_2}} = \frac{4}{15} + \frac{8}{33} = \frac{84}{165}$$

Example 12 : Urn 1 contains 2 white and 4 black balls. Urn II contains 5 white and 7 black balls. A ball is transferred from Urn 1 to II. Find the probability that it will be white.

Let A denote the event that a white ball is drawn and B_1 be the event that a white ball is transferred and B_2, the event that a black ball is transferred. Then by the arguments as in example 11, we have

$A = (AB_1) \cup (AB_2)$

$= P(B_1) P(A/B_1) + P(B_2) P(A/B_2)$

$$= \frac{2}{6} \times \frac{6}{13} + \frac{4}{6} \times \frac{5}{13} = \frac{2}{13} + \frac{10}{39} = \frac{16}{39}$$

Bayes' Formula

We start with an example to give an idea of the type of problem discussed in this section. Suppose two urns I and II contain respectively 1 white, 6 black and 4 white and 3 black balls. One of the urns is selected at random and then a ball is drawn from the selected urn. It happens to be white, what is probability that it came from urn I ? Such Problems are important and arise many a times in practice.

Before the ball was drawn and its colour revealed, the probability that the first urn, would be chosen and had probability 1/2 indication of the colour of the ball drawn changed this probability. It can be seen in the following way.

Let B_1 and B_2 be the events of drawing urn I and urn II respectively and A be the event of drawing a white ball from the selected urn. In this case we have to find the P (B_1/A) i.e. given that the event A has occurred.

Here unconditional probability P (B_1) and P (B_2) are known, each is equal to 1/2. Also known are conditional probability P (A/B_1) and P (A/B_2). The probabilities. P (B_1) and P (B_2) are called prior probabilities and the conditional probabilities P (B_1/A) and P (B_2/A) are known as posterior probabilities, which are the probabilities of hypotheses after event is known to have occured.

To find P (B_1/A) we can use the definition,

$$P(B_1/A) = \frac{P(AB_1)}{P(A)}$$

Now

$$P(A) = P(A/B_1) + P(A/B_2)$$
$$= P(B_1)\,P(A/B_1) + P(B_2)\,P(A/B_2)$$
$$= \frac{1}{2} \times \frac{1}{7} + \frac{1}{2} \times \frac{4}{7} = \frac{5}{14}$$

Also

$$P(AB_1) = P(B_1)\,P(A/B_1)$$
$$= \frac{1}{2} \times \frac{1}{7} = \frac{1}{14}$$

Hence $P(B_1/A) = \frac{1}{14} / \frac{5}{14} = \frac{1}{5}$

Thus we see that once we have the information that a white ball has been drawn, the probability, of B_2 has changed to 1/5 in this case.

Such probabilities can be found by using the Bayes' formula which is given below.

Bayes' Formula : Suppose an event A can occur if and only if one of the events B_1, B_2 ....B_k which are mutually exclusive and exhaustive occurs.

Then the conditional probability of the event B_1/A is given by

$$P(B_1/A) = \frac{P(B_1)P(A/B_1)}{P(B_1)\,P\,(A/B_1) + P(B_2)P(A/B_2) + P(B_3)P(A/B_3) + -- P(B_n)P(A/B_n)}$$

Example 13 : A factory manufacturing razor blades has two machines I and II. Machine I produces 30% and machine II 70% of the items produced. It is known that 5% of the blades are defective in a packet. What is the probability that it is produced by machine I.

Let A, B_1 and B_2 denote the following events :

A = the blade is defective

B_1 = the blade is produced by machine I

B_2 = the blade is produced by machine II

Then, we have to find P(B/A). Using Baye's formula

$$(B_1/A) = \frac{P(B_1)P(A/B_1)}{P(B_1)P(A/B_1) + P(B_2)P(A/B_2)}$$

Clearly

$P(B_1) = 0.30$, $P(B_2) = 0.70$

$P(A/B_1) = 0.05$, $P(A/B_2) = 0.01$

$$P(A/B) = \frac{.30 \times 0.05}{0.30 \times 0.05 + 0.70 \times 0.01} = \frac{0.15}{0.015 + 0.007}$$

$$= \frac{0.015}{0.022} = \frac{15}{22} = 0.68$$

Example 14 : A student taking a true-false test always marks the correct answer when he knows and decides true or false on the basis of flipping a coin when he does not know it.

If the probability that he will know an answer is $\frac{3}{5}$ what is the probability that he knows the answer to a correctly marked question ?

Here the event A is that the student has marked the question as correct and B_1 and B_2 are

B_1 : He knows the answer to the question.

B_2 : He does not know the answer to the question.

It is given in the problem that

$$P(B_1) = \frac{3}{5}$$

and so P (B_2) = 1 - $\frac{3}{5} = \frac{2}{5}$

Now, if he knows the answer of the question, he marks it correct with probability I, and if he does not know the answer, he marks correctly with probability 1/2 only. Thus,

P (A/B_1) = 1 and P (A/B_2) = $\frac{1}{2}$

Hence the required probability is

$$P(B_2 / A) = \frac{P(B_1)P(A / B_1)}{P(B_1)P(A / B_1) + P(B_2)P(A / B_2)}$$

$$= \frac{\frac{3}{5} \times 1}{\frac{3}{5} \times 1 + \frac{2}{5} \times \frac{1}{2}} = \frac{3}{4}$$

Example 15 : A and B are independent witnesses in a case. The probability that A will speak the truth is p and B will speak the truth is q. A and B agree in a certain statement. Find the probability that the statement is true.

Here the event A is that both the witnesses agree in the statement. Let the two events B_1 and B_2 be

B_1 : the statement is true

B_2 : the statement is false

Since nothing is known about the statement, we have

P (B_1) = P (B_2) = ½

Now we find the conditional probabilities.........Given that the statement is true, the event A will happen if both A and B speak truth and probability for this is p.q.i.e.

P $(A.B_1)$ = pq

Similarly given B_2 the event A will happen if both witnesses A and B speak lies and the probability for this is (1-p) (1-q) i.e.

Hence the required probability is

$$P(B_1 / A) = \frac{P(B_1)P(A / B_1)}{P(B_1)P(A / B_1) + P(B_2)P(A / B_2)}$$

$$= \frac{pq}{pq + (1-p)} = \frac{pq}{1-q-p+pq}$$

Some more examples

Example 16 : A bag contains 20 balls number from 1 to 20. One ball is drawn at random. Find the probability that it will be a multiple of 3 or 5. Out of a number of 20 balls, the number of balls which are multiple of 3 are

(3, 6, 9, 12, 15, 18) = 6

The probability of the number being a multiple 3 = p (a multiple of 3) = 6/20

The total number of balls which are multiple of number 5 (5,10,15,20)=4

Hence the probability of the number being a multiple of 5=p (a multiple of 5)=4/20

The number of 15 is both a multiple of 3 and 5, its probability = 1/20

The required probability

$$= \frac{6}{20} + \frac{4}{20} - \frac{1}{20} = \frac{9}{20}$$

Example 17 : Find the chance of throwing more than 15 in one throw with 3 dice P (more than 15) = P (either 16 or 17 or 18)

Total number of events = 6.6.6 = 216

Throwing either 16 or 17 or 18

The number of outcomes favourable to 16 are

(5, 6, 5), (6, 5, 5), (6, 6, 4), (6, 4, 6), (4, 6, 6), (5, 5, 6)

The probability of throwing $16 = \frac{6}{216} = \frac{1}{36}$

The number of outcomes favourable to 18 are (6, 6, 6) =1

The probability of throwing $18 = \frac{1}{216}$

The number of outcomes favourable to 17 are (6, 6, 5), (6, 5, 6), (5, 6, 6)

The probability of throwing 17 are = $\frac{3}{216} = \frac{1}{72}$

Required probability = $\frac{1}{36} + \frac{1}{72} + \frac{1}{216} = \frac{5}{108}$

M.A. (ECONOMICS) PART–I **PAPER–III**

SEMESTER- II **(BASIC QUANTITATIVE METHODS)**

LESSON NO. 1.6 **AUTHOR : DR. VIPLA CHOPRA**

CORRELATION ANALYSIS

Structure

1.6.1 Introduction
1.6.2 Objectives
1.6.3 Meaning of Correlation
1.6.4 Significance of Correlation Analysis
1.6.5 Correlation and Causation
1.6.6 Types of Correlation
    1.6.6.1 Positive and Negative Correlation
    1.6.6.2 Simple and Multiple Correlation
    1.6.6.3 Partial and Total Correlation
    .6.6.4 Linear and Non-Linear Correlation
1.6.7 Methods of Studying Correlation
    1.6.7.1 Scatter Diagram and Graphic Methods
    1.6.7.2 Algebraic or Mathematical Methods
1.6.8 Summary
1.6.9 Key Words
1.6.10 Further Readings
1.6.11 List of Questions
    1.6.11.1Short Questions
    1.6.11.2Long Questions

1.6.1 Introduction:

In the real world, we often encounter situations where data appears as pairs of figures relating to two or more than two variables. A correlation problem considers the joint variation of two variables. Examples of correlation problems are found in the study of the relationship between price and quantity demanded, income and consumption, investment and rate of interest, price and quantity supplied, or relation between height and weight of persons etc. In these examples, both variables are observed as they naturally occur and the extent of relationship can be measured with the help of correlation. The measure of correlation is called correlation coefficient. It means the coefficient of correlation helps us in determining the closeness of the relationship between two or more than two variables. The present lesson studies the closely related problem of correlation or the degree of relationship between variables.

1.6.2 Objective :

- The Meaning of Correlation.
- The Significance of Correlation Analysis.
- Types of Correlation.
- Methods of studying Correlation.

1.6.3 Meaning of Correlation :

If two variables X and Y, vary together in the same or in the opposite directions, then they are said to be correlated. Various experts have defined correlation in their own words and their definitions, broadly speaking, imply that correlation is the degree of association between two or more variables. Some important definitions of correlation are given below :

According to L.R. Cornor, "If two or more quantities vary in sympathy so that movement in the one tend to be accompanied by corresponding movement in the other, then they are said to be correlated."

According to Croxton and Cowdon, "When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as Correlation."

According to A.M. Tuttle, "Correlation is an analysis of the co-variation between two or more variables."

Thus we can say that correlation technique is a statistical technique which shows the relationship between the two variables.

Correlation Co-efficient : It is a numerical measure of the degree of association between two or more variables.

1.6.4 Significance of Correlation analysis :

The use or utility of the study of correlation is clear from the following points :

(i) The degree and extent of the relationship between two variables is, of course, one of the most important problems in statistics. The correlation coefficient helps us in measuring the extent of relationship between two or more than two variables.

(ii) Correlation contributes to economic behaviour. It helps us in knowing the important variables on which others depend.

(iii) It is through correlation that we can predict about the future.

(iv) the predictions made on the basis of correlation analysis are considered to be nearer to reality and hence reliable.

Thus, The technique of correlation co-efficient is the most useful tool in statistical analysis in every discipline.

1.6.5 Correlation and Causation :

Correlation analysis helps us to have an idea about the degree and direction of the relationship between the variables. But it, of course, fails to show the cause and effect relationship between the variables. In a bivariate distribution, if the variables have the cause and effect relationship, there is bound to be high degree of correlation between them. It means causation implies correlation. But the contrawise is not true. The high degree of correlation may be due to the following reasons :

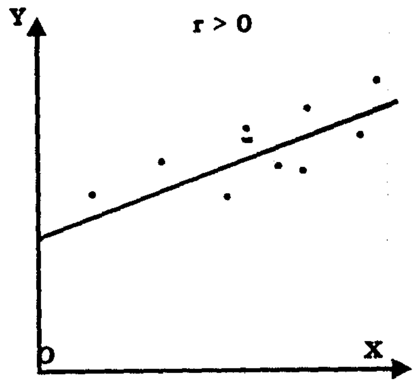
1. Correlation may be due to pure chance or sheer coincidence. Given the data on any two variables, one may obtain a high value of correlation co-efficient, when in fact they do not have any relationship. For example, a high value of correlation co-efficient may be obtained between the income of a person and the size of shoe.
2. The two variables may be acted upon by the outside influences. For instance, the price of sugar and shirt may be correlated due to the fact that these two are related to third variable, i.e. purchasing power.
3. Mutual dependence. Cause and effect relation exists in this case also but it may be very difficult to find out which of the two variables is independent. For instance, if we have data on price of rice and its cost of production, the correlation between them may be very high because higher price of rice may attract farmers to produce more rice and more production of rice may mean higher cost of production, assuming that it is an increasing cost industry. Further, the higher cost of production may in turn raise the price of rice. For the purpose of determining a relationship between two variables in such situations, we can take any one of them as independent variable.

1.6.6 Types of Correlation :

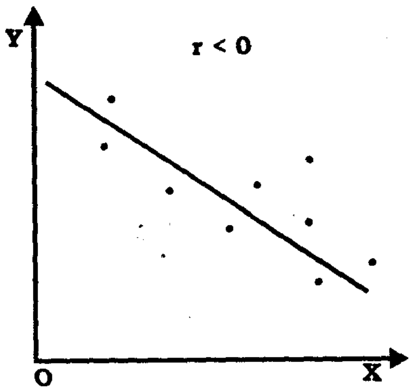
Correlation is described in the following four ways :

1.6.6.1 Positive and Negative Correlation :

Whether correlation is positive or negative would depend upon the direction in which the variables are moving. If both the variables are moving in the same direction, i.e. if with an increase in one variable, the other also increases or with a decrease in one variable, the other also decreases, correlation is said to be positive. On the other hand if they are moving in oppositive directions, correlation is said to be negative.



Positive Correlation



Negative Correlation

Example :

Positive Correlation

X	:	5	6	9	10	12	15
Y	:	10	12	14	16	17	20

Negative Correlation

X	:	5	6	9	10	12	15
Y	:	20	17	16	14	12	10

1.6.6.2 Simple and Multiple Correlation :

The study of correlation for two variables involves application of simple correlation. In multiple correlation three or more variables are studied simultaneously. Multiple correlation consists of the measurement of the relationship between a dependent variable and two or more independent variables.

Example : When we study the relationship between the age in years and both the weight and height of a group of persons, it is a problem of multiple correlation when we study the relationship between income and consumption, it is a problem of simple correlation.

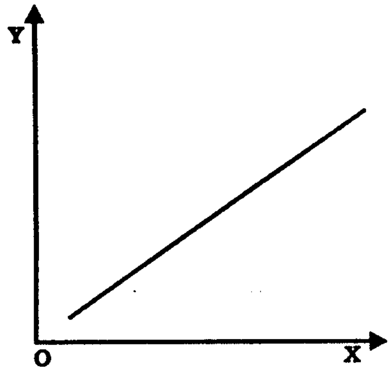
1.6.6.3 Partial and Total Correlation :

Partial Correlation : When there are more than two variables and the relationship between any two of the variables is studied assuming other variable as constant it is a case of partial correlation symbollically if x, y, z are three variable then partial correlation between x and y excluding z will be given by $r_{xy.z}$. Similarly we can calculate $r_{xz.y}$ or $r_{yz.x}$.

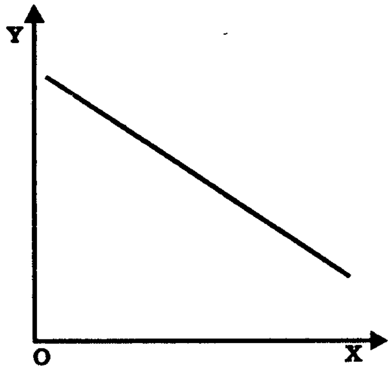
Total Correlation: When the correlation between the variables under study taken together at a time, is worked out, it is called total correlation.

1.6.6.4 Linear and Non-Linear Correlation :

Linear Correlation : The correlation between two variables will be linear if corresponding to a unit change in one variable, there is a constant change in the other variable over the entire range of values.



Linear
Positive Correlation

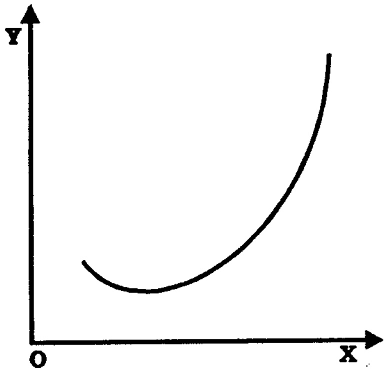


Linear
Negative Correlation

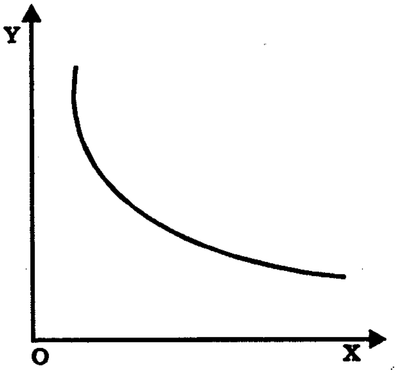
Example :

X	1	2	3	4	5
Y	3	6	9	12	15

Non-Linear Correlation : The relationship between two variables will be non-linear or curvi-linear, if corresponding to a unit change in one variable, the other variable changes at a different rate.



Curvilinear and
Positive Correlation



Curvilinear and
Negative Correlation

Exercise 1

1. What is meant by Correlation ?
2. Value of correlation always lies between

(a) -1 to 0 (b) 0 to 1 (c) -1 to +1 (d) None of these.

3. If two series move in the same direction, the correlation is said to be

(a) positive (b) negative (c) zero (d) none of these.

4. If the amount of change in one variable tends to bear constant ratio to the amount of change in other variable then the correlation is said to be

(a) Linear (b) Non-linear (c) Neither of the two.

1.6.7 Methods of Studying Correlation

We can study the methods of correlation by classifying them as follows

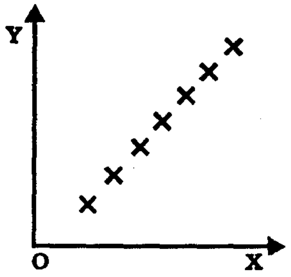
(a) methods based on Graphs and Diagrams

(b) Mathematical Methods

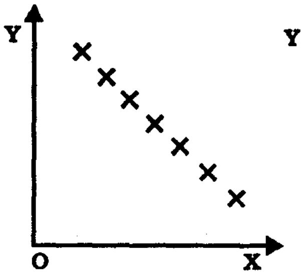
1.6.7.1 Scatter Diagram Method and Graphic Method

Scatter Diagram Method

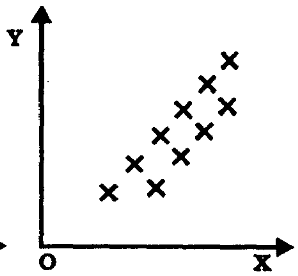
Scatter diagram is one of the simplest method of diagrammatic representation of a bivariate (two variables) distribution. Suppose we are given n number of pairs of values i.e. (x_1, y_1), (x_2, y_2)---------(x_n, y_n) of X and Y variables. Plot these n points in the xy plane. The diagram so obtained will be called "Scatter Diagram". From the diagram, we can make a rough idea about the relationship between the two variables. The following diagrams exhibit the different types of correlation.



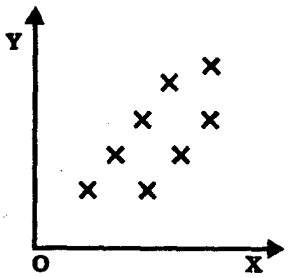
Perfect Positive Correlation



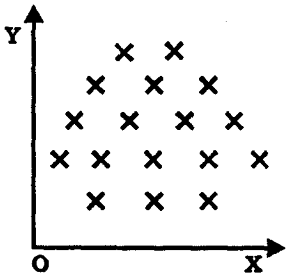
Perfect Negative Correlation



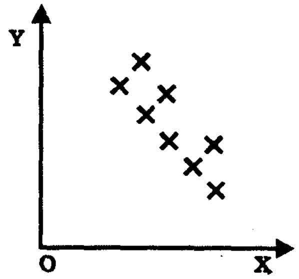
High Degree of Positive Correlation



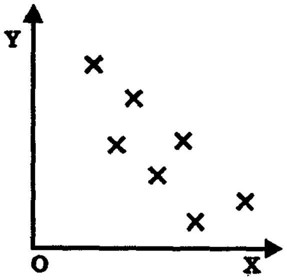
Low Degree of Positive Correlation



No Correlation



High Degree of Negative Correlation



Low Degree of Negative Correlation

Graphic Method :

It is the simplest method to determine the presence of correlation between the two variables. If the values of the two variables are plotted on the graph paper we will get two curves, one for X variable and another for Y variable individually. By looking at the direction and closeness of the two curves, we can judge whether the variables are related or not. If these two curves move in the same direction, correlation is said to be positive. On the other hand, if the curves are moving in the opposite directions, correlation is said to be negative.

1.6.7.2 Algebraic or Mathematical Methods :

The following three methods are generally discussed for studying the correlation mathematically:

1. Karl Pearson's Method
2. Rank Correlation Method
3. Concurrent Deviation Method.

We will discuss the first two methods only.

1. Karl Pearson's or Covariance Method :

Karl Pearson, a reputed statistician, has constructed a well set formula in 1890 based on mathematical treatment. Karl Pearson's method, popularly known as Correlation Coefficient, is most widely used. It is denoted by the symbol "r". Karl Pearson's coefficient of correlation measures both direction and degree of relationship. The value of r lies between ± 1.

(a)	If $r = +1$	: Perfect Degree of Positive Correlation.
(b)	If $r = 0$	: No Correlation
(c)	If $r = -1$	: Perfect Degree of Negative Correlation
(d)	If $0.75 \leq r \leq 1$	: High Degree of Positive Correlation.
(e)	If $0.5 \leq r \leq 0.75$	: Moderate Degree of Positive Correlation
(f)	If $0 \leq r \leq 0.5$	: Low Degree of Positive Correlation.
(g)	If $-0.75 \leq r \leq -1$	: High Degree of Negative Correlation.
(h)	If $-0.5 \leq r \leq -0.75$	: Moderate degree of Negative Correlation.
(i)	If $0 \leq r \leq -0.5$	: Low Degree of Negative Correlation.

However, in practice, such values of r as + 1, -1 and 0 are rare.

The calculation of Karl Pearson's Coefficient of correlation can be done (i) in case of individual series or ungrouped data and (ii) in case of grouped data.

Calculation of Coefficient of Correlation (Ungrouped Data)

There are two methods of measuring correlation coefficient.

(i) Direct Method (Actual Mean Method) : When the deviations are taken from the actual means :

$$r = \frac{\sum xy}{N\sigma_x \times \sigma_y}$$

Where x = x – $\overline{x}$ = Deviations in X series from its actual mean.

y = y – $\overline{y}$ = Deviations in Y series from its actual mean.

N = Number of paired observations.

σ_x = Standard deviation of x series

σ_y = Standard deviation of y series.

Alternatively $r = \frac{\sum xy}{N\sigma_x \times \sigma_y}$

$$r = \frac{\sum xy}{N\sqrt{\frac{\sum x^2}{N}} \times \sqrt{\frac{\sum y^2}{N}}} = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

Example 1: Calculate the coefficient of correlation from the following data:

X :	10	6	9	10	12	13	11	9
Y :	9	4	6	9	11	13	8	4

Solution

x	**x = x –** $\overline{x}$	$\mathbf{x^2}$	**y**	**y = Y –** $\overline{Y}$	$\mathbf{y^2}$	**xy**
10	0	0	9	1	1	0
6	-4	16	4	-4	16	16
9	-1	1	6	-2	4	2
10	0	0	9	+1	1	0
12	2	4	11	3	9	6
13	3	9	13	5	25	15
11	1	1	8	0	0	0
9	-1	1	4	-4	16	4
∑x=80	**0**	**32**	**64**	**0**	**72**	**43**

$$\overline{x} = \frac{80}{8} = 10, \overline{y} = \frac{\sum y}{N} = \frac{64}{8} = 8$$

$$r = \frac{\sum xy}{\sqrt{\sum x^2 . \sum y^2}} = \frac{43}{\sqrt{32 \times 72}} = \frac{43}{\sqrt{2304}} = \frac{43}{48} = +0.896$$

(2) Short-Cut Method (Assumed Mean Method) :

In practice actual means are in fractions. To avoid difficult calculations, deviations are taken from assumed mean. Thus when deviations are taken from an assumed mean the following formula is applicable :

$$r = \frac{\sum dxdy - \frac{\sum dx.\sum dy}{N}}{\sqrt{\sum d^2x - \frac{(\sum dx)^2}{N}}\sqrt{\sum d^2y - \frac{(\sum dy)^2}{N}}}$$

dx = X-A = Deviation of X series from an assumed mean.

dy = Y-A = Deviation of Y series from an assumed mean.

$\sum$dxdy = Sum of the product of the deviations of X and Y series from their assumed means.

$\sum dx^2$ = sum of the squares of the deviations of X series from an assumed mean.

$\sum dy^2$ = sum of the squares of the deviations of Y series from an assumed mean.

Example 2: Calculate the coefficient of correlation from the following data:

Marks in statistics : 20 30 28 17 19 23 35
13 16 38

Marks in Law : 18 35 20 18 25 28 33
18 20 40

Solution

Marks in Statistics	**X=30**		**Marks in Law**	**Y=30**		
X	**dx**	**dx^2**	**y**	**dy**	**dy^2**	**dxdy**
20	-10	100	18	-12	144	120
30	0	0	35	+5	25	0
28	-2	4	20	-10	100	20
17	-13	169	18	-12	144	156
19	-11	121	25	-5	25	55
23	-7	49	28	-2	4	14
35	+5	25	33	+3	9	15
13	-17	289	18	-12	144	204
16	-14	196	20	-10	100	140
38	+8	64	40	+10	100	80
	$\sum dx = -61$	$\sum dx^2 = 1017$		$\sum dy = -45$	$\sum dy^2 = 795$	$\sum dxdy = 804$

$$r = \frac{804 - \frac{(-61)\,(-45)}{10}}{\sqrt{1017 - \frac{(-61)^2}{10}} \times \sqrt{795 - \frac{(-45)^2}{10}}}$$

$$r = \frac{8040 - 3355}{\sqrt{(10170 - 3721)\cdot(7950 - 3025)}}$$

$$r = \frac{4685}{\sqrt{(6449)(4925)}} = \frac{4685}{5635.8} = 0.85$$

Coefficient of Correlation in Grouped Data

In case of large number of observations, the data is classified into two way frequency distribution called bivariate frequency table or correlation table. The class intervals for Y are listed in the columns headings and those for X are listed in the steps at the left of the table.

The formula for calculating coefficient of correlation is

$$r = \frac{\sum \text{fdxdy} - \frac{(\sum \text{fdx})(\sum \text{fdy})}{N}}{\sqrt{\sum \text{fdx}^2 - \frac{(\sum \text{fdx})^2}{N}}\sqrt{\sum \text{fdy}^2 - \frac{(\sum \text{fdy})^2}{N}}}$$

Example 3:

The following table gives the ages of husbands and wives at the time of their marriage. Calculate the correlation coefficient between the ages of husbands and wives.

Ages of Wives→ Ages of Husbands↓	30-40	40-50	50-60	60-70
30-40	20	26	--	--
40-50	8	14	37	--
50-60	--	4	18	3
60-70	--	--	4	6

Solution :

Ages of wives Ages of Husbands			**x** **m** **dx**	**30-40** 35 -10	**40-50** 45 0	**50-60** 55 +10	**60-70** 65 +20				
Y	**m**	**dy**	**d^y d^x**	**-1**	**0**	**+1**	**+2**	**f**	**dy**	**fd^2y**	**fdxdy**
30-40	35	-10	-1	20 $\boxed{20}$	26 $\boxed{0}$			46	-46	46	20
40-50	45	0	0	8 $\boxed{0}$	14 $\boxed{0}$	37 $\boxed{0}$		59	0	0	0
50-60	55	+10	1	--	4 $\boxed{0}$	18 $\boxed{18}$	3 $\boxed{6}$	25	25	25	24
60-70	65	+20	2			4 $\boxed{8}$	6 $\boxed{24}$	10	20	40	32
			f	28	44	59	9	∑f= 140	-1	111	76
			fdx	-28	0	59	18	49			
			fd^2x	28	0	59	36	123			
			fdxdy	20	0	26	30	76			

$$r = \frac{76 - \frac{(49)(-1)}{140}}{\sqrt{123 - \frac{(49)^2}{140}}\sqrt{111 - \frac{(-1)^2}{140}}}$$

$$= \frac{76 - 0.35}{\sqrt{123 - 17.15}\sqrt{111 - 0.007}}$$

$$= \frac{76.35}{10.3 \times 10.54}$$

$$r = \frac{76.35}{108.56} \quad \text{Or} \quad 0.703$$

Algebraic Properties of Pearson's Co-efficient of Correlation

Prof. Karl Pearson's coefficient of correlation has the following algebraic properties :

1. Its value must lie between +1 and -1 i.e. $-1 \leq r \leq 1$
2. It is independent of the change of origin and scale as well.

3. It is independent of the units of measurement.
4. It is independent of the order of comparison of the two variables.

   Symbolically, $r_{xy} = r_{yx}$.

   This is because,

$$r_{xy} = \frac{\sum xy}{N\sigma_x \cdot \sigma_y} = \frac{\sum yx}{N\sigma_y \cdot \sigma_x} = r_{yx}$$

Merits

1. It is the most popular method for expressing the degree and direction of linear association between the two variables.
2. It is based on all observations of the series.
3. Karl Pearson's coefficient of correlation is a pure number independent of units of measurement. Therefore, the comparison between the series can be done easily.

Demerits :

1. Coefficient of correlation does not give any idea about the existence of cause and effect relationship.
2. Coefficient of correlation assumes linear relationship between the variables regardless of the fact whether that assumption is correct or not.
3. Compared with some other methods, this method is more time consuming and cumbersome.
4. Its value is unduly affected by extreme values.

Exercise 2

1. Spell out Pearson's coefficient of Correlation. What are its Limitations ?
2. Find Karl Pearson's Coefficient of Correlation between income and weights for the data :-

Income (Rs):	100	200	300	400	500	600
Weights (Ibs):	110	120	135	140	160	165

(ii) Spearman's Rank Correlation

Prof. Charles Spearman has devised a method of computing coefficient of correlation in 1904. It is based on the ranking of various item-values of the variables.

There are three cases :-

1. When Ranks are given
2. When Ranks are not given
3. Equal Ranks

(i) When Ranks are given :

In Spearman's coefficient of Correlation, we take the differences in ranks, squaring them and finding out the aggregate of the squared differences. Symbolically :

$$r_k = 1 - \frac{6\sum D^2}{N\left(N^2 - 1\right)}$$

r_k = Coefficient of Rank Correlation
D = Rank differences
N = Number of Pairs

Example 4 :

The rankings of ten students in Statistics and Economics are as follows:

statistics :	3	5	8	4	7	10	2	1	6	9
Economics :	6	4	9	8	1	2	3	10	5	7

What is the coefficient of rank correlation ?

Solution :

Ranks:

Ranks in		**Rank**	**Square of**
Statistics	**Economics**	**Differences**	**Rank Differences**
R_1	R_2	R_1-R_2 = D	D^2
3	6	-3	9
5	4	+1	1
8	9	-1	1
4	8	-4	16
7	1	+6	36
10	2	+8	64
2	3	-1	1
1	10	-9	81
6	5	+1	1
9	7	+2	4
	Total	**$\sum$D = 0**	**$\sum D^2$=214**

$$r_k = 1 - \frac{6\sum D^2}{N\left(N^2 - 1\right)} = 1 - \frac{6(214)}{10\left(10^2 - 1\right)}$$

$$= 1 - \frac{1284}{10(99)} = 1 - 1.3 = -0.3$$

When Ranks are not given

When actual values are given then we assign ranks and apply the Spearman's formula for Rank Correlation.

Example 5:

Given X and Y, compute Rank Correlation.

X:	115	134	120	130	124	128
Y:	130	132	128	131	127	125

Solution :-

X	**R_1**	**Y**	**R_2**	**$D=R_1-R_2$**	**D^2**
115	6	130	2	3	9
134	1	132	1	0	0
120	5	128	4	1	1
130	2	131	2	-0	-0
124	4	127	5	-1	1
128	3	125	6	-3	9
		Total			**$\sum D^2 = 20$**

$$r_k = 1 - \frac{6\sum D^2}{N(N^2-1)} = 1 - \frac{6(20)}{(6^3-6)}$$

$$= 1 - \frac{120}{210} = \frac{90}{210} = 0.43$$

Equal Ranks :

In certain cases, we may find equal ranks in case of two or more than two values. In that case, each individual item is given an average rank. For example if two individuals are ranked equal at third place, they are each given the rank $\frac{3+4}{2} = 3.5$, while, if three are ranked at third place then $\frac{3+4+5}{3} = 4$ would be the common rank for third, fourth and fifth place.

Rank correlation is calculated by this formula.

$$r_k = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \ldots\ldots\ldots\ldots\right\}}{N^3 - N}$$

Example 6:

Calculate the Spearman's Rank 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 :

Series A	**Series B**	**Rank of Series A (X)**	**Rank of Series B (Y)**	**D=x-y**	**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
	Total				**$\sum D^2$ =3.00**

$$r_k = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right) + \ldots\ldots\ldots\ldots\right\}}{N^3 - N}$$

$$r_k = 1 - \frac{6\left\{3 + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(2^3 - 2\right)\right\}}{9^3 - 9}$$

$$r_k = 1 - \frac{6\{3+1\}}{9(81-1)} = 1 - \frac{24}{9 \times 80} = 0.967$$

Merits :

1. It is easy to calculate and understand as compared to Pearson's r.
2. When the ranks of different values of the variables are given, it is then the only method left to calculate the degree of correlation.
3. This method is employed usefully when the data is given in a qualitative nature like beauty, honesty etc.

Demerits :

1. This method cannot be employed in a grouped frequency distribution.
2. If the items exceed 30, it is then difficult to calculate.

Concurrent Deviation Method

It.is a very simple and easy to calculate method of correlation. In this method, correlation is calculated the direction of deviations, not their magnitudes. If the deviations of two times series are concurrent, has would move in the same direction and would indicate positive correlation between them. send of concurrent deviation is calculated on this very principle and ordinarily, it indicates the ship between shor time fluctuations only.

Method of Calculating Coefficient of Concurrent Deviation

Method involving the following steps:

1) First of all deviations of both the series are calculated separately. Take a series, the deviation of its every item will depend upon the value of previous item. If the second item is bigger than the first item, then show this by putting '+' sign against the second item in a new colum haeaded by deviation dx. If the second item is smaller then put '-' sign and if equal put '=' sign, which shows no change. This process is continued till the whole series is exhausted.

2) Similarly, we compute deviation in second series and show them in another column headed by deviation dy.

3) Construct another column of the products of dx and dy, i.e. dxdy. This coloum is denoted by concurrent deviation.

4) Find number of pairs of concurrent deviations, i.e., C.

5) Use the following formula:

$$r_C = \pm \sqrt{\pm \frac{2C-N}{N}}$$

Where, r_C = Coefficient of concurrent deviation

C = Number of concurrent deviation

N = Number of pairs of deviations

The use of '+' or '-' sign will depend upon the sign of $\frac{2C-N}{N}$

If it is minus thena minus is placed, if it is plus then plus sign is placed. It is important to note that value of coefficient of concurrent deviation lies between + 1 and -1.

Example 5 Find coefficient of correlation from the following data by the concurrent deviation method

X:	85	91	56	72	95	76	89	51	59	90
Y:	18.3	20.8	16.9	15.7	19.2	18.1	17.5	14.9	18.9	13.8

Solution

X	Deviation in X Series (dx)	Y	Deviation in Y Series (dy)	Concurrent Deviation dxdy
85		18.3		
91	+	20.8	+	+
56	-	16.9	-	+
72	+	15.7	-	-
95	+	19.2	+	+
76	-	18.1	-	+
89	+	17.5	-	-
51	-	14.9	-	+
59	+	18.9	+	+
90	+	15.4	-	-
	N=9		N=9	C=6

Putting these values in the formula:

$$r_c = \pm\sqrt{\pm\frac{2C-N}{N}}$$

$$r_c = +\sqrt{+\frac{2*6-9}{9}}$$

$$r_c = +\sqrt{+\frac{3}{9}} = \sqrt{0.333} = 0.577$$

1.6.8 Summary:

In this lesson the concept of correlation or the association between two variables has been discussed. The significance of correlation analysis as the most useful tool in stastical analysis in every discipline has been studied. Among the different methods of studying correlation, a scatter plot the variables may suggest that two variables are related but the value of the Pearson Correlation Coefficient r quantifies this association. The correlation coefficient r may assume values between –1 and +1. The sign indicates whether the association is direct (+ve) or inverse (–ve). A numerical value of r equal to unity indicates perfect association while a value of zero indicates no association.

Spearman's Rank Correlation for data with ranks is outlined.

Key Words :-

1. **Correlation :** Degree of Association between two variables.
2. **Correlation Coefficient :** A number lying between –1 (Perfect Negative Correlation) and +1 (Perfect Positive Correlation) to quantify the association between two variables.
3. **Scatter Diagram :** An ungrouped plot of two variables, on the X and Y axes.
4. **Linear Correlation :** The relationship between two variables will be linear, if corresponding to a unit change in one variable, the other variable changes at the same rate.
5. **Positive Correlation :** It the two variables move in the same direction, the correlation is said to be positive.
6. **Coefficient of Determination :** Coefficient of determination is the square of the coefficient of correlation.
7. **Negative correlation :** If the two variables move in the opposite direction the correlation, is said to be negative.

1.6.10. Further Readings

W.J. Stevenson	:	Business statistics – Concepts and Applications
S.C.Gupta and V.K. Kapoor	:	Fundamental of Mathematical Statistics.
Murray R. Spiegel	:	Statistics.
S.P. Gupta	:	Statistical Methods
Suranjan Saha	:	Practical Business Mathematics and Statistics

1.6.11 List of Questions

1.6.11.1 Short questions

1. What is Correlation? State its importance.
2. State Karl Pearson's Coefficient of Correlation. Give its formula. What are its limitations ?
3. What is Linear and Non-Linear Correlation ?
4. What is meant by Rank Correlation ?
5. What is meant by Coefficient of Concurrent Deviation ?
6. Discuss the merits and demerits of Rank Correlation.
7. If the difference of Ranks in each pair is zero, what is the Rank Correlation Coefficient.
8. Discuss the various types of Correlation.
9. Distinguish between Positive and Negative Correlation.

10. What are the limits of value 'r' ? What do positive, negative and zero value of 'r' indicate ?

1.6.11.2 Long Questions

1. Calculate Karl Pearson's coefficient of correlation between Imports and Exports.

Imports (Rs. billion)	42	44	58	55	89	98	66
Exports (Rs. billion)	56	49	53	58	65	76	58

2. From the following distribution find the correlation between age and playing habits of students and regular players.

Age :	15-16	16-17	17-18	18-19	19-20	20-21
No. of Students:	200	270	340	360	400	300
Regular Players:	150	162	170	180	180	90

3. Find out coefficient of correlation between x and y from the following table.

x→ Y↓	0-20	20-40	40-50	50-60	Total
10-20	4	2	2	—	8
20-30	5	4	6	4	19
30-40	6	8	10	11	35
40-50	4	6	8	4	22
50-60	—	2	4	4	10
60-70	—	2	3	1	6
Total	**19**	**24**	**33**	**24**	**100**

4. Calculate Rank Correlation Coefficient for the following :

Marks in English:	29	28	17	15	20	26	27	25	34	19
Marks in Maths:	31	32	25	29	42	15	43	32	20	40

5. Two Judges gave the following Ranks to a series of one act plays in a Drama Competition. Examine the relationship between their judgement:

Judge A :	8	7	6	3	2	1	5	4
Judge B :	7	5	4	1	3	2	6	8

6. Calculate coefficient of Correlation

X :	150	154	160	172	160	164	180
Y :	200	180	170	160	190	180	172

7. Given $N = 10$, $\sum x = 140$, $\sum y = 150$, $\sum (x-10)^2 = 1800$
$\sum (y-15)^2 = 215$, $\sum (x-10)\ (y-15) = 60$
Find Karl Pearson's coefficient of correlation and estimate the value of y when x=15

8. Define Karl Pearson's Coefficient of Correlation and discuss the methods of studying it.

9. The mean of two series is 40 and 50 respectively. The number of items of the series is 30 each having S.D. 6 and 7. If the total of products of deviations of two series from their respective means be 360. Find the coefficient of correlation between the two.

10. What is meant by rank correlation? When it is used and what are its uses?

M.A. (ECONOMICS) PART–I **PAPER–III**

SEMESTER-II **(BASIC QUANTITATIVE METHODS)**

LESSON NO. 1.7 **AUTHOR : DR. VIPLA CHOPRA**

REGRESSION ANALYSIS

Structure

1.7.1 Introduction

1.7.2 Objectives

1.7.3 Definitions

1.7.4 Regression Lines

  1.7.4.1 Principle of Least Squares

1.7.5 Regression Equations

1.7.6 Methods of Studying Regression Analysis

  1.7.6.1 Graphic Method

  1.7.6.2 Algebraic Method

1.7.7 Difference between Correlation and Regression

1.7.8 Uses of Regression Analysis

1.7.9 Summary

1.7.10 Key Words

1.7.11 Further Readings

1.7.12 List of Questions

  1.7.12.1 Short Questions

  1.7.12.2 Long Questions

1.7.1 Introduction

In the previous lesson on Correlation analysis we have studied the extent and degree of relationship between two variables. Similarly, we can estimate or predict the value of a variable given the value of another variable on the basis of functional relationship between them. The statistical technique of estimating or predicting the unknown value of a dependent variable from the known value of an independent variable is called regression analysis. The literal or dictionary meaning of the word 'Regression' is 'stepping back or returning to the average value.' The term was first used by British biometrician Sir Francis Galton in the later part of the 19th century in connection with some studies he made on estimating the extent to which the stature of the sons of tall parents reverts or regresses back to the mean stature of the population. He studied the relationship between the heights of about one thousand fathers and sons and published the results in a paper 'Regression towards Mediocrity in Hereditary Stature.' The features of this study were :

(i) The tall fathers have tall sons and short fathers have short sons.

85

(ii) The average height of the sons of a group of tall fathers is less than that of the fathers and the average height of the sons of a group of short fathers is more than that of the fathers.

The line showing this tendency to go back was called by Galton a "Regression Line". The modern statistician use the term 'estimating line' instead of regression line as this concept is more classificatory now.

It is clear from above that regression analysis is a statistical device with the help of which we are in a position to estimate (or predict) the unknown values of one variable from known values of another variable. The variable whose value is influenced or is to be predicted is called dependent variable and the variable which influences the values or is used for prediction, is called independent variable. In regression analysis independent variable is also known as regressor or predictor or explanator while the dependent variable is also known as regressed or explained variable.

1.7.2 Objectives :

The major objectives of the present lesson are :

* To study the functional relationship between dependent and independent variables from given data and thereby provide a mechanism for prediction.
* To draw two regression lines based on least square assumption.
* Different methods of finding out regression equations.
* Properties of regression coefficients.
* Differences between correlation and regression.
* Utility and limitations of regression analysis.

1.7.3 Definitions

According to **Taro Yamane,** "One of the most frequently used techniques in economics and business research, to find a relation between two or more variables that are related causally, is regression analysis."

Ya Lun Choue defines it as, "Regression analysis attempts to establish the nature of relationship between variables----and thereby provide a mechanism for prediction, or fore casting."

Morris Hamburg claims, "The term regression analysis refers to the methods by which estimates are made of the values of a variable from a knowledge of the values of one or more other variables and to the measurement of the errors involved in this estimation process."

Thus, from the above definitions, we can derive that the regression analysis is an absolute measure with the help of which we can estimate or predict the unknown values of a variable from the known values of another variable.

1.7.4 Regression Lines

A regression line is a graphic technique to show the functional relationship between the two variables X and Y i.e. dependent and independent variables. In case of two variables x and y, we shall have two lines of regression; one of Y on X and the other of X on Y.

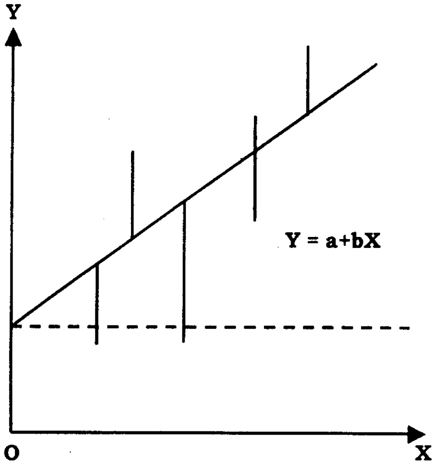
Definition : Line of regression of Y on X is the line which gives the best estimate for the value of Y for any specified value of X.

Similarly, line of regression of X on Y is the line which gives the best estimate for the value of X for any specified value of Y.

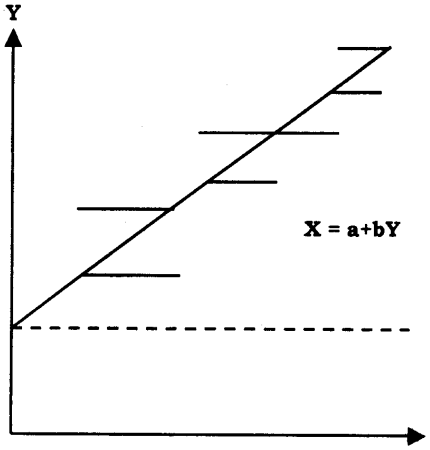
1.7.4.1 Principle of Least Squares :

The lines of regression are drawn on least square assumption. According to the least square method the line of regression should be drawn through the plotted points in such a way that the sum of the squares of the deviations of the actual Y values from the computed 'Y' values is the minimum or the least. The line which fits the points in the best manner should have $\sum(Y-Y_C)^2$ as minimum. A line fitted by this method is called the line of best fit. One of the characteristics of the line of best fit is that the deviations above the line are equal to the deviations below the line. It implies that the total of the positive and negative deviations is zero, i.e. $\sum(Y-Y_C)=0$. The line of best fit or the straight line goes through the over all mean of the data i.e. $\overline{X},\overline{Y}$.

The regression line of Y on X is drawn in such a way that it minimises total of squares of the vertical deviations and the regression line of X on Y minimises the total of the squares of the horizontal deviations.



Regression line of Y on X
$\sum(Y-Y_c)^2$ is Minimum
Fig. 1



Regression line of X on Y
$\sum(X-X_c)^2$ is Minimum
Fig. 2

1.7.5 Regression Equations/Estimating Lines

Regression lines are based on regression equations. These are also known as estimating equations. These are algebraic expression of regression lines. As there are two regression lines, so there are two regression equations i.e. the regression equation of X on Y which shows the variation in the values of X for given changes in Y and the regression equation of Y on X.

The regression equation of Y on X is expressed as follows :

$$Y = a + b X$$

Here Y is dependent variable. X is independent variable. 'a' is "Y-intercept" because its value is the point at which the regression line crosses the Y-axis, that is, the vertical axis. 'b' is the slope of the line. It represents change in Y variable for a unit change in X variable. Regression equation of X on Y is expressed as follows :

$$X = a' + b'Y$$

Here Y is Independent variable and X is dependent variable.

a' is the "X-intercept" because its value is the point at which the regression line crosses the X-axis, i.e. the horizontal axis. b' is the slope of the line It represents change in X variable for a unit change in Y variable.

Exercise 1

1. Explain regression.
2. State clearly the concept of regression lines.

1.7.6 Methods of Studying Regression Analysis

Broadly speaking there are two different methods of studying simple regression between two related variables. They are :

1. Graphic Method and
2. Algebraic Method.

1.7.6.1 Graphic Method :

Under this method, one or two regression lines are drawn on a graph paper to estimate the values of one variable say X on the basis of the given value of another variable say, Y. If there is a perfect correlation (i.e. r =1) between the two variables, only one regression line can be drawn for in that case both the regression lines of X on Y and Y on X will coincide. In case the correlation between the two variables is not perfect, two lines of regression are to be drawn on the graph paper one of which will be the regression line of X on Y and the other will be the regression line of Y on X. If there is no correlation between the two variables, then the two lines of regression will be perpendicular to each other.

The different forms of the regression lines under different state of correlation are exhibited here as under :

(i) Where there is perfect positive correlation i.e. r = +1

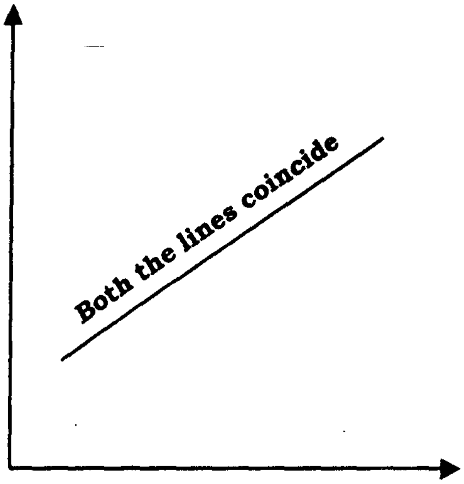


Fig.3

(ii) When there is perfect negative correlation i.e. r = -1

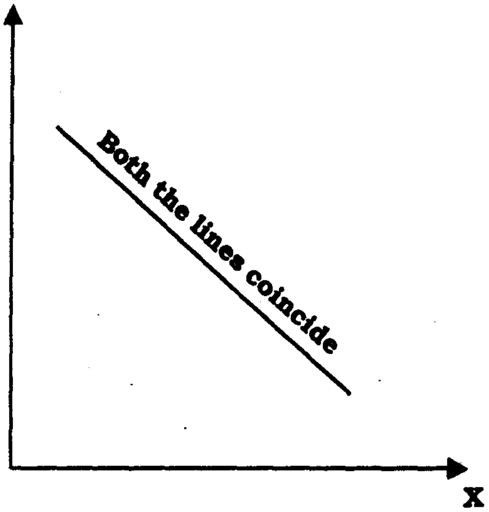


Fig.4

(iii) When there is no correlation i.e. r = 0

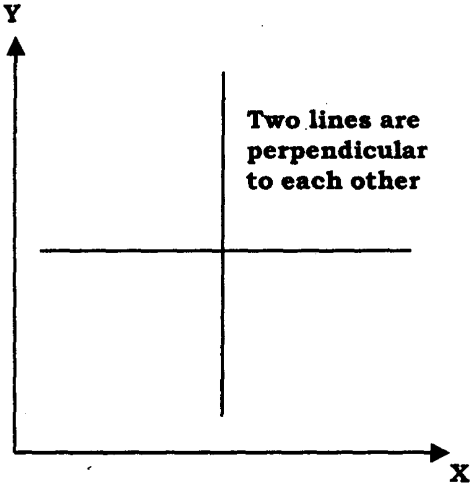


Fig. 5

(iv) When there is a high degree of correlation

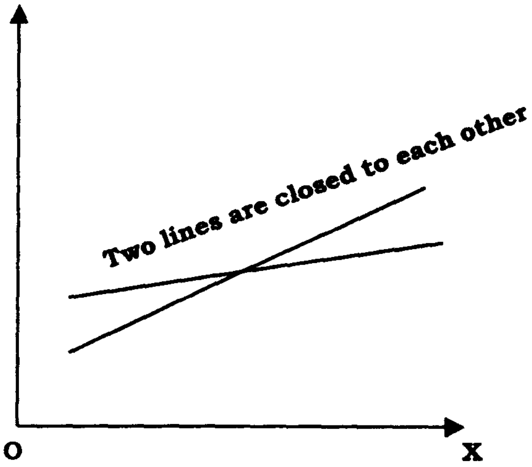


Fig. 6

(v) When there is a low degree of correlation

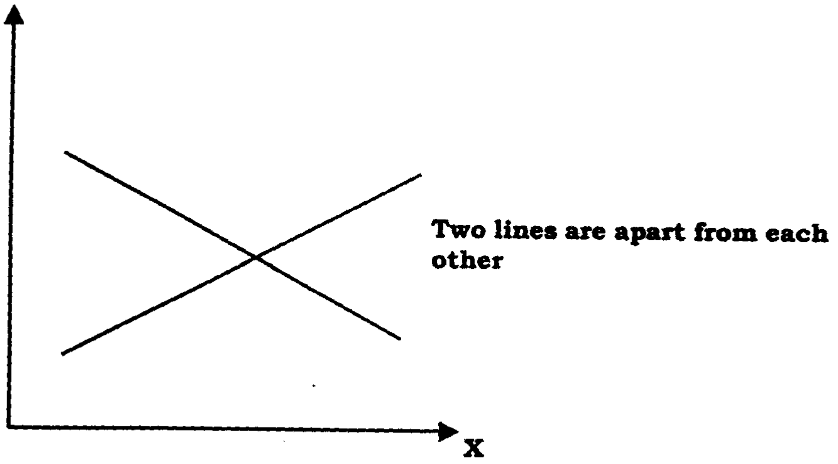


Fig. 7

There are two methods of drawing regression lines

(a) Scatter diagram method (b) Method of Least Squares.

(a) Scatter diagram method

Under this method a graph paper is taken on which the independent variable say, X is represented along the Horizontal axis and the dependent variable say, Y is represented along the vertical axis. The points are then plotted on the graph paper representing the various pair of values of both the variables X and Y which give the picture of a scatter diagram with several points scattered around. After this two free-hand straight lines are drawn across the Scattered points in such a manner that the sum of the deviations of the points on one side of a line is equal to sum of the deviations of the points on its other side. The line which is drawn in between such vertical deviations is represented as the regression line of Y on X and the line which is drawn in between such horizontal deviations is represented as the line of regression of X on Y. The point at which both these regression lines cut each other represents the **Mean** of the two variables.

Example 1 : Given the following pairs of values of variable X and Y.

X	2	3	5	6	8	9
Y	6	5	7	8	12	11

By graphic inspection draw an estimating line.

Solution : Taking X on the X-axis and Y on the Y-axis, pairs of observations are plotted on a graph paper. Then a free hand estimating line is drawn in such a manner that the sum of the positive and negative deviations on either side of the line is zero. (Fig.8)

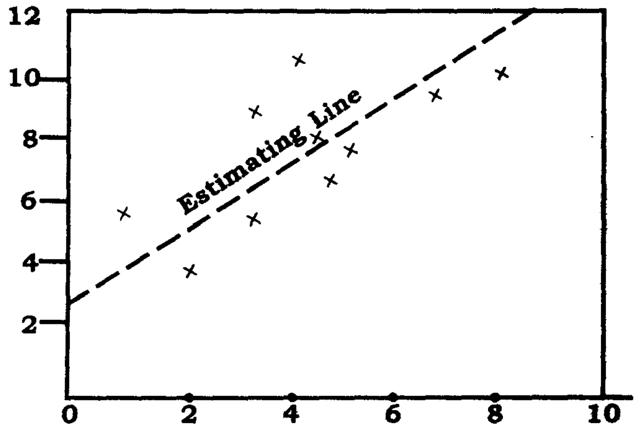


Fig.8

Merits and Limitations of the Method

This method is very simple and easy. It does not take much time to draw the estimating line. As it is drawn free-hand, different persons may draw different lines for the same data. With practice, it is not difficult to draw an approximately correct line.

(b) Method of Least Square

The other method of drawing a line of regression is the method of least squares. The method of least square has been explained earlier in section 22.4.

Exercise 2

1. Explain the relationship between correlation and regression lines.
2. Explain the scatter diagram method of drawing regression lines.

1.7.6.2 Algebraic Method

Under this method the two regression equations are formulated to represent the two regression lines or the lines of estimates respectively viz.,

(i) The regression line of X on Y and

(ii) The regression line of Y on X.

To obtain such equations we are to apply any of the following algebraic methods :

(i) Normal equation method

(ii) Method of deviation from the actual means; and

(iii) Method of deviation from the assumed Means.

(i) Normal Equation Method

The two main equations generally used in regression analysis are : (i) Y on X, (ii) X on Y.

for Y on X, the equation is $Y_c = a + b X$

For X on Y, the equation is $X_c = a + b Y$

a and b are constant values and 'a' is called the intercept. 'b' represents the slope of the line.

Regression Equation of Y on X

$Y_c = a + b X$

We can arrive at two normal equations as follows

Given $Y = a + b X$

$\sum Y = Na + b\sum X$ ....(i)

$\sum XY = a\sum X + b\sum X^2$ ....(ii)

Equations (i) and (ii) are called normal equations.

Regression Equation of X on Y

The regression equation of X on Y is expressed as

$X = a + b Y$

For determining the values of 'a' and 'b' we determine two normal equations which can be solved simultaneously.

$\sum X = Na + b\sum Y$ ....(iii)

$\sum XY = a\sum Y + b\sum Y^2$ ....(iv)

Equations (iii) and (iv) are normal equations.

Example 2 : Given the bivariate data :

X :	1	5	3	2	1	1	7	3
Y :	6	1	0	0	1	2	1	5

(a) Fit the regression line of Y on X and hence predict Y if X = 10.

(b) Fit the regression line of X on Y and hence predict X, if Y = 2.5.

Solution : Regression line of Y on X is given by

$Y = a + b X$

To find the values of a and b two normal equations are required

$\sum Y = Na + b\sum X$

$\sum XY = a\sum X + b\sum X^2$

Substituting the values, we get

$16 = 8a + 23b$ ....(i)

$36 = 23a + 99b$ ....(ii)

Table 1

Computation of Regression Equations

X	Y	X^2	Y^2	XY
1	6	1	36	6
5	1	25	1	5
3	0	9	0	0
2	0	4	0	0
1	1	1	1	1
1	2	1	4	2
7	1	49	1	7
3	5	9	25	15
$\sum X = 23$	$\sum Y = 16$	$\sum X^2=99$	$\sum Y^2=68$	$\sum XY=36$

Multiply (i) by 23 and (ii) by 8, we get

$368 = 184\ a + 529\ b$ ....(iii)

$288 = 184\ a + 792\ b$ ....(iv)

Subtracting (iii) from (iv) we get

$263\ b = -\ 80$

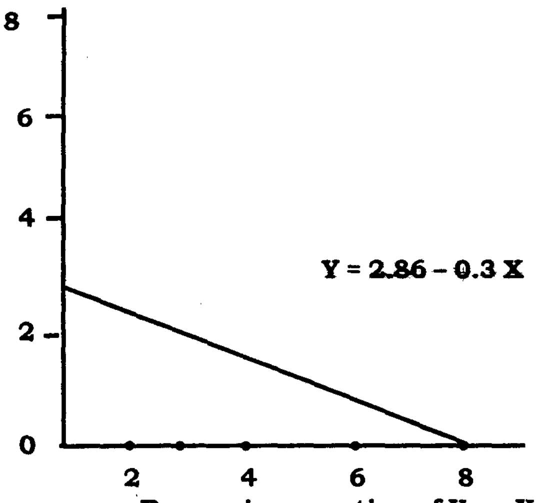
$b = -0.304 = -0.30$ approx.

Substituting $b = -0.30$ in equation (i), we get

$16 = 8a + (23)\ (-0.3) = 81\ -6.9$

$-8a = -16\ -\ 6.9$

$a = 2.86$



Regression equation of Y on X

Fig.9

Thus the regression equation of Y on X is

$Y = 2.86 - 0.30X$

Now the regression equation of X on Y is

$X = a + b\,Y$

The two normal equations are

$\sum X = Na + b\sum Y$ ....(v)

$\sum XY = a\sum Y + b\sum Y^2$ ....(vi)

Substituting the values in equation (v) and (vi), we get

$23 = 8\,a + 16\,b$ ....(v)

$36 = 16\,a + 68\,b$ ....(vi)

Multiplying (v) by 2

$46 = 16\,a + 32\,b$ ....(vii)

Now deducting (vii) from (vi)

$36\,b = -10$

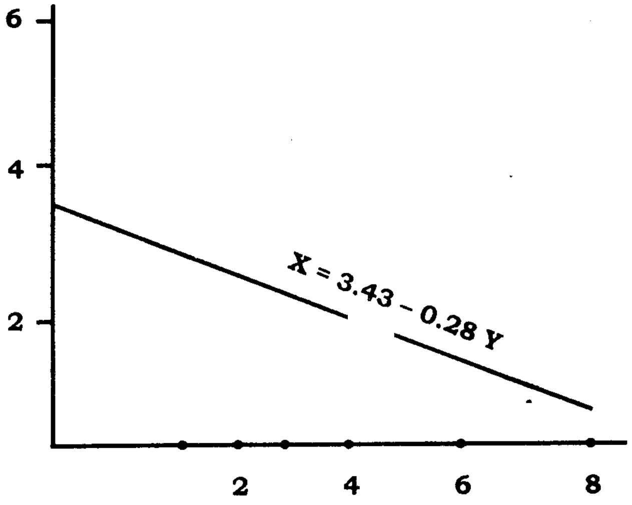
$b = -0.28$

Substituting b = -0.28 in equation in (v), we get

$23 = 8\,a + 16\,(-0.28)$

$23 = 8\,a - 4.48$

$a = 3.43$



Regression of X on Y

Fig.10

The regression of X on Y is

X = 3.43 - 0.28 Y σy

Table II **Table III**

Y = 2.86 - 0.30 X σx **X = 3.43 - 0.28 Y**

X	Y	Y_c	$Y-Y_c$	Y	X	X_c	$X-X_c$
	(Actual)		(Deviation		(Actual)		
1	6	2.56	+3.44	6	1	1.75	-0.75
5	1	1.36	-0.36	1	5	3.15	+1.85
3	0	1.96	-1.96	0	3	3.43	-0.43
2	0	2.26	-2.26	0	2	3.43	-1.43
1	1	2.56	-1.56	1	1	3.15	-2.15
1	2	2.56	-0.58	2	1	2.87	-1.87
7	1	0.76	+0.24	1	7	3.15	.385
3	5	1.96	+3.04	5	3	2.03	+0.97

The points of two regression lines are arrived in Tables II and III.

(ii) Method of Deviation from the actual Means

Under this method two regression equations are modified as under :

(i) Regression equation of X on Y. This is given by

$$X = \overline{X} + b_{xy}(Y - \overline{Y}) \quad \text{or} \quad X - \overline{X} = b_{xy}\left(Y - \overline{Y}\right)$$

(ii) Regression equation of Y on X. This is given by

$$Y = \overline{Y} + b_{yx}\left(X - \overline{X}\right) \quad \text{or} \quad Y - \overline{Y} = b_{yx}(X - \overline{X})$$

$\overline{X}$ = arithmetic average of the X variable.

$\overline{Y}$ = arithmetic average of the Y variable.

b_{xy} = regression coefficient of X on Y = $r \cdot \frac{\sigma x}{\sigma y}$

$$b_{xy} = r \cdot \frac{\sigma x}{\sigma y}$$

Where r = Correlation Coefficient,

σx = Standard deviation of X variable.

σy = Standard deviation of Y variable.

and b_{yx} = regression coefficient of Y on X = $r \frac{\sigma x}{\sigma y}$

$$b_{YX} = r \frac{\sigma x}{\sigma y}$$

To simplify the process of finding the above two regression coefficients the following formulae may be substituted in place of the given ones :

(i) b_{xy} or $b_1 = \frac{\sum xy}{\sum y^2}$;and

(ii) b_{yx} or $b_2 = \frac{\sum xy}{\sum x^2}$

Properties of regression Coefficients

The regression coefficients have a number of valuable properties which may be cited as under :

1. The geometric mean of the two regression coefficients gives the coefficient of correlation i.e.,

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

Proof : $b_{xy} = r\frac{6_x}{6_y}, b_{yx} = r\frac{6_y}{6_x}$

$$b_{xy}.b_{yx} = \frac{\sigma x}{\sigma y}\frac{5_x}{5_y}. - \frac{6_y}{6_x} = r^2$$

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

2. Both the regression Coefficients must have same algebraic signs i.e. both of them will have either + or - sings.
3. If the regression Coefficients are positive the correlation coefficient will be positive; and if the regression coefficients are negative them the correlation coefficient will be negative.
4. If one of the regression coefficients is greater than unity (1), the other must be less than unity. This is because r can never be greater than one.
5. From the regression coefficients we can find out the value of any factor forming part of it, if the value of the other 3 factors are given.

6. Regression coefficients are independent of change of origin but not of scale.

Example 3 :

Using the method of deviations from the actual Means, from the data given below find

(i) the two regression equations

(ii) the correlation coefficient and

(iii) the most probable value of Y when X = 30.

X :	25	28	35	32	31	36	29	38	34	32
Y :	43	46	49	41	36	32	31	30	33	39

Solution :

X	**Y**	**x** **(X-32)**	**y** **(Y-38)**	**x^2**	**y^2**	**xy**
25	43	-7	5	49	25	-35
28	46	-4	8	16	64	-32
35	49	3	11	9	121	33
32	41	0	3	0	9	0
31	36	-1	-2	1	4	2
36	32	4	-6	16	36	-24
29	31	-3	-7	9	49	21
38	30	6	-8	36	64	-48
34	33	2	-5	4	25	-10
32	39	0	1	0	1	0
$\sum X=320$	$\sum Y=380$	$\sum x=0$	$\sum y=0$	$\sum x^2=140$	$\sum y^2=398$	$\sum xy=-93$

(a) (i) Regression equation of X on Y

$$X = \overline{X} + r\frac{6_x}{6_y}\left(Y = \overline{Y}\right)$$

where $\overline{X} = \frac{\sum X}{N} = \frac{320}{10} = 32$

$$\overline{Y} = \frac{\sum Y}{N} = \frac{380}{10} = 38$$

$$6_x = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\frac{140}{10}} = 3.74 \text{ approx.}$$

$$6_y = \sqrt{\frac{\sum y^2}{N}} = \sqrt{\frac{398}{10}} = 6.31 \text{ approx.}$$

$$r = \frac{\sum xy}{N6_x 6_y} = \frac{-93}{10 \times 3.74 \times 6.31}$$

$$= \frac{-93}{10 \times 23.5994} = \frac{-93}{235.99} = -0.394$$

Putting the respective values in the above equation, we get

$$X = 32 + -0.394 \times \frac{3.74}{6.31}(Y\text{-}38)$$

$= 32 - 0.2337 (Y\text{-}38)$

$= 32 + 8.8806 - 0.2337\ Y$

$X = 40.8806 - 0.2337\ Y$

(ii) Regression equation of Y on X

$$Y = \overline{Y} + r\frac{6_y}{6_x}(X - \overline{X})$$

$$Y = 38 - 0.394 \times \frac{6.31}{3.74}(X - 32)$$

$= 38 - 0.6643 (X\text{-}32)$

$Y = 38 + 21.2576 - 0.6643\ X$

$= 59.2576 = 0.6643\ X$

$\therefore\ Y = 59.2576 - 0.6643\ X$

(b) Coefficient of Correlation

The coefficient of correlation between the two variables, X and Y is given by

$$r_{xy} = \frac{\sum xy}{N6_x 6_y} = \frac{-93}{10 \times 3.74 \times 6.31} = \frac{-93}{235.994}$$

$r_{xy} = -0.394$

(c) Probable value of Y when X = 30

This will be determined by the regression equation of Y on X as follows :

$$Y = 59.2576 - 0.6643\ X$$

Thus, when X = 30, Y = 59.2576 - 0.6643 (30)

$$= 59.2576 - 19.929 = 39.3286$$

Exercise 3

1. Obtain the equations of the two lines of regression (using Normal Equation Method) for the data given below :

X :	1	2	3	4	5	6	7	8	9
Y :	9	8	10	12	11	13	14	16	15

2. The following are the marks obtained by 8 students in Mathematics and Statistics. Find the regression line of marks in statistics on marks in Mathematics.

Marks in Mathematics (x)	:	50	40	60	46	50	48	59	47
Marks in Statistics (y)	:	30	37	42	32	35	45	40	35

(iii) Method of Deviation from Assumed Mean/Short-Cut Method :

Under this method the regression between any two related variables is studied on the basis of the deviations of the items from their respective assumed Means rather than their actual values. In practice we get means in fractions and for simplicity we take deviations from assumed means. When the deviations are taken from the assumed means, the procedure for finding regression equations remains the same. The value of $r\frac{6_x}{6_y}$ will now be obtained as follows :

$$r\frac{6_x}{6_y} = \frac{\sum d_x d_y - \frac{\sum d_x \times \sum d_y}{N}}{\sum d_y^2 - \frac{\left(\sum d_y\right)^2}{N}} = b_{xy}$$

d_x= X – A and d_y = Y – A

Similarly $$r\frac{6_y}{6_x} = \frac{\sum d_x d_y - \frac{\sum d_x \times \sum d_y}{N}}{\sum d_x^2 - \frac{\left(\sum d_x\right)^2}{N}} = b_{yx}$$

d_x= X – A and d_y = Y – A

Example 4 : Given the bivariate data.

X :	2	6	4	3	2	3	8	4
Y :	7	2	1	1	2	3	2	6

Obtain regression equations taking deviations from 5 in case of X and 4 in case of Y.

Solution :

X	$\mathbf{d_x}$**=X-5**	$\mathbf{d_x^2}$	**Y**	$\mathbf{d_y}$**=Y–4**	$\mathbf{d_y^2}$	$\mathbf{d_x d_y}$
2	-3	9	7	3	9	-9
6	+1	1	2	-2	4	-2
4	-1	1	1	-3	9	+3
3	-2	4	1	-3	9	+6
2	-3	9	2	-2	4	+6
3	-2	4	3	-1	1	+2
8	3	9	2	-2	4	-6
4	-1	1	6	+2	4	2
$\sum X=32$	$\sum d_x= -8$	$\sum d_x^2=38$	$\sum Y=24$	$\sum d_y= -8$	$\sum d_y^2=44$	$\sum d_x d_y= -2$

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

$$\text{Since } b_{xy}=\frac{\sum d_x d_y-\frac{\sum d_x \sum d_y}{N}}{\sum d_y^2-\frac{\left(\sum d_y\right)^2}{N}}$$

$$b_{xy}=\frac{-2-\frac{(-8)(-8)}{8}}{44-\frac{(-8)^2}{8}}, \qquad b_{xy}=\frac{-2-8}{44-8}=\frac{-10}{36}=-0.28$$

± ± ± ±

$$\overline{X}=\frac{\sum X}{N}=\frac{32}{8}=4$$

$$\overline{Y} = \frac{\sum Y}{N} = \frac{24}{8} = 3$$

X-4 = -0.28 (Y-3)

(X-4) = -0.28 Y + 0.84

X = 4.84 - 0.28 Y

Regression equations of Y on X is

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

$$b_{yx} = \frac{\sum d_x d_y - \frac{(\sum d_x)(\sum d_y)}{N}}{\sum d_x^2 - \frac{(\sum d_x)^2}{N}}$$

$$b_{yx} = \frac{-2 - \frac{(-8)(-8)}{8}}{38 - \frac{(-8)^2}{8}} = \frac{-2-8}{38-8} = \frac{-10}{30} = -0.33$$

Thus the regression equation of Y on X will be

Y-3 = -0.33 (X-4)

Y = 1.32 - 0.33 X + 3

Y = 4.32 - 0.33 X

Example 5 : Find the coefficient of correlation from the following two regression equations:

3Y - 2X - 10 = 0,

2Y - 50 - X = 0

Also find the estimated value of Y when X = 0.

Solution : The two regression equations are

3Y - 2X = 10

- 2X = 10 - 3Y

$X = -5 + \frac{3}{2}Y$ ....(i) Reg. Equation of X on Y

2Y - 50 - X = 0

2Y = 50 + X

$Y = 25 + \frac{1}{2}X$ ....(ii) Reg. Equation of Y on X.

$$b_{xy} = \frac{3}{2}, b_{yx} = \frac{1}{2}$$

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{\frac{3}{2} \times \frac{1}{2}} = \sqrt{0.75} = 0.87$$

1.7.7 Difference between Correlation and Regression

1. The correlation analysis tests the closeness of the variables, where as regression analysis measures the extent of change in dependent variable due to change in the independent variable.
2. In regression analysis, the casual relationship in variables moving in the same or opposite direction is studied while in correlation analysis, the study is made by taking into consideration the cause and effect relationship between the variables.
3. Correlation has very limited scope of application. But the scope of applicability of regression analysis is very wide. It can be covered under linear as well as non-linear relationship between variables.

1.7.8 Uses of Regression Analysis

The technique of regression is considered to be the most useful statistical tool applied in various fields of sociological and scientific disciplines. It is helpful in making quantitative predictions in business in the behaviour of the related variables. Following are some of the main uses of regression analysis :

1. The regression analysis technique is very useful in predicting the probable value of an unknown variable in response to some known related variable. The estimation or prediction of future production, consumption, prices, investments, sales, profits, income, etc., are of paramount importance to a businessman or economist. Population estimates and population projections are indispensable for efficient planning of an economy. Regression analysis is one of the very scientific techniques for making such predictions.
2. The regression technique is useful in establishing the nature of the relationship between two variables.
3. The predictions made on the basis of estimated inter-relationship through the techniques of regression analysis provide sound basis for policy formulation in socio-economic fields.

1.7.9 Summary

In this lesson fundamentals of linear regression have been highlighted. The statistical technique of estimating or predicting the unknown value of a dependent variable from the known value of an independent variable is called regression analysis. The estimation of the parameters of this regression analysis is accomplished by the least squares criterion which tries to minimise the sum of squares of the errors for all the data points. Two methods of studying regression analysis have also been illustrated through various examples. In the end the differences between correlation and regression and uses of regression analysis have been given.

1.7.10 Key Words

Dependable Variable : The variable of interest or focus which is influenced by one or more independent variable (s).

Estimate : A value obtained from data for a certain parameter of the assumed model or a forecast value obtained from the model.

Independent Variable : A variable that can be set either to a desirable value or takes values that can be observed but not controlled.

Linear Regression : When dependent variable moves in a fixed proportion of the unit movement of independent variable, it is called a linear regression.

Regression Line : A graphic technique to show the functional relationship between dependent and independent variables.

Regression equations : Regression equations are algebraic expression of regression lines.

Least Square Criterion : According to the least squares criteria the line should be drawn through the plotted points in such a way that the sum of squares of the deviations of the actual Y values from the computed 'Y' values is the least.

1.7.11. Further Readings

1. S.C. Gupta	:	Fundamentals of Mathematical Statistics
2. S.P. Gupta	:	Statistics Methods
3. Digamber Patri	:	Statistics Methods
4. W.J. Stevenson	:	Business Statistics - Concepts and Applications
5. Murray R. Spiegal	:	Statistics.

1.7.12 Lists of Questions

1.7.12.1 Short Questions

1. State clearly the concept of regression.
2. Distinguish between correlation and regression analysis.
3. State the important properties of regression coefficients.
4. Define regression analysis. Explain its utility and limitations.
5. What are regression coefficients ? Show that $r^2 = b_{xy} \times b_{yx}$
6. When one regression coefficient in negative the other would be
   (a) Negative (b) Positive
   (c) Zero(d) None of these
7. The regression lines cut each other at the point of
   (a) Average of X and Y (b) Average of X only
   (c) Average of Y only (d) None of these
8. Find the most likely production corresponding to a rainfall 40" from the following data :

	Rain	Production
Average	30"	500 kg
Standard Deviation	5"	100 kg

   Coefficient of Correlation = 0.8

9. Form the following two regression equations, state which one is of X on Y and which one is of Y on X ;
   $2X + 4Y = 10$
   $4X + 6Y = 8$
10. Given $b_{xy} = 0.85$, $b_{yx} = 0.89$ and $\sigma_x = 6$, find the value of r and σ_x.

1.7.12.2 Long Questions

1. Obtain the regression equations for the following :

X :	15	27	27	30	34	38	46
Y :	120	140	150	170	180	200	250

2. Fit a regression equation of Y on X for the following data and estimate the value of Y when X is 40.

X :	18	20	22	25	30
Y :	24	28	30	35	44

3. Calculate the regression coefficients for the data given below :

X :	8	6	4	7	5
Y :	9	8	5	6	2

4. Two random variables have the least square regression lines with equations $3x + 2y = 26 = 0$ and $6x + y - 31 = 0$.
   Find the mean value and the correlation coefficient between X and Y.
5. Find out two regression equations of X and Y variables from the following information
   $r = .6$, $\sigma_x = 1.5$, $\sigma_y = 2$, $\overline{X} = 10$, $\overline{Y} = 20$
6. Given the data below, find the lines of regression using least square technique.

X :	20	11	15	10	17	17
Y :	5	15	14	17	8	9