ਜਗਤ ਗੁਰੂ ਨਾਨਕ ਦੇਵ
ਪੰਜਾਬ ਸਟੇਟ ਓਪਨ ਯੂਨੀਵਰਸਿਟੀ
ਪਟਿਆਲਾ

The Motto of Our University
(SEWA)
SKILL ENHANCEMENT
EMPLOYABILITY
WISDOM
ACCESSIBILITY

JAGAT GURU NANAK DEV

PUNJAB STATE OPEN UNIVERSITY, PATIALA

(Established by Act No. 19 of 2019 of the Legislature of State of Punjab)

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SEMESTER I

SARM 2: DESCRIPTIVE STATISTICS

Head Quarter: C/28, The Lower Mall, Patiala-147001
Website: www.psou.ac.in

The Study Material has been prepared exclusively under the guidance of Jagat Guru Nanak Dev Punjab State Open University, Patiala, as per the syllabi prepared by Committee of experts and approved by the Academic Council.

COURSE COORDINATOR AND EDITOR:

DR. Pinky Sra

Assistant Professor

Jagat Guru Nanak Dev Punjab State Open University, Patiala.

JAGAT GURU NANAK DEV PUNJAB STATE OPEN UNIVERSITY, PATIALA
(Established by Act No. 19 of 2019 of the Legislature of State of Punjab)

PREFACE

Jagat Guru Nanak Dev Punjab State Open University, Patiala was established in December 2019 by Act 19 of the Legislature of State of Punjab. It is the first and only Open University of the State, entrusted with the responsibility of making higher education accessible to all, especially to those sections of society who do not have the means, time or opportunity to pursue regular education.

In keeping with the nature of an Open University, this University provides a flexible education system to suit every need. The time given to complete a programme is double the duration of a regular mode programme. Well-designed study material has been prepared in consultation with experts in their respective fields.

The University offers programmes which have been designed to provide relevant, skill-based and employability-enhancing education. The study material provided in this booklet is self- instructional, with self-assessment exercises, and recommendations for further readings. The syllabus has been divided in sections, and provided as units for simplification.

The University has a network of 110 Learner Support Centres/Study Centres, to enable students to make use of reading facilities, and for curriculum-based counseling and practicals. We, at the University, welcome you to be a part of this institution of knowledge.

Prof. G.S.Batra
Dean Academic Affairs

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SARM 2: DESCRIPTIVE STATISTICS

Max. Marks: 100
External: 70
Internal: 30
Pass: 40%
Credits: 6

OBJECTIVES

Descriptive statistics summarize and organize characteristics of a data set, a collection of observations from a sample or entire population.

INSTRUCTIONS FOR THE PAPER SETTER/EXAMINER:

1. The syllabus prescribed should be strictly adhered to.
2. The question paper will consist of three sections: A, B, and C. Sections A and B willhave four questions from the respective sections of the syllabus and will carry 10 marks each. The candidates will attempt two questions from each section.
3. The Question paper will contain 60 percent theory and 40 percent numerical proportion.
4. Section C will have fifteen short answer questions covering the entire syllabus. Eachquestion will carry 3 marks. Candidates will attempt any ten questions from this section.
5. The examiner shall give a clear instruction to the candidates to attempt questions only at one place and only once. Second or subsequent attempts, unless the earlier oneshave been crossed out, shall not be evaluated.
6. The duration of each paper will be three hours.

INSTRUCTIONS FOR THE CANDIDATES:

Candidates are required to attempt any two questions each from the sections Aand B of the question paper and any ten short questions from Section C. They have to attempt questions only at one place and only once. Second or subsequent attempts, unless the earlier ones have been crossed out, shall not beevaluated. Simple calculator can be use in the examination.

SECTION A

Unit 1: Dispersion - Objectives and significance of Good Measures, Measures of Dispersion -

Range, Quartile Deviation, Mean Deviation and Standard Deviation (ungrouped data).

Unit 2: Co-efficient of variation (CV), Lorenz Curve, Meaning and Measures of skewness kurtosis, Moments

Unit 3: Correlation: Meaning, Properties and Types.

Unit 4: Methods of Correlation: Scatter Diagram, Karl Pearson's Correlation Co-efficient & Spearman's Rank, Correlation Co-efficient.

SECTION B

Unit 5: Regression- Meaning, Properties, Types, Meaning of Line of Correlation, Difference between correlation and regression.

Unit 6: Measurement of Regression equations X on Y and Yon X

Unit 7: Index Numbers: Meaning and Uses and Types of Index Numbers, problems in the construction, Methods of Index Numbers: Laspayer's, Paasche and Fisher.

Unit 8: Tests of consistency of Index Number Formulae, Chain index or Chain Base Index Numbers, Base Shifting, Splicing and Deflation. Limitations of Index Numbers.

Note: Statistical analysis should also be taught with the help of MS Excel, SPSS or any other related software tool.

Suggested Readings

A.M Goon, M.K Gupta and B. Dasgupta, fundamental of statistics Vol-I, World press Calcutta

Gupta SC: Fundamental of statistics, S. Chand & Company. New Delhi

Gupta, SP: Statistical Methods, S. Chand & Company. New Delhi

Monga, GS: Mathematics and Statistics for Economics, Vikas Publishing House, New Delhi.

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SARM 1: INTRODUCTION TO STATISTICS

EDITOR AND COURSE CO-ORDINATOR- DR. PINKY SRA

SECTION A

UNIT NO.	UNIT NAME
Unit 1	Dispersion - Objectives and significance of Good Measures, Measures of Dispersion - Range, Quartile Deviation, Mean Deviation and Standard Deviation (ungrouped data).
Unit 2	Co-efficient of variation (CV), Lorenz Curve, Meaning and Measures of skewness kurtosis, Moments
Unit 3	Correlation: Meaning, Properties and Types.
Unit 4	Methods of Correlation: Scatter Diagram, Karl Pearson's Correlation Co-efficient & Spearman's Rank, Correlation Co-efficient.

SECTION B

UNIT NO.	UNIT NAME
Unit 5	Regression- Meaning, Properties, Types, Meaning of Line of Correlation, Difference between correlation and regression.
Unit 6	Measurement of Regression equations X on Y and Yon X
Unit 7	Index Numbers: Meaning and Uses and Types of Index Numbers, problems in the construction, Methods of Index Numbers: Laspayer’s, Paasche and Fisher.
Unit 8	Tests of consistency of Index Number Formulae, Chain index or Chain Base Index Numbers, Base Shifting, Splicing and Deflation. Limitations of Index Numbers

CERTIFICATE/DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SARM 2: DESCRIPTIVE STATISTICS

SEMESTER I

UNIT 1: DISPERSION - OBJECTIVES AND SIGNIFICANCE OF GOOD MEASURES, MEASURES OF DISPERSION

STRUCTURE

1.0 Learning Objectives

1.1 Introduction and Meaning of Dispersion

1.2 Benefit / Uses of Dispersion

1.3 Features of Good Measure of Dispersion

1.4 Absolute and Relative Measure of Dispersion

1.5 Measure of Dispersion - Range

- **1.5.1 Range in individual series**
- **1.5.2 Range in discrete series**
- **1.5.3 Range in continuous series**
- **1.5.4 Merits and Limitations of Range**

1.6 Measure of Dispersion – Quartile Deviations

- **1.6.1 Quartile Deviations in Individual Series**
- **1.6.2 Quartile Deviations in Discrete Series**
- **1.6.3 Quartile Deviations in Continuous Series**
- **1.6.4 Merits and Limitations of Quartile Deviations**

1.7 Measure of Dispersion – Mean Deviation

- **1.7.1 Mean Deviation in Individual Series**
- **1.7.2 Mean Deviation in Discrete Series**
- **1.7.3 Mean Deviation in continuous series**
- **1.7.4 Merits and Limitations of Mean Deviation**

1.8 Measure of Dispersion – Standard Deviation

1.8.1 Standard Deviation in Individual Series

1.8.2 Standard Deviation in Discrete Series

1.8.3 Standard Deviation in Continuous Series

1.8.4 Combined Standard Deviation

1.8.5 Properties of Standard Deviation

1.8.6 Merits and Limitations of Standard Deviation

1.9 Let us Sum Up

1.10 Key Terms

1.11 Questions for Practice

1.12 Further Readings

1.0 LEARNING OBJECTIVES

After studying the Unit, students will be able to:

- Explain the meaning of Dispersion
- Compare absolute and relative measures of Dispersion
- Understand features of a good measure of Dispersion
- Calculate the Range and Quartile Deviation
- Measure the Dispersion using Mean and Standard Deviation
- Compare the variation of the two series.

1.1 INTRODUCTION AND MEANING

Statistics is a tool that helps us in the extraction of information from a large pool of data. There are many tools in statistics that help us in the extraction of data. Central tendency of data is one such tool. A good measure of central tendency is one that could represent the whole data. However, many a time we find that the average is not representing it data. The following example will make this clear:

Series X	Series Y	Series Z
100	94	1
100	105	2

100 100 100	101 98 102	3 4 490
$\sum X = 500$	$\sum Y = 500$	$\sum Z = 500$
$\bar{X} = \frac{\sum X}{N} = \frac{500}{5} = 100$	$\bar{Y} = \frac{\sum Y}{N} = \frac{500}{5} = 100$	$\bar{Z} = \frac{\sum Z}{N} = \frac{500}{5} = 100$

We can see that in all the above series the average is 100. However, in the first series average fully represents its data as all the items in the series are 100 and average is also 100. In the second series, the items are very near to its average which is 100, so we can say that average is a good representation of the series. But in case of third series, average does not represent its data as there is a lot of difference between items and the average. In order to understand the nature of data it is very important to see the difference between items and the data. This could be done by using dispersion.

Dispersion is a very important statistical tool that helps us in PROGRESS the nature of data. Dispersion shows the extent to which individual items in the data differ from its average. It is a measure of the difference between data and the individual items. It indicates how that are lacks uniformity. Following are some of the definitions of Dispersion.

According to Simpson and Kafka, "The measures of the scatterness of a mass of figures in a series about an average is called a measure of variation, or dispersion".

According to Spiegel, "The degree to which numerical data lend to spread about an average value is called the variation, or dispersion of the data".

As the dispersion gives the average difference between items and its Central tendency, it is also known as the average of second order.

1.2 BENEFITS / USES OF DISPERSION

Benefits of Dispersion analysis are outlined as under:

1. **To examine reliability of Central tendency:** We have already discussed that a good measure of Central tendency is one which could represent its series. Dispersion gives us the idea that whether average is in a position to represent its series or not. On the basis of this we can calculate reliability of the average.

2. **To compare two series**: In case there are two series and we want to know which series has more variation, we can use dispersion as its tool. In such cases normally we use relative measure of dispersion for comparing two series.
3. **Helpful in quality control**: Dispersion is a tool that is frequently used in quality control by business houses. Every manufacturer wants to maintain same quality and reduce the variation in production. Dispersion can help us in finding the deviations and removing the deviations in quality.
4. **Base of further statistical analysis:** Dispersion is a tool that is used in a number of statistical analyses. For example, we use dispersion while calculating correlation, Regression, Skewness and Testing the Hypothesis, etc.

1.3 FEATURES OF GOOD MEASURE OF DISPERSION

A good measure of dispersion has a number of features which are mentioned below:

1. A good tool of dispersion must be easy to understand and simple to calculate.
2. A good measure of dispersion must be based on all the values in the data.
3. It should not be affected by presence of extreme values in the data.
4. A good measure is one which is rigidly defined.
5. A good measure of dispersion must be capable of further statistical analysis.
6. A good measure must not be affected by the sampling size.

1.4 ABSOLUTE AND RELATIVE MEASURE OF DISPERSION

There are two measures of dispersion: absolute measure and relative measure

1. **Absolute measure:** the absolute measure of dispersion is one that is expressed in the same statistical unit in which the original values of that data are expressed. For example, if original data is represented in kilograms, the dispersion will also be represented in kilograms. Similarly, if data is represented in rupees the dispersion will also be represented in rupees. However, this measure is not useful when we have to compare two or more series that have different units of measurement or belongs to different population.
2. **Relative measure of Dispersion**: The relative measure of dispersion is independent of unit of measurement and is expressed in pure numbers. Normally it is a ratio of the dispersion to

the average of the data. It is very useful when we have to compare two different series that are having different unit of measurement or belongs to a different population.

Absolute Measure of Dispersion

- Range
- Quartile Deviation
- Mean Deviation
- Standard Deviation

Relative Measure of Dispersion

- Coefficient of Range
- Coefficient of Quartile Deviation
- Coefficient of Mean Deviation
- Coefficient of Standard Deviation

1.5 MEASURE OF DISPERSION - RANGE

Range is one of the simplest and oldest measures of Dispersion. We can define Range as the difference between highest value of the data and the lowest value of the data. The more is the difference between highest and the lowest value, more is the value of Range which shows high dispersion. Similarly, less is the difference between highest and lowest value, less is value of Range that shows less dispersion. Following is formula for calculating the value of range:

$$\textbf{Range = Highest Value - Lowest Value}$$
$$\textbf{R = H – L}$$

Coefficient of Range: Coefficient of Range is relative measure of Range and can be calculated using the following formula.

$$\textbf{Coefficient of Range} = \frac{\boldsymbol{Highest\ Value - Lowest\ Value}}{\boldsymbol{Highest\ Value + Lowest\ Value}} = \frac{\boldsymbol{H - L}}{\boldsymbol{H + L}}$$

1.5.1 Range in Individual Series:

Example 1. Following are daily wages of workers, find out value of Range and Coefficient of Range.

Wage (Rs.)	330	300	470	500	410	380	425	360

Solution:

Range = Highest Value − Lowest Value

= 500 – 300 = 200

Coefficient of Range = $\frac{Highest\ Value - Lowest\ Value}{Highest\ Value + Lowest\ Value}$

= $\frac{500 - 300}{500 + 300}$ = .25

1.5.2 Range in Discrete Series:

Example 2. Following are daily wages of workers, find out value of Range and Coefficient of Range.

Wage (Rs.)	300	330	360	380	410	425	470	500
No. of Workers	5	8	12	20	18	15	13	9

Solution: Range = Highest Value − Lowest Value

= 500 – 300 = 200

Coefficient of Range = $\frac{Highest\ Value - Lowest\ Value}{Highest\ Value + Lowest\ Value}$

= $\frac{500 - 300}{500 + 300}$ = .25

1.5.3 Range in Continuous Series:

Example 3. Following are daily wages of workers, find out value of Range and Coefficient of Range.

Wage (Rs.)	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of Workers	5	8	12	20	18	15	13	9

Solution: Range = Highest Value − Lowest Value

= 90 – 10 = 80

Coefficient of Range = $\frac{Highest\ Value - Lowest\ Value}{Highest\ Value + Lowest\ Value}$

= $\frac{90 - 10}{90 + 10}$ = .80

1.5.4 Merits and Limitations of Range

1. Range is one of the easiest and simplest method of dispersion.
2. The range a measure that is rigidly defined.
3. This method gives broad picture of variation in the data.
4. Range is very useful in various fields of business such as quality control and checking the difference between share prices in the stock exchange.
5. Range is also useful in forecasting.

Limitations of range

1. Range is not exact measure of depreciation at only gives vague picture.
2. It is not based on all the values of data.
3. It is affected by the extreme values of the data.
4. It is also affected by fluctuations in the sample.
5. In case of open-ended series range cannot be calculated.

TEST YOUR PROGRESS (A)

1. Compute for the following data Range and Coefficient of Range

28	110	27	77	19	94	63	25	111

2. Given below is heights of students of two classes. Compare Range of the heights:

Class I	167	162	155	180	182	175	185	158
Class II	169	172	168	165	177	180	195	167

3. Find Range and coefficient of Range

X	5	10	15	20	25	30	35	40
f	6	4	12	7	24	21	53	47

4. Calculate coefficient of Range:

X;	10-20	20-30	30-40	40-50	50-60
F:	8	10	12	8	4

Answers

1. 92, 0.7	3. 35, 0.778
2. .088, .083	4. .714

1.6 MEASURE OF DISPERSION – QUARTILE DEVIATION

Range is simple to calculate but suffers from limitation that it takes into account only extreme values of the data and gives a vague picture of variation. Moreover, it cannot be calculated in case of open-end series. In such case we can use another method of Deviation that is Quartile Deviation or Quartile Range. Quartile Range is the difference between Third Quartile and First Quartile of the data. Following is formula for calculating Quartile Range.

$$\textbf{Quartile Range} = Q_3 - Q_1$$

Quartile Deviation: Quartile deviation is the Arithmetic mean of the difference between Third Quartile and the First Quartile of the data.

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

Coefficient of Quartile Deviation: Coefficient of Quartile Deviation is a relative measure of Quartile Deviation and can be calculated using the following formula.

$$\textbf{Coefficient of Range} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

1.6.1 Quartile Deviation in Individual Series:

Example 4. Following are daily wages of workers, find out value of Quartile Range, Quartile Deviation and Coefficient of Quartile Deviation.

Wage (Rs.)	300	330	380	410	425	470	500

Solution:

4th = Value of $\frac{N+1}{4}$ th item = Value o $\frac{7+1}{4}$ th item

= Value of 2nd item

= 330

4th$\frac{3(N+1)}{4}$ th item = Value of $\frac{3(7+1)}{4}$ th item

= Value of 6th item

= 470

Quartile Range = $Q_3 - Q_1$

= 470 – 330 = 140

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

$$= \frac{470 - 330}{2} \qquad = 70$$

$$\text{Coefficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{470 - 330}{470 + 330} \qquad = .175$$

1.6.2 Quartile Deviation in Discrete Series:

Example 5. Following are daily wages of workers, find out value of Quartile Range, Quartile Deviation and Coefficient of Quartile Deviation.

Wage (Rs.)	300	330	380	410	425	470	500
No. of Workers	5	8	12	20	18	15	13

Solution:

Calculation of Quartile

Wage (Rs.) (X)	**No. of Workers (f)**	**Cumulative Frequency (cf)**
300	5	5
330	8	13
380	12	25
410	20	45
425	18	63
470	15	78
500	13	91

Q_1 = Value of $\frac{N+1}{4}$ th item = Value of $\frac{91+1}{4}$ th item

= Value of 23rd item

= 380

Q3 = Value of $\frac{3(N+1)}{4}$ th item = Value of $\frac{3(91+1)}{4}$ th item

= Value of 69th item

= 470

Quartile Range = $Q_3 - Q_1$

= 470 – 380 = 90

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

$$= \frac{470 - 380}{2} \qquad = 45$$

Coefficient of Quartile Deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

$= \frac{470 - 380}{470 + 380}$ = .106

1.6.3 Quartile Deviation in Continuous Series:

Example 6. Following are daily wages of workers, find out value of Quartile Range, Quartile Deviation and Coefficient of Quartile Deviation.

Wage (Rs.)	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of Workers	5	8	12	20	18	15	13	9

Solution:

Calculation of Quartile

Wage (Rs.) (X)	**No. of Workers (f)**	**Cumulative Frequency (cf)**
10-20	5	5
20-30	8	13
30-40	12	25
40-50	20	45
50-60	18	63
60-70	15	78
70-80	13	91
80-90	9	100

Calculation of Q_1

Q_1 Class = Value of $\frac{N}{4}$ th item = Value of $\frac{100}{4}$ th item

Q_1 Class = Value of 25th item

Q_1 Class = 30-40

$Q_1 = L_1 + \frac{\frac{n}{4} - cf}{f} \times c$

Where $L_1 = 30$, n = 100; cf = 13; f = 12; c = 10

$Q_1 = 30 + \frac{\frac{100}{4} - 13}{12} \times 10 = 40$

Calculation of Q_3

Q_3 Cass = Value of $\frac{3N}{4}$ th item = Value of $\frac{300}{4}$ th item

Q_3 Class = Value of 75th item

Q_3 Class = 60-70

$$Q_3 = L_1 + \frac{\frac{3n}{4} - cf}{f} \times c$$

Where $L_1 = 60$, $n = 100$; $cf = 63$; $f = 15$; $c = 10$

$$Q_1 = 60 + \frac{\frac{3(100)}{4} - 63}{15} \times 10 \quad = 68$$

Calculation of Quartile Range, Quartile Deviation and Coefficient of Quartile Deviation

Quartile Range = $Q_3 - Q_1$

$= 68 - 40 \qquad = 28$

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

$= \frac{68 - 40}{2} \qquad = 14$

Coefficient of Quartile Deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

$= \frac{68 - 40}{68 + 40} \qquad = .259$

1.6.4 Merits of Quartile Deviation

1. Quartile deviation is a tool that is easy to calculate and understand.
2. Quartile deviation is the best tool of dispersion in case of open-ended series.
3. This method of dispersion is better than range.
4. Unlike the range, it is not affected by the extreme values.
5. This method of dispersion is rigidly defined.
6. This method is very useful especially when we want to know the variability of middle half of the data. Under this method first 25% of items that are less than Q_1 and upper 25% of items that are more than Q_3 are excluded and only middle 50% of items are taken.

Limitations of Quartile Deviation

1. Quartile deviation considers only middle 50% items of the data and ignores rest of the items.
2. It is not possible to make any further algebraic treatment of the quartile deviation.
3. It is not based on all the items.
4. Quartile deviation is highly affected by fluctuation in the sample.
5. It is comparatively difficult to calculate quartile deviation than range.

TEST YOUR PROGRESS (B)

1. Find Quartile deviation and coefficient of Quartile Deviation:

X: 59, 60, 65, 64, 63, 61, 62, 56, 58, 66

2. Find Quartile deviation and coefficient of Quartile Deviation:

X	58	59	60	61	62	63	64	65	66
F	15	20	32	35	33	22	20	10	8

3. Find Quartile deviation and coefficient of Quartile Deviation

X	0-100	100-200	200-300	300-400	400-500	500-600	600-700
F:	8	16	22	30	24	12	6

4. Calculate the quartile Range, Q.D and coefficient of Q.D

X	0-10	10-20	20-30	30-40	0-500	50-60	60-70	70-80	80-90
F:	11	18	25	28	30	33	22	15	22

Answers

1. 2.75, 0.0447	3. 113.54, 0.335
2. 1.5, .024	4. 34.84, 17.42, .3769

1.7 MEASURE OF DISPERSION – MEAN DEVIATION

Both Range and Quartile Deviation are positional method of Dispersion and takes into consideration only two values. Range considers only highest and lowest values while calculating Dispersion, while Quartile Deviation considers on First and Third Quartile for calculating Dispersion. Both these methods are not based on all the values of the data and are considerably affected by the sample unit. A good measure of Dispersion considers all the values of data.

Mean Deviation is a tool for measuring the Dispersion that is based on all the values of Data. Contrary to its name, it is not necessary to calculate Mean Deviation from Mean, it can also be calculated using the Median of the data or Mode of the data. In the Mean deviation we calculated deviations of the items of data from its Average (Mean, Median or Mode) by taking positive signs only. When we divide the sum of deviation by the number of items, we get the value of Mean Deviation.

In simple words: “Mean Deviation is the value obtained by taking arithmetic mean of the deviations obtained by deducting average of data whether Mean, Median or Mode from values of data, ignoring the signs of the deviations.”

1.7.1 Mean Deviation in case of Individual Series:

As we have already discussed Mean Deviation can be calculated from Mean, Median or Mode. Following is the formula for calculating Mean Deviation in case of Individual series.

$$\textbf{Mean Deviation from Mean } (MD_{\bar{X}}) = \frac{\sum|X - \bar{X}|}{n} = \frac{\sum|D_{\bar{X}}|}{n}$$

$$\textbf{Mean Deviation from Median } (MD_{M}) = \frac{\sum|X - M|}{n} = \frac{\sum|D_{M}|}{n}$$

$$\textbf{Mean Deviation from Mode } (MD_{Z}) = \frac{\sum|X - Z|}{n} = \frac{\sum|D_{Z}|}{n}$$

In case we want to calculate Coefficient of Mean Deviation, it can be done using following formulas.

$$\textbf{Coefficient of Mean Deviation from Mean } (MD_{\bar{X}}) = \frac{MD_{\bar{X}}}{\bar{X}}$$

$$\textbf{Coefficient of Mean Deviation from Median } (MD_{M}) = \frac{MD_{M}|}{M}$$

$$\textbf{Coefficient of Mean Deviation from Mode } (MD_{Z}) = \frac{MD_{Z}}{Z}$$

Example 7. Following are the marks obtained by Students of a class in a test. Calculated Mean Deviation from (i) Mean (ii) Median (iii) Mode. Also calculate Coefficient of Mean Deviation.

Wage (Rs.)	5	7	8	8	9	11	13	14	15

Solution: Let us calculate Mean Median and Mode

Mean ($\bar{X}$) = $\frac{5+7+8+8+9+11+13+14+15}{9} = \frac{90}{9} = 10$

M2thn (M) = Value of $\frac{N+1}{2}$ th item = Value of $\frac{9+1}{2}$ th item

= Value of 5th item = 9

Mode = Item having maximum frequency i.e., 8.

Calculation of Deviations

Marks	$D_{\bar{X}} = \lvert X - \bar{X} \rvert$	$D_{M} = \lvert X - M \rvert$	$D_{Z} = \lvert X - Z \rvert$

X	(Where $\bar{X}$ = 10)	(Where M = 9)	(Where Z = 8)
5	5	4	3
7	3	2	1
8	2	1	0
8	2	1	0
9	1	0	1
11	1	2	3
13	3	4	5
14	4	5	6
15	5	6	7
	$\sum D_{\bar{X}} = 26$	$\sum D_M = 25$	$\sum D_Z = 26$

$D_{\bar{X}}) = \frac{\sum| X - \bar{X} |}{n} = \frac{\sum| D_{\bar{X}} |}{n} = \frac{26}{9} = 2.88$

Coefficient of Mean Deviation from Mean ($MD_{\bar{X}}$) = $\frac{MD_{\bar{X}}}{\bar{X}} = \frac{2.88}{10} = .288$

2. Mean Deviation from Median (MD_M) = $\frac{\sum| X - M |}{n} = \frac{\sum| D_M |}{n} = \frac{25}{9} = 2.78$

Coefficient of Mean Deviation from Median (MD_M) = $\frac{MD_M|}{M} = \frac{2.78}{9} = .309$

3. Mean Deviation from Mode (MD_Z) = $\frac{\sum| X - Z |}{n} = \frac{\sum| D_Z |}{n} = \frac{26}{9} = 2.88$

Coefficient of Mean Deviation from Mode (MD_Z) = $\frac{MD_Z}{Z} = \frac{2.88}{8} = .36$

1.7.2 Mean Deviation in case of Discrete Series:

Following is the formula for calculating Mean Deviation in case of Discrete series.

$$\textbf{Mean Deviation from Mean } (MD_{\bar{X}}) = \frac{\sum f| X - \bar{X} |}{n} = \frac{\sum f| D_{\bar{X}} |}{n}$$

$$\textbf{Mean Deviation from Median } (MD_M) = \frac{\sum f| X - M |}{n} = \frac{\sum f| D_M |}{n}$$

$$\textbf{Mean Deviation from Mode } (MD_Z) = \frac{\sum f| X - Z |}{n} = \frac{\sum f| D_Z |}{n}$$

Example 8. Following are the wages of workers that are employed in a factory. Calculate Mean Deviation from (i) Mean (ii) Median (iii) Mode. Also calculate Coefficient of Mean Deviation.

Wage (Rs.)	300	330	380	410	425	470	500
No. of Workers	6	8	15	25	18	15	13

Solution: Let us calculate Mean Median and Mode

X	**f**	**fX**	**cf**
300	5	1500	5
330	8	2640	13
380	15	5700	28
410	26	10660	54
425	18	7650	72
470	15	7050	87
500	13	6500	100
		$\sum$X = 41700	

Mean ($\overline{X}$) = $\frac{\sum X}{n} = \frac{41700}{100} = 417$

Median (M) = Value of $\frac{N+1}{2}$ th item = Value of $\frac{100+1}{2}$ th item

= Value of 50.5 item

= 410

Mode = Item having maximum frequency i.e., 410.

Calculation of Deviations

X	f	$D_{\overline{X}} = \lvert X - \overline{X} \rvert$ ($\overline{X} = 417$)	$fD_{\overline{X}}$	$D_M = \lvert X - M \rvert$ (M = 410)	fD_M	$D_Z = \lvert X - Z \rvert$ (Z = 410)	fD_Z
300	5	117	585	110	550	110	550
330	8	87	696	80	640	80	640
380	15	37	555	30	450	30	450
410	26	7	182	0	0	0	0
425	18	8	144	15	270	15	270
470	15	53	795	60	900	60	900
500	13	83	1079	90	1170	90	1170
			$\sum fD_{\overline{X}}$ = 4036		$\sum fD_M$ = 3980	$\sum D_Z$ = 26	$\sum fD_Z$ = 3980

1. Mean Deviation from Mean ($MD_{\bar{X}}$) = $\frac{\sum f|\, X-\bar{X}\,|}{n} = \frac{\sum f|\, D_{\bar{X}}\,|}{n} = \frac{4036}{100} = 40.36$

Coefficient of Mean Deviation from Mean ($MD_{\bar{X}}$) = $\frac{MD_{\bar{X}}}{\bar{X}} = \frac{40.36}{417} = .097$

2. Mean Deviation from Median (MD_M) = $\frac{\sum f|\, X-M\,|}{n} = \frac{\sum f|\, D_M\,|}{n} = \frac{3980}{100} = 39.80$

Coefficient of Mean Deviation from Median (MD_M) = $\frac{MD_M|}{M} = \frac{39.80}{410} = .097$

3. Mean Deviation from Mode (MD_Z) = $\frac{\sum f|\, X-Z\,|}{n} = \frac{\sum f|\, D_Z\,|}{n} = \frac{3980}{100} = 39.80$

Coefficient of Mean Deviation from Mode (MD_Z) = $\frac{MD_Z}{Z} = \frac{39.80}{410} = .097$

1.7.3 Mean Deviation in case of Continuous Series:

In case of calculation of Mean Deviation in continuous series, the formula will remain same as we have done in Discrete Series but only difference is that instead of taking deviation from Data, we take deviations from mid value of the data. Further in case of continuous series also the Mean Deviation can be calculated from Mean, Median or Mode. However, in most of the cases it is calculated from Median. Following formulas are used for continuous series:

$$\textbf{Mean Deviation from Mean } (MD_{\bar{X}}) = \frac{\sum f|\, X-\bar{X}\,|}{n} = \frac{\sum f|\, D_{\bar{X}}\,|}{n}$$

$$\textbf{Mean Deviation from Median } (MD_M) = \frac{\sum f|\, X-M\,|}{n} = \frac{\sum f|\, D_M\,|}{n}$$

$$\textbf{Mean Deviation from Mode } (MD_Z) = \frac{\sum f|\, X-Z\,|}{n} = \frac{\sum f|\, D_Z\,|}{n}$$

Example 9. Following are daily wages of workers, find out value of Mean Deviation and Coefficient of Mean Deviation.

Wage (Rs.)	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of Workers	5	8	12	20	18	15	13	9

Solution:

Wage (Rs.) (X)	**No. of Workers (f)**	**Cumulative Frequency (Cf)**	**Mid Value (m)**	**| D_M | | m – M |**	**| f D_M |**

10-20	5	5	15	37.78	188.9
20-30	8	13	25	27.78	222.24
30-40	12	25	35	17.78	213.36
40-50	20	45	45	7.78	155.6
50-60	18	63	55	2.22	39.96
60-70	15	78	65	12.22	183.3
70-80	13	91	75	22.22	288.86
80-90	9	100	85	32.22	289.98
	N = 100				**$\sum \lvert f D_M \rvert$ = 1582.2**

Calculation of Median

Median Class = Value of $\frac{N}{2}$ th item = Value of $\frac{100}{2}$ th item

Median Class = Value of 50th item

Median Class = 50-60

$$M = L_1 + \frac{\frac{n}{2} - cf}{f} \times c$$

Where L_1 = 50, n = 100; cf = 45; f = 18; c = 10

$$M = 50 + \frac{\frac{100}{2} - 45}{18} \times 10 = 52.78$$

Calculation of Mean Deviation from Median

Mean Deviation from Median (MD_M) = $\frac{\sum f \lvert X - M \rvert}{n} = \frac{\sum f \lvert D_M \rvert}{n} = \frac{1582.2}{100} = 15.82$

Coefficient of Mean Deviation from Median (MD_M) = $\frac{MD_M}{M} = \frac{15.82}{52.78} = .30$

1.7.4 Merits and Limitations of Mean Deviation

1. We can calculate mean deviation very easily.
2. Mean deviation is based on all the items of the Data. Change in any value of the data is also going to affect mean deviation.
3. As it is based on all the items of the data, it is not affected by the extreme values of the data.
4. Mean deviation can be calculated from Mean, Median or Mode.
5. Mean deviation is a rigidly defined method of measuring dispersion.

6. Mean deviation can be used for comparison of two different series.

Limitations of Mean Deviation

1. While calculating the mean deviation, we consider only positive sign and ignore the negative sign.
2. In case mean deviation is calculated from mode, it is not a reliable measure of dispersion as mode is not a true representative of the series.
3. It is very difficult to calculate Mean Deviation in case of open-ended series.
4. Mean deviation is not much capable of further statistical calculations.
5. In case we have Mean Deviation of two different series, we cannot calculate combined mean deviation of the data.
6. In case value of Mean, Median or Mode is in fraction, it is difficult to calculate mean deviation.

TEST YOUR PROGRESS (C)

1. Calculate Mean Deviation from i) Mean, ii) Median, iii) Mode

X: 7, 4, 10, 9, 15, 12, 7, 9, 7

2. With Median as base calculate Mean Deviation of two series and compare variability:

Series A:	3484	4572	4124	3682	5624	4388	3680	4308
Series B:	487	508	620	382	408	266	186	218

3. Calculate Co-efficient of mean deviation from Mean, Median and Mode from the following data

X:	4	6	8	10	12	14	16
f:	2	1	3	6	4	3	1

4. Calculate Co-efficient of Mean Deviation from Median.

X;	20-25	25-30	30-40	40-45	45-50	50-55	55-60	60-70	70-80
F:	7	13	16	28	12	9	7	6	2

5. Calculate M.D. from Mean and Median

X	0-10	10-20	20-30	30-40	40-50
f	6	28	51	11	4

6. Calculate Co-efficient of Mean Deviation from Median.

X	16-20	21-25	26-30	31-35	36-40	41-45	46-50	51-55	56-60
f	8	13	15	20	11	7	3	2	1

Answers

1. 2.35, 2.33, 2.56	3. 0.239, 0.24, 0.24	5. M.D. (Mean) 6.572, Coefficient of M.D. (Mean) 0.287, M.D. (Median) 6.4952, Coefficient of M.D. (Median) 0.281
2. 11.6%, 30.73%	4. 0.214	6.0.22

1.8 MEASURE OF DISPERSION – STANDARD DEVIATION

Standard deviation is assumed as best method of calculating deviations. This method was given by great statistician Karl Pearson in the year 1893. In case of Mean deviation, when we take deviations from actual mean, the sum of deviations is always zero. In order to avoid this problem, we have to ignore the sign of the deviations. However, in case of Standard Deviation this problem is solved by taking the square of the deviations, because when we take a square of the negative sign, it is also converted into the positive sign. Then after calculating the Arithmetic mean of the deviations, we again take square root, to find out standard deviation. In other words, we can say that "Standard Deviation is the square root of the Arithmetic mean of the squares of deviation of the item from its Arithmetic mean."

The standard deviation is always calculated from the Arithmetic mean and is an absolute measure of finding the dispersion. We could also find a relative measure of standard deviation which is known as coefficient of standard deviation.

Coefficient of Standard Deviation – Coefficient of Deviation is the relative measure of the standard deviation and can be calculated by dividing the Value of Standard Deviation with the Arithmetic Mean. The value of coefficient always lies between 0 and 1, where 0 indicates no Standard Deviation and 1 indicated 100% standard deviation. Following is the formula for calculating coefficient of Standard Deviation.

$$\textbf{Coefficient of Standard Deviation} = \frac{SD}{\bar{X}}$$

Coefficient of Variation – Coefficient of Variation is also relative measure of the standard deviation, but unlike Coefficient of Standard Deviation it is not represented in fraction rather it is represented in terms of % age. It can be calculated by dividing the Value of Standard Deviation with the Arithmetic Mean and then multiplying resulting figure with 100. The value of coefficient always lies between 0 and 100. Following is the formula for calculating coefficient of Standard Deviation. Low Coefficient of Variation implies less variation, more uniformity and reliability. Contrary to this higher Coefficient of Variation implies more variation, less uniformity and reliability.

$$\textbf{Coefficient of Standard Deviation} = \frac{\textbf{SD}}{\bar{\textbf{X}}} \times 100$$

Variance – Variance is the square of the Standard Deviation. In other words, it is Arithmetic mean of square of Deviations taken from Actual Mean of the data. This term was first time used by R. A. Fischer in 1913. He used Variance in analysis of financial models. Mathematically:

$$\textbf{Variance} = (\textbf{Standard Deviation})^2 \text{ or } \sigma^2$$

1.8.1 Standard Deviation in case of Individual Series

Following are the formula for calculating Standard Deviation in case of the Individual Series:

1. **Actual Mean Method –** In this method we take deviations from actual mean of the data.

$$\textbf{Standard Deviation (SD or } \sigma\textbf{)} = \sqrt{\frac{\sum x^2}{n}}$$

Where x = X - $\bar{X}$, n = Number of Items.

2. **Assumed Mean Method -** In this method we take deviations from assumed mean of the data. Any number can be taken as assumed mean, however for sake of simplicity it is better to take whole number as assumed mean.

$$\textbf{Standard Deviation (SD or } \sigma\textbf{)} = \sqrt{\frac{\sum dx^2}{n} - \left(\frac{\sum dx}{n}\right)^2}$$

Where dx = X – A, n = Number of Items.

3. **Direct Methods -** In this method we don't take deviations and standard deviation is calculated directly from the data.

$$\textbf{Standard Deviation (SD or } \boldsymbol{\sigma}\textbf{)} = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2}$$

Example 10. Following are the marks obtained by Students of a class in a test. Calculate Standard Deviation using (i) Actual Mean (ii) Assumed Mean (iii) Direct Method. Also calculate Coefficient of Standard Deviation.

Marks	5	7	11	16	15	12	18	12

Solution:

1. Standard Deviation using Actual Mean

Marks X	x **= X -** $\bar{X}$ **(Where** $\bar{X}$ **= 12)**	x^2
5	-7	49
7	-5	25
11	-1	01
16	4	16
15	3	09
12	0	00
18	6	36
12	0	00
$\sum$**X = 96**		$\sum x^2$ **= 136**

Mean ($\bar{X}$) = $\frac{\sum X}{n} = \frac{96}{8} = 12$

Standard Deviation (SD or $\boldsymbol{\sigma}$) = $\sqrt{\frac{\sum x^2}{n}} = \sqrt{\frac{136}{8}} = \sqrt{17}$ =4.12

Coefficient of Standard Deviation = $\frac{\text{SD}}{\bar{X}} = \frac{4.12}{12}$ = .34

2. Standard Deviation using Assumed Mean

Marks X	dx **= X - A (Where A = 11)**	dx^2
5	-6	36
7	-4	16
11	0	00
16	5	25
15	4	16
12	1	01

18 12	7 1	49 01
$\sum$X = 96	**$\sum dx$ = 8**	**$\sum dx^2$ = 144**

Mean ($\bar{X}$) =A+ $\frac{\sum dX}{n} = 11 + \frac{8}{8} = 12$

Standard Deviation ($\boldsymbol{\sigma}$) = $\sqrt{\frac{\sum dx^2}{n} - \left(\frac{\sum dx}{n}\right)^2} = \sqrt{\frac{144}{8} - \left(\frac{8}{8}\right)^2} = \sqrt{18-1} = \sqrt{17} = 4.12$

Coefficient of Standard Deviation = $\frac{\text{SD}}{\bar{X}} = \frac{4.12}{12} = .34$

3. Standard Deviation by Direct Method

Marks **X**	**X^2**
5	25
7	49
11	121
16	256
15	225
12	144
18	324
12	144
$\sum$X = 96	**$\sum$X2 = 1288**

Mean ($\bar{X}$) = $\frac{\sum X}{n} = \frac{96}{8} = 12$

Standard Deviation ($\boldsymbol{\sigma}$) = $\sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2} = \sqrt{\frac{1288}{8} - \left(\frac{96}{8}\right)^2} = \sqrt{161-144} = \sqrt{17} = 4.12$

Coefficient of Standard Deviation = $\frac{\text{SD}}{\bar{X}} = \frac{4.12}{12} = .34$

Example 11. Two Players scored following scores in 10 cricket matches. On base of their performance find out which is better scorer and also find out which player is more consistent.

Player X	26	24	28	30	35	40	25	30	45	17
Player Y	10	15	24	26	34	45	25	31	20	40

Solution: Mean and Standard Deviation of Player X

Score **X**	**x = X - $\bar{X}$** **(Where $\bar{X}$ = 30)**	**x^2**
26	-4	16

24	-6	36
28	-2	2
30	0	0
35	5	25
40	10	100
25	-5	25
30	0	0
45	15	225
17	-13	169
$\sum$X = 300		**$\sum x^2$ = 600**

Mean ($\bar{X}$) = $\frac{\sum X}{n} = \frac{300}{10} = 30$

Standard Deviation (SD or **σ**) = $\sqrt{\frac{\sum x^2}{n}} = \sqrt{\frac{600}{10}} = \sqrt{60} = 7.746$

Coefficient of Variation = $\frac{\text{SD}}{\bar{X}} \times 100 = \frac{7.746}{30} \times 100 = 25.82\%$

Mean and Standard Deviation of Player Y

Score Y	**y = Y - $\bar{Y}$ (Where $\bar{Y}$ = 27)**	**y^2**
10	-17	289
15	-12	144
24	-3	9
26	-1	1
34	7	49
45	18	324
25	-2	4
31	4	16
20	-7	49
40	13	169
$\sum$X = 270		**$\sum x^2$ = 1054**

Mean ($\bar{Y}$) = $\frac{\sum Y}{n} = \frac{270}{10} = 27$

Standard Deviation (SD or **σ**) = $\sqrt{\frac{\sum y^2}{n}} = \sqrt{\frac{1054}{10}} = \sqrt{105.40} = 10.27$

Coefficient of Variation = $\frac{\text{SD}}{\bar{Y}} \times 100 = \frac{10.27}{27} \times 100 = 38.02\%$

Conclusion:

1. As average score of Player X is more than Player Y, he is better scorer.
2. As Coefficient of Variation of Player X is less than Player Y, he is more consistent also.

1.8.2 Standard Deviation in case of Discrete Series

Following are the formula for calculating Standard Deviation in case of the Discrete Series:

1. **Actual Mean Method –** In this method we take deviations from actual mean of the data.

$$\textbf{Standard Deviation (SD or } \sigma \textbf{)} = \sqrt{\frac{\sum fx^2}{n}}$$

Where x = X - $\bar{X}$, f = Frequency, n = Number of Items.

2. **Assumed Mean Method -** In this method we take deviations from assumed mean of the data.

$$\textbf{Standard Deviation (SD or } \sigma \textbf{)} = \sqrt{\frac{\sum fdx^2}{n} - \left(\frac{\sum fdx}{n}\right)^2}$$

Where dx = X – A, n = Number of Items.

3. **Direct Methods -** In this method we don't take deviations and standard deviation is calculated directly from the data.

$$\textbf{Standard Deviation (SD or } \sigma \textbf{)} = \sqrt{\frac{\sum fX^2}{n} - \left(\frac{\sum fX}{n}\right)^2}$$

Example 12. Following are the marks obtained by Students of a class in a test. Calculate Standard Deviation using (i) Actual Mean (ii) Assumed Mean (iii) Direct Method.

Marks	5	10	15	20	25	30	35
Frequency	2	7	11	15	10	4	1

Solution:1. Standard Deviation using Actual Mean

Marks X	**f**	**fX**	x **= X -** $\bar{X}$ **(**$\bar{X}$ **= 19)**	x^2	fx^2
5	2	10	-14	196	392
10	7	70	-9	81	567
15	11	165	-4	16	176

20	15	300	1	1	15
25	10	250	6	36	360
30	4	120	11	121	484
35	1	35	16	256	256
	N = 50	**$\sum$fX = 950**			**$\sum x^2$ = 2250**

Mean ($\bar{X}$) = $\frac{\sum fX}{n} = \frac{950}{50} = 19$

Standard Deviation (SD or **σ**) = $\sqrt{\frac{\sum fx^2}{n}} = \sqrt{\frac{2250}{50}} = \sqrt{45} = 6.708$

2. Standard Deviation using Assumed Mean

Marks X	**f**	**dx = X - A (A = 20)**	**dx^2**	**fdx**	**fdx^2**
5	2	-15	225	-30	450
10	7	-10	100	-70	700
15	11	-5	25	-55	275
20	15	0	0	0	0
25	10	5	25	50	250
30	4	10	100	40	400
35	1	15	225	15	225
	N = 50			**$\sum$fdx = -50**	**$\sum fdx^2$ = 2300**

Standard Deviation (**σ**) $= \sqrt{\frac{\sum fdx^2}{n} - \left(\frac{\sum fdx}{n}\right)^2}$

$= \sqrt{\frac{2300}{50} - \left(\frac{-50}{50}\right)^2} = \sqrt{46 - 1} = \sqrt{45} = 6.708$

3. Standard Deviation using Direct Method

Marks X	**f**	**X^2**	**fX**	**fX^2**
5	2	25	10	125
10	7	70	70	700
15	11	225	165	2475
20	15	400	300	6000
25	10	625	250	6250
30	4	900	120	3600
35	1	1225	35	1225
	N = 50		**$\sum$fX = 950**	**$\sum fX^2$ = 20300**

Standard Deviation ($\boldsymbol{\sigma}$) $= \sqrt{\frac{\sum fX^2}{n} - \left(\frac{\sum fX}{n}\right)^2}$

$= \sqrt{\frac{20300}{50} - \left(\frac{950}{50}\right)^2} = \sqrt{406 - 361} = \sqrt{45} = 6.708$

1.8.3 Standard Deviation in case of Continuous Series

In case of continuous series, the calculation will remain same as in case of discrete series but the only difference is that instead of taking deviations from data, deviations are taken from Mid value of the data. Formulas are same as discussed above for discrete series.

Example 13. Following are the marks obtained by Students of a class in a test. Calculate Standard Deviation using (i) Actual Mean (ii) Assumed Mean (iii) Direct Method. Also calculate coefficient of variation and Variance.

Marks	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	2	7	11	15	10	4	1

Solution:1. Standard Deviation using Actual Mean

Marks X	**m**	**f**	**fX**	$\boldsymbol{x}$ **= m -** $\boldsymbol{\bar{X}}$ **(**$\boldsymbol{\bar{X}}$ **= 21.5)**	$\boldsymbol{x^2}$	$\boldsymbol{fx^2}$
5-10	7.5	2	15	-14	196	392
10-15	12.5	7	87.5	-9	81	567
15-20	17.5	11	192.5	-4	16	176
20-25	22.5	15	337.5	1	1	15
25-30	27.5	10	275	6	36	360
30-35	32.5	4	130	11	121	484
35-40	37.5	1	37.5	16	256	256
		N = 50	**∑fX = 1075**			$\boldsymbol{\sum x^2}$ **= 2250**

Mean ($\bar{X}$) = $\frac{\sum fX}{n} = \frac{1075}{50} = 21.5$

Standard Deviation (SD or $\boldsymbol{\sigma}$) = $\sqrt{\frac{\sum fx^2}{n}} = \sqrt{\frac{2250}{50}} = \sqrt{45} = 6.708$

2. Standard Deviation using Assumed Mean

Marks	**m**	**f**	$\boldsymbol{dx}$ **= X - A**	$\boldsymbol{dx^2}$	**fdx**	$\boldsymbol{fdx^2}$

X			(A = 22.5)			
5-10	7.5	2	-15	225	-30	450
10-15	12.5	7	-10	100	-70	700
15-20	17.5	11	-5	25	-55	275
20-25	22.5	15	0	0	0	0
25-30	27.5	10	5	25	50	250
30-35	32.5	4	10	100	40	400
35-40	37.5	1	15	225	15	225
		N = 50			**∑fdx = -50**	**$\sum fdx^2$ = 2300**

Standard Deviation (**σ**) $= \sqrt{\frac{\sum fdx^2}{n} - \left(\frac{\sum fdx}{n}\right)^2}$

$= \sqrt{\frac{2300}{50} - \left(\frac{-50}{50}\right)^2} = \sqrt{46 - 1} = \sqrt{45} = 6.708$

3. Standard Deviation using Direct Method

Marks X	m	f	X^2	fX	fX^2
5-10	7.5	2	56.25	15	112.5
10-15	12.5	7	156.25	87.5	1093.75
15-20	17.5	11	306.25	192.5	3368.75
20-25	22.5	15	506.25	337.5	7593.75
25-30	27.5	10	756.25	275	7562.5
30-35	32.5	4	1056.25	130	4225
35-40	37.5	1	1406.25	37.5	1406.25
		N = 50		**∑fX = 1075**	**$\sum fX^2$ = 25366.5**

Standard Deviation (**σ**) $= \sqrt{\frac{\sum fX^2}{n} - \left(\frac{\sum fX}{n}\right)^2}$

$= \sqrt{\frac{25366.5}{50} - \left(\frac{1075}{50}\right)^2} = \sqrt{507.25 - 462.25} = \sqrt{45} = 6.708$

Coefficient of Standard Deviation = $\frac{\text{SD}}{\bar{X}} \times 100 = \frac{6.708}{21.5} \times 100 = 31.2\%$

Variance =$(\text{Standard Deviation})^2$ or $\sigma^2 = (6.708)^2 = 45$

1.8.4 Combined Standard Deviation

The main benefit of standard deviation is that if we know the mean and standard deviation of two or more series, we can calculate combined standard deviation of all the series. This feature is not available in other measures of dispersion. That's why we assume that standard deviation is best

measure of finding the dispersion. Following formula is used for this purpose:

$$\sigma_{123} = \sqrt{\frac{n_1\,\sigma_1^2 \;+\; n_2\,\sigma_2^2 \;+ n_3\,\sigma_3^2 \;+ n_1\,d_1^2 \;+\; n_2\,d_2^2 \;+ n_3\,d_3^2}{n_1 + \;n_2 + n_3}}$$

Where

n_1, n_2, n_3 = number of items in series 1, 2 and 3

σ_1, σ_2, σ_3 = standard deviation of series 1, 2 and 3

d_1, d_2, d_3 = difference between mean of the series and combined mean for 1, 2 and 3.

Example14. Find the combined standard deviation for the following data

	Firm A	***Firm B***
No. of Wage Workers	**70**	**60**
Average Daily Wage $(\boldsymbol{Rs.})$	**40**	**35**
S.D of wages	***8***	***10***

Solution: Combined mean wage of all the workers in the two firms will be

$$\overline{X_{12}} = \frac{N_1\overline{X_1} + N_2\overline{X_2}}{N_1 + N_2}$$

Where $N_1 =$ Number of workers in Firm A

$N_2 =$ Number of workers in Firm B

$\overline{X_1} =$ Mean wage of workers in Firm A

and $\overline{X_2} =$ Mean wage of workers in Firm B

We are given that

$$N_1 = 70 \qquad N_2 = 60$$

$$\overline{X_1} = 40 \qquad \overline{X_2} = 35$$

$\therefore$ Combined Mean, $\overline{X_{12}}$

$$= \frac{(70\times40)+(60\times35)}{70+60}$$

$$= \frac{4900}{130}$$

$$= Rs.\,37.69$$

Combined Standard Deviation =

$$\sigma_{123} = \sqrt{\frac{n_1\,\sigma_1^2 \;+\; n_2\,\sigma_2^2 \;+ n_1\,d_1^2 \;+\; n_2\,d_2^2}{n_1 + \;n_2}}$$

$d_1 = 40 - 37.69 = 2.31$

$d_2 = 35 - 37.69 = -2.69$

$\sigma_{123} = \sqrt{\frac{70\,(8)^2 + 60\,(10)^2 + 70\,(2.31)^2 + 60\,(-2.69)^2}{70 + 60}} = 9.318$

Example 15. Find the missing values

	Firm A	***Firm B***	***Firm C***	***Combined***
No. of Wage Workers	**50**	***?***	**90**	**200**
Average Daily Wage $(\boldsymbol{Rs.})$	**113**	?	**115**	**116**
S.D of wages	***6***	***7***	***?***	***7.746***

Solution: Combined $n = n_1 + n_2 + n_3$

$200 = 50 + n_2 + 90$

$N_2 = 60$

Now Combined mean wage of all the workers in the two firms will be

$$\overline{X_{12}} = \frac{N_1\overline{X_1} + N_2\overline{X_2} + N_3\overline{X_3}}{N_1 + N_2 + N_3}$$

We are given that

$N_1 = 50 \qquad N_2 = 60 \qquad N_3 = 90$

$\overline{X_1} = 113 \qquad \overline{X_2} = ? \qquad \overline{X_3} = 115 \qquad \overline{X_{123}} = 116$

∴ Combined Mean, $\overline{X_{12}}$

$$116 = \frac{(50\times113) + (60\times\overline{X_2}) + (90\times115)}{50+60+90}$$

$$116 = \frac{565 + (60\times\overline{X_2}) + 1035}{50+60+90}$$

$2320 = 1600 + 6\,\overline{X_2}$

$\overline{X_2} = 120$

Combined Standard Deviation =

$$\sigma_{123} = \sqrt{\frac{n_1\,\sigma_1^2 \;+\; n_2\,\sigma_2^2 \;+\; n_3\,\sigma_3^2 \;+ n_1\,d_1^2 \;+\; n_2\,d_2^2 \;+ n_3\,d_3^2}{n_1 + \;n_2 + n_3}}$$

$d_1 = 113 - 116 = -3$

$d_2 = 120 - 116 = 4$

$d_3 = 115 - 116 = -1$

$$\sigma_{123} = \sqrt{\frac{50\,(6)^2 + 60\,(7)^2 + 90\,(\sigma_3)^2 + 50\,(-3)^2 + 60\,(4)^2 + 90\,(-1)^2}{50 + 60 + 90}} = 7.746$$

Squaring the both sides

$$60 = \frac{180 + 294 + 9\sigma_3^2 45 + 96 + 9}{200}$$

$1200 = 9\,\sigma_3^2 + 624$

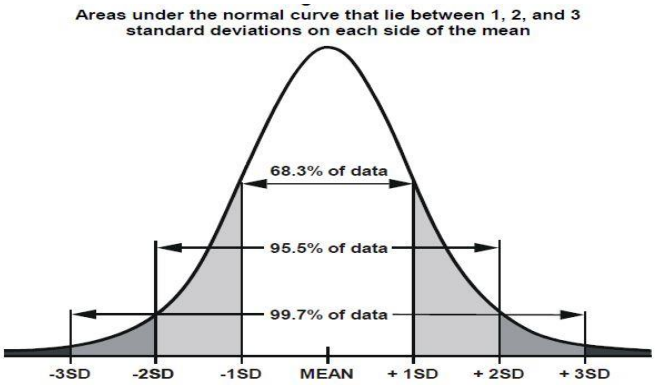
$\sigma_3 = 8$

1.8.5 Properties of Standard Deviation

1. Standard Deviation of first 'n' natural numbers is $\sqrt{\frac{n^2 - 1}{12}}$.
2. It is independent of change in origin it means it is not affected even if some constant is added or subtracted from all the values of the data.
3. It is not independent of change in scale. So if we divide or multiply all the values of the data with some constant, Standard Deviation is also multiplied or divided by same constant.
4. We can calculate combined Standard Deviation by following formula:

$$\sigma_{123} = \sqrt{\frac{n_1\,\sigma_1^2 + n_2\,\sigma_2^2 + n_3\,\sigma_3^2 + n_1\,d_1^2 + n_2\,d_2^2 + n_3\,d_3^2}{n_1 + n_2 + n_3}}$$

5. In case of normal distribution following results are found:



68.27% item lies within the range of: $\overline{X} \pm \sigma$

95.45% item lies within the range of: $\overline{X} \pm 2\,\sigma$

99.73% item lies within the range of: $\overline{X} \pm 3\sigma$

6. In case of normal distribution there is relation between Quartile Deviation, Mean Deviation and Standard Deviation which is as follows:

   6 (Q.D.) = 5 (M.D.) = 4 (S.D.)

7. In perfect symmetric distribution following result follows:

   Range = 6 (S.D.)

8. When we take square of Standard Deviation it is called Variance.

   Variance = $(S.D)^2$

1.8.6 Merits and Limitations of Standard Deviation

1. It is rigidly defined.
2. It is best measure to find out deviations.
3. It is based on arithmetic mean.
4. It is based on all the values.
5. We can find combined standard deviation of different series under this.
6. It is capable of further algebraic treatment.
7. By finding coefficient of variation, we can compare two different series.

Limitations of Standard Deviation

1. It is comparatively difficult to calculate.
2. It is mostly affected by extreme values.
3. Common people are not aware about the concept of standard deviation.

TEST YOUR PROGRESS (D)

1. Calculate Standard Deviation and find Variance:

X:	5	7	11	16	15	12	18	12

2. Two Batsmen X and Y score following runs in ten matches. Find who is better Scorer and who is more consistent?

X:	26	24	28	30	35	40	25	30	45	17
Y:	10	15	24	26	34	45	25	31	20	40

3. Calculate S.D, coefficient of SD, coefficient of Variation:

X	15	25	35	45	55	65
f	2	4	8	20	12	4

4. Find Standard Deviation.

X;	5-10	10-15	15-20	20-25	25-30	30-35
F:	2	9	29	24	11	6

5. Find Standard Deviation and coefficient of variation.

X;	30-39	40-49	50-59	60-69	70-79	80-89	90-99
F:	1	4	14	20	22	12	2

6. Find Standard Deviation.

X;	0-50	50-100	100-200	200-300	300-400	400-600
F:	4	8	10	15	9	7

7. Find combined Mean and Combined Standard Deviation:

Part	No. of Items	Mean	S.D.
1	200	25	3
2	250	10	4
3	300	15	5

8. Find missing information:

	Group I	Group II	Group III	Combined
No. of Items	200	?	300	750
Mean	?	10	15	16
S.D	3	4	?	7.1924

Answers

1. 4.12, 16.97	3. 11.83, 0.265, 26.5%	5. 12.505, 18.36%	7. 16, 7.2
2. X is better and consistent, X mean 30 CV 25.82%, Y mean 27 CV 38.02%	4. 5.74	6. 141.88	8. 250, 25, 5

1.9 LET US SUM UP

- Dispersion shows that whether average is a good representative of the series or not.
- High dispersion means values differ more than its average.

- There are two measures of dispersion, Absolute measure and relative measure.
- There are four methods that can be used for measuring the dispersion namely, Range, Quartile Deviation, Mean Deviation and Dispersion.
- Range is simples' method of dispersion.
- Mean deviation can be calculated from Mean, Median or Mode
- Standard Deviation is the best measure of Dispersion.
- If we know standard deviation of two series, we can calculate combined standard deviation.

1.10 KEY TERMS

- **Dispersion:** Dispersion shows the extent to which individual items in the data differs from its average. It is a measure of difference between data and the individual items. It indicates that how that are lacks the uniformity.
- **Range:** Range is the difference between highest value of the data and the lowest value of the data. The more is the difference between highest and the lowest value, more is the value of Range which shows high dispersion.
- **Quartile Deviation:** Quartile deviation is the Arithmetic mean of the difference between Third Quartile and the First Quartile of the data.
- **Mean Deviation:** Mean Deviation is the value obtained by taking arithmetic mean of the deviations obtained by deducting average of data whether Mean, Median or Mode from values of data, ignoring the signs of the deviations.
- **Standard Deviation:** Standard Deviation is the square root of the Arithmetic mean of the squares of deviation of the item from its Arithmetic mean.
- **Variance:** It is square of Standard Deviation.
- **Absolute Measure:** Absolute measure of dispersion is one which is expressed in the same statistical unit in which the original values of that data are expressed. For example, if original data is represented in kilograms, the dispersion will also be represented in kilogram.
- **Relative Measure:** The relative measure of dispersion is independent of unit of measurement and is expressed in pure number. Normally it is a ratio of the dispersion to the average of the data.
- **Coefficient of Standard Deviation:** Coefficient of Deviation is the relative measure of the standard deviation and can be calculated by dividing the Value of Standard Deviation with the

Arithmetic Mean. The value of coefficient always lies between 0 and 1, where 0 indicates no Standard Deviation and 1 indicated 100% standard deviation.

1.11 QUESTIONS FOR PRACTICE

1. What is Dispersion? Explain its uses.
2. What are features of good measure of Dispersion?
3. What are absolute and relative measure of dispersion?
4. What is range? Give its merits and limitations.
5. What are Quartile deviations? Give its merits and limitations.
6. What is mean deviation. How it is calculated.
7. What is standard deviation? How it is calculated.
8. How combined standard deviation can be calculated.
9. Give properties of standard deviation.

1.12 FURTHER READINGS

- J. K. Sharma, *Business Statistics*, Pearson Education.
- S.C. Gupta, *Fundamentals of Statistics*, Himalaya Publishing House.
- S.P. Gupta and Archana Gupta, *Elementary Statistics*, Sultan Chand and Sons, New Delhi.
- Richard Levin and David S. Rubin, *Statistics for Management*, Prentice Hall of India, New Delhi.
- M.R. Spiegel, *Theory and Problems of Statistics*, Schaum's Outlines Series, McGraw Hill Publishing Co.

CERTIFICATE/DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SARM 2: DESCRIPTIVE STATISTICS

SEMESTER I

UNIT 2: CO-EFFICIENT OF VARIATION (CV), LORENZ CURVE, MEANING AND MEASURES OF SKEWNESS, KURTOSIS, MOMENTS

STRUCTURE

2.0 Learning Objectives

2.1 Introduction and Concept of Coefficient of Variation (CV)

2.2 Importance and applications in various fields

2.4 Calculation of Coefficient of Variation

2.6 Interpreting CV Values and Their Significance

2.7 Comparing Coefficient of Variation (CV) Values across Different Datasets or Populations

2.8 Evaluating the Relative Variability and Risk Associated with CV

2.9 Advantages of using Coefficient of Variation (CV) over other measures of dispersion

2.10 Limitations of using Coefficient of Variation (CV)

2.11 Discussing scenarios where CV is useful and informative

2.12 Lorenz curve introduction

- **2.12.1 Examples of Lorenz curves**
- **2.12.2 Lorenz Dominance**

2.13 Meaning and Measures of skewness

2.14 Meaning and Measures of kurtosis

2.15 Meaning and Measures of Moments

2.16 Sum Up

2.17 Questions for Practice

2.18 MCQs

2.19 Suggested Readings

2.0 LEARNING OBJECTIVES

After studying the Unit, Learner will be able to know:

- concept of Coefficient of Variation (CV)
- applications in various fields
- calculation of Coefficient of Variation
- advantages of using Coefficient of Variation (CV) over other measures of dispersion
- limitations of using Coefficient of Variation (CV)
- meaning of skewness, kurtosis, Moments

2.1 INTRODUCTION AND CONCEPT OF COEFFICIENT OF VARIATION (CV)

The Coefficient of Variation (CV) is a statistical measure that quantifies the relative variability or dispersion of a dataset in relation to its mean. It provides a standardized measure of variability, allowing for comparisons between datasets with different scales or units of measurement. The CV is widely used in various fields, including finance, economics, healthcare, quality control, and social sciences.

The Coefficient of Variation is defined as the ratio of the standard deviation (SD) to the mean (μ) of a dataset, expressed as a percentage. It is calculated using the following formula:

CV = (SD / μ) * 100

Where: CV: Coefficient of Variation, SD: Standard Deviation, μ: Mean

The CV represents the relative variability in a dataset when compared to the mean. It allows for the comparison of datasets with different units of measurement or scales. By expressing the CV as a percentage, it provides a standardized measure that is independent of the scale or magnitude of the data. The concept of CV can be understood by considering two datasets with the same mean but different standard deviations. Dataset A has a smaller standard deviation compared to Dataset B. The CV for Dataset A will be smaller, indicating that the variability in Dataset A is relatively lower compared to its mean. Conversely, the CV for Dataset B will be larger, indicating higher

relative variability. The CV is a useful measure as it allows researchers and analysts to assess and compare the variability of different datasets, providing insights into the consistency or volatility within the data. A lower CV suggests a more stable and less variable dataset, while a higher CV indicates greater variability or dispersion.

In practical terms, the CV helps in decision-making processes, risk assessment, and evaluating the performance of systems or processes. For example, in finance, a lower CV may indicate a less risky investment, while a higher CV may suggest a more volatile investment. In healthcare, the CV can be used to assess the consistency of treatment outcomes or the spread of diseases across different populations. Overall, the Coefficient of Variation is a valuable statistical measure that provides a standardized assessment of variability, enabling comparisons between datasets and aiding in data-driven decision-making.

2.2 IMPORTANCE AND APPLICATIONS IN VARIOUS FIELDS

The Coefficient of Variation (CV) holds significant importance in a wide range of fields due to its ability to provide insights into the relative variability of datasets. Here are some key areas where the CV finds applications:

1. **Finance and Risk Assessment:** In finance, the CV is utilized to assess the risk associated with investments. A lower CV indicates a more stable investment with less variability, while a higher CV suggests greater volatility and potential risk. Portfolio managers use the CV to compare the risk and return profiles of different investment options and make informed decisions. Risk analysts employ the CV to evaluate the volatility of financial markets, securities, and investment portfolios.
2. **Economics and Economic Indicators:** Economic researchers use the CV to analyze economic indicators, such as inflation rates, GDP growth, or unemployment rates, across different regions or time periods. The CV helps in understanding the variability in economic data and provides insights into the stability or volatility of economic trends. It aids in comparing the economic performance of countries or sectors, guiding policy decisions and investment strategies.
3. **Healthcare and Medicine:** In medical research and clinical studies, the CV is employed to assess the variability in treatment outcomes or patient responses to different interventions. Epidemiologists use the CV to analyze the spread and variability of diseases

across different populations, helping in the assessment of disease control measures. Quality control professionals utilize the CV to evaluate the consistency of medical test results, ensuring reliability and precision.

4. **Quality Control and Process Stability:** The CV is widely used in quality control to assess the consistency and stability of manufacturing processes. It helps in monitoring process variations, identifying sources of variability, and implementing corrective actions. The CV aids in comparing the performance of different manufacturing sites or production lines, facilitating process optimization and quality improvement.
5. **Social Sciences and Behavioral Variables:** Researchers in social sciences use the CV to analyze and compare behavioral variables such as income distribution, crime rates, educational performance, or survey responses. It provides insights into the variability and inequality within social systems, helping policymakers and social scientists understand and address societal issues.
6. **Environmental Studies and Natural Resource Management:** The CV is applied in environmental studies to assess the variability of ecological indicators, such as water quality, air pollution levels, or biodiversity indices. It helps in identifying regions or ecosystems with high variability and focusing conservation efforts or resource management strategies accordingly.
7. **Experimental Sciences and Laboratory Research:** The CV is useful in experimental sciences to evaluate the precision and consistency of laboratory measurements. It assists in comparing the reliability of different measurement techniques or instruments, optimizing experimental protocols, and interpreting research findings.

2.4 CALCULATION OF COEFFICIENT OF VARIATION

This section delves into the calculation of the CV using the formula: $CV = (SD / \mu) * 100$, where SD is the standard deviation and μ is the mean.

The numerator and denominator components of the formula are explained in detail, ensuring a clear understanding of the calculation process.

Step-by-step calculation examples are provided to illustrate the application of the CV formula.

The Coefficient of Variation (CV) is calculated using a straightforward formula that involves the standard deviation (SD) and the mean (μ) of a dataset. The CV formula is as follows:

$CV = (SD / \mu) * 100$

Where: CV: Coefficient of Variation, SD: Standard Deviation, μ: Mean

The CV formula is designed to provide a measure of relative variability by expressing the standard deviation as a percentage of the mean. It standardizes the variability metric, allowing for comparisons between datasets with different units of measurement or scales.

Let's further explain the components of the CV formula:

Standard Deviation (SD):

The standard deviation is a measure of the dispersion or variability of a dataset.

It quantifies how far individual data points deviate from the mean.

A higher standard deviation indicates greater variability in the dataset, while a lower standard deviation suggests less variability.

Mean (μ):

The mean is the average value of a dataset.

It represents the central tendency or the typical value around which the data points cluster.

The mean is calculated by summing all the values in the dataset and dividing by the total number of observations.

Coefficient of Variation (CV):

The CV is the ratio of the standard deviation to the mean, expressed as a percentage.

By multiplying the ratio by 100, the CV is converted into a percentage value.

The CV quantifies the relative variability of the dataset in relation to its mean, providing a standardized measure.

The CV is a dimensionless measure since it represents a ratio of two quantities with the same units. It is often expressed as a percentage to enhance its interpretability and make comparisons more intuitive.

For example, let's consider a dataset of exam scores where the mean is 80 and the standard deviation is 10. The CV can be calculated as follows:

CV = (10 / 80) * 100

CV = 12.5%

In this case, the CV of 12.5% indicates that the dataset has a relative variability of approximately 12.5% with respect to the mean. This implies that the scores exhibit moderate variability around the average performance.

Calculating the CV allows for a standardized assessment of variability, enabling comparisons between datasets with different means and standard deviations. It provides a useful measure to evaluate the relative consistency or dispersion within a dataset, supporting data analysis and decision-making processes.

To understand the Coefficient of Variation (CV) formula, it is important to grasp the meaning and significance of its numerator (standard deviation) and denominator (mean) components. Let's delve into these components in more detail:

Numerator Component: Standard Deviation (SD)

The numerator of the CV formula involves the standard deviation (SD) of the dataset. The standard deviation is a measure of the dispersion or variability of the data points from the mean. It quantifies how much individual data points deviate from the average value.

A higher standard deviation indicates greater variability in the dataset, signifying that the data points are more spread out from the mean. Conversely, a lower standard deviation suggests less variability, indicating that the data points are closer to the mean.

The standard deviation is calculated using the following steps:

Compute the mean (μ) of the dataset.

Calculate the difference between each data point and the mean.

Square each difference.

Calculate the mean of the squared differences.

Take the square root of the mean squared differences to obtain the standard deviation.

The numerator (SD) in the CV formula captures the extent of variability in the dataset, serving as a measure of dispersion.

Denominator Component: Mean (μ)

The denominator of the CV formula involves the mean (μ) of the dataset. The mean represents the average value of the dataset and serves as a measure of central tendency. It is obtained by summing all the data points and dividing the sum by the total number of observations.

The mean is a crucial component in the CV formula as it provides a reference point around which the data points are evaluated for their variability. By dividing the standard deviation by the mean, the CV expresses the variability in relation to the average value. This ratio allows for the comparison of datasets with different means and scales, making the CV a standardized measure of variability.

The CV formula, combining the standard deviation (numerator) and the mean (denominator), provides a relative measure of variability. By expressing the standard deviation as a percentage of the mean, the CV allows for meaningful comparisons across datasets.

Overall, the numerator (standard deviation) captures the dispersion or variability within the dataset, while the denominator (mean) provides a reference point for evaluating the relative variability. The combination of these components in the CV formula enables a standardized assessment of variability, facilitating comparisons and analysis across different datasets.

Example 1: Consider a dataset of monthly sales figures for a retail store over a year:

50,000, 48,000, 52,000, 55,000, 49,000, 51,000, 53,000, 50,000, 54,000, 52,000, 47,000, 50,000

Step 1: Calculate the mean (μ) of the dataset.

μ = (50,000 + 48,000 + 52,000 + 55,000 + 49,000 + 51,000 + 53,000 + 50,000 + 54,000 + 52,000 + 47,000 + 50,000) / 12

μ = 50,333.33

Step 2: Calculate the standard deviation (SD) of the dataset.

Calculate the squared difference between each data point and the mean.

Sum up the squared differences.

Divide the sum by the total number of observations (12 in this case).

Take the square root of the result.

SD = √[((50,000 - 50,333.33)^2 + (48,000 - 50,333.33)^2 + ... + (50,000 - 50,333.33)^2) / 12]

SD = √[8,666,666.67 / 12]

SD = √[722,222.22]

SD ≈ 849.84

Step 3: Calculate the CV using the formula: CV = (SD / μ) * 100

CV = (849.84 / 50,333.33) * 100

CV ≈ 1.69%

The Coefficient of Variation (CV) for this dataset of monthly sales figures is approximately 1.69%. It indicates a relatively low level of variability in sales when compared to the mean.

Example 2:

Consider a dataset of daily temperature readings in Celsius for a week:

18, 17, 16, 20, 19, 18, 17

Step 1: Calculate the mean (μ) of the dataset.

μ = (18 + 17 + 16 + 20 + 19 + 18 + 17) / 7

μ = 17.86

Step 2: Calculate the standard deviation (SD) of the dataset.

Calculate the squared difference between each data point and the mean.

Sum up the squared differences.

Divide the sum by the total number of observations (7 in this case).

Take the square root of the result.

SD = √[((18 - 17.86)^2 + (17 - 17.86)^2 + ... + (17 - 17.86)^2) / 7]

SD = √[0.48 / 7]

SD ≈ 0.30

Step 3: Calculate the CV using the formula: CV = (SD / μ) * 100

CV = (0.30 / 17.86) * 100

CV ≈ 1.68%

The Coefficient of Variation (CV) for this dataset of daily temperature readings is approximately 1.68%. It suggests a relatively low level of variability in temperature across the week when compared to the mean.

These examples illustrate the step-by-step calculation process of the Coefficient of Variation (CV) for different datasets, showcasing how the CV captures the relative variability in relation to the mean.

2.6 INTERPRETING COEFFICIENT OF VARIATION (CV) VALUES AND THEIR SIGNIFICANCE

When interpreting Coefficient of Variation (CV) values, it is important to consider the magnitude of the CV and its implications for the dataset under analysis. Here are some general guidelines for understanding the significance of CV values:

Low Coefficient of Variation (CV):

A low CV indicates a relatively low level of variability in the dataset compared to the mean.

It suggests that the data points are clustered closely around the mean, indicating a higher level of consistency or stability.

In practical terms, a low CV implies that the dataset is relatively homogeneous or has a narrow range of values.

Examples where a low CV may be desirable include quality control processes, where consistency and precision are crucial.

Moderate Coefficient of Variation (CV):

A moderate CV suggests a moderate level of variability in the dataset compared to the mean.

It indicates that the data points have some dispersion around the mean, but not to a significant extent.

In practical terms, a moderate CV implies that there is a certain degree of diversity or spread in the dataset, but it is not excessively high.

Examples where a moderate CV may be observed include economic indicators, where some fluctuation is expected but not extreme.

High Coefficient of Variation (CV):

A high CV indicates a relatively high level of variability in the dataset compared to the mean.

It suggests that the data points are spread out from the mean, indicating a higher level of dispersion or volatility.

In practical terms, a high CV implies that the dataset is heterogeneous or has a wide range of values.

Examples where a high CV may be observed include financial markets, where significant fluctuations and risks are present.

It is important to note that the interpretation of CV values depends on the context and the nature of the dataset being analyzed. What constitutes a low, moderate, or high CV may vary across different fields, industries, or research domains. It is crucial to compare CV values within the specific context or against relevant benchmarks or standards.

Additionally, it is important to consider other factors and characteristics of the dataset alongside the CV. For example, the presence of outliers, data distribution, sample size, and specific domain knowledge may provide additional insights into the variability and its implications.

Ultimately, the interpretation of CV values should be done in conjunction with the goals of the analysis, the nature of the dataset, and the specific context in which the data is being examined. It is advisable to consider the CV alongside other statistical measures and domain-specific knowledge for a comprehensive understanding of the dataset's variability.

2.7 COMPARING COEFFICIENT OF VARIATION (CV) VALUES ACROSS DIFFERENT DATASETS OR POPULATIONS

The Coefficient of Variation (CV) is a useful measure for comparing the relative variability of datasets or populations, especially when they have different means or scales. When comparing CV values across different datasets, consider the following points:

Magnitude of CV:

Compare the magnitude of the CV values between datasets. A higher CV indicates greater relative variability compared to the mean, while a lower CV suggests lower relative variability.

Keep in mind that the interpretation of "high" or "low" CV values may vary depending on the context and the specific field or industry under consideration.

Consider the Means:

Take into account the means of the datasets being compared. Even if two datasets have similar CV values, their actual levels of variability may differ if their means are substantially different.

Consider the absolute values of the means and how they relate to the CV values to get a comprehensive understanding of the variability.

Sample Size:

Be mindful of the sample sizes of the datasets being compared. Smaller sample sizes may result in more variability and less reliable CV estimates.

Larger sample sizes generally provide more stable and accurate estimates of variability.

Domain-specific Considerations:

Consider the specific characteristics and requirements of the field or domain being studied. What may be considered as acceptable or desirable levels of variability can differ depending on the context.

Look for domain-specific benchmarks, standards, or prior research to compare the CV values against.

Similarity of Data:

Ensure that the datasets being compared are measuring similar variables or phenomena. Comparing CV values across unrelated or dissimilar datasets may not yield meaningful insights.

If possible, ensure that the datasets are collected or measured using similar methods or under comparable conditions for more accurate comparisons.

Statistical Significance:

Consider conducting appropriate statistical tests to determine if the differences in CV values between datasets are statistically significant.

Hypothesis testing or confidence intervals can help assess whether the observed differences in CV are due to chance or reflect true differences in variability.

Remember that the CV is just one measure of relative variability, and it should be considered alongside other statistical measures and domain-specific knowledge. Additionally, always exercise caution when comparing CV values, as their interpretation can be influenced by various factors.

2.8 EVALUATING THE RELATIVE VARIABILITY AND RISK ASSOCIATED WITH COEFFICIENT OF VARIATION (CV)

The Coefficient of Variation (CV) provides a measure of relative variability in a dataset, and it can be used to assess the associated risk. Here are some considerations for evaluating the relative variability and risk associated with CV:

Comparison to Benchmarks: Compare the CV of the dataset to relevant benchmarks or standards within the specific field or industry.

Look for established norms or guidelines that indicate acceptable levels of variability or risk.

If the CV value exceeds established benchmarks, it may indicate higher relative variability or risk compared to the expected or desired range.

Impact on Decision-Making: Assess how the variability captured by the CV may impact decision-making or outcomes.

Consider the context and consequences of higher variability. For example, in financial investments, a higher CV may imply greater risk and potential for larger fluctuations in returns.

Sensitivity to Changes: Evaluate the sensitivity of the dataset or system to changes in the CV value.

Higher CV values suggest a greater sensitivity to variability, indicating that small changes can have a significant impact.

Consider whether the dataset or system can tolerate or manage such variability without adverse effects.

Stability and Consistency: Examine the stability and consistency of the dataset over time or across similar conditions.

If the CV value fluctuates significantly, it may indicate an unstable or inconsistent dataset, which could introduce risks in decision-making or planning.

Risk Assessment: Consider the specific risks associated with the dataset and how the CV relates to those risks.

Higher CV values generally suggest higher risk, as they indicate greater variability or dispersion in the data.

Assess the potential consequences or impact of that variability on the desired outcomes or objectives.

Domain-specific Factors: Take into account domain-specific knowledge, expertise, and guidelines for evaluating the relative variability and associated risk.

Different fields or industries may have different thresholds for acceptable levels of variability or risk. It's important to note that the evaluation of risk associated with CV should be done in conjunction with other relevant measures, statistical analyses, and domain-specific considerations. The CV provides a relative measure of variability, and its interpretation in terms of risk will depend on the specific context, objectives, and domain of the analysis.

2.9 ADVANTAGES OF USING COEFFICIENT OF VARIATION (CV) OVER OTHER MEASURES OF DISPERSION

Standardizes Comparison: One of the key advantages of CV is that it standardizes the comparison of variability across datasets or populations with different means or scales. By expressing the standard deviation as a percentage of the mean, the CV allows for meaningful comparisons and facilitates the understanding of relative variability.

Scale-Invariant: The CV is scale-invariant, which means it remains the same regardless of the units or scales used in the dataset. This makes it particularly useful when comparing datasets that have different measurement units or scales, as it eliminates the influence of the scale and focuses solely on relative variability.

Useful for Heterogeneous Datasets: CV is particularly advantageous when dealing with datasets that have different means or distributions but exhibit similar levels of relative variability. It helps identify datasets with comparable levels of dispersion, even if their absolute values differ significantly.

Interpretability: CV is relatively easy to interpret compared to other measures of dispersion, such as standard deviation or range. The percentage representation of CV allows for a straightforward understanding of the relative variability in relation to the mean.

Useful for Risk Assessment: CV is commonly used in risk assessment and decision-making processes. It provides a standardized measure of variability that can aid in assessing the potential risk or uncertainty associated with certain datasets or populations. Higher CV values often indicate higher levels of risk or volatility.

2.10 LIMITATIONS OF USING COEFFICIENT OF VARIATION (CV)

Sensitivity to Outliers: CV can be sensitive to extreme values or outliers in the dataset, especially when they have a significant impact on the standard deviation. Outliers can distort the variability measure and lead to inaccurate interpretations of CV. It is important to carefully examine and consider the presence of outliers before relying solely on CV.

Dependence on Mean: CV is dependent on the mean of the dataset. If the mean is close to zero or very small, the CV value can become large, making it difficult to interpret the relative variability accurately. In such cases, caution should be exercised when interpreting CV.

Reliability with Small Sample Sizes: CV estimates can be less reliable with small sample sizes. As the sample size decreases, the variability estimate becomes less precise, which can affect the accuracy of the CV calculation. Larger sample sizes tend to provide more stable and reliable CV estimates.

Limited to Continuous Variables: CV is most suitable for continuous variables and may not be as applicable for categorical or ordinal variables. It relies on the assumption of a continuous distribution and may not provide meaningful results for discrete or categorical data.

Lack of Information on Data Distribution: CV does not provide information about the specific distribution or shape of the data. It focuses solely on relative variability and does not capture the complete picture of the data distribution.

It is important to consider these advantages and limitations when deciding to use CV as a measure of dispersion and to supplement its interpretation with other relevant statistical measures and domain-specific knowledge.

2.11 DISCUSSING SCENARIOS WHERE CV IS USEFUL AND INFORMATIVE

The coefficient of Variation (CV) is a useful and informative measure in various scenarios. Here are some scenarios where CV can provide valuable insights:

Comparative Analysis: CV is particularly useful when comparing the variability of different datasets or populations. It allows for a standardized comparison of relative variability, regardless of the scales or units of measurement. For example, when comparing the performance of different investment portfolios, CV can help identify the portfolio with the most consistent returns relative to the mean.

Quality Control and Process Monitoring: In industries where consistency and precision are critical, such as manufacturing or healthcare, CV can be used to assess the variation in process outputs. A low CV indicates a more stable and reliable process, while a high CV suggests greater variability and potential quality issues. Monitoring the CV over time can help identify variations and guide process improvement efforts.

Financial Risk Assessment: CV is commonly employed in financial analysis to evaluate the risk associated with investments or portfolios. Higher CV values indicate greater variability in returns, reflecting higher risk and potential for larger fluctuations. It helps investors and financial analysts assess the volatility and relative riskiness of different assets or investment strategies.

Research and Experimental Studies: In scientific research and experimental studies, CV can be used to measure and compare the variability in experimental results. It provides insights into the consistency or reproducibility of the outcomes. Researchers can use CV to determine the stability of measurements and identify factors contributing to variability.

Biomedical and Health Sciences: CV is applicable in various areas of biomedical and health sciences. For instance, in clinical studies, CV can be used to evaluate the variability of treatment effects or patient responses. It helps assess the consistency of outcomes and the potential impact of variability on clinical decision-making.

Environmental Monitoring: In environmental studies, CV can be used to assess the variability of ecological or environmental parameters. It provides information about the stability and predictability of natural systems, such as water quality measurements, air pollutant concentrations, or biodiversity indices. Comparing CV values across different locations or time periods can help identify areas of concern or potential impacts.

Overall, CV is particularly useful in situations where the focus is on comparing relative variability, assessing risk, monitoring processes, or evaluating the consistency of outcomes. Its ability to standardize comparisons and provide a percentage-based measure makes it informative across a wide range of fields and industries. However, it is essential to consider the limitations of CV and supplement its interpretation with other statistical measures and domain-specific knowledge for a comprehensive analysis.

2.12 LORENZ CURVE

The Lorenz curve is a graphical representation of income inequality or wealth distribution within a population. It was developed by Max O. Lorenz, an American economist, in 1905. The curve is widely used in economics and sociology to analyze and compare the distribution of resources in different societies or regions.

The Lorenz curve is created by plotting the cumulative percentage of total income or wealth received by a given percentage of the population. The x-axis represents the cumulative percentage of the population, ranked by ascending order of income or wealth, while the y-axis represents the cumulative percentage of total income or wealth held by that portion of the population.

A perfectly equal distribution of income or wealth would be represented by a 45-degree line (known as the line of perfect equality) from the origin (0,0) to the point (100,100) on the graph. In this case, each percentage of the population would receive exactly the same percentage of income or wealth.

However, in reality, income and wealth are typically distributed unequally. The Lorenz curve will generally lie below the line of perfect equality, indicating the extent of income or wealth inequality. The greater the deviation from the line of perfect equality, the more unequal the distribution.

The Lorenz curve can also be summarized by a single numerical measure called the Gini coefficient. It is calculated as the area between the Lorenz curve and the line of perfect equality, divided by the total area under the line of perfect equality. The Gini coefficient ranges between 0 and 1, where 0 represents perfect equality and 1 represents maximum inequality.

The Lorenz curve and Gini coefficient provide valuable insights into the distribution of income or wealth and help policymakers and researchers assess the level of economic inequality within a society.

2.12.1 EXAMPLES OF LORENZ CURVES

The Lorenz curve is a graphical representation of income inequality or wealth distribution within a population. It plots the cumulative percentage of income or wealth on the y-axis against the cumulative percentage of the population ranked by income or wealth on the x-axis. Here are a few examples of Lorenz curves to illustrate different income distributions:

a) **Perfect Equality:** In a perfectly equal society, where every individual has the same income or wealth, the Lorenz curve would be a straight line at a 45-degree angle from the origin (0,0) to (100,100). This indicates that the cumulative percentage of income/wealth is equal to the cumulative percentage of the population at every point.
b) **High-Income Inequality:** In a society with high-income inequality, the Lorenz curve will be concave, indicating that a small percentage of the population holds a large percentage of income or wealth. The curve will lie far below the line of perfect equality. This suggests that a significant portion of the population has relatively low income or wealth, while a small fraction holds a disproportionately large share.
c) **Moderate Income Inequality:** In a society with moderate income inequality, the Lorenz curve will still be concave but closer to the line of perfect equality. This indicates that while there is some concentration of income or wealth among a portion of the population, it is not as extreme as in the case of high-income inequality.
d) **Low Income Inequality:** In a society with low-income inequality, the Lorenz curve will be convex, meaning that income or wealth is more evenly distributed across the population. The curve will lie above the line of perfect equality, suggesting that a larger percentage of the population holds a proportionate share of the income or wealth.

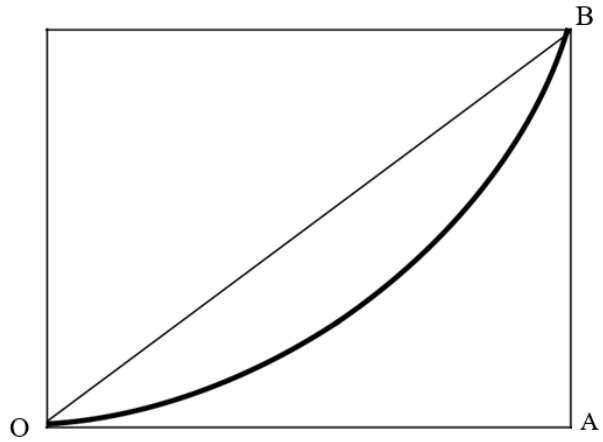
Please note that these are general examples, and actual Lorenz curves may vary depending on the specific income or wealth distribution within a population. The shape of the curve provides insights into the level of inequality present and can be used to compare different societies or track changes in income distribution over time.

Lorenz curve express the relation between the cumulative proportion of people with income at least equal to some specific value and the cumulative proportion of income received by these people. Lorenz curve is represented by a function L(P), which corresponds to a fraction received

by the p-th lower fraction of the population when it is ordered by increasing income. The curve slope is always positive and convex, so L (0) = 0 and L (1) = 1.

The line L(p)=p is the line of perfect equity, corresponding to the OB line in the graph below. It is a situation in which everybody receives the same amount of income. The line of extreme inequity corresponds to the lines AO and AB. It is a situation in which everybody receives zero income except the richest person, who accumulates the total income.

Lorenz curve is always between the line of perfect equity and the line of extreme inequity. When nearest to the line of perfect equity, the more egalitarian is the income distribution.



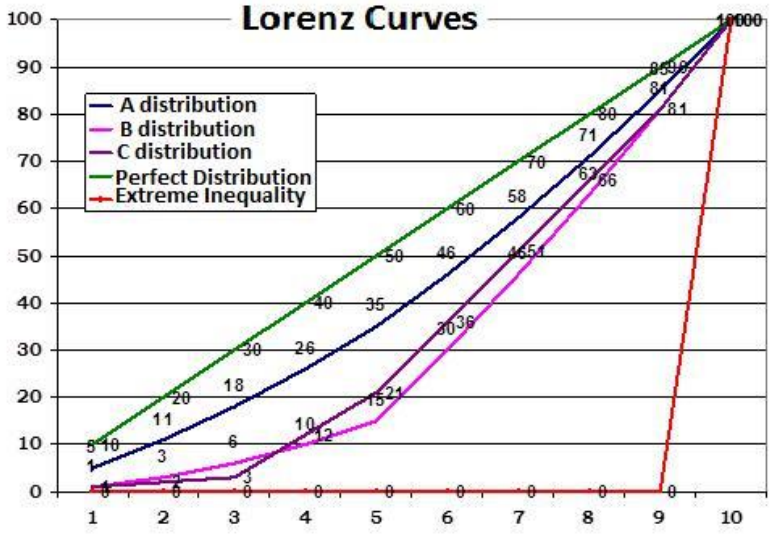
2.12.2 LORENZ DOMINANCE

We say that the Lorenz curve of distribution A dominates distribution of B if curve A is above curve B in all points of the distribution. In this case one can say A is more equal than B. And if both distributions have the same mean, A is preferable to B.

If there is an intersection point between two Lorenz curves, we can only make statements about stretches of the distribution. In this case, we can always find welfare functions that rank the distribution differently. In the example below, distribution of A dominates distribution of B and C, but distributions of B and C do not have any type of dominance when compared.

								Perfect		
	A Distribution		**B Distribution**		**C Distribution**			**Distribution**	**Extreme Inequity**	

		Accumulated		Accumulated		Accumulated		Accumulated		Accumulated
1	**5**	5	**1**	1	**1**	1	**10**	10	**0**	0
2	**6**	11	**2**	3	**1**	2	**10**	20	**0**	0
3	**7**	18	**3**	6	**1**	3	**10**	30	**0**	0
4	**8**	26	**4**	10	**9**	12	**10**	40	**0**	0
5	**9**	35	**5**	15	**9**	21	**10**	50	**0**	0
6	**11**	46	**15**	30	**15**	36	**10**	60	**0**	0
7	**12**	58	**16**	46	**15**	51	**10**	70	**0**	0
8	**13**	71	**17**	63	**15**	66	**10**	80	**0**	0
9	**14**	85	**18**	81	**15**	81	**10**	90	**0**	0
10	**15**	100	**19**	100	**19**	100	**10**	100	**100**	100



2.13 MEANING AND MEASURES OF SKEWNESS

Skewness is a statistical measure that helps to assess the asymmetry or lack of symmetry in a probability distribution of a random variable. It indicates the degree to which the values in a dataset are skewed or deviate from a symmetric distribution. Skewness can take positive or negative values or even zero, each indicating a different type of skewness:

a) Positive Skewness: If the distribution has a long tail on the right side and the majority of the data is concentrated on the left side, it is said to have positive skewness. The right tail is stretched out, and the mean is typically greater than the median.

b) Negative Skewness: If the distribution has a long tail on the left side and the majority of the data is concentrated on the right side, it is said to have negative skewness. The left tail is stretched out, and the mean is typically smaller than the median.

c) Zero Skewness: If the distribution is perfectly symmetrical, it has zero skewness. This means that the data is equally distributed on both sides of the mean, and the mean and median are equal.

There are various measures of skewness used to quantify the extent of skewness in a dataset. Some common measures include:

Pearson's First Coefficient of Skewness (moment skewness): It is defined as the third standardized moment of a distribution. The formula for Pearson's first coefficient of skewness is:

Skewness = (3 * (Mean - Median)) / Standard Deviation

Here, Mean refers to the arithmetic mean, Median is the median of the data, and Standard Deviation is the standard deviation of the dataset. Bowley's Skewness Coefficient: It is a measure of skewness based on quartiles. The formula for Bowley's skewness coefficient is:

Skewness = (Q1 + Q3 - 2 * Median) / (Q3 - Q1)

Here, Q1 and Q3 are the first and third quartiles, respectively.

Sample Skewness: It is a measure of skewness based on moments. The formula for sample skewness is:

Skewness = (1 / n) * ∑[(xi - Mean) / Standard Deviation]^3

Here, n is the sample size, xi represents each observation in the dataset, Mean is the arithmetic mean, and Standard Deviation is the standard deviation.

These are some commonly used measures of skewness, and each provides a different perspective on the skewness of the data. It's important to consider multiple measures and examine the data distribution to gain a comprehensive understanding of skewness.

2.14 MEANING AND MEASURES OF KURTOSIS

Kurtosis is a statistical measure that describes the shape and distribution of a probability distribution or a dataset. It provides information about the presence and nature of outliers and the relative peakedness or flatness of the distribution compared to a normal distribution.

The term "kurtosis" originates from the Greek word "kurtos," which means "curved" or "arched." Kurtosis measures the curvature of the distribution's tails, indicating whether the distribution has more or fewer outliers or extreme values than a normal distribution.

Positive kurtosis indicates heavy tails and a distribution that is more peaked than the normal distribution, while negative kurtosis indicates light tails and a flatter distribution compared to the normal distribution.

There are different ways to measure kurtosis, but the most commonly used measures are Pearson's kurtosis (or excess kurtosis) and Fisher's kurtosis.

a) Pearson's Kurtosis (Excess Kurtosis):

Pearson's kurtosis measures the kurtosis relative to the normal distribution. It subtracts 3 from Fisher's kurtosis, which is the kurtosis of a normal distribution. The formula for Pearson's kurtosis is:

Kurtosis = (M4 / M2^2) - 3

Where:

M4 is the fourth moment about the mean.

M2 is the second moment about the mean (variance).

If the calculated kurtosis is greater than 0, it indicates positive kurtosis (heavier tails than a normal distribution), and if it is less than 0, it indicates negative kurtosis (lighter tails than a normal distribution).

b) Fisher's Kurtosis:

Fisher's kurtosis measures the kurtosis without subtracting 3, meaning it does not account for the kurtosis of a normal distribution. The formula for Fisher's kurtosis is:

Kurtosis = M4 / (M2^2)

Fisher's kurtosis can be positive or negative, depending on whether the tails of the distribution are heavier or lighter than those of a normal distribution.

Interpreting kurtosis values:

A kurtosis of 0 indicates the same kurtosis as a normal distribution (mesokurtic).

Positive kurtosis values indicate heavy tails (leptokurtic), indicating more outliers or extreme values than a normal distribution.

Negative kurtosis values indicate light tails (platykurtic), indicating fewer outliers or extreme values than a normal distribution.

It is important to note that kurtosis alone does not provide a complete description of a distribution. It should be considered in conjunction with other statistical measures and graphical representations to understand the overall characteristics of the data distribution.

2.15 MEANING AND MEASURES OF MOMENTS

In statistics, moments are mathematical quantities that describe the shape, location, and variability of a probability distribution or a set of data. They provide a way to summarize and analyze the properties of a distribution.

The "n-th moment" of a distribution refers to the expected value of the n-th power of the random variable. The moments are calculated by taking the weighted average of the powers of the variable, where the weights are given by the probability density function (pdf) or the probability mass function (pmf) of the distribution.

The moments of a distribution can be used to derive various statistical measures that provide insights into the characteristics of the data. The first four moments, in particular, are commonly used and have specific interpretations:

a) First Moment: The first moment is the mean or the expected value of the distribution. It represents the center or location of the distribution and provides information about its average value.
b) Second Moment: The second moment is the variance, which measures the spread or variability of the distribution. It quantifies how far the values of the random variable deviate from the mean. The square root of the variance is the standard deviation.
c) Third Moment: The third moment is called the skewness. It describes the asymmetry of the distribution. A positive skewness indicates a longer right tail, while a negative skewness indicates a longer left tail.
d) Fourth Moment: The fourth moment is called the kurtosis. It characterizes the shape of the distribution's tails and peaks. High kurtosis implies heavy tails and sharp peaks, while low kurtosis indicates light tails and flat peaks.

Higher-order moments beyond the fourth can also be calculated, but their interpretations become more complex and specialized. For example, the fifth moment is related to skewness in a more refined manner, and the sixth moment is related to a measure called excess kurtosis.

In summary, moments are statistical descriptors that summarize the properties of a distribution, and the first four moments (mean, variance, skewness, and kurtosis) are commonly used to provide insights into the shape, location, and variability of the data.

2.16 SUM UP

The coefficient of variation (CV) is the ratio of the standard deviation to the mean. The higher the coefficient of variation, the greater the level of dispersion around the mean. The Lorenz curve is a graphical representation of the distribution of income or wealth in a society. The farther the curve moves from the baseline, represented by the straight diagonal line, the higher the level of inequality. Skewness is a measure of asymmetry or distortion of symmetric distribution. It measures the deviation of the given distribution of a random variable from a symmetric distribution, such as a normal distribution. A normal distribution is without any skewness, as it is symmetrical on both sides. Kurtosis is a measure of whether the data are heavy-tailed or light-

tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails or outliers. Data sets with low kurtosis tend to have light tails or lack of outliers. Moments are measures of the shape and variability of a data set. They are used to describe the location and dispersion of the data. Several types of moments can be calculated, each providing different information about the data set.

2.17 QUESTIONS FOR PRACTICE

A. LONG ANSWER QUESTIONS

Q1. Why is it important to consider the distribution of data when interpreting the Coefficient of Variation (CV)? How does the presence of outliers affect the Coefficient of Variation (CV) and its interpretation?

Q2. In what situations would you prefer to use the Coefficient of Variation (CV) over other measures of dispersion, such as the standard deviation? Can you provide an example scenario where the Coefficient of Variation (CV) would be useful in making a comparison or decision?

Q3. What are the limitations of the Lorenz curve as a measure of income inequality? How has the Lorenz curve been used to analyze wealth distribution and social disparities?

Q4. Can you provide examples of real-world applications where the Lorenz curve has been utilized for policy or research purposes? How does skewness relate to other measures of central tendency, such as mean and median?

Q5. Are there any real-world examples where knowledge of skewness is particularly useful? Does skewness have any practical implications in finance, economics, or other fields?

Q6. How does skewness affect the interpretation of a dataset's variability or dispersion? What are some common misconceptions or pitfalls when interpreting skewness? Are there any distributions that are inherently symmetric and have zero skewness?

Q7. What are the characteristics of a mesokurtic distribution? How does kurtosis relate to the tails of a distribution?

Q8. How can kurtosis be interpreted in the context of financial markets? What are some common misconceptions about kurtosis?

Q9. How are higher-order moments, such as skewness and kurtosis, calculated? What insights can be gained from the skewness of a distribution?

Q10. How does kurtosis measure the "peakedness" or "flatness" of a distribution? Are there any limitations or assumptions when using moments to analyze data?

Q11. Can moments be used to compare distributions from different datasets? How can moments be used in statistical modeling and parameter estimation?

B. SHORT ANSWER QUESTIONS

Q1. How is the Coefficient of Variation (CV) calculated?

Q2. What does a high Coefficient of Variation (CV) indicate about a dataset?

Q3. When is the Coefficient of Variation (CV) particularly useful in statistical analysis?

Q4. How does the Coefficient of Variation (CV) allow for the comparison of datasets with different means?

Q5. What are the limitations of using the Coefficient of Variation (CV) as a measure of variability?

Q6. Can the Coefficient of Variation (CV) be used to compare datasets with different units of measurement?

Q7. What is the Lorenz curve and how is it used in economics?

Q8. Can you explain the concept of income inequality and its relationship to the Lorenz curve?

Q9. How is the Lorenz curve constructed and what does it represent?

Q10. What are the key properties of the Lorenz curve?

Q11. How can the Lorenz curve be used to measure and compare income distribution across different countries or regions?

Q12. What does it mean when the Lorenz curve is concave or convex?

Q13. How is the Gini coefficient derived from the Lorenz curve?

Q14. What is skewness and how is it defined?

Q15. How does skewness measure the asymmetry of a distribution?

Q16. What are the possible values of skewness and what do they indicate about the shape of a distribution?

Q17. How can positive skewness be interpreted in terms of a distribution's tail?

Q18. How does negative skewness relate to the skewness of a distribution's tail?

Q19. Can a distribution have zero skewness? If so, what does it imply about the symmetry of the distribution?

Q20. What are some statistical measures or tests used to assess the skewness of a dataset?

Q21. How does sample size impact the assessment of skewness in a dataset?

Q22. What is kurtosis and how is it defined?

Q23. What is the difference between leptokurtic and platykurtic distributions?

Q24. How does kurtosis measure the shape of a distribution?

Q25. What are the implications of positive kurtosis in a distribution?

Q26. Can you provide an example of a distribution with negative kurtosis?

Q27. How is excess kurtosis calculated?

Q28. Can skewness be used as a standalone measure to describe the shape of a distribution?

Q29. What is the definition of a moment in statistics?

Q30. How are moments used to describe the shape of a distribution?

Q31. What is the difference between the mean and the first moment of a distribution?

Q32. Can you explain the concept of variance and its relationship to moments?

C. NUMERALS TO SOLVE

1) Compute the Karl Pearson's coefficient of skewness from the following data:

Daily Expenditure (Rs.): 0-20 20-40 40-60 60-80 80-100

No. of families: 13 25 27 19 16

2) The following figures relate to the size of capital of 285 companies:

Capital (in Rs. Lacs.) 1-5 6-10 11-15 16-20 21-25 26-30 31-35

Total

No. of companies 20 27 28 38 48 54 70

Compute the Bowley's and Kelly's coefficients of skewness and interpret the results.

3) The following measures were computed for a frequency distribution.

Mean = 50, coefficient of Variation = 35% and

Karl Pearson's Coefficient of Skewness = - 0.25.

Compute the Standard Deviation, Mode and Median of the distribution.

4)Calculate the first four moments about the mean for the following

distribution. Also calculate Q, and comment upon the nature of

skewness.

Marks:	0-20	20-40	40-60	60-80	80-100
Frequency:	8	28	35	17	12

5)The first three moments of a distribution about the value 3 of a variable are 2, 10 and 30 respectively. Obtain, p 2, p, and hence, Comment upon the nature of skewness.

6) Compute the first four central moments from the following data. Also find the two beta coefficients.

V due: 5 10 15 20 25 30 35

Frequency: 8 15 20 32 23 17 5

7)The first four moments of distribution are 1, 4, 10, and 46 respectively. Compute the moment coefficients of skewness and kurtosis and comment upon the nature of the distribution.

2.18 MCQ's

1: What does the coefficient of variation measure?

A) The dispersion of data

B) The central tendency of data

C) The correlation between variables

D) The probability of an event

Answer: A) The dispersion of data

2: How is the coefficient of variation calculated?

A) Standard deviation divided by mean

B) Mean divided by standard deviation

C) Range divided by mean

D) Mean multiplied by standard deviation

Answer: A) Standard deviation divided by mean

3: What does a higher coefficient of variation indicate?

A) Higher dispersion of data

B) Lower dispersion of data

C) Higher correlation between variables

D) Lower correlation between variables

Answer: A) Higher dispersion of data

4: Which of the following is true regarding the coefficient of variation?

A) It is expressed as a percentage.

B) It is always a positive value.

C) It can be negative.

D) It measures the mean of the data.

Answer: A) It is expressed as a percentage

5: When comparing two sets of data, which of the following statements is true?

A) The set with a lower coefficient of variation has less dispersion.

B) The set with a higher coefficient of variation has less dispersion.

C) The sets cannot be compared using the coefficient of variation.

D) The sets have the same dispersion if their coefficients of variation are equal.

Answer: A) The set with a lower coefficient of variation has less dispersion

6. The Lorenz curve is a graphical representation that shows the relationship between:

a) Income and wealth

b) Income and consumption

c) Wealth and consumption

d) Population and income distribution

Answer: a) Income and wealth

7. The Lorenz curve is commonly used to measure:

a) Economic growth

b) Income inequality

c) Poverty rates

d) Unemployment levels

Answer: b) Income inequality

8. In a perfectly equal society, the Lorenz curve would be:

a) A straight line at a 45-degree angle

b) A horizontal line

c) A vertical line

d) A concave curve

Answer: a) A straight line at a 45-degree angle

9. The Gini coefficient is derived from the Lorenz curve and represents:

a) Income inequality

b) Poverty rates

c) Economic growth

d) Unemployment levels

Answer: a) Income inequality

10. A society with high-income inequality would have a Lorenz curve that is:

a) Close to the 45-degree line

b) Close to the horizontal axis

c) Close to the vertical axis

d) A steep concave curves

Answer: d) A steep concave curves

11. The area between the Lorenz curve and the 45-degree line represents:

a) Income inequality

b) Economic growth

c) Poverty rates

d) Unemployment levels

Answer: a) Income inequality

12. The Lorenz curve is widely used in which field of study?

a) Sociology

b) Medicine

c) Mathematics

d) Environmental science

Answer: a) Sociology

13. The Lorenz curve was developed by economist Max O. Lorenz in which year?

a) 1939

b) 1945

c) 1950

d) 1971

Answer: a) 1939

14. Skewness is a measure of:

a) The dispersion of data points

b) The symmetry of the distribution

c) The concentration of data around the mean

d) The relationship between two variables

Answer:14. b) The symmetry of the distribution

15. A positively skewed distribution has:

a) A long-left tail

b) A long right tail

c) An equal-length tail on both sides

d) No tails

Answer:15. b) A long right tail

16. Kurtosis measures:

a) The variability of the data

b) The symmetry of the data

c) The peakedness of the distribution

d) The spread of the data

Answer:16. c) The peakedness of the distribution

17. A leptokurtic distribution has:

a) A high peak and heavy tails

b) A low peak and light tails

c) A symmetric shape

d) A flat shape with no tails

Answer:17. a) A high peak and heavy tails

18. Moments in statistics are used to describe:

a) The shape of the distribution

b) The central tendency of the data

c) The spread or dispersion of the data

d) The relationship between two variables

Answer:18. a) The shape of the distribution

19. The first moment of a distribution is also known as:

a) Skewness

b) Kurtosis

c) Mean

d) Standard deviation

Answer:19. c) Mean

20. The second moment of a distribution is also known as:

a) Skewness

b) Kurtosis

c) Variance

d) Median

Answer:20. c) Variance

21. The third moment of a distribution is related to:

a) Skewness

b) Kurtosis

c) Mean

d) Mode

Answer:21. a) Skewness

22. The fourth moment of a distribution is related to:

a) Skewness

b) Kurtosis

c) Median

d) Mode

Answer:22. b) Kurtosis

23. A distribution with a skewness value of 0 indicates:

a) A symmetric distribution

b) A positively skewed distribution

c) A negatively skewed distribution

d) No information about the shape of the distribution

Answer:23. a) A symmetric distribution

24. Which of the following statements about skewness is true?

a) Positive skewness indicates a longer right tail.

b) Negative skewness indicates a longer left tail.

c) Skewness is always zero for symmetric distributions.

d) All of the above.

Answer:24. d) All of the above.

25. Which of the following statements is true?

a) Skewness and kurtosis are measures of central tendency.

b) Skewness and kurtosis are measures of dispersion.

c) Skewness and kurtosis are measures of shape.

d) Skewness and kurtosis are measures of variability.

Answer: 25. c) Skewness and kurtosis are measures of shape.

2.19 SUGGESTED READINGS

- Daniel, Wayne W., Bio-statistics: A Foundation for Analysis in the Health Sciences. John Wiley (2005).
- Das, M. N. &Giri, N. C.: Design and analysis of experiments. John Wiley.
- Dunn, O.J Basic Statistics: A primer for the Biomedical Sciences. (1964, 1977) by John Wiley.
- Goldstein, A Biostatistics-An introductory text (1971). The Macmillion New York.
- Goon, A.M., Gupta M.K. & Das Gupta, Fundamentals of statistics, Vol.-I & II (2005).
- Gupta, S. C. and Kapoor, V.K. (2008): Fundamentals of Applied Statistics, 4th Edition (Reprint), Sultan Chand &Sons
- Hogg, R. V., Mckean, J. and Craig, A. T. Introduction to Mathematical Statistics, Prentice Hall.
- Mood, A.M., Graybill, F.A. & Bose, D.C. Introduction to the Theory of Statistics, Mc Graw-Hill.

COURSE NAME: STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SARM 2: DESCRIPTIVE STATISTICS

SEMESTER I

UNIT 3: CORRELATION: MEANING, PROPERTIES, TYPES AND SCATTER DIAGRAM

STRUCTURE

3.0 Learning Objectives

3.1 Introduction

3.2 Meaning of Correlation

3.3 Uses of Correlation

3.4 Types of Correlation

- **3.4.1 Positive, Negative and Zero Correlation**
- **3.4.2 Simple and Multiple Correlation**
- **3.4.3 Total and Partial Correlation**
- **3.4.4 Linear and Non-Linear Correlation**

3.5 Degrees of Correlation

3.6 Scatter Diagram Method

3.7 Properties of Correlation

3.8 Sum Up

3.9 Questions for Practice

3.10 MCQs

3.11 Suggested Readings

3.0 LEARNING OBJECTIVES

After studying the Unit, students will be able to:

- Define what is Correlation
- Distinguish between different types of correlation
- Understand the benefits of correlation

- Find correlation using the graphic method
- Plot a scatter diagram

3.1 INTRODUCTION

When we study measurement of central tendency, dispersion analysis, skewness analysis etc., we study the nature and features of data in which only one variable is involved. However, In our daily life we come across a number of things in which two or more variables are involved and such variables may be related to each other. As these variables are related to each other, it is important to understand the nature of such a relation and its extent. Identification of such relations helps us in solving a number of problems of daily life. This is not only helpful in our daily life but also helpful in solving many business problems. For example, if a businessman knows the relation between income and demand, Price and Demand, etc., it will help him in the formulation of business plans. Correlation is one such statistical technique that helps us in understanding relation between two or more variables.

3.2 MEANING OF CORRELATION

Correlation is a statistical technique that studies the relationship between two or more variables. It studies how to variables are related to each other. It studies how the change in value of one variable affects the other variable, for example in our daily life we will find the relation between income and expenditure, income and demand, Price and Demand age of husband-and-wife etcetera correlation helps in understanding such relations of different variables two variables are said to be related to each other when a change in the value of one variable so results in to change in the value of other variables.

Therefore, when X and Y are related to each other, then it has four possibilities:

(a) X may be causing Y

(b) Y may be causing X

(c) X and Y both are bidirectionally related, i.e., X is causing Y and Y is causing X

(d) X and Y are related to each other through some third variable

However, correlation has nothing to do with causation. It simply attempts to find the degree of mutual association between them. It is possible that two variables might be found highly correlated, but they are not causing the change in each other. There may be a correlation due to pure chance. For example, we may find a high degree of correlation between the number of trees

in a city and number of drug addicts. However, there is no theoretical base that relates these variables together Such correlation is known as Spurious Correlation or Non-sense Correlation.

According to Croxton and Cowden, "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 W.I. King, "Correlation means that between two series or groups of data, there exists some causal connection."

3.3 USES OF CORRELATIONS

1. It helps us in understanding the extent and direction of the relation between two variables. It shows, whether two variables are positively correlated or negatively correlated. It also shows whether relation between two variables is high or low.
2. Correlation also helps in the prediction of future, for example, if we know relation between monsoon and agricultural produce, we can predict that what will be the level of produce on basis of monsoon prediction. We can also predict price of Agricultural Products depending on level of produce.
3. With the help of correlation, we can find the value of one variable when the value of other variable is known. This can be done by using the statistical technique called regression analysis.
4. Correlation also helps in business and Commerce. A businessman can fix price of its product using the correlation analysis. Correlation also helps him in deciding business policy.
5. Correlation also helps government in deciding its economic policy. With the help of correlation government can study relation of various economic variables, thus government can decide their economic policies accordingly.
6. Correlation is also helpful in various statistical Analysis. Many Statistical techniques use correlation for further analysis.

3.4 TYPES OF CORRELATION

3.4.1 Positive, Negative and No Correlation

a. **Positive correlation**: It is a situation in which two variables move in the same direction. In this case, if the value of one variable increases the value of the other variable also increases. Similarly, if the value of one variable decrease, the value of other variables also decreases. So,

when both the variables either increase or decrease, it is known as a positive correlation. For example, we can find a Positive correlation between Income and Expenditure, Population and Demand for food products, Incomes and Savings, etc. The following data shows positive correlation between two variables:

Height of Persons: X	158	161	164	166	169	172	174
Weight of Person: Y	61	63	64	66	67	69	72

b. Negative or Inverse Correlation: When two variables move in opposite directions from each other, it is known as negative or inverse correlation. In other words, we can say that when the value of one variable increases value of other variable decreases, it is called negative correlation. In our life we find a negative correlation between a number of variables, for example, there is a negative correlation between Price and Demand, the Number of Workers and Time required to complete the work, etc. The following data shows the negative correlation between two variables:

Price of Product: X	1	2	3	4	5
The demand of Product: Y	50	45	40	35	30

c. Zero or No Correlation: When two variables does not show any relation, it is known as zero or no correlation. In other words, we can say that in the case of zero correlation, the change in value of one variable does not affect the value of other variables. In this case two variables are independent from each other. For example, there is zero correlation between the height of the student and the marks obtained by the student.

3.4.2 Simple and Multiple Correlation:

a. Simple Correlation: When we study relation between two variables only, it is known as simple correlation. For example, relation between income and expenditure, Price and Demand, are situations of simple correlation.

b. Multiple Correlation: Multiple correlation is a situation in which more than two variables are involved. Here relation between more than two variables is studied together, for example if we are studying the relation between income of the consumer, price if the product and

demand for the product, it is a situation of multiple correlations.

3.4.3 Total and Partial Correlation:

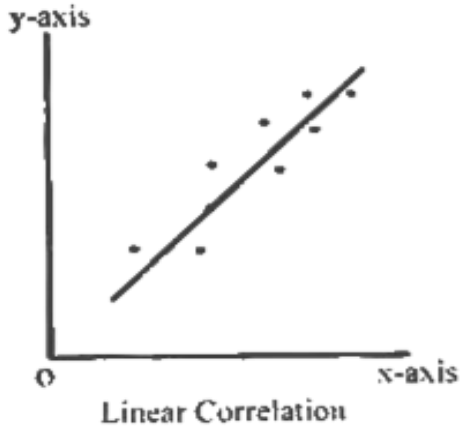
a. Total Correlation: In case we study relation of more than two variables and all the variables are taken together, it is a situation of total correlation. For example, if we are studying the relationship between the income of the consumer, price of the product and demand of the product, taking all the factors together it is called total correlation.

b. Partial Correlation: In case of partial correlation more than two variables are involved, but while studying the correlation we take only two factors into consideration assuming that the value of other factors is constant. For example, while studying the relationship between income of the consumer, price of the product and demand for the product, we take into consideration only relation between price of the product and demand for the product assuming that income of the consumer is constant.

3.4.4 Linear and Non-Linear Correlation:

a. Linear Correlation: When the change in value of one variable results in constant ratio of change in the value of other variable, it is called linear correlation. In such case if we draw the values of two variables on the graph paper, all the points on the graph paper will fall on a straight line. For example, every change in income of consumer by Rs. 1000 results in increase in consumption by 10 kg., which is known as linear correlation. Following data shows example of linear correlation:

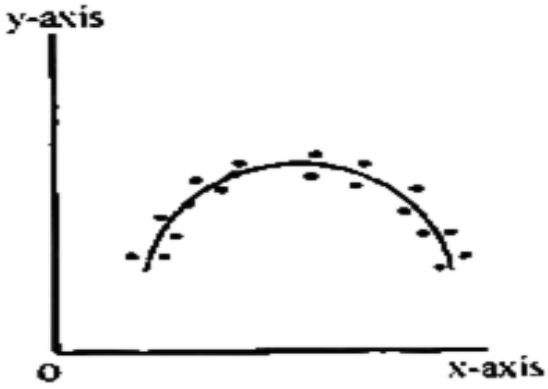
Price of Product: X	1	2	3	4	5
Demand for Product: Y	50	45	40	35	30



Linear Correlation

b. Non - Linear Correlation: When the change in value of one variable does not result in constant ratio of change in the value of other variable, it is called nonlinear correlation. In such case, if we draw the value of two variables on the graph paper all the points will not fall in the straight line on the graph. Following data shows nonlinear correlation between two variables:

Price of Product: X	1	2	3	4	5
Demand of Product: Y	50	40	35	32	30



Non Linear Correlation

3.5 DEGREES OF CORRELATION

Here degrees of correlation shown in the following table:

Degrees of Correlation	Positive	Negative
1. Perfect Degree	+1	-1
2. Very High Degree	+0.9 0and more	-0.9 0and more
3. High Degree	+0.75 to .90	-0.75 to .90
4. Moderate Degree	+0.50 to 0.75	-0.50 to 0.75
5. Low degree	+0.25 to 0.50	-0.25 to 0.50
6. Very Low Degree	+Less than 0.25	-Less than 0.25
7. Zero Degree	0	0

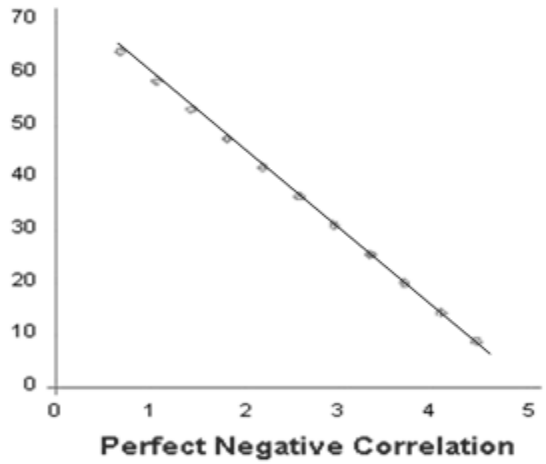
3.6 SCATTER DIAGRAM METHOD

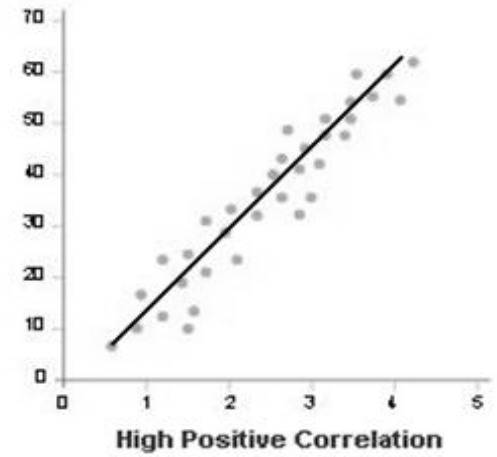
Scatter Diagram is one of the oldest and simple methods of measuring the correlation. This is a graphic method of measuring the correlation. This method uses diagram representation of bivariate data to find out degree and direction of correlation. Under this method, values of the data are plotted on a graph paper by taking one variable on the x-axis and other variable on the y-axis.

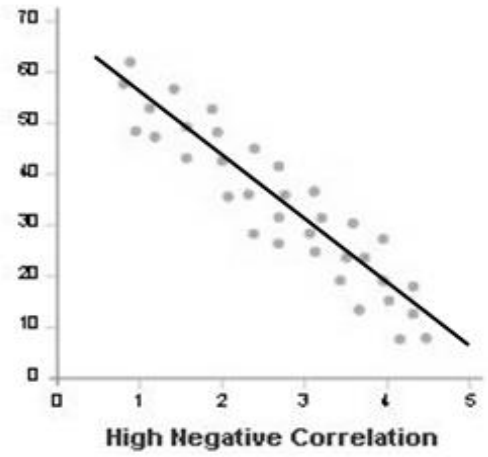
Normally independent variable is shown on x-axis whereas the value of the dependent variable is taken on the y-axis. Once all the values are drawn on the graph paper, we can find out degree of correlation between two variables by looking at direction of dots on the graph. Scatter Diagram shows whether two variables are co-related to each other or not. It also shows the direction of correlation whether positive or negative and the shows extent of correlation whether high or low. The following situations are possible in the scatter diagram.

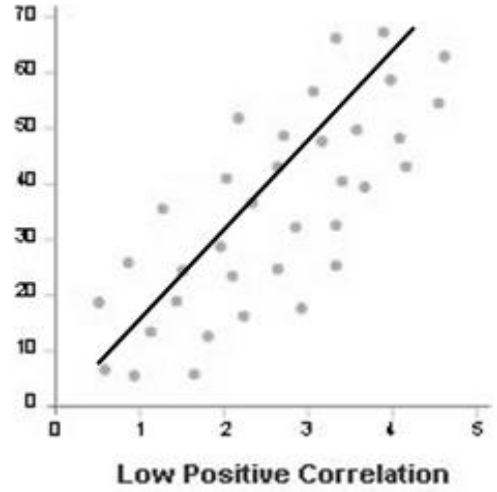
1. **Perfect Positive Correlation**: After we plot two variables on the graph, if the points of graph fall in a straight line that moves from lower left-hand side to the upper corner on the right-hand side, then it is assumed that there is perfect positive correlation between the variables.
2. **Perfect Negative Correlation:** After drawing the variables on the graph, if all the points fall in a straight line but direction of the points is downward from right-hand corner to left-hand side corner, then it is assumed that there is perfect negative correlation between the variable.
3. **High Degree of Positive Correlation**: If we draw two variables on the graph and we find that the points move in upward direction from left-hand corner to the right-hand corner but not in a straight line, rather these are in narrow band, we can assume that there is high degree of positive correlation between the variables.
4. **High Degree of Negative Correlation**: After plotting the dots on a graph, if we find that all the dots move downward from left-hand corner to the right-hand side corner but not in a straight line rather in a narrow band, we can say that there is high degree of negative correlation between the variables.
5. **Low Degree of Positive Correlation**: In case the dots drawn on a graph paper moves upward from left side to right side but the dots are widely scattered, it can be said that there is low degree of positive correlation between the variables.
6. **Low Degree of Negative Correlation**: In case the points drawn on a graph are in downward direction from left side to right side but the points are widely scattered, it is the situation of low degree of negative correlation between the variables.
7. **Zero or No Correlation**: Sometime find that the dots drawn on a graph paper do not move in any direction and are widely scattered in the graph paper, we can assume that there is no correlation between the two variables.

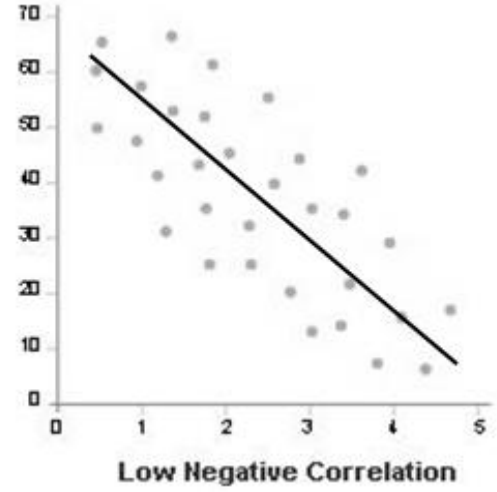

70
60
50
40
30
20
10
0
0 1 2 3 4 5
Perfect Positive Correlation

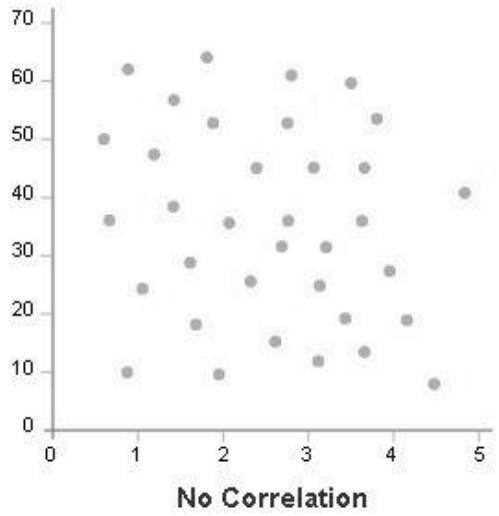

70
60
50
40
30
20
10
0
0 1 2 3 4 5
Perfect Negative Correlation


70
60
50
40
30
20
10
0
0 1 2 3 4 5
High Positive Correlation


70
60
50
40
30
20
10
0
0 1 2 3 4 5
High Negative Correlation


70
60
50
40
30
20
10
0
0 1 2 3 4 5
Low Positive Correlation


70
60
50
40
30
20
10
0
0 1 2 3 4 5
Low Negative Correlation



NOTE:

- Correlation coefficient shows the linear relationship between X and Y. Thus, even if there is a strong non-linear relationship between X and Y, the correlation coefficient may be low.
- Correlation coefficient is independent of scale and origin. If we subtract some constant from one (or both) of the variables, the correlation coefficient will remain unchanged. Similarly, if we divide one (or both) of the variables by some constant, the correlation coefficient will not change.
- Correlation coefficient varies between -1 and +1. This means r cannot be smaller than -1 and cannot be greater than +1.

The existence of a linear relationship between two variables is not to be interpreted to mean a cause-effect relationship between the two.

3.7 PROPERTIES OF CORRELATION

1. Range: The coefficient of Correlation always lies between -1 to +1.
2. Degree Of Measurement: Correlation Coefficient is independent of units of measurement.
3. Direction: The sign of Correlation is positive (+ve) if the values of variables move in the same direction, if -ve then the opposite direction.
4. Symmetry: Correlation Coefficient deals with the property of symmetry. It means $r_{xy}=r_{yx}$,
5. Geometric Mean: The coefficient of Correlation is also the geometric mean of two regression coefficients Rxy= bxy. byx
6. If x and y are independent then $r_{xy}=0$

7. Change of Origin: The correlation coefficient is independent of change of origin
8. Change of Scale: The correlation coefficient is independent of change in Scale
9. Coefficient of determination: The square of the correlation coefficient (r_{xy}) is known as the coefficient of determination

3.8 LET US SUM UP

- Correlation shows the relation between two or more variables.
- The value of the coefficient of correlation always lies between -1 and +1.
- Correlation may be positive or negative.
- Correlation may be linear or nonlinear.

3.9 QUESTIONS FOR PRACTICE

1. What is Correlation?
2. What are uses of measuring correlation?
3. Explain the properties of Correlation Coefficient.
4. Give different types of correlation.
5. What are the various degrees of correlation coefficient?
6. What do you mean by scatter diagram?

3.10 MCQ

Q1: The relation of three or more variables is called:

(a) simple correlation

(b) partial correlation

(c) multiple correlation

(d) none of these

Answer: C

Q 2: The coefficient of correlation lies between +0.25 and + 0.75, it is called:

(a) perfect degree of correlation

(b) high degree of correlation

(c) moderate degree of correlation

(d) low degree of correlation

Answer: C

Q 3: The coefficient of correlation lies always between:

(a) 0 and +1

(b)-1 and 0

(c)-1 and+1

(d) none of these

Answer: C

Q 4: When two variables change in a constant proportion, it is called:

(b) non-linear correlation

(c) partial correlation

(d) none of these

Answer: A

Q 5: A scatter diagram:

(a) Is a statistical test

(b) Must be linear

(c) Must be curvilinear

(d) is a graph of x and y values

Answer: D

Q 6: Maximum value of correlation coefficient is:

(a) 0

(b) +1

(c)-1

(d) None of these

Answer: B

Q 7: The correlation coefficient will be -1 if the slope of the straight line in a scatter diagram is:

(a) Positive

(b) Negative

(c) Zero

(d) None of these

Answer: B

Q 8: In a 'negative' relationship

(a) As x increases, y increases

(b) As x decreases, y decreases

(c) As x increases, y decreases

(d) Both (a) and (b)

Answer: C

Q 9: Scatter diagram helps us to:

(a) Find the nature of correlation between two variables

(b) Obtain the mathematical relationship between two variables

(c) Compute the extent of correlation between two variables

(d) Both (a) and (c)

Answer: A

Q 10: The lowest strength of association is reflected by which of the following correlation coefficients?

(a) 0.95

(b)-0.60

(c)-0.35

(d) 0.29

Answer: D

Q 11: The relation between price and demand is:

(a) positive

(b) negative

(c) one to one

(d) no relationship

Answer: B

Q 12: When r = 1, all the points in a scatter diagram would lie:

(a) On a straight line directed from lower left to upper right

(b) On a straight line directed from upper left to lower right

(c) On a straight line

(d) Both (a) and (b)

Answer: A

Q 13: The highest strength of association is reflected by which of the following correlation coefficients?

(a)-1.0

(b)-0.95

(c) 0.1

(d) 0.85

Answer: A

Q 14: When two variables change in the same direction, then such a correlation

(a) negative

(b) positive

(c) no correlation

(d) all of these

Answer: B

Q 15: Question: Correlation coefficient is ----- of the units of measurement.

(a) Independent

(b) Dependent

(c) Both (a) and (b)

(d) None of these

Answer: A

Q 16: The correlation between sale of cold drinks and day temperature is:

(a) Positive

(b) Negative

(c) Zero

(d) None of these

Answer: A

Q 17: If all the plotted points in a scatter diagram lie on a single line, then the correlation is:

(a) Perfect positive

(b) Perfect negative

(c) Both (a) and (b)

(d) Either (a) or (b)

Answer D

Q 18: Correlation coefficient is dependent on the choice of both origin and the scale of observations,

(a) True

(b) False

(c) Both (a) and (b)

(d) none of these

Answer: B

Q19: limits of the correlation coefficient?

(a) No limit

(b) 0 and 1, including the limits

(c) - 1 and 1

(d)-1 and 2

Answer: C

Q 20. The correlation coefficient is used to determine:

(a) A specific value of the y-variable given a specific value of the x-variable

(b) A specific value of the x-variable given a specific value of the y-variable

(c) The strength of the relationship between the x and y variables

(d) None of these

Answer: C

3.11 SUGGESTED READINGS

- J. K. Sharma, *Business Statistics,* Pearson Education.
- S.C. Gupta, *Fundamentals of Statistics,* Himalaya Publishing House.
- S.P. Gupta and Archana Gupta, *Elementary Statistics,* Sultan Chand and Sons, New Delhi.
- Richard Levin and David S. Rubin, *Statistics for Management,* Prentice Hall of India, New Delhi.
- M.R. Spiegel, *Theory and Problems of Statistics,* Schaum's Outlines Series, McGraw HillPublishing Co.

CERTIFICATE/DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SARM 2: DESCRIPTIVE STATISTICS

SEMESTER I

UNIT 4: KARL PEARSON'S CORRELATION CO-EFFICIENT & SPEARMAN'S RANK, CORRELATION COEFFICIENT

STRUCTURE

4.0 Learning Objectives

4.1 Meaning Karl Pearsons's Coefficient of Correlation

4.2 Features of Karl Pearsons's Coefficient of Correlation

4.3 Different methods to calculate Coefficient of Correlation

- **4.3.1 Direct Method of Karl Pearsons's Coefficient of Correlation**
- **4.3.2 Actual Mean Method of Karl Pearsons's Coefficient of Correlation**
- **4.3.3 Assumed Mean Method of Karl Pearsons's Coefficient of Correlation**
- **4.3.4 Step Deviation Method of Karl Pearsons's Coefficient of Correlation**
- **4.3.5 Karl Pearsons's Coefficient of Correlation from Standard Deviation**
- **4.3.6 Assumptions of Karl Pearsons's Coefficient of Correlation**
- **4.3.7 Limitations of Karl Pearsons's Coefficient of Correlation**

4.4 Spearman's Rank Correlation

- **4.4.1 Features of Spearman's Rank Correlation**
- **4.4.2 Spearman's Rank Correlation when Ranks are given**
- **4.4.3 Spearman's Rank Correlation when Ranks are not given**
- **4.4.4 Spearman's Rank Correlation when there is repetition in Ranks**
- **4.4.5 Merits of Spearman's Rank Correlation**
- **4.4.6 Limitations of Spearman's Rank Correlation**

4.5 Let us Sum Up

4.6 Key Terms

4.7 Questions for Practice

4.8 Suggested Readings

4.0 LEARNING OBJECTIVES

After studying the Unit, students will be able to:

- Understand the meaning of correlation
- Features of correlation
- Calculate correlation by Karl Pearson Method
- Measure correlation using Rank correlation method
- Merits and demerits of the methods

4.1 MEANING: KARL PEARSONS'S COEFFICIENT OF CORRELATION

Karl Peason's Coefficient of Correlation is the most important method of measuring the correlation. Karl Peason's Coefficient of correlation is also denoted as 'Product Moment Correlation' also. The coefficient of correlation given by Karl Pearson is denoted as a symbol 'r'. It is the relative measure of finding the correlation. According to Karl Pearson we can determine correlation by dividing the product of deviations taken from mean of the data.

4.2 FEATURES OF KARL PEARSON'S COEFFICIENT OF CORRELATION

1. Karl Pearson's method is an algebraic method of finding correlation.
2. The coefficient of correlation may be positive or negative.
3. This method is based on the arithmetic mean of the data and the standard deviation of the data.
4. The value of the coefficient of correlation always lies between -1 and + 1.

   -1 refers to 100 % negative correlation, plus one refers to 100% positive correlation, and zero refers to no correlation between the items.
5. This method is based on all the items of the Data.

4.3 DIFFERENT METHODS TO CALCULATE COEFFICIENT OF CORRELATION

4.3.1 Direct method of calculating Correlation

Correlation can be calculated using the direct method without taking any mean. The following are the steps:

1. Take two series X and Y.

2. Find the sum of these two series denoted as $\sum X$ and $\sum Y$.
3. Take the square of all the values of the series X and series Y.
4. Find the sum of the square so calculated denoted by $\sum X^2$ and $\sum Y^2$.
5. Multiply the corresponding values of series X and Y and find the product.
6. Sum up the product so calculated denoted by $\sum$ X Y.
7. Apply the following formula for calculating the correlation.

Coefficient of Correlation, $\mathbf{r} = \frac{\mathbf{N\sum XY - (\sum X)(\sum Y)}}{\sqrt{\mathbf{N\sum X^2 - (\sum X)^2}}\sqrt{\mathbf{N\sum Y^2 - (\sum Y)^2}}}$

Example 1: Find coefficient of correlation

X	**Y**
2	**4**
3	**5**
1	**3**
5	**4**
6	**6**
4	**2**

Solution:

X	Y	X^2	Y^2	XY
2	4	4	16	8
3	5	9	25	15
1	3	1	9	3
5	4	25	16	20
6	6	36	36	36
4	2	16	4	8
$\sum X = 21$	$\sum Y = 24$	$\sum X^2 = 91$	$\sum Y^2 = 106$	$\sum XY = 90$

$N = 6$

Coefficient of Correlation, $r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$

$= \frac{6\times 90 - 21\times 24}{\sqrt{6\times 91 - (21)^2}\sqrt{6\times 106 - (24)^2}}$

$= \frac{540 - 504}{\sqrt{546 - 441}\sqrt{636 - 576}}$

$= \frac{36}{\sqrt{105}\sqrt{60}}$ $= \frac{36}{10.246\times 7.7459}$

$$= \frac{36}{79.31} = 0.4539$$

$\Rightarrow$ $r = 0.4539$

4.3.2 Actual Mean method of calculating Correlation

Under this Correlation is calculated by taking the deviations from actual mean of the data. The following are the steps:

1. Take two series X and Y.
2. Find the mean of both the series X and Y, denoted by $\overline{\mathbf{X}}$ and $\overline{\mathbf{Y}}$.
3. Take deviations of series X from it mean and it is denoted by 'x'.
4. Take deviations of series Y from it mean and it is denoted by 'y'.
5. Take square of deviation of series X denoted by x^2.
6. Sum up square of deviations of series X denoted by $\sum x^2$.
7. Take square of deviation of series Y denoted by y^2.
8. Sum up square of deviations of series Y denoted by $\sum y^2$.
9. Find the product of x and y and it is denoted by xy.
10. Find the sum of 'xy 'it is denoted by $\sum xy$
11. Apply the following formula for calculating the correlation.

$$\mathbf{r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}}$$

Example 2. Calculate Karl Pearson's coefficient of correlation

X	**50**	**50**	**55**	**60**	**65**	**65**	**65**	**60**	**60**	**50**
Y	**11**	**13**	**14**	**16**	**16**	**15**	**15**	**14**	**13**	**13**

Solution: When deviations are taken from actual arithmetic mean, 'r' is given by

$$r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where $x = X - \overline{X}$ = Deviation from A. M. of X series

$y = Y - \overline{Y}$ = Deviation from A. M. of Y series

X	Y	$x = (X - \overline{X})$	x^2	$y = (Y - \overline{Y})$	y^2	xy
50	11	-8	64	-3	9	24
50	13	-8	64	-1	1	8
55	14	-3	9	0	0	0
60	16	2	4	2	4	4

65	16	7	49	2	4	14
65	15	7	49	1	1	7
65	15	7	49	1	1	7
60	14	2	4	0	0	0
60	13	2	4	−1	1	−2
50	13	−8	64	−1	1	8
$\sum X = 580$	$\sum Y = 140$		$\sum x^2 = 360$		$\sum y^2 = 22$	$\sum xy = 70$

Here, $N = 10$

A. M. of X series, $\overline{X} = \frac{\sum X}{N} = \frac{580}{10} = 58$

A. M. of Y series, $\overline{Y} = \frac{\sum Y}{N} = \frac{140}{10} = 14$

Coefficient of Correlation, $r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}} = \frac{70}{\sqrt{360 \times 22}} = \frac{70}{\sqrt{7920}} = 0.7866$

$\Rightarrow \quad r = 0.7866$

4.3.3 Assumed Mean method of calculating Correlation

Under this Correlation is calculated by taking the deviations from assumed mean of the data. Following are the steps:

1. Take two series X and Y.
2. Take any value as assumed mean for series X.
3. Take deviations of series X from its assumed mean and it is denoted by ‘dx’.
4. Find sum of deviations denoted by ∑dx.
5. Take square of deviation of series X denoted by dx^2
6. Sum up square of deviations of series X denoted by $\sum dx^2$.
7. Take any value as assumed mean for series Y.
8. Take deviations of series Y from its assumed mean and it is denoted by ‘dy’.
9. Find sum of deviations of series Y denoted by ∑dy.
10. Take square of deviation of series Y denoted by dy^2
11. Sum up square of deviations of series Y denoted by $\sum dy^2$.
12. Find the product of dx and dy and it is denoted by dxdy.
13. Find the sum of ‘dxdy’ it is denoted by ∑ dxdy
14. Apply the following formula for calculating the correlation.

$$\mathbf{r = \frac{N\sum dxdy - (\sum dx)(\sum dy)}{\sqrt{N\sum dx^2 - (\sum dx)^2}\sqrt{N\sum dy^2 - (\sum dy)^2}}}$$

Example 3. Compute coefficient of correlation from the following figures

City	A	B	C	D	E	F	G
Population (in '000)	**78**	**25**	**16**	**14**	**38**	**61**	**30**
Accident Rate (Per million)	**80**	**62**	**53**	**60**	**62**	**69**	**67**

Solution: Here, $N = 7$

Coefficient of Correlation, r is given by

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

Where dx = Deviations of terms of X series from assumed mean $A_X = X - A_X$

dy = Deviations of terms of Y series from assumed mean $A_Y = Y - A_Y$

X	Y	$dx = X - A_X$ $A_X = 38$	$dy = Y - A_Y$ $A_Y = 67$	dx^2	dy^2	$dxdy$
70	80	32	13	1024	169	416
25	62	-13	-5	169	25	65
16	53	-22	-14	482	196	308
14	60	-24	-7	576	49	168
38	62	0	-5	0	25	0
61	69	23	2	529	4	46
30	67	-8	0	64	0	0
		$\sum dx = -12$	$\sum dy = -16$	$\sum dx^2 = 2846$	$\sum dy^2 = 468$	$\sum dxdy = 1003$

Here, $N = 7$

$\therefore$ Coefficient of Correlation, $r = \frac{7\times1003 - (-12)(-16)}{\sqrt{7\times2846 - (-12)^2}\sqrt{7\times468 - (-16)^2}}$

$$= \frac{7021 - 192}{\sqrt{19{,}922 - 144}\sqrt{3276 - 256}}$$

$$= \frac{6829}{\sqrt{19{,}778}\sqrt{3020}} = 0.8837$$

$$r = 0.8837$$

4.3.4 Step Deviation method of calculating Correlation

Under this method assumed mean is taken but the difference is that after taking the deviation, these are divided by some common factor to get the step deviations. Following are the steps:

1. Take two series X and Y.
2. Take any value as assumed mean for series X.
3. Take deviations of series X from its assumed mean and it is denoted by 'dx'.
4. Divide the value of 'dx' so obtained by some common factor to get $\mathbf{dx'}$
5. Find sum of deviations denoted by $\sum \mathbf{dx'}$.
6. Take square of deviation of series X denoted by $\mathbf{dx'}^{2}$
7. Sum up square of deviations of series X denoted by $\sum \mathbf{dx'}^{2}$.
8. Take any value as assumed mean for series Y.
9. Take deviations of series Y from its assumed mean and it is denoted by 'dy'.
10. Divide the value of 'dy' so obtained by some common factor to get $\mathbf{dy'}$
11. Find sum of deviations of series Y denoted by $\sum \mathbf{dy'}$.
12. Take square of deviation of series Y denoted by $\mathbf{dy'}^{2}$
13. Sum up square of deviations of series Y denoted by $\sum \mathbf{dy'}^{2}$.
14. Find the product $\mathbf{dx'}$ of and $\mathbf{dy'}$ and it is denoted by $\mathbf{dx'\,dy'}$.
15. Find the sum of 'dxdy' it is denoted by $\sum \mathbf{dx'\,dy'}$
16. Apply the following formula for calculating the correlation.

$$\textbf{Coefficient of Correlation, } \mathbf{r} = \frac{\mathbf{N}\sum \mathbf{dx'dy'} - (\sum \mathbf{dx'})(\sum \mathbf{dy'})}{\sqrt{\mathbf{N}\sum \mathbf{dx'^2} - (\sum \mathbf{dx'})^2}\sqrt{\mathbf{N}\sum \mathbf{dy'^2} - (\sum \mathbf{dy'})^2}}$$

Example 4. Find the coefficient of correlation by Karl Pearson's method

Price (Rs.)	**5**	**10**	**15**	**20**	**25**
Demand (kg)	**40**	**35**	**30**	**25**	**20**

Solution:

X	Y	$dx = X - A$ $A = 15$	$dx' = \frac{dx}{C_1}$ $C_1 = 5$	$dy = Y - B$ $B = 30$	$dy' = \frac{dy}{C_1}$ $C_2 = 5$	dx'^2	dy'^2	$dx'dy'$
5	40	-10	-2	10	2	4	4	-4
10	35	-5	-1	5	1	1	1	-1
15	30	0	0	0	0	0	0	0
20	25	5	1	-5	-1	1	1	-1
25	20	10	2	-10	-2	4	4	-4

			$\sum dx' = 0$		$\sum dy' = 0$	$\sum dx'^2 = 10$	$\sum dy'^2 = 10$	$\sum dx'dy' = -10$

Here, $N = 5$

Coefficient of Correlation, $r = \frac{N\sum dx'dy' - (\sum dx')(\sum dy')}{\sqrt{N\sum dx'^2 - (\sum dx')^2}\sqrt{N\sum dy'^2 - (\sum dy')^2}}$

$$= \frac{5\times(-10) - 0\times 0}{\sqrt{5\times 10 - 0^2}\sqrt{5\times 10 - 0^2}}$$

$$= \frac{-50}{\sqrt{50}\times\sqrt{50}} = -1$$

$\Rightarrow \quad r = -1$

4.3.5 Calculating Correlation with help of Standard Deviations

Under this method assumed mean is taken but the difference is that after taking the deviation, these are divided by some common factor to get the step deviations. Following are the steps:

1. Take two series X and Y.
2. Find the mean of both the series X and Y, denoted by $\bar{X}$ and $\bar{Y}$.
3. Take deviations of series X from it mean and it is denoted by 'x'.
4. Take deviations of series Y from it mean and it is denoted by 'y'.
5. Find the product of x and y and it is denoted by xy.
6. Find the sum of 'xy 'it is denoted by $\sum xy$
7. Calculate the standard deviation of both series X and Y.
8. Apply the following formula for calculating the correlation.

$$\mathbf{r = \frac{\sum xy}{N\sigma_X\sigma_Y}}$$

Example 5. Given

No. of pairs of observations $= 10$

$\sum xy = 625$

X Series Standard Deviation $= 9$

Y Series Standard Deviation $= 8$

Find 'r'.

Solution: We are given that

$N = 10, \quad \sigma_X = 9 \quad \sigma_Y = 8 \quad$ and $\quad \sum xy = 625$

Now $r = \frac{\sum xy}{N\sigma_X\sigma_Y}$

$$= \frac{625}{10\times 9\times 8} \quad = \frac{625}{720} \quad = 0.868$$

$\Rightarrow \quad r = +.868$

Example 6. Given

No. of pairs of observations $= 10$

X Series Arithmetic Mean $= 75$

Y Series Arithmetic Mean $= 125$

X Series Assumed Mean $= 69$

Y Series Assumed Mean $= 110$

X Series Standard Deviation $= 13.07$

Y Series Standard Deviation $= 15.85$

Summation of products of corresponding deviation of X and Y series $= 2176$

Find 'r'.

Solution: We are given that

$N = 10, \quad \overline{X} = 75, \quad A_X = 69, \quad \sigma_X = 13.07$

$\overline{Y} = 125, \quad A_Y = 110, \quad \sigma_Y = 15.85$

and $\sum xy = 2176$

Now $r = \frac{\sum xy - N(\overline{X} - A_X)(\overline{Y} - A_Y)}{N\sigma_X\sigma_Y}$

$= \frac{2176 - 10(75-69)(125-110)}{10 \times 13.07 \times 15.85} = \frac{2176 - 900}{2071.595}$

$= 0.6159 \approx 0.616$

$\Rightarrow \quad r = +0.616$

Example 7. A computer while calculating the coefficient of correlation between the variables X and Y obtained the values as

$$N = 6, \quad \sum X = 50, \quad \sum X^2 = 448$$

$$\sum Y = 106, \quad \sum Y^2 = 1896, \quad \sum XY = 879$$

But later on, it was found that the computer had copied down two pairs of observations as

X	Y
5	15
10	18

While the correct values were

X	Y
6	18
10	19

Find the correct value of correlation coefficient.

Solution**:** Incorrect value of $\sum X = 50$

$\therefore$ Correct value of $\sum X = 50 - 5 - 10 + 6 + 10$

$= 51$

Incorrect value of $\sum Y = 106$

$\therefore$ Correct value of $\sum Y = 106 - 15 - 18 + 18 + 19$

$= 110$

Incorrect value of $\sum X^2 = 448$

$\therefore$ Correct value of $\sum X^2 = 448 - 5^2 - (10)^2 + 6^2 + (10)^2$

$= 459$

Incorrect value of $\sum Y^2 = 1896$

$\therefore$ Correct value of $\sum Y^2 = 1896 - 15^2 - (18)^2 + (18)^2 + 19^2$

$= 2032$

Incorrect value of $\sum XY = 879$

$\therefore$ Correct value of $\sum XY = 879 - (5 \times 15) - (10 \times 18) + (6 \times 18) + (10 \times 19)$

$= 952$

Thus, the corrected value of coefficient of correlation

$$= \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

$$= \frac{6\times 952 - 51\times 110}{\sqrt{6\times 459 - (51)^2}\sqrt{6\times 2032 - (110)^2}}$$

$$= \frac{5712 - 5610}{\sqrt{2754 - 2601}\sqrt{12{,}192 - 12{,}100}}$$

$$= \frac{102}{\sqrt{153}\sqrt{92}} \qquad = \frac{102}{12.369\times 9.59}$$

$$= \frac{102}{118.618} = 0.8599$$

$\Rightarrow$ $r = +0.8599$

Example 8. Find Correlation between daily wage and food expenditure.

Daily Wage

Food Expenditure	100-150	150-200	200-250	250-300	300-350
0 – 10	5	4	5	2	4
10 – 20	2	7	3	7	1
20 – 30	-	6	-	4	5
30 – 40	8	-	4	-	8
40 – 50	-	7	3	5	10

Solution: Assumed Mean of series X = 225

Assumed Mean of series Y = 25

Class interval of series X = 50

Class interval of series Y = 10

Value of dx is calculated by applying the formula = $\frac{m-A}{c}$

Value of dy is calculated by applying the formula = $\frac{m-A}{c}$

Calculation of Karl Pearson's coefficient of correlation

X / **Y**	**Mid Point**	**100 - 150**	**150 – 200**	**200 – 250**	**250 – 300**	**300 – 350**	**f**	**dy**	**fdy**	**fdy^2**	**$fdxdy$**
0 – 10	**5**	**20** **5**	**8** **4**	**0** **5**	**-4** **2**	**-16** **4**	20	-2	-40	80	8
10 – 20	**15**	**4** **2**	**7** **7**	**0** **3**	**-7** **7**	**-2** **1**	20	-1	-20	20	2
20 – 30	**25**	**-** **-**	**0** **6**	**-** **-**	**0** **4**	**0** **5**	15	0	0	0	0
30 – 40	**35**	**-16** **8**	**-** **-**	**0** **4**	**-** **-**	**16** **8**	20	1	20	20	0
40 – 50	**45**	**-** **-**	**-14** **7**	**0** **3**	**10** **5**	**40** **10**	25	2	50	100	36
F		15	24	15	18	28	**100 = N**		**10**	**220**	**46**
Dx		-2	-1	0	1	2			**∑fdy**	**$\sum fdy^2$**	**∑fdxdy**
Fd_x		-30	-24	0	18	56	20	**$\sum fd_x$**			
fd_x^2		60	24	0	18	112	214	**$\sum fd_x^2$**			
Fd_xdy		8	1	0	-1	38	46	**∑fdxdy**			

Coefficient of Correlation, r =

$$\frac{N\sum fdx'dy' - (\sum fdx')(\sum fdy')}{\sqrt{N\sum fdx'^2 - (\sum fdx')^2}\sqrt{N\sum fdy'^2 - (\sum fdy')^2}}$$

$$= \frac{100\times46 - 20\times10}{\sqrt{100\times214 - (20)^2}\sqrt{100\times(220) - (10)^2}}$$

$$= \frac{4600-200}{\sqrt{21400-400}\ \sqrt{22000-100}}$$

$$= \frac{4400}{\sqrt{21000}\sqrt{21900}} \qquad = \frac{4400}{21445.28} = .2052$$

4.3.6 Assumptions of Karl Pearson's Coefficient of Correlation

1. There exists linear relation between two variables.
2. Relation between variables is not mere chance rather there is cause and effect relation between the variables.
3. Data is taken as a normal series.
4. There is no error in measurement of the data.

4.3.7 Limitations of Karl Pearson's Coefficient of Correlation

1. It is comparatively difficult to calculate.
2. It is time consuming method.
3. It is based on unrealistic assumptions.
4. It is affected by extreme values.
5. It cannot be applied on qualitative data.

TEST YOUR UNDERSTANDING (A)

1. From the following data of prices of product X and Y draw scatter diagram.

	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Price of X	60	65	65	70	75	75	80	85	80	100
Price of Y	120	125	120	110	105	100	100	90	80	60

2. Calculate Karl Pearson's coefficient of correlation

X	21	22	23	24	25	26	27	28	29	30
Y	46	42	38	34	30	26	22	18	14	10

3. Calculate Karl Pearson's coefficient between X and Y

X	42	44	58	55	89	98	66
Y	56	49	53	58	65	76	58

4. Find correlation between marks of subject A Subject B

Subject A	24	26	32	33	35	30
Subject B	15	20	22	24	27	24

5. Find correlation between Height of Mother and Daughter

Height of Mother (Inches)	54	56	56	58	62	64	64	66	70	70
Height of Daughter (Inches)	46	50	52	50	52	54	56	58	60	62

6. What is the Karl Pearson's coefficient of correlation if $\sum xy = 40$, $n = 100$, $\sum x^2 = 80$ and $\sum y^2 = 20$.

7. Calculate the number of items for which r = 0.8, $\sum xy = 200$. Standard deviation of y = 5 and $\sum x^2 = 100$ where x and y denote the deviations of items from actual means.

8. Following values were obtained during calculation of correlation:

N = 30; $\sum X = 120$ $\sum X^2 = 600$ $\sum Y = 90$ $\sum Y^2 = 250$ $\sum XY = 335$

Later found that two pairs were taken wrong which are as follows:

pairs of observations as:	(X, Y):	(8, 10)	(12, 7)
While the correct values were:	(X, Y):	(8, 12)	(10, 8)

Find correct correlation.

9. From the data given below calculate coefficient of correlation.

	X series	Y series
Number of items	8	8
Mean	68	69
Sum of squares of deviation from mean	36	44
Sum, of product of deviations x and y from means	24	24

10. Find the correlation between age and playing habits from the following data :

Age	15	16	17	18	19	20
No of students	20	270	340	360	400	300
Regular players	150	162	170	180	180	120

11. From the table given below find the correlation coefficient between the ages of husbands and wives

Age of Wives Y	Ages of Husbands X					
	$20-30$	$30-40$	$40-50$	$50-60$	$60-70$	Total
$15-25$	5	9	3	–	–	17
$25-35$	–	10	25	2	–	37
$35-45$	–	1	12	2	–	15
$45-55$	–	–	4	16	5	25
$55-65$	–	–	–	4	2	6
Total	5	20	44	24	7	$100/100$

Answers

2) -1	4) .92	6) 1,	8) -.4311	10) -.94
3) .9042	5) .95	7) 25	9) .603	11) .795

4.4 SPEARMAN'S RANK CORRELATION

Karl Peason's Coefficient of Correlation is very useful if data is quantitative, but in case of qualitative data it is a failure. Spearman's Rank correlation is a method that can calculate correlation both from quantitative and qualitative data if the data is ranked like in singing contest we rank the participants as one number, two number or three number etc. This method was given by Charles Edward Spearman in 1904. In this method we give Rank to the data and with help of such ranks, correlation is calculated.

4.4.1 Features of Spearman's Rank correlation

1. The coefficient of correlation may be positive or negative.
2. The value of coefficient of correlation always lies between -1 and + 1. -1 refers to 100% negative correlation, plus one refers to 100% positive correlation, and zero refers to no correlation between the items.
3. This method is based ranks of the data.
4. Sum of difference between ranks in this method is always zero i.e., $\sum D = 0$.
5. There is no assumption of normal distribution in this method.
6. In case all the ranks of the two series are same the value of $\sum D^2 = 0$, it shows that there is perfect positive correlation between the data.

4.4.2 Spearman's Rank Correlation when ranks are given

1. Calculate the difference between ranks of both the series denoted by $\sum D$.
2. Take square of deviations and calculate the value of D^2.
3. Calculate sum of square of deviations denoted by $\sum D^2$.
4. Apply following formula.

Example 9. Following are given the ranks of 8 pairs. Find 'r'

Rank X	6	4	8	2	7	5	3	1
Rank Y	4	8	7	3	6	5	1	2

Solution:

Rank X	Rank Y	Difference of Ranks D	D^2
6	4	$+2$	4
4	8	-4	16
8	7	-1	1
2	3	-1	1

7	6	+1	1
5	5	0	0
3	1	+2	4
1	2	−1	1
N = 8			$\sum D^2 = 28$

Coefficient of Rank Correlation, $r = 1 - \frac{6\sum D^2}{N(N^2-1)}$

$= 1 - \frac{6\times 28}{8(8^2-1)}$

$= 1 - \frac{168}{8(64-1)} \quad = 1 - \frac{168}{8(63)}$

$= 1 - \frac{168}{504} = 1 - 0.33 = 0.67$

⇒ Rank Correlation Coefficient $= 0.67$

Example 10. In a beauty contest, three judges gave the following ranks to 10 contestants. Find out which pair of judges agree or disagree the most.

Judge 1	**5**	**1**	**6**	**3**	**8**	**7**	**10**	**9**	**2**	**4**
Judge 2	**9**	**7**	**10**	**5**	**8**	**4**	**3**	**6**	**1**	**2**
Judge 3	**6**	**4**	**7**	**10**	**5**	**3**	**1**	**9**	**2**	**8**

Solution:

Ranks by			$D_1 = R_1 - R_2$	D_1^2	$D_2 = R_2 - R_3$	D_2^2	$D_3 = R_1 - R_3$	D_3^2
Judge 1 R_1	Judge 2 R_2	Judge 3 R_3						
5	9	6	−4	16	3	9	−1	1
1	7	4	−6	36	3	9	−3	9
6	10	7	−4	16	3	9	−1	1
3	5	10	−2	4	−5	25	−7	49
8	8	5	0	0	3	9	3	9
7	4	3	+3	9	1	1	4	16
10	3	1	+7	49	2	4	9	81
9	6	9	+3	9	−3	9	0	0
2	1	2	+1	1	−1	1	0	0
4	2	8	+2	4	−6	36	−4	16
				$\sum D_1^2 = 144$		$\sum D_2^2 = 112$		$\sum D_3^2 = 182$

Now $r_{12} = 1 - \frac{6\sum D_1^2}{N(N^2-1)}$

$$= 1 - \frac{6\times144}{10(10^2-1)}$$

$$= 1 - \frac{864}{10(100-1)} \quad = 1 - \frac{864}{10(99)}$$

$$= 1 - \frac{864}{990} = 1 - 0.873 = 0.127$$

$\therefore$ $r_{12} = +0.127 \Rightarrow$ Low degree +ve correlation

$$r_{23} = 1 - \frac{6\sum D_2{}^2}{N(N^2-1)}$$

$$= 1 - \frac{6\times112}{10(10^2-1)} \quad = 1 - \frac{672}{10(100-1)}$$

$$= 1 - \frac{672}{10(99)} \quad = 1 - \frac{672}{990}$$

$$= 1 - 0.679 = 0.321$$

$\therefore$ $r_{23} = +0.321 \Rightarrow$ Moderate degree +ve correlation

Similarly, $$r_{31} = 1 - \frac{6\sum D_3{}^2}{N(N^2-1)}$$

$$= 1 - \frac{6\times182}{10(10^2-1)} \quad = 1 - \frac{1092}{10(100-1)}$$

$$= 1 - \frac{1092}{10(99)} \quad = 1 - \frac{1092}{990}$$

$$= 1 - 1.103 = -0.103$$

$\therefore$ $r_{31} = -0.103 \Rightarrow$ Low degree −ve correlation

$\Rightarrow$ Since r_{23} is highest, so 2nd and 3rd judges agree the most.

Also, r_{31} being lowest, 3rd and 1st judges disagree the most.

4.4.3 Spearman's Rank Correlation when ranks are not given

1. Assign the ranks in descending order to series X by giving first rank to highest value and second rank to value lower than higher value and so on.
2. Similarly assign the ranks to series Y.
3. Calculate the difference between ranks of both the series denoted by $\sum D$.
4. Take square of deviations and calculate the value of D^2.
5. Calculate sum of square of deviations denoted by $\sum D^2$.
6. Apply following formula.

Example 11. Following are the marks obtained by 8 students in Maths and Statistics. Find the Rank Correlation Coefficient.

Marks in Maths	**60**	**70**	**53**	**59**	**68**	**72**	**50**	**54**
Marks in stats	**44**	**74**	**54**	**64**	**84**	**79**	**53**	**66**

Solution:

X	Ranks R_1	Y	Ranks R_2	Difference of Ranks $D = R_1 - R_2$	D^2
60	4	44	8	−4	16

70	2	74	3	-1	1
53	7	54	6	$+1$	1
59	5	64	5	0	0
68	3	84	1	$+2$	4
72	1	79	2	-1	1
50	8	53	7	$+1$	1
54	6	66	4	$+2$	4
					$\sum D^2 = 28$

Here $N = 8$

$\Rightarrow$ Rank Coefficient of Correlation, $r = 1 - \frac{6\sum D^2}{N(N^2-1)}$

$$= 1 - \frac{6\times 28}{8(8^2-1)}$$

$$= 1 - \frac{168}{8(64-1)}$$

$$= 1 - \frac{168}{8(63)}$$

$$= 1 - \frac{168}{504}$$

$$= 1 - 0.33 = 0.67$$

$\Rightarrow$ Rank Correlation Coefficient $= 0.67$

4.4.4 Spearman's Rank Correlation when there is repetition in ranks

1. Assign the ranks in descending order to series X by giving first rank to highest value and second rank to value lower than higher value and so on. If two items have same value, assign the average rank to both the item. For example, two equal values have ranked at 5th place than rank to be given is 5.5 to both i.e., mean of 5th and 6th rank. ($\frac{5+6}{2}$).
2. Similarly assign the ranks to series Y.
3. Calculate the difference between ranks of both the series denoted by $\sum D$.
4. Take square of deviations and calculate the value of D^2.
5. Calculate sum of square of deviations denoted by $\sum D^2$.
6. Apply following formula.

$$\mathbf{r = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2)\right\}}{N(N^2-1)}}$$

Where m = no. of times a particular item is repeated.

Example 12. Find the Spearman's Correlation Coefficient for the data given below

X	110	104	107	82	93	93	115	95	93	113
Y	80	78	90	75	81	70	87	78	73	85

Solution: Here, in X series the value 93 occurs thrice ($m_1 = 3$), i. e. at 7^{th}, 8^{th} and 9^{th} rank. So, all the three values are given the same average rank, i. e. $\frac{7+8+9}{3} = 8^{th}$ rank.

Similarly, in Y series the value 78 occurs twice ($m_2 = 2$), i. e. at 6^{th} and 7^{th} rank. So, both the values are given the same average rank, i. e. $\frac{6+7}{2} = 6.5^{th}$ rank.

X	Ranking of X R_1	Y	Ranking of Y R_2	Difference of Ranks $D = R_1 - R_2$	D^2
110	3	80	5	-2	4
104	5	78	6.5	-1.5	2.25
107	4	90	1	$+3$	9
82	10	75	8	$+2$	4
93	8	81	4	$+4$	16
93	8	70	10	-2	4
115	1	87	2	-2	1
95	6	78	6.5	-0.5	0.25
93	8	73	9	-1	1
113	2	85	3	-1	1
					$\sum D^2 = 42.5$

Here $N = 10$

Spearman's Rank Correlation Coefficient, $r = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2)\right\}}{N(N^2-1)}$

i. e. $r = 1 - \frac{6\left\{42.50 + \frac{1}{12}(3^3-3) + \frac{1}{12}(2^3-2)\right\}}{10(10^2-1)}$

$= 1 - \frac{6\left\{42.50 + \frac{24}{12} + \frac{6}{12}\right\}}{10(100-1)} = 1 - \frac{6\left\{42.50+2+\frac{1}{2}\right\}}{10\times 99}$

$= 1 - \frac{6\{42.5+2.5\}}{990} \quad = 1 - \frac{6\times 45}{990}$

$= 1 - 0.2727 = 0.7273$

$\Rightarrow$ Rank Correlation Coefficient $= 0.7273$

Example 13. The rank correlation coefficient between the marks obtained by ten students in Mathematics and Statistics was found to be 0.5. But later on, it was found that the difference in ranks in the two subjects obtained by one student was wrongly taken as 6 instead of 9. Find the correct rank correlation.

Solution: Given $N = 10$, Incorrect $r = 0.5$

We know that

Rank Correlation Coefficient, $r = 1 - \frac{6\sum D^2}{N(N^2-1)}$

$\Rightarrow \quad 0.5 = 1 - \frac{6\sum D^2}{10(10^2-1)} = 1 - \frac{6\sum D^2}{10\times 99}$

$\Rightarrow$ Incorrect $\sum D^2 = \frac{990}{6} \times 0.5 = 82.5$

$\therefore$ The corrected value of $\sum D^2 = 82.5 - 6^2 + 9^2$

$= 82.5 - 36 + 81 = 127.5$

$\therefore$ Correct Rank Correlation Coefficient, $r = 1 - \frac{6\times 127.5}{10(10^2-1)}$

$= 1 - \frac{765}{10(100-99)}$

$= 1 - \frac{765}{10\times 99}$

$= 1 - \frac{765}{990}$

$= 1 - 0.7727$

$= 0.2273$

4.4.5 Merits of Spearman's Rank Correlation

1. This is easy to understand.
2. It can calculate correlation from qualitative data also.
3. It does not put condition of normal series of data.
4. It can deal with quantitative data also.
5. It is not affected by presence of extreme values.

4.4.6 Limitations of Spearman's Rank Correlation

1. It cannot deal with grouped data.
2. If large data is there, it is difficult to apply this method.
3. It cannot be applied further algebraic treatment.
4. Combined correlation cannot be calculated.
5. It gives only approximate correlation; it is not based on actual values.

TEST YOUR UNDERSTANDING (B)

1. Find Rank correlation on base of following data.

X	78	36	98	25	75	82	90	62	65	39
Y	84	51	91	60	68	62	86	58	53	47

In Dance competition following ranks were given by 3 judges to participants. Determine which two judges have same preference for music:

1st Judge	1	6	5	10	3	2	4	9	7	8

2nd Judge	3	5	8	4	7	10	2	1	6	9
3rd Judge	6	4	9	8	1	2	3	10	5	7

3. Find Rank correlation on base of following data.

X	25	30	38	22	50	70	30	90
Y	50	40	60	40	30	20	40	70

4. Find Rank correlation on base of following data.

X	63	67	64	68	62	66	68	67	69	71
Y	66	68	65	69	66	65	68	69	71	70

Answers

1) .82	2) I and II -.2121, II and III -.297, I and III .6364, so judge I and III	3) 0	4) 0.81

4.5 LET US SUM UP

- Karl Person's coefficient of correlation is the most popular method of correlation.
- It can deal only with quantitative data.
- Spearman's Rank correlation calculated correlation on the basis of ranks given to data.
- It can deal with qualitative data also.

4.6 KEY TERMS

- **Correlation:** Correlation is a statistical technique which studies the relation between two or more variables. It studies that how to variables are related to each other.
- **Positive correlation**: It is a situation in which two variables move in the same direction. In this case if the value of one variable increases the value of other variable also increase. Similarly, if the value of one variable decrease, the value of other variable also decrease.
- **Negative or Inverse Correlation:** When two variables move in opposite direction from each other, it is known as negative or inverse correlation. In other words, we can say that when the value of one variable increase value of other variable decrease, it is called negative correlation.
- **Linear Correlation:** When the change in value of one variable results into constant ratio of change in the value of other variable, it is called linear correlation. In such case if we draw the values of two variables on the graph paper, all the points on the graph paper will fall on a straight line.

- **Non - Linear Correlation:** When the change in value of one variable does not result into constant ratio of change in the value of other variable, it is called non-linear correlation. In such case, if we draw the value of two variables on the graph paper all the points will not fall in the straight line on the graph.
- **Simple Correlation:** When we study relation between two variables only, it is known as simple correlation. For example, relation between income and expenditure, Price and Demand, are situations of simple correlation.
- **Multiple Correlation:** Multiple correlation is a situation in which more than two variables are involved. Here relation between more than two variables are studied together, for example if we are studying the relation between income of the consumer, price if the product and demand of the product, it is a situation of multiple correlation.

4.7 QUESTIONS FOR PRACTICE

1. Give Karl Persons method of calculating correlation.
2. Give Karl Pearson's coefficient of correlation in case of actual and assumed mean.
3. What are merits and limitations of Karl Pearson's method.
4. What is Spearman's Rank correlation. How it is determined.
5. In case of repeated ranks how would you determine Spearman's Rank correlation.
6. What are the limitations of Spearman's Rank correlation

4.8 SUGGESTED READINGS

- J. K. Sharma, Business Statistics, Pearson Education.
- S.C. Gupta, Fundamentals of Statistics, Himalaya Publishing House.
- S.P. Gupta and Archana Gupta, Elementary Statistics, Sultan Chand and Sons, New Delhi.
- Richard Levin and David S. Rubin, Statistics for Management, Prentice Hall of India, New
  Delhi.
- M.R. Spiegel, Theory and Problems of Statistics, Schaum's Outlines Series, McGraw HillPublishing Co.

CERTIFICATE/DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SARM 2: DESCRIPTIVE STATISTICS

SEMESTER I

UNIT 5: REGRESSION- MEANING, PROPERTIES, TYPES, DIFFERENCE BETWEEN CORRELATION AND REGRESSION

STRUCTURE

5.0 Learning Objectives

5.1 Introduction

5.2 History of Regression Analysis

5.3 Meaning of Regression Analysis

5.4 Benefits of Regression Analysis

5.5 Properties of Regression Coefficients

5.6 Limitations of Regression Analysis

5.7 Different Types of Regression

5.8 Relationship between Correlation and Regression

5.9 Sum Up

5.10 Key Terms

5.11 Questions for Practice

5.12 Suggested Readings

5.0 LEARNING OBJECTIVES

After studying the Unit, leaner will be able to:

- Describe what is regression
- Distinguish between different types of Regression
- Understand the benefits of Regression
- Show how correlation and regression are related

- Understand the properties of regression coefficients

5.1 INTRODUCTION

Statistics has many applications in our life whether it's business life or our routine life. There are many techniques in statistics that can help us in prediction. Regression is one such technique. In the literary meaning the term ‘Regression’ is 'going back', or 'stepping down’. So, regression analysis is a tool in statistics that can help in the prediction of one variable when the value of other variable is known if there exists any close relation between two or more variables, though such relation may be positive or negative. The technique of Regression can be widely used as a powerful tool in almost all fields whether science, social science, Business, etc. However, particularly, in the fields of business and management this technique is very useful for studying the relationship between different variables such as Price and Demand, Price and Supply, Production and Consumption, Income and Consumption, Income and Savings, etc.

When we find a regression between two or more variables, we try to understand the behavior of one variable with help movement of the other variable in a particular direction. For example, if the correlation coefficient between value of sales and amount spent on advertisement say +0.9, it means that if advertisement expenditure is increased, Sale is also likely to increase, as there is a very high positive relation between the two variables. However, correlation only tells the relation between two variables, but it does not tell the extent to which a change in one variable will affect the change in other variables. For this purpose, we have to calculate the co-efficient of Regression. The regression Coefficient is a statistical measure that tries to find out the value of one variable known as the dependent variable when the value of another variable known as an independent variable is known. Thus, in the case of two variables, like Advertisement expenditure and amount of Sales, we can estimate the likely amount of Sales if the value of Advertisement expenditure is given. Similarly, we can predict the value of Advertisement expenditure required, to achieve a particular amount of Sales. This can be done using the two regression coefficients

5.2 HISTORY OF REGRESSION ANALYSIS

The technique of Regression analysis was developed by the British Biometrician Sir Francis Galton in 1877 while he was studying the relationship between the heights of fathers and the heights of their sons. The term ‘regression' was first time used by him in his paper 'Regression towards Mediocrity in Hereditary Stature" in which he said:

(i) Tall fathers will most probably have tall sons, and short fathers will most probably have short sons, and the average height of the sons of tall fathers' will mostly be less than the average height of their fathers;

(ii) He also said that the average height of the sons of short fathers' is most likely to be more than the average height of their fathers; and

(iii) That the deviations of the mean height of the sons are most likely to be less than the deviations of the mean height of the, or that when the fathers' height moves above or below the mean, the sons' height tends to go back (regress) towards the mean.

Professor Galton in his study analyzed the relationship between the two variables that is the heights of the fathers and the heights of the sons using the graphical technique and named the line describing the relationship between the height of the father and height of the son as the 'Line of Regression'.

5.3 MEANING OF REGRESSION ANALYSIS

Many experts have defined the term Regression in their own way. Some of these definitions are given below:

According to Sir Francis Galton, the term regression analysis is defined as "the law of regression that tells heavily against the full hereditary transmission of any gift, the more bountifully the parent is gifted by nature, the rarer will be his good fortune if he begets a son who is richly endowed as himself, and still more so if he has a son who is endowed yet more largely."

In the words of Ya Lun Chou, "Regression analysis attempts to establish the nature of the relationship between variables that is to study the functional relationship between the variables and thereby provide a mechanism for prediction or forecasting".

5.4 BENEFITS / USES OF REGRESSION ANALYSIS

The benefits of Regression analysis are outlined as under:

1. **Forecasting or Prediction** – Regression provides a relationship between two or more variables that are related to each other. So, with the help of this technique, we can easily forecast the values of one variable that is unknown from the values of another variable that is known.

2. **Cause and Effect Relationship –** This analysis helps in finding the cause-and-effect relationship between two or more variables. It is a powerful tool for measuring the cause-and-effect relationship among economic variables. In the field of economics, it is very beneficial in the estimation of Demand, Production, Supply, etc.
3. **Measuring Error in Estimation –** Regression helps in measuring errors in estimates made through the regression lines. In case the point of the Regression line is less scattered around the relevant regression line, it means there are less chances of error but if there are more scattered around a line of regression, it means there are more chances of error.
4. **Finding Correlation Coefficient between two variables –** Regression provides a measure of the coefficient of correlation between the two variables. We can calculate correlation by taking the square root of the product of the two regression coefficients.
5. **Usefulness in Business and Commerce** – Regression is a very powerful tool of statistical analysis in the field of business and commerce as it can help a businessman in prediction of various values such as demand, production etc.
6. **Useful in day-to-day life -** This technique is very useful in our daily life as it can predict various factors such as birth rate, death rate, etc.
7. **Testing Hypothesis –** The technique of regression can be used in testing the validity of an economic theory or testing any hypothesis.

5.5 LIMITATIONS OF REGRESSION ANALYSIS

Though Regression is a wonderful statistical tool, still it suffers from some limitations. The following are the limitations of Regression analysis:

1. Regression analysis assumes that there exists cause and effect relationship between the variables and such a relation is not changeable. This assumption may not always hold good and thus could give misleading results.
2. Regression analysis is based on some limited data available. However, as the values are based on limited data it may give misleading results.
3. Regression analysis involves very lengthy and complicated steps of calculations and analysis. A layman may not be in a position to use this technique.
4. Regression analysis can be used in case of quantitative data only. It cannot be used where data is of a qualitative nature such as hard work, beauty, etc.

5.6 DIFFERENT TYPES OF REGRESSION ANALYSIS

1. Simple and Multiple Regression

- **Simple Regression**: When there are only two variables under study it is known as a simple regression. For example, we are studying the relationship between Sales and Advertising expenditure. If we consider sales as Variable X and advertising as variable Y, then the X = a+b Y is known as the regression equation of X on Y where X is the dependent variable and Y and the independent variable. In other words, we can find the value of variable X (Sales) if the value of Variable Y (Advertising) is given.
- **Multiple Regression:** The study of more than two variables at a time is known as multiple regression. Under this, only one variable is taken as a dependent variable and all the other variables are taken as independent variables. For example, if we consider sales as Variable X, advertising as variable Y and Income as Variable Z, then using the functional relation X = *f* (Y, Z), we can find the value of variable X (Sales) if the value of Variable Y (Advertising) and the value of variable Z (Income) is given.

2. Total and Partial Regression

a. Total Regression: Total regression analysis is one in which we study the effect of all the variables simultaneously. For example, when we want to study the effect of advertising expenditure of business represented by variable Y, income of the consumer represented by variable Z, on the amount of sales represented by variable X, we can study impact of advertising and income simultaneously on sales. This is a case of total regression analysis. In such cases, the regression equation is represented as follows:

X = *f* (Y, Z),

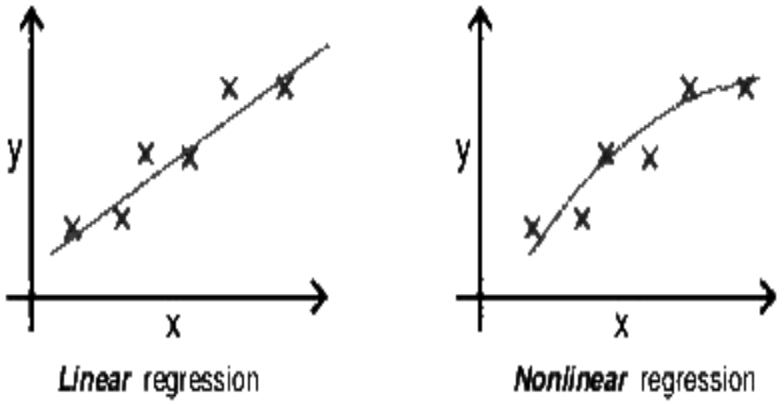
a. Partial Regression: In the case of Partial Regression one or two variables are taken into consideration and the others are excluded. For example, when we want to study the effect of advertising expenditure of business represented by variable Y, income of the consumer represented by variable Z, on the amount of sales represented by variable X, we will not study impact of both income and advertising simultaneously, rather we will first study effect of income on sales keeping advertising constant and then effect of advertising on sales keeping income constant. Partial regression can be written as

X=f (Y not Z)

3. Linear and Non-Linear Regression

a. **Linear Regression:** When the functional relationship between X and Y is expressed as the first-degree equations, it is known as linear regression. In other words, when the points plotted on a scatter diagram concentrate around a straight line it is the case of linear regression.

b. **Non-linear Regression**: On the other hand, if the line of regression (in the scatter diagram) is not a straight line, the regression is termed curved or non-linear regression. The regression equations of non-linear regression are represented by equations of a higher degree. The following diagrams show the linear and non-linear regressions:



5.7 PROPERTIES OF REGRESSION COEFFICIENTS

The regression coefficients discussed above have a number of properties which are given as under:

1. The geometric Mean of the two regression coefficients gives the coefficients of correlation i.e., r = $\sqrt{bxy * byx}$
2. Both the regression coefficients must have the same sign i.e., in other words, either both coefficients will have + signs or both coefficients will have - signs. This is due to the fact that in the first property, we have studied the geometric means of both coefficients will give us value of correlation. If one sign will be positive and other will be negative, the product of both signs will be negative. And it is not possible to find out correlation of negative value.
3. The signs of regression coefficients will give us signs of coefficient of correlation. This means if the regression coefficients are positive the correlation coefficient will be positive, and if the regression coefficients are negative then the correlation coefficient will be negative.

4. If one of the regression coefficients is greater than unity or 1, the other must be less than unity. This is because the value of the coefficient of correlation must be between ± 1. If both the regression coefficients are more than 1, then their geometric mean will be more than 1 but the value of correlation cannot exceed 1.
5. The arithmetic mean of the regression coefficients is either equal to or more than the correlation coefficient $\frac{bxy+byx}{2} \geq \sqrt{bxy * byx}$
6. If the regression coefficients are given, we can calculate the value of standard deviation by using the following formula.

   bxy = $r \frac{\sigma x}{\sigma y}$ or byx = $r \frac{\sigma y}{\sigma x}$
7. Regression coefficients are independent of change of origin but not of scale. This means that if the original values of the two variables are added or subtracted by some constant, the values of the regression coefficients will remain the same. But if the original values of the two variables are multiplied, or divided by some constant (common factors) the values of the regression equation will not remain the same.

5.8 RELATIONSHIP BETWEEN CORRELATION AND REGRESSION

1. Correlation is a quantitative tool that measure of the degree of relationship that is present between two variables. It shows the degree and direction of the relation between two variables. Regression helps us to find the value of a dependent variable when the value of independent variable is given.
2. Correlation between two variables is the same. For example, if we calculate the correlation between sales and advertising or advertising and sales, the value of correlation will remain the same. But this is not true for Regression. The regression equation of Advertising on sales will be different from regression equation of Sales on advertising.
3. If there is a positive correlation, the distance between the two lines will be less. That means the two regression lines will be closer to each other- Similarly, if there is a low correlation, the lines will be farther from each other. A positive correlation implies that the lines will be upward-sloping whereas a negative correlation implies that the regression lines will be downward sloping.
4. Correlation between two variables can be calculated by taking the square root of the product of the two regression coefficients.

Following are some of the differences between Correlation and Regression:

1. Correlation measures the degree and direction of relationship between two variables. Regression measures the change in value of a dependent variable given the change in value of an independent variable.
2. Correlation does not depict a cause-and-effect relationship. Regression depicts the causal relationship between two variables.
3. Correlation is a relative measure of linear relationship that exists between two variables. Regression is an absolute measure that measures the change in value of a variable.
4. Correlation between two variables is the same. In other words, Correlation between the two variables is the same. rxy =ryx. Regression is not symmetrical in formation. So, the regression coefficients of X on Y and Yon X are different.
5. Correlation is independent of Change in origin or scale. Regression is independent of Change in origin but not of scale.
6. Correlation is not capable of any further mathematical treatment. Regression can be further treated mathematically.
7. The coefficient of correlation always lies between -1 and +1. The regression coefficient can have any value.

5.9 SUM UP

- Regression is a useful tool for forecasting.
- With the help of regression, we can predict the value of can find the value of X if the value of Y is given or the value of Y if value of X is given.
- It creates the mathematical linear relation between two variables X and Y, out of which one variable is dependent and other is independent.

5.10 KEY TERMS

- **Regression:** Regression creates the mathematical linear relation between two variables X and Y, out of which one variable is dependent and the other is independent.
- **Simple Regression**: When there are only two variables under study it is known as a simple regression. For example, we are studying the relationship between Sales and Advertising expenditure.

- **Multiple Regression:** The study of more than two variables at a time is known as multiple regression. Under this, only one variable is taken as a dependent variable and all the other variables are taken as independent variables.
- **Total Regression:** Total regression analysis is one in which we study the effect of all the variables simultaneously.
- **Partial Regression:** In the case of Partial Regression one or two variables are taken into consideration and the others are excluded.
- **Linear Regression:** When the functional relationship between X and Y is expressed as the first-degree equations, it is known as linear regression. In other words, when the points plotted on a scatter diagram concentrate around a straight line it is the case of linear regression.
- **Non-linear Regression**: On the other hand, if the line of regression (in scatter diagram) is not straight, the regression is termed curved or non-linear regression.
- **Least Square method**: According to the Least Square method, regression line should be plotted in such a way that sum of squares of the difference between actual value and estimated value of the dependent variable should be least or minimum possible.

5.11 QUESTIONS FOR PRACTICE

1. What is Regression? What are the uses of Regression?
2. Discuss the properties of regression analysis.
3. What is the relationship between Regression and correlation?
4. Explain different types of regressions.

5.12 SUGGESTED READINGS

- J. K. Sharma, *Business Statistics,* Pearson Education.
- S.C. Gupta, *Fundamentals of Statistics,* Himalaya Publishing House.
- S.P. Gupta and Archana Gupta, *Elementary Statistics,* Sultan Chand and Sons, New Delhi.
- Richard Levin and David S. Rubin, *Statistics for Management,* Prentice Hall of India, New Delhi. Hill Publishing Co.

CERTIFICATE/DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SARM 2: DESCRIPTIVE STATISTICS

SEMESTER I

UNIT 6: MEASUREMENT OF REGRESSION EQUATIONS

STRUCTURE

6.0 Learning Objectives

6.1 Meaning of Regression Lines

6.2 Least Square Method of fitting Regression lines

6.3 Direct Method of Estimating Regression Equations

6.4 Other Methods of Estimating Regression Equations

6.5 Sum Up

6.6 Questions for Practice

6.7 Suggested Readings

6.0 LEARNING OBJECTIVES

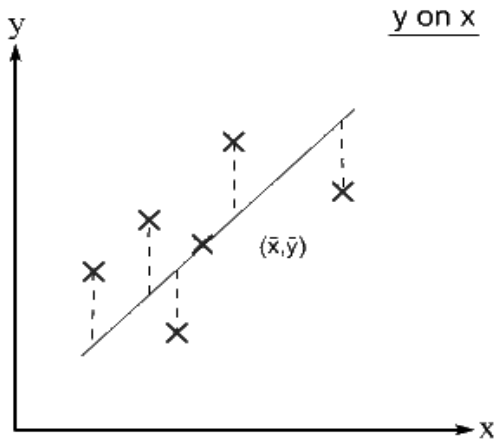
After study of this unit, learners can learn about the:

- Meaning of line of regression X on Y
- Meaning of line of regression Y on X
- Different methods of estimation of regression coefficient

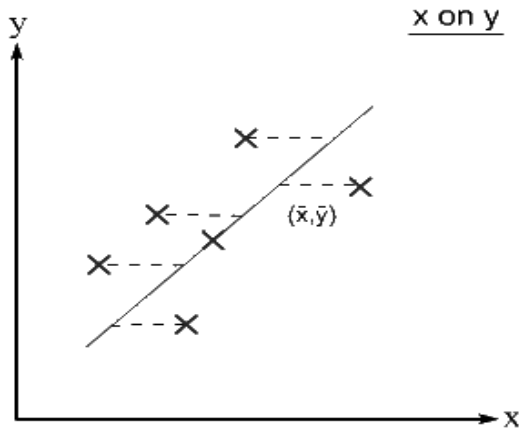
6.1 MEANING OF REGRESSION LINES

The lines that are used in Regression for the purpose of estimation are called regression lines. In other words, the lines that are used to study the dependence of one variable on the other variable are called regression lines. If we have two variables X and Y then there.

a. Regression Line of Y on X: Regression Line Y on X measures the dependence of Y on X and we can estimate the value of Y for the given values of X. In this line, Y is the dependent variable and X is the independent variable.

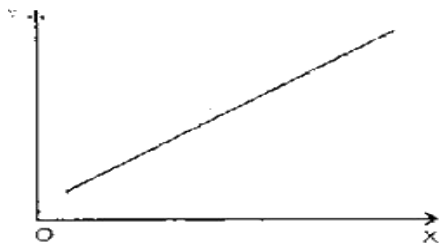


b. Regression Line of X on Y: Regression Line X on Y measures the dependence of X on Y and we can estimate the value of X for the given values of Y. In this line X is dependent variable and Y is independent variable.



The direction of two regression equations depends upon the degree of correlation between two variables. Following can be the cases of correlation between two variables:

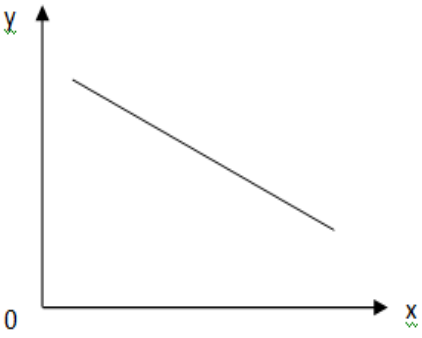
1. Perfect positive correlation: If there is a perfect positive correlation between two variables (i.e., r = +1), both the lines will coincide with each other and will be having a positive slope. Both the lines X on Y and Yon X will be same in this case. In other words, in that case, only one regression line can be drawn as shown in the diagram. The slope of the line



Both the line will coincide with positive slope

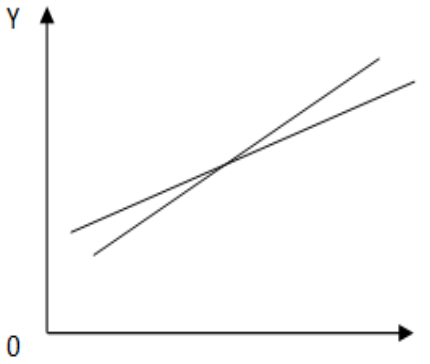
will be upward.

2. Perfect negative correlation: If there is a perfect negative correlation between two variables (i.e., r = -1), both the lines will coincide with each other and will in such case these lines will be having negative slope. Both the lines X on Y and Yon X will be same but downward sloping. In other words, in that case only one regression line can be drawn as shown in the diagram. The slope of the line will be upward.



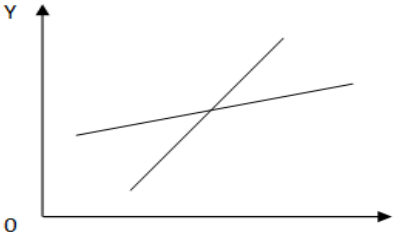
Both the line will coincide with negative slope

3. High degree of correlation: If there is a high egree of correlation between two variables, both the lines will be near to each other. In other words, these lines will be closer to each other but the lines will not coincide with each other. Both the lines will be separate. Further the direction of lines depends upon the positive or negative correlation.



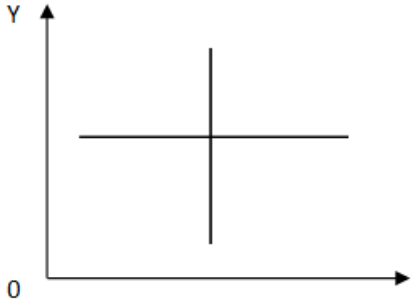
Both the line will be close to each other

4. Low degree of correlation: If there is a low degree of correlation between two variables, both the lines will be having more distance from each other. In other words, these lines will be farther to each other, that is the gap between the two lines will be more. Both the lines will be separate. Further the direction of lines depends upon the positive or negative correlation.



Both the line will be farther from each other

5. No correlation: If there is a no correlation between two variables (i.e., r = 0), both the lines will be perpendicular to each other. In other words, these lines will cut each other at $90^{0.}$ This diagram depicts the perpendicular relation between the two regression lines when there is absolutely zero correlation between the two variables under the study.



Both the line will be perpendicular to each other

6.2 LEAST SQUARE METHOD OF FITTING REGRESSION LINES

Under this method, the lines of best fit are drawn as the lines of regression. These lines of regression are known as the lines of the best fit because, with the help of these lines we can make the estimate of the values of one variable depending on the value of other variables. According to the Least Square method, regression line should be plotted in such a way that sum of the square of the difference between actual value and the estimated value of the dependent variable should be the least or minimum possible. Under this method, we draw two regression lines that are

a. **Regression line Y on X** – it measures the value of Y when the value of X is given. In other words, it assumes that X is an independent variable whereas the other variable Y is dependent variable. Mathematically this line is represented by

   Y = a + bX

   Where Y – Dependent Variable

   X – Independent Variable

   a & b – Constants

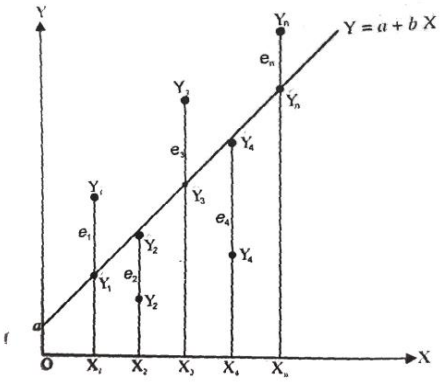
b. **Regression line X on Y** – it measures the value of X when value of Y is given. In other words, it assumes that Y is an independent variable whereas the other variable X is dependent variable. Mathematically this line is represented by

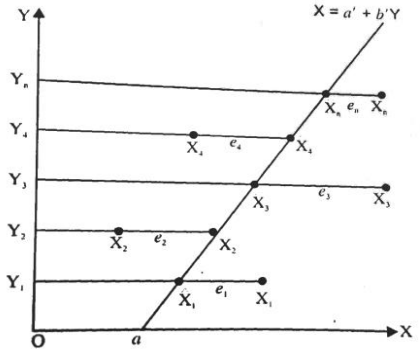
   X = a + bY

   Where X – Dependent Variable

   Y – Independent Variable

a & b – Constants





Equation Y on X Equation X on Y

In the above two regression lines, there are two constants represented by “a” and “b”. The constant “b” is also known as regression coefficient, which are denoted as “byx” and “bxy”, Where “byx” represent regression coefficient of equation Y on X and “bxy” represent regression coefficient of equation X on Y. When the value of these two variables “a” and “b” is determined we can find out the regression line.

6.3 DIRECT METHODS TO ESTIMATE REGRESSION EQUATION

The regression equations can be obtained by 'Normal Equation Method" as follows:

1. **Regression Equation of Y on X:** The regression equation Y on X is in the format of Y = a + bx, where Y is a Dependent Variable and X is an Independent Variable. To estimate this regression equation, the following normal equations are used:

   $\Sigma Y = na + b_{yx} \Sigma X$

   $\Sigma XY = a \Sigma X + b_{yx} \Sigma X^2$

   With the help of these two equations the values of ‘a' and 'b' are obtained and by putting the values of ‘a' and 'b' in the equation Y = a + b X we can predict or estimate value of Y for any value of X.

2. **Regression Equation of X on Y:** The regression equation X on Y is in the format of X = a + bY, where X is a Dependent Variable and Y is an Independent Variable. To estimate this regression equation, following normal equations are used:

$$\Sigma X = na + b_{xy} \Sigma Y$$

$$\Sigma XY = a \Sigma Y + b_{xy} \Sigma Y^2$$

With the help of these two equations the values of 'a' and 'b' are obtained and by putting the values of 'a' and 'b' in the equation X = a + bY we can predict or estimate value of Y for any value of X.

Example 1. Find out the two regression lines for the data given below using the method of least square.

Variable X:	**5**	**10**	**15**	**20**	**25**
Variable Y:	**20**	**40**	**30**	**60**	**50**

Determination of the regression lines by the method of least square. Also, find out

a. **Value of Y when value of X is 40**
b. **Value of X when value of Y is 80**

Solution:

X	Y	X2	Y2	XY
5	20	25	400	100
10	40	100	1600	400
15	30	225	900	450
20	60	400	3600	1200
25	50	625	2500	1250
XX = 75	XY = 200	XX2=1375	XY2=9,000	XXY =3400

(i) Regression line of Y on X

This is given by $Y = a + bX$

where *a* and *b* are the two constants that are found by solving simultaneously the two normal equations as follows:

$\Sigma Y = na + b_{yx} \Sigma X$

$\Sigma XY = a \Sigma X + b_{yx} \Sigma X^2$

Substituting the given values in the above equations we get,

$200 = 5a + 75b$ …….. (i)

3400 =75a + 1375b …….. (ii)

Multiplying the eqn. (i) by 15 we get

3000 = 75a + 1125b…….. (iii)

Subtracting the equation (iii) from equation (ii) we get,

3400 =75a + 1375b

-3000 = -75a - 1125b

400 = 250b

or b = 1.6

Putting the above value of b in the eqn. (i) we get,

200 = 5a+ 75(1.6) or

5a =200- 120 or

a = 16

Thus, a = 16, and b = 1.6

Putting these values in the equation Y = *a* + bX we get

Y = 16+ 1.6X

So, when X is 40, the value of Y will be

Y = 16+ 1.6(40) = 80

(ii) Regression line of X on Y

This is given by X = *a* + *bY*

where *a* and *b* are the two constants that are found by solving simultaneously the two normal equations as follows:

$\Sigma X = na + b_{xy}\Sigma Y$

$\Sigma XY = a \Sigma Y + b_{xy}\Sigma Y^2$

Substituting the given values in the above equations we get,

75 = 5a+200b …….. (i)

3400 = 200a + 9000b …….. (ii)

Multiplying the eqn. (i) by 40 we get

3000 = 200a + 8000b…….. (iii)

Subtracting the equation (iii) from equation (ii) we get,

3400 = 200a + 9000b

-3000 = -200a + -8000b

400 = 1000b

or b = .4

Putting the above value of b in the eqn. (i) we get,

75 = 5a + 200(.4) or

5a = -5 or

a = -1

Thus, a = -1, and b = .4

Putting these values in the equation X = $a + bY$ we get

X = -1+ .4Y

So, when Y is 80, the value of X will be

X = -1+ .4(80) = 31

6.4 OTHER METHODS OF ESTIMATING REGRESSION EQUATION

This method discussed above is known as direct method. This is one of the popular methods of finding the regression equation. But sometimes this method of finding regression equations becomes cumbersome and lengthy especially when the values of X and Y are very large. In this case, we can simplify the calculation by taking the deviations of X and Y than dealing with the actual values of X and Y. In such case

Regression equation Y on X

Y = a + bX

will be converted to $(Y - \bar{Y}) = byx\ (X - \bar{X})$

Similarly, Regression equation X on Y:

X = a + bY

will be converted into $(X - \bar{X}) = bxy\ (Y - \bar{Y})$

Now when we are using these regression equations, the calculations will become very simple as now we have to calculate value of only one constant which is the value of "b" which is our regression coefficient. As there are two regression equations, we need to calculate two regression coefficients that are Regression Coefficient X on Y, which is symbolically denoted as "bxy" and

similarly Regression Coefficient Y on X, which is denoted as "byx". However, these coefficients can also be calculated using different methods. As we take deviations under this method, we can take deviations using actual mean, assumed mean or we can calculate it by not taking the deviations. The following formulas are used in such cases:

Method	Regression Coefficient X on Y	Regression Coefficient Y on X
When deviations are taken from actual mean	$bxy = \frac{\sum xy}{\sum y^2}$	$byx = \frac{\sum xy}{\sum x^2}$
When deviations are taken from assumed mean	$bxy = \frac{N\sum dxdy - \sum dx \sum dy}{N\sum dy^2 - (\sum dy)^2}$	$byx = \frac{N\sum dxdy - \sum dx \sum dy}{N\sum dx^2 - (\sum dx)^2}$
Direct Method: Using sum of X and Y	$bxy = \frac{N\sum XY - \sum X \sum Y}{N\sum Y^2 - (\sum Y)^2}$	$byx = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2}$
Using the correlation coefficient (r) and standard deviation (σ)	$bxy = r \cdot \frac{\sigma x}{\sigma y}$	$byx = r \cdot \frac{\sigma y}{\sigma x}$

Example 2. From the information given below obtain two regression lines X on Y and Yon X using

1. **Actual Mean Method**
2. **Assumed Mean Method**
3. **Direct Method (Without taking Mean)**

Number of Hrs. Machine Operated	**7**	**8**	**6**	**9**	**11**	**9**	**10**	**12**
Production (Units in 000):	**4**	**5**	**2**	**6**	**9**	**5**	**7**	**10**

Solution:

1. **<u>Actual Mean Method</u>**

Calculation of Regression Equation

X	Y	$x = X - \bar{X}$	x^2	$y = Y - \bar{Y}$	y^2	xy
7	4	-2	4	-2	4	4
8	5	-1	1	-1	1	1
6	2	-3	9	-4	16	12
9	6	0	0	0	0	0
11	9	2	4	3	9	6
9	5	0	0	-1	1	0

10	7	1	1	1	1	1
12	10	3	9	4	16	12
$\sum X = 72$	$\sum Y = 48$		$\sum x^2 = 28$		$\sum y^2 = 48$	$\sum xy = 36$

$\bar{X} = \frac{\sum X}{N} = \frac{72}{8} = 9$

$\bar{Y} = \frac{\sum Y}{N} = \frac{48}{8} = 6$

Regression equation of X on Y:

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

Where $bxy = \frac{\sum xy}{y^2}$

$= \frac{36}{48}$

$= .75$

So $(X - 9) = .75 (Y - 6)$

$X - 9 = .75Y - 4.5$

$X = 4.5 + .75Y$

Regression equation of Y on X:

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

Where $byx = \frac{\sum xy}{x^2}$

$= \frac{36}{28}$

$= 1.286$

So $(Y - 6) = 1.286 (X - 9)$

$Y - 6 = 1.286X - 11.57$

$Y = -5.57 + 1.286X$

2. <u>Assumed Mean Method</u>

Calculation of Regression Equation

X	**Y**	**dx =X – A (A = 8)**	**dx^2**	**dy =Y – A (A = 5)**	**dy^2**	**dx *dy**
7	4	-1	1	-1	1	1
8	5	0	0	0	0	0
6	2	-2	4	-3	9	6

9	6	1	1	1	1	1
11	9	3	9	4	16	12
9	5	1	1	0	0	0
10	7	2	4	2	4	4
12	10	4	16	5	25	20
$\sum X = 72$	**$\sum Y = 48$**	**$\sum dx = 8$**	**$\sum dx^2 = 36$**	**$\sum dy = 8$**	**$\sum dy^2 = 56$**	**$\sum xy = 44$**

$\bar{X} = \frac{\sum X}{N} = \frac{72}{8} = 9$

$\bar{Y} = \frac{\sum Y}{N} = \frac{48}{8} = 6$

Regression equation of X on Y:

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

Where $bxy = \frac{N\sum dxdy - \sum dx \sum dy}{N\sum dy^2 - (\sum dy)^2}$

$= \frac{8(44) - (8)(8)}{8(56) - (8)^2}$

$= \frac{352 - 64}{448 - 64} = \frac{288}{384} = .75$

So $(X - 9) = .75 (Y - 6)$

$X - 9 = .75Y - 4.5$

$X = 4.5 + .75Y$

Regression equation of Y on X:

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

Where $byx = \frac{N\sum dxdy - \sum dx \sum dy}{N\sum dx^2 - (\sum dx)^2}$

$= \frac{8(44) - (8)(8)}{8(36) - (8)^2} = \frac{288}{224}$

$= 1.286$

So $(Y - 6) = 1.286 (X - 9)$

$Y - 6 = 1.286X - 11.57$

$Y = -5.57 + 1.286X$

3. <u>Direct Method (Without taking Mean)</u>

Calculation of Regression Equation

X	**Y**	**X^2**	**Y^2**	**XY**

7	4	49	16	28
8	5	64	25	40
6	2	36	4	12
9	6	81	36	54
11	9	121	81	99
9	5	81	25	45
10	7	100	49	70
12	10	144	100	120
$\sum X = 72$	**$\sum Y = 48$**	**$\sum X^2 = 676$**	**$\sum Y^2 = 336$**	**$\sum XY = 468$**

$\bar{X} = \frac{\sum X}{N} = \frac{72}{8} = 9$

$\bar{Y} = \frac{\sum Y}{N} = \frac{48}{8} = 6$

Regression equation of X on Y:

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

Where $bxy = \frac{N\sum XY - \sum X \sum Y}{N\sum Y^2 - (\sum Y)^2}$

$= \frac{8\,(468) - (72)\,(48)}{8\,(336) - (48)^2}$

$= \frac{3744 - 3456}{2688 - 2304} = \frac{288}{384}$

$= .75$

So $(X - 9) = .75\,(Y - 6)$

$X - 9 = .75Y - 4.5$

$X = 4.5 + .75Y$

Regression equation of Y on X:

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

Where $byx = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2}$

$= \frac{8\,(468) - (72)\,(48)}{8\,(676) - (72)^2}$

$= \frac{3744 - 3456}{5408 - 5184} = \frac{288}{224}$

$= 1.286$

So $(Y - 6) = 1.286\,(X - 9)$

$Y - 6 = 1.286X - 11.57$

$Y = -5.57 + 1.286X$

Example 3. Find out two Regression equations on basis of the data given below:

	X	Y
Mean	60	80
Standard Deviation (S.D.)	16	20
Coefficient of Correlation	.9	

Also find value of X when Y = 150 and value of Y when X = 100.

Solution: Regression equation of X on Y:

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

Where $bxy = r \frac{\sigma x}{\sigma y}$

$= .9 \frac{16}{20}$

$= .72$

So $(X - 60) = .72 (Y - 80)$

$X - 60 = .72Y - 57.6$

$X = 2.4 + .72Y$

When Y = 150 than X = 2.4 + .72(150) = 110.4

Regression equation of Y on X:

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

Where $byx = r \frac{\sigma y}{\sigma x}$

$= .9 \frac{20}{16}$

$= 1.125$

So $(Y - 80) = 1.125 (X - 60)$

$Y - 80 = 1.125X - 67.5$

$Y = 12.5 + 1.125 X$

When X = 100 than Y = 12.5 + 1.125 (100) = 125

Example 4. From the following data find out two lines of regression land also find out value of correlation.

$\sum X = 250$; $\sum Y = 300$; $\sum XY = 7900$;

$\sum X^2 = 6500$; $\sum Y^2 = 10000$; $n = 10$

Solution:

$\overline{X} = \frac{\sum X}{N} = \frac{250}{10} = 25$

$\overline{Y} = \frac{\sum Y}{N} = \frac{300}{10} = 30$

Regression equation of Y on X:

$(Y - \overline{Y}) = b_{xy} (X - \overline{X})$

Where $byx = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2}$

$= \frac{10\,(7900) - (250)\,(300)}{10\,(6500) - (250)^2}$

$= \frac{79000 - 75000}{65000 - 62500} = \frac{4000}{2500} = 1.6$

So $(Y - 30) = 1.6\,(X - 25)$

$Y - 30 = 1.6X - 40$

$Y = -10 + 1.6\,X$

Regression equation of X on Y:

$(X - \overline{X}) = b_{xy} (Y - \overline{Y})$

Where $bxy = \frac{N\sum XY - \sum X \sum Y}{N\sum Y^2 - (\sum Y)^2}$

$= \frac{10\,(7900) - (250)\,(300)}{10\,(10000) - (300)^2}$

$= \frac{79000 - 75000}{100000 - 90000}$

$= \frac{4000}{10000} = .4$

So $(X - 25) = .4\,(Y - 30)$

$X - 25 = .4Y - 12$

$X = 13 + .4Y$

Coefficients of Correlation

$r = \sqrt{bxy * byx}$

$r = \sqrt{1.6 * 0.4}$

$r = \sqrt{.64}$

$r = .8$

Example 5. From the following data find out two lines of regression land also find out value of correlation. Also find value of Y when X = 30

$\sum X = 140$; $\sum Y = 150$; $\sum (X - 10)(Y - 15) = 6$;

$\sum (X - 10)^2 = 180$; $\sum (Y - 15)^2 = 215$; $n = 10$

Solution:

Let's take assumed mean of Series X = 10 and Series Y = 15.

$\sum dx = \sum (X - 10) = \sum X - 10n = 140 - 100 = 40$

$\sum dy = \sum (Y - 15) = \sum Y - 15n = 150 - 150 = 0$

$\sum dx^2 = \sum (X - 10)^2 = 180$

$\sum dy^2 = \sum (Y - 15)^2 = 215$

$\sum dxdy = \sum (X - 10)(Y - 15) = 6$

So,

$\overline{X} = A + \frac{\sum X}{N} = 10 + \frac{40}{10} = 14$

$\overline{Y} = A + \frac{\sum Y}{N} = 15 + \frac{0}{10} = 15$

Regression equation of Y on X:

$(Y - \overline{Y}) = b_{xy} (X - \overline{X})$

Where $byx = \frac{N\sum dxdy - \sum dx \sum dy}{N\sum dx^2 - (\sum dx)^2}$

$= \frac{10(6) - (40)(0)}{10(180) - (40)^2}$

$= \frac{60}{200} = .3$

So $(Y - 15) = .3 (X - 14)$

$Y - 15 = .3X - 4.2$

$Y = 10.8 + .3X$

When X = 30 than Y = 10.8 + .3(30) = 19.8

Regression equation of Y on X:

$(Y - \overline{Y}) = b_{xy} (X - \overline{X})$

Where $byx = \frac{N\sum dxdy - \sum dx \sum dy}{N\sum dx^2 - (\sum dx)^2}$

$= \frac{10(6) - (40)(0)}{10(25) - (0)^2}$

$= \frac{60}{250} = .24$

So $(Y - 15) = .24 (X - 14)$

Y – 15 = .24X – 3.36

Y = 11.64+ .24X

Coefficients of Correlation

r = $\sqrt{bxy * byx}$ = $\sqrt{.3 * .24}$

r = $\sqrt{.072}$

r = .268

Example 5. From the following data find out which equation is equation X on Y and which equation is equation Y on X. Also find $\overline{X}$, $\overline{Y}$ and r.

3X + 2Y – 26 = 0

6X + Y – 31 = 0

Solution:

To find $\overline{X}$ and $\overline{Y}$, we will solve following simultaneous equations

3X + 2Y = 26 ……. (i)

6X + Y = 31 ……. (ii)

Multiply equation (i) with 2, we get

6X + 4Y = 52 ……. (iii)

Deduct equation (ii) from equation (iii)

6X + 4Y = 52

-<u>6X - Y = -31</u>

3Y = 21

Y = 7

Or $\overline{Y}$ = 7.

Put the value of Y in Equation (i), we get

3X + 2(7) = 26

3X + 14 = 26

3X = 12

X = 4

or $\overline{X}$ = 4

Let 3X + 2Y = 26 be regression equation X on Y

$3X = 26 - 2Y$

$X = \frac{26}{3} - \frac{2}{3} Y$

So $b_{xy} = -\frac{2}{3}$

Let $6X + Y = 31$ be regression equation Y on X

$Y = 31 - 6X$

So $b_{yx} = -6$

As $r = \sqrt{bxy * byx}$

$r = -\sqrt{-(\frac{2}{3}) \times (-.6)}$

$r = -2$, but this is not possible as value of r always lies between $-1 \; and \; +1$. So, our assumption is wrong and equation are reverse.

Let $6X + Y = 31$ be regression equation X on Y

$6X = 31 - Y$

$X = \frac{31}{6} - \frac{1}{6} Y$

So, $b_{xy} = -\frac{1}{6}$

Let $3X + 2Y = 26$ be regression equation Y on X

$2Y = 26 - 3X$

$Y = \frac{26}{2} - \frac{3}{2} X$

So $b_{yx} = -\frac{3}{2}$

As $r = \sqrt{bxy * byx}$

$r = -\sqrt{-(\frac{1}{6}) \times -(\frac{3}{2})}$

$r = -.5$, which is possible. So, our assumption is right.

So, $\overline{Y} = 7$; $\overline{X} = 4$;

X on Y is $X = \frac{31}{6} - \frac{1}{6} Y$

Y on X is $Y = \frac{26}{2} - \frac{3}{2} X$

$r = -.5$

TEST YOUR UNDERSTANDING

1. Find both regression equations:

X	6	2	10	4	8
Y	9	11	5	8	7

2. From following estimate the value of Y when X = 30 using regression equation.

X	25	22	28	26	35	20	22	40	20	18	19	25
Y	18	15	20	17	22	14	15	21	15	14	16	17

3. Fit two regression lines:

X	30	32	38	35	40
Y	10	14	16	20	15

Find X when Y = 25 and find Y when X = 36.

4. Find out two Regression equations on basis of the data given below:

	X	Y
Mean	65	67
Standard Deviation (S.D.)	2.5	3.5
Coefficient of Correlation	.8	

5. In a data the Mean values of X and Y are 20 and 45 respectively. Regression coefficient b_{yx} = 4 and b_{xy} 1/9. Find

a. coefficient of correlation
b. Standard Deviation of X, if S.D. of Y = 12
c. Find two regression lines

6. You are supplied with the following information. Variance of X = 36

$12X - 51Y + 99 = 0$

$60X - 27Y = 321.$

Calculate:

(a) The average values of X and Y

(b) The standard deviation of Y and

7. The lines of regression of Y on X and X on Y are Y = X + 5 and 16X = 9Y + 4 respectively

Also, σy =4 Find $\bar{X}$, $\bar{Y}$, σx and r.

8. Given:

$$\sum X = 56, \sum Y = 40, \sum X^2 = 524$$
$$\sum Y^2 = 256, \sum XY = 364, N = 8$$

(a) find the regression equatiopn of X on Y

Answers

1. X = 16.4 – 1.3Y, Y = 11.9 - .65X
2. 18.875
3. Y = .46X – 1.1, X = .6Y + 26, Value of Y = 15.46, Value of X = 40.25
4. Y = 1.12X – 5.8, X = .57Y +26,81
5. .67, 2, Y= 4X – 35 and X = 1/9 Y +15
6. Mean of X = 13, Mean of Y= 17, S.D of Y = 8
7. Mean of X = 7, Mean of Y= 12, S.D of X = 3, r= .75
8. X = 1.5Y - 0.5, r = .977

6.5 LET US SUM UP

- there are two regression equations X on Y and Y on X.
- Regression can be linear or nonlinear.
- It can be simple or multiple.
- Regression is based on the principle of Least Squares.
- We can also find out correlation coefficient with help of regression coefficients.

6.6 QUESTIONS FOR PRACTICE

Q1. How two regression lines are determined under direct method.

Q2. Explain various methods of finding regression equations.

Q3. What are limitations of regression analysis.

Q4. What are properties of regression coefficients.

6.7 FURTHER READINGS

- J. K. Sharma, *Business Statistics,* Pearson Education.
- S.C. Gupta, *Fundamentals of Statistics,* Himalaya Publishing House.

- S.P. Gupta and Archana Gupta, *Elementary Statistics*, Sultan Chand and Sons, New Delhi.
- Richard Levin and David S. Rubin, *Statistics for Management*, Prentice Hall of India, New Delhi. Hill Publishing Co.

CERTIFICATE/DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SARM 2: DESCRIPTIVE STATISTICS

SEMESTER I

UNIT 7: INDEX NUMBERS: MEANING AND USES AND TYPES OF INDEX NUMBERS, PROBLEMS IN THE CONSTRUCTION, METHODS OF INDEX NUMBERS

STRUCTURE

7.0 Learning Objectives

7.1 Introduction and Meaning of Index Numbers

7.2 Features of Index Numbers

7.3 Uses of Index Numbers

7.5 Problems in the Construction of Index Numbers

7.6 Different Types of Index Numbers

7.7 Different Methods of Index Numbers

- **7.7.1 Simple Index Number**
  - **7.7.1.1 Simple Aggregative Method**
  - **7..7.1.2 Simple Price Relative Method**
- **7.7.2 Weighted Index Number**
  - **7.7.2.1 Weighted Aggregative Method**
    - **a) Laspeyre's Method**
    - **b) Paasche's Method**
    - **c) Dorbish and Bowley Method**
    - **d) Fisher's Method**
    - **e) Mashal Edgeworth Method**
  - **7.7.2.2 Weighted Price Relative Method**

7.9 Sum Up

7.10 Key Terms

7.11 Questions for Practice

7.12 Further Readings

7.0 LEARNING OBJECTIVES

After studying the Unit, the learner will be able to learn about:

- Meaning of Index Numbers
- Uses of Index Numbers
- Understand how index numbers are prepared
- Problems to construct Index Numbers
- Different methods to construct Index Number

7.1 INTRODUCTION AND MEANING OF INDEX NUMBER

Human life is dynamic and hardly there is anything that remains the same over a period of time. whether it is the price of goods, Population of the country, Industrial Production, imports and Exports of the country, everything changes with the passage of time. It is the tendency of humans that he wants to measure the changes that are taking place over a period of time. Now questions arise about how we can measure these changes that are taking place. Index number is one such statistical tool that can help us in measuring these changes.

An index number is a statistical tool that measures the changes in the data over a period of time. Index number is not a new tool used in statistics, rather the use of index numbers is very old. As per available records, the index number was first time constructed in the year 1764 by an Italian named Carli. In his index number, Carli compared the prices of the Year 1750 with the price level of the year 1500. Though normally index numbers are used for measuring the change in price over a period of time, hardly there is any area in Economics or Commerce where Index numbers are not used. There are different types of index numbers that are used in economics such as Industrial Production Index, Agricultural Production Index and Population Index, etc.

An index number is a device with the help of which we can measure the relative change in one variable over some time. Normally while preparing the index number, we compare the current prices of a product with the price of some past period known as the base year. The index number of the base year is mostly taken as 100. A few definitions of index numbers given by different experts are as follows:

According to Croxton and Cowden, "Index numbers are devices for measuring differences in the magnitude of a group of related variables."

According to A.L. Bowley, "A series of index numbers reflects in its trend and fluctuations the movements of some quantity to which it is related."

According to Spiegel, "An Index number is a statistical measure designed to show changes in a variable, or a group of related variables with respect to time, geographic location, or other characteristics such as income, profession etc."

7.2 FEATURES OF INDEX NUMBER

1. Index numbers are specialized type of average. Normally used measures of average like Mean Median and Mode can be used for two or more different series, if their units are same. In case units of two series are different, these cannot be represented by normal average, However, Index numbers can help in this situation.
2. Normally index numbers are represented in percentages. However, the % sign is not used while showing index numbers.
3. Index numbers give the effect of change over some time or the change that is taking place in two different locations.
4. Index numbers measure those changes that are not capable of measurement normally in quantitative figures. For example, we cannot measure the change in the cost of living directly, but Index numbers can help us in this situation.

7.3 USES OF INDEX NUMBERS

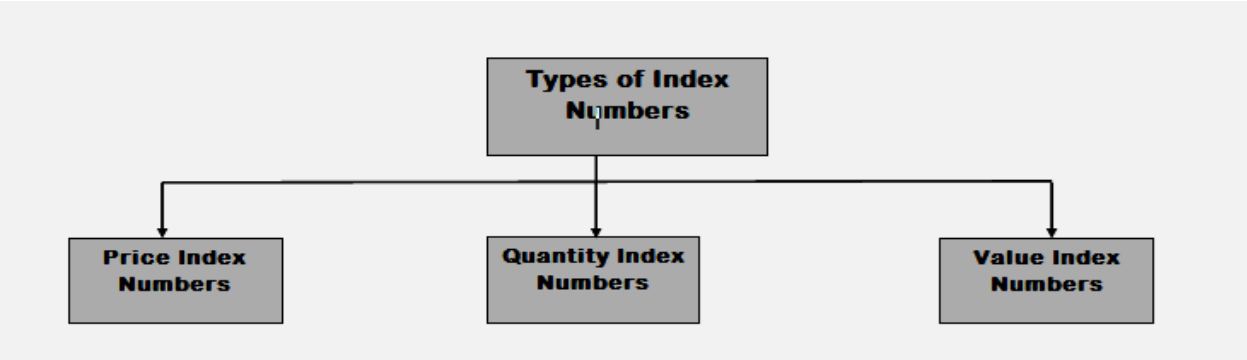
1. Index number is a very powerful tool for economic and business analysis. We often call index number 'Barometer of the Economy'. With the help of Index Numbers, we can see pulse of the economy.
2. Index number is a very helpful tool in planning activities and formulation of business policy.
3. With the help of index numbers, economists try to find out trends in prices, production, import and exports, etc.
4. Index number shows the cost of living over a period of time. This also helps the government in fixing the wage rate of labour.
5. Index number also helps us in the calculation of Real National Income of the country.

7.5 PROBLEMS IN CONSTRUCTION OF INDEX NUMBERS

1. **Purpose of Index Numbers**: The first step in construction of index numbers is to decide the purpose of preparing the Index Number. As there is no single purpose index, so we must decide the objective of index very carefully.
2. **Selection Of Base Year**: The selection of base period is most important step in preparation of index number. The best base period is the period with which we can find accurate change in the variable. Following are some of the guidelines that must be kept in mind while selecting the base period.
   - The period selected as base period should be normal one. There must not be any problems like War, Flood, Earthquake, Economic Depression etc. in the base period.
   - The difference between base period and the current period should not be very large
   - Only that period should be taken as base period full data is available.
3. **Selection of Number of Items or Commodities**: The next major problem in preparation for index number is to select the number of items that will form the Index number. The following points must be kept in mind while deciding the number of items in the index numbers.
   - Only those items should be selected that represent the habits and tastes of majority of customers.
   - The number of items selected should not be very large or very small.
   - Only those items should be selected that are available in standard quantity.
   - Only those items must be selected that were available in base period as well as current period.
4. **Selection of Source of Data:** As in index numbers, we compare current variables with the variables of past periods, the source from which data is collected must be authentic. In case of non-authentic data, it will give wrong picture.
5. **Price Quotations:** The prices of the commodities differ from place to place, It is very important to select the price which represents majority of places. Further while preparing index numbers, we may take wholesale prices or retail prices in consideration.
6. **Selection of the Average**: There are different types of averages, like Arithmetic Mean, Geometric Mean, Harmonic Mean, Median and Mode that can be used in preparation for index numbers. One must select appropriate averages in preparation of index numbers based on our objective.

7. **Selection of Appropriate Weight:** The next major problem in preparation of index numbers is to assign weight to the different items. All the items of the data under consideration are not equally important, some items may be more important and some items may be less important. So, more weight must be assigned to important items while preparing the index number. Now the problem is how to assign weights to the items. Normally we take quantity of the items consumed as weight in Index Number.
8. **Selection of appropriate formula:** There are several formulas that can be used for preparing index numbers. for example, Laspeyer's method. Bowle Method and Fisher Method etc. Each method has its advantages and limitations. so must be selected very carefully.

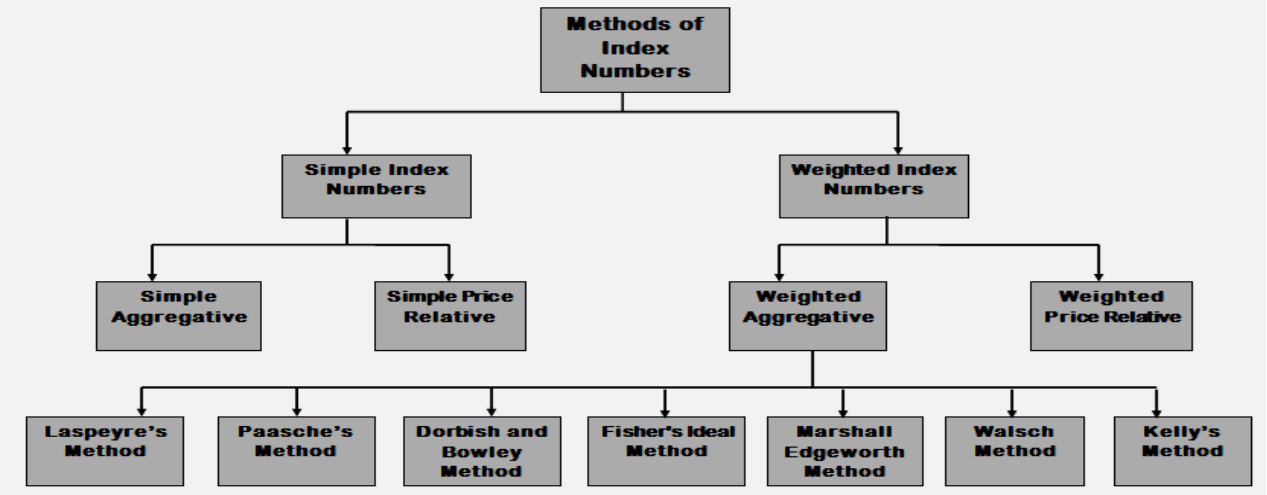
7.6 DIFFERENT TYPES OF INDEX NUMBERS



1. **Price Index Numbers**: These index numbers are used for measuring the change in prices of the commodities over a period of time. In other words, we can say that these index numbers find the change in value of money over a period of time. These index numbers are most popular index numbers. These Index numbers may be based on Wholesale Price Index or Retail Price Index.
2. **Quantity Index Numbers**: The Quantity or Volume Index Numbers measure the change in quantities used by people over a period of time. under these index numbers, we calculate change in physical quantity of goods produced, consumed or sold over a period of time. There are different types of quantity index numbers such as Agricultural Production Index Number, Industrial Production Index Number, Export Import Index Number etc.
3. **Value Index Numbers**: Value Index Numbers compare the change in total value over period of time. These index numbers take into consideration both prices and quantity of the product while finding the change over a period of time. These Index Numbers are very useful in finding consumption habits of the consumers.

7.7 DIFFERENT METHODS OF INDEX NUMBERS

As we have already discussed, Index number is a device that shows changes in price over a period of time. Now a question arises that how to calculate the index number. There are several methods for preparing the index numbers. The following chart shows various methods of preparing index numbers.



7.7.1 SIMPLE INDEX NUMBER

This further divided into the simple aggregative and simple price relative

7.7.1.1 SIMPLE AGGREGATIVE METHOD

This is one of the old and simple methods of finding the index number. Under this method we calculate the index number of a given period by dividing the aggregate of all the prices of the current year by the aggregate of all the prices of the base year. After that we multiply the resultant figure with 100 to find the index number. The following are the steps:

1. Decide the base year.
2. Add all the prices of base year for all available commodities, it is denoted by $\sum P_0$.
3. Add all the prices of base year for all available commodities, it is denoted by $\sum P_1$.
4. Use following the formula for calculating index number under this method:

$$P_{01} = \frac{\sum P_1}{\sum P_0} \times 100$$

Where,

P_{01} – Price Index Number of Current Year

$\sum P_1$ – Aggregate of Prices of Current Year

$\sum P_0$ – Aggregate of Prices of Base Year

Example 1. Construct Simple Aggregative Index number of the year 2020 by taking the base as prices of 2015.

Commodity	**Price of the Year 2015**	**Price of the Year 2020**
Wheat	**20**	**26**
Sugar	**40**	**34**
Oil	**60**	**120**
Pulses	**80**	**140**

Solution: Price Index (The year 2015 taken as the base year)

Commodity	Price of the Year 2015 P_0	Price of the Year 2020 P_1
Wheat	20	26
Sugar	40	34
Oil	60	120
Pulses	80	140
	$\sum P_0 = 200$	$\sum P_1 = 320$

Price Index (P_{01}) = $\frac{\sum P_1}{\sum P_0} \times 100$ = $\frac{320}{200} \times 100$ = 160

Price index shows that prices have increased by 60% in 2020 than 2015.

Merits of Simple Aggregative Method

1. This method is simple to calculate.
2. This method is very simple to understand.
3. This method does not need many mathematical calculations

Limitations of Simple Aggregative Method

1. This method does not give change in price over a period of time.
2. Prices of different commodities are measured in different units some are measured in Kilograms where as others in Meters etc. It creates problems in calculation.
3. This method is influenced by unit of measurement.
4. This method ignores the relative importance of the item.
5. This method uses only Arithmetic mean as a tool for calculating index number. Other measures of average like Geometric mean or median etc. cannot be used in this method.

6. Index number in this method is influenced by magnitude of the price.

7.7.1.2 SIMPLE PRICE RELATIVE METHOD

This method is a bit improved over the simple aggregative method. The simple aggregative method is affected by the magnitude of the price of the item. However, this method is not affected by magnitude of the price of item. Further, in this method it is not necessary to use Arithmetic mean as average rather we can use any method of finding average, such as Arithmetic mean, Geometric Mean, Median, Mode etc. However, normally we prefer to use Arithmetic mean in this case. Following are the steps of this method:

1. Decide the base year.
2. Calculate the price relative to current year for each commodity by dividing current Prices (P_1) with base year price (P_0) using the following formula $\frac{P_1}{P_0}. \times 100$
3. Find sum of all the price relatives so calculated.
4. Divide the sum or price relatives by number of items to get index number by using the following formula:

$$P_{01} = \frac{\sum \frac{P_1}{P_0} \times 100}{N}$$

Merits of Simple Price Relative Method:

1. This method is very simple to calculate and understand.
2. This method is not affected by the magnitude of price of a particular item.
3. This method is not affected by unit of measurement of the item.
4. This method is not necessarily based on Arithmetic Mean, we can use other averages like Geometric Mean, median etc also.
5. Equal weights are provided for each item.

Limitations of Simple Price Relative Method:

1. Selection of average is a difficult task in this method.
2. If it is to be calculated using Geometric Mean, then a calculation is very difficult.
3. It does not consider which item is used more and provides equal weights to all items.

Example 2. Construct Simple Price Relative Index number of the year 2020 by taking the

base as prices of 2015.

Commodity	Price of the Year 2015	Price of the Year 2020
Wheat	**20**	**26**
Sugar	**40**	**34**
Oil	**60**	**120**
Pulses	**80**	**140**

Solution: Price Index (Year 2015 taken as the base year)

Commodity	Price of the Year 2015 P_0	Price of the Year 2020 P_1	Price Relative $\frac{P_1}{P_0}. \times 100$
Wheat	20	26	$\frac{26}{20}. \times 100 = 130$
Sugar	40	34	$\frac{34}{40}. \times 100 = 85$
Oil	60	120	$\frac{120}{60}. \times 100 = 200$
Pulses	80	140	$\frac{140}{80}. \times 100 = 175$
			$\sum\frac{P_1}{P_0}. \times 100 = 590$

Price Index (P_{01}) = $\frac{\sum\frac{P_1}{P_0} \times 100}{N} = \frac{590}{4} = 147.50$

Price index shows that prices have increased by 47.5% in 2020 than 2015.

TEST YOUR PROGRESS (A)

1. Calculate Index number for 2015 taking 209 as base using Simple Aggregative Method and Simple Average of Relatives Method:

Items	Price 2011	Price 2015
A	350	510
B	45	40
C	77	156
D	37	47
E	10	12

2. Find index using simple average of price relative using 2017 as base.

Items	Price 2017	Price 2019
A	15	30
B	18	24

C	16	20
D	14	21
E	25	35
F	40	30

3. Find simple aggregative index

Items	P_0	P_1
Oil	60	70
Pulses	70	60
Rice	50	40
Sugar	40	40

Answers

1) 147.4, 132.84, 2) 137.22, 3) 95.45

7.7.2 WEIGHTED INDEX NUMBER

Which is further divided into weighted aggregative and weighted price relative method

7.7.2.1 WEIGHTED AGGREGATIVE PRICE INDEX

Simple Aggregative methods of Index Numbers assume that all the items of Index Number are equally important. There is no item that is more important than others. So, this method provides equal weightage to all items. However, in practical life it is not true. Some items carry more importance than other items, for example in human's life expenditure on food carries more importance than expenditure on entertainment. So, we have weighted method of index numbers which considers relative importance of the item also.

The weighted Aggregative Method is one such method. This method is more or less same as Simple Aggregative Method but main difference is that it also considers relative weights of the items also. Generally, the quantity of the item consumed is considered as weight in this case. There are many methods of calculating Weighted Aggregative Price Index which are discussed as follows:

a) Laspeyre's Method:

This method was suggested by Mr. Laspeyre in 1871. Under this method base year quantities of the various products are assumed as weight for preparing the index numbers. The following steps may be used:

1. Multiply Prices of the base year (P_0) with the quantities of the base year (Q_0) for every commodity.
2. Add the values calculated in step 1, the sum is denoted as $\sum P_0 Q_0$
3. Multiply Prices of the current year (P_1) with the quantities of the base year (Q_0) for every commodity.
4. Add the values calculated in step 3, the sum is denoted as $\sum P_1 Q_0$.
5. Use following formula for calculating index number:

$$P_{01} = \frac{\sum P_1 Q_0}{\sum P_0 Q_0} \times 100$$

b) Paasche's Method:

This method was suggested by Mr. Paasche in 1874. Under this method current year quantities of the various products are assumed as weight for preparing the index numbers. The following steps may be used:

1. Multiply Prices of the base year (P_0) with the quantities of the current year (Q_1) for every commodity.
2. Add the values calculated in step 1, the sum is denoted as $\sum P_0 Q_1$
3. Multiply Prices of the current year (P_1) with the quantities of the current year (Q_1) for every commodity.
4. Add the values calculated in step 3, the sum is denoted as $\sum P_1 Q_1$.
5. Use the following formula for calculating index number:

$$P_{01} = \frac{\sum P_1 Q_1}{\sum P_0 Q_1} \times 100$$

c) Dorbish and Bowley's Method:

This method is based on both Lasypeyre's Method and Paasche's Method, that's why this method is also known as L-P formula. Under this method, we calculate the index number by taking the arithmetic mean of the formula given by Laspeyre and Paasche. So, following formula is used in case of the Dorbish and Bowley Method:

$$P_{01} = \frac{\frac{\sum P_1 Q_0}{\sum P_0 Q_0} + \frac{\sum P_1 Q_1}{\sum P_0 Q_1}}{2} \times 100$$

d) Fisher's Ideal Index Method:

This method was suggested by Prof Irving Fisher and it is assumed as one of the best methods of

constructing the Index Number. That's why this method is also called Ideal Index Number. This method is based on both Lasypeyre's Method and Paasche's Method, but instead of taking arithmetic mean of both formulas, Fisher used the geometric mean of the formula given by Laspeyre and Paasche. So, following formula for calculating Fisher's ideal Index number:

$$\sqrt{\frac{\sum P_1 Q_0}{\sum P_0 Q_0} \times \frac{\sum P_1 Q_1}{\sum P_0 Q_1}} \times 100$$

Fisher's Method is called ideal index number due to following reasons:

1. This method uses geometric mean as base which is perhaps best average for constructing Index numbers.
2. This method considers both quantities of base year as well as current year as weight.
3. This method satisfies both the time reversal and factor reversal tests.
4. It is comprehensive method and covers all values of data i.e., P_0, Q_0, P_1, Q_1 etc.

e) Marshal Edgeworth Index Method:

Like Fisher's method, this method also uses the quantities of base as well as current year as weight. Under this method arithmetic mean of the quantity of base and current year is assumed as weight. This method is comparatively simple than Fisher's method as it does not use complex concept of Geometric mean. Following is the formula of this method.

$$P_{01} = \frac{\sum P_1 (Q_0 + Q_1)}{\sum P_0 (Q_0 + Q_1)} \times 100 \quad \text{or}$$

$$\frac{\sum P_1 Q_0 + \sum P_1 Q_1}{\sum P_0 Q_0 + \sum P_0 Q_1} \times 100$$

Example 3. Construct Weighted Aggregative Index number of the year 2020 by taking the base as prices of 2015 using Laspeyre, Paasche, Dorbish & Bowley, Fisher, Marshal Edgeworth and Kelly's method.

Item	Price of the The year 2015	Quantity of the The year 2015	Price of the Year 2020	Quantity of the Year 2020
A	6	50	10	56
B	2	100	2	120
C	4	60	6	60
D	10	30	12	24

E	8	40	12	36

Solution:

Item	P_0	Q_0	P_1	Q_1	P_0Q_0	P_0Q_1	P_1Q_0	P_1Q_1
A	6	50	10	56	300	336	500	560
B	2	100	2	120	200	240	200	240
C	4	60	6	60	240	240	360	360
D	10	30	12	24	300	240	360	288
E	8	40	12	36	320	288	480	432
					$\sum P_0Q_0$ = **1360**	$\sum P_0Q_1$ = **1344**	$\sum P_1Q_0$ = **1900**	$\sum P_1Q_1$ = **1880**

1. Laspeyre's Method:

$$P_{01} = \frac{\sum P_1Q_0}{\sum P_0Q_0} \times 100 = \frac{1900}{1360} \times 100 = 139.71$$

2. Paasche's Method:

$$P_{01} \quad \frac{\sum P_1Q_1}{\sum P_0Q_1} \times 100 = \frac{1880}{1344} \times 100 = 139.88$$

3. Dorbish and Bowley's Method:

$$P_{01} = \frac{\frac{\sum P_1Q_0}{\sum P_0Q_0} + \frac{\sum P_1Q_1}{\sum P_0Q_1}}{2} \times 100 = \frac{\frac{1900}{1360} + \frac{1880}{1344}}{2} \times 100 = \frac{2.796}{2} = 139.79$$

4. Fisher's Ideal Index Method:

$$\sqrt{\frac{\sum P_1Q_0}{\sum P_0Q_0} \times \frac{\sum P_1Q_1}{\sum P_0Q_1}} \times 100 = \sqrt{\frac{1900}{1360} \times \frac{1880}{1344}} \times 100 = \sqrt{1.9543} \times 100 = 139.79$$

5. Marshal Edgeworth Index Method: $\frac{\sum P_1Q_0 + \sum P_1Q_1}{\sum P_0Q_0 + \sum P_0Q_1} \times 100$

$$= \frac{1900 + 1880}{1360 + 1344} \times 100 = \frac{3780}{2704} \times 100 = 139.79$$

Example 4. Construct Weighted Aggregative Index number using Laspeyre, Paasche, Dorbish & Bowley and Fisher, methods.

Item	Price of the Base Year	Expenditure of the Base Year	Price of the Current Year	Expenditure of the Current Year
A	2	40	5	75
B	4	16	8	40
C	1	10	2	24

D	5	25	10	60

Solution:

We know that Expenditure = Price × Quantity

So, Quantity = $\frac{Expenditure}{Price}$

Item	P_0	Q_0	P_1	Q_1	P_0Q_0	P_0Q_1	P_1Q_0	P_1Q_1
A	2	20	5	15	40	30	100	75
B	4	4	8	5	16	20	32	40
C	1	10	2	12	10	12	20	24
D	5	5	10	6	25	30	50	60
					$\sum P_0Q_0$ = **91**	$\sum P_0Q_1$ = **92**	$\sum P_1Q_0$ = **202**	$\sum P_1Q_1$ = **199**

1. Laspeyre's Method:

$$P_{01} = \frac{\sum P_1Q_0}{\sum P_0Q_0} \times 100 = \frac{202}{91} \times 100 = 221.98$$

2. Paasche's Method:

$$P_{01} \quad \frac{\sum P_1Q_1}{\sum P_0Q_1} \times 100 = \frac{199}{92} \times 100 = 216.39$$

3. Dorbish and Bowley's Method:

$$P_{01} = \frac{\frac{\sum P_1Q_0}{\sum P_0Q_0} + \frac{\sum P_1Q_1}{\sum P_0Q_1}}{2} \times 100$$

$$= \frac{\frac{202}{91} + \frac{199}{92}}{2} \times 100 = \frac{4.3828}{2} = 219.14$$

4. Fisher's Ideal Index Method:

$$\sqrt{\frac{\sum P_1Q_0}{\sum P_0Q_0} \times \frac{\sum P_1Q_1}{\sum P_0Q_1}} \times 100$$

$$= \sqrt{\frac{202}{191} \times \frac{199}{92}} \times 100 = \sqrt{4.8015} \times 100 = 219.12$$

7.7.2.2 WEIGHTED PRICE RELATIVE METHOD

This method almost similar to simple price-relative method. However, simple price relative gives equal importance to all items under consideration. But in our life, all items do not carry equal importance. Some items are more important or on some items we spend more amount. Changes in price of some items affect us more than changes in price of some other items. So, we have weighted

price relative method. This method is similar to simple price relative method but also assigns weights to the items. Further, in this method it is not necessary to use Arithmetic mean as average rather we can use any method of finding average, such as Arithmetic mean, Geometric mean, etc. However, normally we prefer to use Arithmetic mean in this case. Following are the steps of this method:

1. Decide the base year.
2. Calculate the price relative to current year for each commodity by dividing current Prices (P_1) with base year price (P_0) using the following formula $\frac{P_1}{P_0}. \times 100.$
3. Find the weights of the items to be assigned.
4. Multiply price relative so calculated with the weights and find out the product of both.
5. Find sum of product so calculated.
6. Find sum of the weights assigned.
7. Divide the sum of the weighted price relatives by sum of weights to get index number by using the following formula:

$$P_{01} = \frac{\sum W \frac{P_1}{P_0} \times 100}{\sum W}$$

Merits of Weighted Price Relative Method:

1. This method is very simple to calculate and understand.
2. This method is not affected by the magnitude of price of a particular item.
3. This method is not affected by unit of measurement of the item.
4. This method is not necessarily based on Arithmetic Mean, we can use other averages like Geometric Mean, median etc also.
5. Weights are assigned according to importance of the items.

Limitations of Weighted Price Relative Method:

1. Selection of average is a difficult task in this method.
2. If it is to be calculated using Geometric Mean, then calculation is very difficult.
3. Selection of weights is a difficult task.

Example 5. Construct Weighted Price Relative Index number of the year 2020 by taking the base as prices of 2015.

Commodity	Price of the Year 2015	Price of the Year 2020	Weights
Wheat	**20**	**26**	**40**
Sugar	**40**	**34**	**5**
Oil	**60**	**120**	**3**
Pulses	**80**	**140**	**2**

Solution: Price Index (Year 2015 taken as the base year)

Commodity	Price of the Year 2015 P_0	Price of the Year 2020 P_1	Price Relative $\frac{P_1}{P_0}. \times 100$	Weights (W)	Weighted Price Relatives $W\frac{P_1}{P_0}. \times 100$
Wheat	20	26	$\frac{26}{20}. \times 100 = 130$	40	5200
Sugar	40	34	$\frac{34}{40}. \times 100 = 85$	5	424
Oil	60	120	$\frac{120}{60}. \times 100 = 200$	3	600
Pulses	80	140	$\frac{140}{80}. \times 100 = 175$	2	350
				$\sum W = 50$	$\sum W\frac{P_1}{P_0}. \times 100 = 6575$

Price Index (P_{01}) = $\frac{\sum W\frac{P_1}{P_0} \times 100}{\sum W} = \frac{6575}{50} = 131.50$

Price index shows that prices have increased by 31.5% in 2020 than 2015.

7.9 SUM UP

- Index number shows change in variable over a period of time.
- Price index shows change in price in current year in comparison to base year.
- Normally the base of index is taken as 100.
- There are different types of indexes like price index, quantity index, value index.
- Index number can be prepared without assigning weights or after assigning weights.
- Popular weighted aggregative index is Laspeyre, Paasche, Bowley, Fisher, Marshal Edgeworth and Kelly.

7.10 KEY TERMS

- **Index Numbers:** An index number is a device with the help of which we can measure the relative change in one variable over a period of time. Normally while preparing the index number, we compare the current year variable with the variable of as base year. The index number of the base year is mostly taken as 100
- **Price Index Numbers**: These index numbers are used for measuring the change in prices of the commodities over a period of time. In other words, we can say that these index numbers find the change in value of money over a period of time. These index numbers are most popular index number. These Index numbers may be based on Wholesale Price Index or Retail Price Index.
- **Quantity Index Numbers**: The Quantity or Volume Index Numbers measure the change in quantities used by people over a period of time. Under these index numbers, we calculate change in physical quantity of goods produced, consumed or sold over a period of time. There are different types of quantity index numbers such as Agricultural Production Index Number, Industrial Production Index Number, Export Import Index Number etc.
- **Value Index Numbers**: Value Index Numbers compare the change in total value over period of time. These index numbers take into consideration both prices and quantity of the product while finding the change over a period of time. These Index Numbers are very useful in finding consumption habits of the consumers.

7.11 QUESTIONS FOR PRACTICE

Q1.What are index numbers? What are its uses?

Q2.Explain problems faced in construction of index numbers.

Q3.What are different types of Index numbers.

Q4.Explain different steps in construction of index numbers.

Q5.What are different methods of construction of index numbers?

Q6.Explain Simple Aggregative Index numbers. What are its Merits and Limitations?

Q7.What is Simple Price Relative Index numbers? What are its Merits and Limitations?

Q8.Explain Weighted Aggregative Index numbers. What are its Merits and Limitations?

Q9.What is Weighted Price Relative Index numbers? What are its Merits and Limitations?

7.12 SUGGESTED READINGS

- J. K. Sharma, *Business Statistics*, Pearson Education.
- S.C. Gupta, *Fundamentals of Statistics*, Himalaya Publishing House.
- S.P. Gupta and Archana Gupta, *Elementary Statistics*, Sultan Chand and Sons, New Delhi.
- Richard Levin and David S. Rubin, *Statistics for Management*, Prentice Hall of India, New Delhi.

CERTIFICATE/DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SARM 2: DESCRIPTIVE STATISTICS

SEMESTER I

UNIT 8: TESTS OF CONSISTENCY OF INDEX NUMBER FORMULAE, CHAIN INDEX OR CHAIN BASE INDEX NUMBERS, BASE SHIFTING, SPLICING AND DEFLATION

STRUCTURE

8.0 Learning Objectives

8.1 Introduction

8.2 Test of Consistency for Index Numbers

- **8.2.1 Unit Test**
- **8.2.2 Time Reversal Test**
- **8.2.3 Factor Reversal Test**
- **8.2.4 Circular Test**

8.3 Chain Index or Chain Base Index Numbers

8.4 Base Shifting

8.5 Splicing and Deflation

8.6 Limitations of Index Numbers

8.7 Applications of Index Number

8.8 Key Terms

8.9 Sum Up

8.10 Suggested Readings

8.0 LEARNING OBJECTIVES

After studying this unit, the learner will able to know about:

- Test of Consistency for Index Numbers
- Chain Index or Chain Base Index Numbers

- Base Shifting
- Splicing and Deflation

8.1 INTRODUCTION

Many of you must also be aware of the Stock Exchange Share Price Index – commonly referred to as BSE SENSEX or, more recently, NSE SENSEX. In fact, these various types of index series have come to be used in many activities such as industrial production, export, prices, etc. In this Unit, you will study and understand the meaning and uses of index numbers, various problems resulting from the incorrect use of index numbers, methods for construction of various index numbers, and their limitations. There are a number of methods through which index numbers can be calculated. Each method has its own merits and demerits. Now the question is which of these methods can be treated as best. In order to find out which method is better than the others, there are four tests. If any index number satisfies these tests, we may consider the index number to be the ideal one.

8.2 TESTS OF CONSISTENCY FOR INDEX NUMBERS

8.2.1 Unit Test

Unit test says that any index number can be treated as ideal only if it is free from the unit in which quantity is measured. Whether prices are quoted for a single item or for dozen items, the index number must not be affected by the same. Only a simple average of price relative method satisfies this condition.

8.2.2 Time Reversal Test

This test was suggested by Fisher. According to this test, an ideal index number is one that works both ways that i.e., backward and forward. So, if index is prepared by taking old period as base year and new period as current year it comes to be 200 it means prices in current period are doubled. Now say reverse is done, new period is taken as base and old period is taken as current year, this test says that index should be 50 which means earlier prices were half of current prices. In other words, we can say that the following conditions should be satisfied

$$\boxed{P_{01} \times P_{10} = 1}$$

Following is the formula of the time reversal test in different cases:

1. **Laspeyre's Method:**

$$P_{01} \times P_{10} = \frac{\sum P_1Q_0}{\sum P_0Q_0} \times \frac{\sum P_0Q_1}{\sum P_1Q_1} \neq 1$$

This method does not satisfy time reversal test.

2. Paasche's Method:

$$P_{01} \times P_{10} = \frac{\sum P_1Q_1}{\sum P_0Q_1} \times \frac{\sum P_0Q_0}{\sum P_1Q_0} \neq 1$$

This method does not satisfy time reversal test.

3. Dorbish and Bowley's Method:

$$P_{01} \times P_{10} = \frac{\frac{\sum P_1Q_0}{\sum P_0Q_0} + \frac{\sum P_1Q_1}{\sum P_0Q_1}}{2} \times \frac{\frac{\sum P_0Q_1}{\sum P_1Q_1} + \frac{\sum P_0Q_0}{\sum P_1Q_0}}{2} \neq 1$$

This method does not satisfy time reversal test.

4. Fisher's Ideal Index Method:

$$P_{01} \times P_{10} = \sqrt{\frac{\sum P_1Q_0}{\sum P_0Q_0} \times \frac{\sum P_1Q_1}{\sum P_0Q_1}} \times \sqrt{\frac{\sum P_0Q_1}{\sum P_1Q_1} \times \frac{\sum P_0Q_0}{\sum P_1Q_0}} = 1$$

This method satisfies time reversal test.

5. Marshal Edgeworth Index Method:

$$P_{01} \times P_{10} = \frac{\sum P_1Q_0 + \sum P_1Q_1}{\sum P_0Q_0 + \sum P_0Q_1} \times \frac{\sum P_0Q_1 + \sum P_0Q_0}{\sum P_1Q_1 + \sum P_1Q_0} = 1$$

This method satisfies time reversal test.

8.2.3 Factor Reversal Test (F.R.T.)

This test was also suggested by Fisher. According to this test, an ideal index number is one which does not give inconsistent result if we change price with quantity and quantity with price. According to this test when we multiply change in price with change in quantity the ratio must be equal to total change in value.

$$\boxed{P_{01} \times Q_{01} = \frac{\sum P_1Q_1}{\sum P_0Q_0}}$$

Following is the formula of factor reversal test in different cases:

1. Laspeyre's Method:

$$P_{01} \times Q_{10} = \frac{\sum P_1Q_0}{\sum P_0Q_0} \times \frac{\sum Q_1P_0}{\sum Q_0P_0} \neq \frac{\sum P_1Q_1}{\sum P_0Q_0}$$

This method does not satisfy factor reversal test.

2. Paasche's Method:

$$P_{01} \times Q_{10} = \frac{\sum P_1 Q_1}{\sum P_0 Q_1} \times \frac{\sum Q_1 P_1}{\sum Q_0 P_1} \neq \frac{\sum P_1 Q_1}{\sum P_0 Q_0}$$

This method does not satisfy factor reversal test.

3. Dorbish and Bowley's Method:

$$P_{01} \times Q_{10} = \frac{\frac{\sum P_1 Q_0}{\sum P_0 Q_0} + \frac{\sum P_1 Q_1}{\sum P_0 Q_1}}{2} \times \frac{\frac{\sum Q_1 P_0}{\sum Q_0 P_0} + \frac{\sum Q_1 P_1}{\sum Q_0 P_1}}{2} \neq \frac{\sum P_1 Q_1}{\sum P_0 Q_0}$$

This method does not satisfy factor reversal test.

4. Fisher's Ideal Index Method:

$$P_{01} \times Q_{10} = \sqrt{\frac{\sum P_1 Q_0}{\sum P_0 Q_0} \times \frac{\sum P_1 Q_1}{\sum P_0 Q_1}} \times \sqrt{\frac{\sum Q_1 P_0}{\sum Q_0 P_0} \times \frac{\sum Q_1 P_1}{\sum Q_0 P_1}} = \frac{\sum P_1 Q_1}{\sum P_0 Q_0}$$

This method satisfies factor reversal test.

5. Marshal Edgeworth Index Method:

$$P_{01} \times Q_{10} = \frac{\sum P_1 Q_0 + \sum P_1 Q_1}{\sum P_0 Q_0 + \sum P_0 Q_1} \times \frac{\sum Q_1 P_0 + \sum Q_1 P_1}{\sum Q_0 P_0 + \sum Q_0 P_1} = \frac{\sum P_1 Q_1}{\sum P_0 Q_0}$$

This method does not satisfy factor reversal test.

8.2.4 Circular Test: Circular test was given by Wester Guard. This test is like Time Reversal test but applied to more number of years. According to this test if data of the different periods is compared by shifting the base, we should be able to get the index of any period by correlating the different base periods used.

Symbolically, $\frac{P_1}{P_0} * \frac{P_2}{P_1} * \frac{P_3}{P_2} = 1$

Only Simple Aggregative, Simple Geometric Mean of price relatives and Kelly's index meet this criterion.

Example 1. Construct Weighted Aggregative Index number using Laspeyre, Paasche, and Fisher method also check whether these satisfy T.R.T. and F.R.T or not?

Item	Price of the Base Year	Qty. of the Base Year	Price of the Current Year	Qty. of the Current Year
A	**30**	**7**	**40**	**5**
B	**40**	**12**	**60**	**8**
C	**60**	**10**	**50**	**15**
D	**30**	**15**	**20**	**18**

Solution: We know that Expenditure = Price × Quantity

So, Quantity = $\frac{Expenditure}{Price}$

Item	P_0	Q_0	P_1	Q_1	P_0Q_0	P_0Q_1	P_1Q_0	P_1Q_1
A	30	7	40	5	210	150	280	200
B	40	12	60	8	480	320	720	480
C	60	10	50	15	600	900	500	750
D	30	15	20	18	450	540	300	360
					$\sum P_0Q_0$ = **1740**	$\sum P_0Q_1$ = **1910**	$\sum P_1Q_0$ = **1800**	$\sum P_1Q_1$ = **1790**

1. Laspeyre's Method:

$$P_{01} = \frac{\sum P_1Q_0}{\sum P_0Q_0} \times 100 = \frac{1800}{1740} \times 100 = 103.45$$

Time Reversal Test

$$\frac{\sum P_1Q_0}{\sum P_0Q_0} \times \frac{\sum P_0Q_1}{\sum P_1Q_1} = \frac{1800}{1740} \times \frac{1910}{1790} \neq 1$$

It does not satisfy time reversal test.

Factor Reversal Test

$$\frac{\sum P_1Q_0}{\sum P_0Q_0} \times \frac{\sum Q_1P_0}{\sum Q_0P_0} = \frac{1800}{1740} \times \frac{1910}{1740} \neq \frac{1790}{1740}$$

It does not satisfy facto reversal test.

2. Paasche's Method:

$$P_{01} \quad \frac{\sum P_1Q_1}{\sum P_0Q_1} \times 100 = \frac{1790}{1910} \times 100 = 93.72$$

Time Reversal Test

$$= \frac{\sum P_1Q_1}{\sum P_0Q_1} \times \frac{\sum P_0Q_0}{\sum P_1Q_0} = \frac{1790}{1910} \times \frac{1740}{1800} \neq 1$$

It does not satisfy time reversal test.

Factor Reversal Test

$$\frac{\sum P_1Q_1}{\sum P_0Q_1} \times \frac{\sum Q_1P_1}{\sum Q_0P_1} = \frac{1790}{1910} \times \frac{1790}{1800} \neq \frac{1790}{1740}$$

It does not satisfy factor reversal test.

3. Fisher's Ideal Index Method:

$$\sqrt{\frac{\sum P_1Q_0}{\sum P_0Q_0} \times \frac{\sum P_1Q_1}{\sum P_0Q_1}} \times 100$$

$$= \sqrt{\frac{1800}{1740} \times \frac{1790}{1910}} \times 100 = \sqrt{.96948} \times 100 = 98.462$$

Time Reversal Test

$$= \sqrt{\frac{\sum P_1Q_0}{\sum P_0Q_0} \times \frac{\sum P_1Q_1}{\sum P_0Q_1}} \times \sqrt{\frac{\sum P_0Q_1}{\sum P_1Q_1} \times \frac{\sum P_0Q_0}{\sum P_1Q_0}} \neq 1$$

$$= \sqrt{\frac{1800}{1740} \times \frac{1790}{1910}} \times \sqrt{\frac{1910}{1790} \times \frac{1740}{1800}} = 1$$

It satisfies time reversal test.

Factor Reversal Test

$$\sqrt{\frac{\sum P_1Q_0}{\sum P_0Q_0} \times \frac{\sum P_1Q_1}{\sum P_0Q_1}} \times \sqrt{\frac{\sum Q_1P_0}{\sum Q_0P_0} \times \frac{\sum Q_1P_1}{\sum Q_0P_1}} = \frac{\sum P_1Q_1}{\sum P_0Q_0}$$

$$\sqrt{\frac{1800}{1740} \times \frac{1790}{1910}} \times \sqrt{\frac{1910}{1740} \times \frac{1790}{1800}} = \frac{1790}{1740}$$

It satisfies facto reversal test.

TEST YOUR PROGRESS

1. Using the following data, construct Fisher's Ideal index and show how it satisfies Factor Reversal Test and Time Reversal Test?

Commodity	Price in Rupees Per Unit		Number of Units	
	Base year	Current Year	Base year	Current Year
A	6	10	50	56
B	2	2	100	120
C	4	6	60	60
D	10	12	50	24
E	8	12	40	36

2. Apply Laspeyre, Paasche and Fisher Method on the following data and check whether these methods satisfy Time Reversal and Factor Reversal Test or not?

Item	P_0	Q_0	P_1	Q_1
A	5	15	5	5
B	7	5	4	3
C	8	6	6	10
D	3	8	3	4

3. Calculate Fisher's index number to the following data. Also show that it satisfies Time Reversal Test.

Commodity	2016		2017	
	Price (Rs.)	Quantity (Kg)	Price (Rs.)	Quantity (Kg)
A.	40	12	65	14
B.	72	14	78	20
C.	36	10	36	15
D.	20	6	42	4
E.	46	8	52	6

Answer:

1. Fisher's IN = 138.5, TRT= 1, FRT = 1880/1560
2. L= 85.165, P = 86.232, F = 85.697, only Fisher method satisfy both tests.
3. Fisher's IN = 103 TRT = 1

8.3 CHAIN INDEX OR CHAIN BASE INDEX NUMBERS

When the data are available for more than two years, the method available is the chain base method. Here, figures for each year expressed as a percentage of the previous year, which is known as Link Relatives. Link relative is a price (or quantity) relative with the condition that the base year is the preceding year. Whenever more than one commodity is considered, the link relatives of all the commodities are averaged (simple or weighted).

In other words, the link relatives as well as their averages are index numbers in which for each year the preceding year is the base year. These averages of link relatives show the conditions of the different years in comparison with their preceding years and are found to be of great use by businessmen and industrialists.

Steps in the construction of Chain Index Numbers

1. Calculate the link relatives by expressing the figures as the percentage of the preceding year. Thus,

   $$\text{Link Relatives of current year} = \frac{\text{price of current year}}{\text{price of previous year}} \times 100$$

2. Calculate the chain index by applying the following formula:

$$\text{Chain Index} = \frac{\text{Current year link relative} * \text{preceding year chain index}}{}$$

As long as the base year is common, the chain base indices are likely to be same as the fixed base indices. Sometimes, we may wish to convert chain base indices (C.B.I.) to fixed base indices (F.B.I.) (where in the bases become different) or vice versa.

So, we then need to chain them together by successive multiplication to form a chain index. Thus, unlike fixed base methods, in this method, the base year changes every year. Hence, for the year 2001, it will be 2000, for 2002 it will be 2001, and so on. Let us now study this method step by step.

Advantages of Chain Index Numbers Method

- It allows the addition or introduction of the new items in the series and also the deletion of obsolete items.
- Management in an organisation typically contrasts the current period with the one that occurred right before it rather than any other period in the past. This method is more practical for management because the base year fluctuates every year.

Disadvantages of Chain Index Numbers Method

- If the data for any one of the years is not available then we cannot compute the chain index number for the succeeding period. This is so because we need to calculate the link relatives, is based on previous one.
- In the event that one of the link relatives is calculated incorrectly, all following link relatives will likewise be calculated incorrectly due to the compounding effect of the inaccurate information. As a result, the entire series will provide an inaccurate picture.

8.4 BASE SHIFTING

Base shifting means the changing of the given base period (year) of a series of index numbers and recasting them into a new series based on some recent new base period. This step is quite often necessary under the following situations:

- When the base year is too old or too distant from the current period to make meaningful and valid comparisons. The base year should be normal year of economic stability not too far distant from the given year.

- we want to compare series of index numbers with different base periods, to make quick and valid comparisons both the series must be expressed with a common base period.

Base shifting requires the recomputation of the entire series of the index numbers with the new base However, this is a very difficult and time-consuming job. A relatively much simple, though approximate method consists in taking the index number of the new base year as 100 and then expressing the given series of index numbers as a percentage of the index number of the time period selected as the new base year. Thus, the series of index numbers, recast with a new base is obtained by the formula

Recast Index number of any year = $\frac{Old\ Index\ No.of\ the\ Year}{Index\ no.of\ new\ base\ Year} \times 100$

Reconstruct the following Index using 1980 as base year

Years	1977	1978	1979	1980	1981	1982	1983
Index no.	110	130	150	175	180	200	220

Solution:

Year	Index No.	Index No. (Base 1980=100)
1977	110	$\frac{100}{175} \times 110 = 62.85$
1978	130	$\frac{100}{175} \times 130 = 74.28$
1979	150	$\frac{100}{175} \times 150 = 85.71$
1980	175	100
1981	180	$\frac{100}{175} \times 180 = 102.85$
1982	200	$\frac{100}{175} \times 200 = 114.28$
1983	220	$\frac{100}{175} \times 220 = 125.71$

8.5 SPLICING AND DEFLATION

Splicing is a technique where we link the two or more index number series which contain the same items and a common overlapping year but with different base year to form a continuous series. It may be forward splicing or backward splicing.

Sometimes, Splicing is a specific situation may arise for shifting the base period of an index number series to some recent period. For instance, in course of time a few commodities which are

being considered for constructing indices may get replaced with new commodities, as a result their relative weightage may also change. In some cases, the weights may have become outdated and we may take into account the revised weights. Consequently, whatever be the reasons, index number series loses continuity and now we have two different index number series with different base periods which are not directly comparable. It is, therefore, essential to connect these two different series of indices into one continuous series.

Therefore, the statistical procedure involved in connecting these two series of indices to make continuity is termed as 'Splicing'. Thus, splicing means reducing two overlapping series of indices with different base periods into a continuous index number series. In equation form, we can say,

$$\text{Spliced Index Numbers} = \times \frac{\text{New Index No.of current period} \times \textit{Old Index No.of New base Period}}{100}$$

The following example would illustrate the procedure of splicing:

Given below are two sets of indices, one with 1990 as base and the other with 1993 as base.

Year	**1990**	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**
Consumer Price Index (1990=Base) (Old Index No. Series)	100	130	170	200			
Consumer Price Index (1994=base) (New Index No. Series)				100	120	115	125

Splicing the New Series of Indices into the Old Series of Indices (Backward)

Year	Consumer Price Index (1990=Base) (Old Index No. Series)	Consumer Price Index (1994=base) (New Index No. Series)	Spliced Consumer Index (New Index	
1990	100			100
1991	130			130
1992	170			170
1993	200	100	$\frac{100\times200}{100} = 200$	200
1994		120	$\frac{120\times200}{100} = 240$	240

1995		115	$\frac{115\times200}{100}=230$	230
1996		125	$\frac{125\times200}{100}=250$	250

Splicing the New Series of Indices into the Old Series of Indices (Forward)

Year	Consumer Price Index (1990=Base) (Old Index No. Series)	Consumer Price Index (1994=base) (New Index No. Series)	Spliced Consumer Index (New Index 200/100)	
1990	100		$\frac{100\times100}{100}=50$	50
1991	130		$\frac{130\times100}{100}=65$	65
1992	170		$\frac{170\times100}{100}=85$	85
1993	200	100	$\frac{200\times100}{100}=100$	100
1994		120		120
1995		115		115
1996		125		125

As far as deflating is concerned, we know the fact that there is an indirect relationship between price of good and purchasing power, as the price of goods gradually increases, as a result the purchasing power of money (value of money) decreases. Therefore, the real wages become less than the money wage. In such a situation the real wage may be obtained by reducing the money wage to the extent the price level has risen. Thus, the process of finding out the real wage by applying appropriate price indices to the money wages so as to allow for the changes in the price level is called 'deflating'.

The formula for real wage is:

$$\frac{Money\ Wage}{Price\ Index}\times 100$$

The formula for Real Wage Index Number is

$$\frac{\text{Current period's real wage}}{\text{base period's real wage}}\times 100$$

Below mentioned example consisting of the following data related to wages and price index of different years. It would illustrate the procedure of constructing real wage index numbers.

Example: Construction of Real Wage Index

Year	**Wages (Rs.)**	**Price Index**	**Real Wage (Deflated Income)**	**Real Wage Index (1990=100)**
1990	200	100	$\frac{200}{100} \times 100 = 200$	$\frac{200}{200} \times 100 = 100$
1991	280	130	$\frac{280}{130} \times 100 = 215.38$	$\frac{215.38}{200} \times 100 = 107.69$
1992	280	135	$\frac{280}{135} \times 100 = 207.41$	$\frac{207.41}{200} \times 100 = 103.70$
1993	340	150	$\frac{340}{150} \times 100 = 226.67$	$\frac{226.67}{200} \times 100 = 113.33$
1994	360	160	$\frac{360}{160} \times 100 = 225$	$\frac{225}{200} \times 100 = 112.5$
1995	380	160	$\frac{380}{160} \times 100 = 237.5$	$\frac{237.5}{200} \times 100 = 118.75$

8.6 LIMITATIONS OF INDEX NUMBERS

1. It estimates relative changes only, which may or may not approximate indicators. It represents the generalized truth based on the overall average of the items. Therefore, it does not apply to specific units.
2. It does not consider every item as based on the sample items. Thus, in case of an inadequate sample or a sample selected through a faulty process, there will be inaccurate results.
3. It does not pay any attention to the qualitative changes in the product while constructing quantitative changes in the form of the price or production. An increase in the price possibly results from the improvement in the quality of the product. But it is neglected in the index numbers.
4. It can be created in a way that allows for manipulation. This manipulation can be made in a selection of a particular base year, a particular group of commodities, a specific set of prices, etc.
5. Index number not fully true because they are simply rough indications (approximations) of the relative changes. The index number simply indicate arithmetical tendency of the temporal changes in the variable. The choice of representative commodities may lead to mistaken conclusions as they are based on samples. If there are errors in the choice of base periods or weights, etc. then result is not reliable.

6. The choice of variables included in an index can introduce bias if it does not accurately represent the entire population or if certain items are overrepresented or underrepresented.
7. International comparisons are difficult in index numbers on the account of different bases, different sets of commodities or difference in their quality or quantity. Hence index numbers do not help international comparisons.
8. Comparison between different times is not easy. Over long periods, some popular commodities are replaced by others or consumption habits of people got changed. With the passage of time, it is difficult to make comparison of index number. Thus, comparisons of changes in variables over long periods are not reliable.
9. Most of the index numbers are prepared on the basis of wholesale prices. But in real life, retail prices are most relevant, but it is difficult to collect retail prices.

8.7 APPLICATIONS OF INDEX NUMBER

1. It used in the fields of commerce, metrology, labour industry etc.
2. Used to study the difference between the comparable categories of animals, people or items.
3. Used for industrial production to measure the change in the level of industrial production in the country.
4. Index number used for import prices and export prices are used to measure the changes in the trade of a country.
5. Index numbers used to measure the fluctuation during intervals of time, group differences of geographical position of degree etc.
6. It used to compare to total variations in the prices of different commodities in which the unit of measurements differs with time and price.
7. It used to measure the purchasing power of money.
8. Helpful in forecasting future economic trends.

8.7 LET US SUM UP

- There is test to check consistency of the index number.
- Only Fisher index satisfy Time Reversal and Factor Reversal tests.
- Consumer price index shows change in cost of living of the consumer.

8.8 KEY TERMS

- **Time Reversal Test (T.R.T.):** This test was suggested by Fisher. According to this test, an ideal index number is one which works both ways that i.e., backward and forward. In other words, we can say that following condition should be satisfied: $\mathbf{P_{01} \times P_{10} = 1}$
- **Factor Reversal Test (F.R.T.):** This test was also suggested by Fisher. According to this test, an ideal index number is one which does not give inconsistent result if we change price with quantity and quantity with price. According to this test when we multiply change in price with change in quantity the ratio must be equal to total change in value.

$$\mathbf{P_{01} \times Q_{01}} = \frac{\sum P_1 Q_1}{\sum P_0 Q_0}$$

- **Splicing:** The process of connecting the series of index numbers of old base period with the series of index numbers of new base period is called splicing. Splicing may be done in two ways: one is splicing the new series of indices with old series of indices. Another is splicing the old series of indices with new series of indices.
- **Deflating:** Deflating means the process of finding out the real wage by applying appropriate price indices to the money wage so as to allow for the changes in the price level.

8.9 QUESTIONS FOR PRACTICE

Q1. What are tests of consistency of index numbers. Give various tests of consistency.

Q2. Explain Time Reversal and Factor Reversal Test.

Q3. Why Fisher's Index is known as Ideal Index Number.

Q4. Briefly explain different methods for construction of indices and their limitations.

Q5. Why do we consider Fisher's index as an ideal index?

Q6. Write short notes on:

a) Price Index

b) Quantity Index

c) Splicing of Indices

d) Deflating of Indices

8.10 FURTHER READINGS

- J. K. Sharma, *Business Statistics,* Pearson Education.
- S.C. Gupta, *Fundamentals of Statistics,* Himalaya Publishing House.

- S.P. Gupta and Archana Gupta, *Elementary Statistics*, Sultan Chand and Sons, New Delhi.
- Richard Levin and David S. Rubin, *Statistics for Management*, Prentice Hall of India, New Delhi.
- Hooda, R.P, 2001. *Statistics for Business and Economics*, Macmillan India Ltd.
- Gupta, S.P., *Statistical Methods*, 2000, Sultan Chand and Sons.
- Gupta, C.B. and Vijay Gupta, 2001. *An Introduction to Statistical Methods*, Vikas Publishing House Pvt. Ltd., New Delhi.