

Basic Statistics

BA - Economics

I Year

Paper Code: BAEC 1914



Pondicherry University
(A Central University)

Directorate of Distance Education
R.V. Nagar, Kalapet, Puducherry – 605 014

Advisory Committee

1. Prof. Gurmeet Singh
*Vice-Chancellor,
Pondicherry University.*
2. Prof. Rajeev Jain
*OSD, C&CR,
Pondicherry University.*
3. Prof. C.K. Ramaiah
*Director, Directorate of Distance Education
Pondicherry University.*

Reviewer Committee

1. Prof. C.K. Ramaiah
*Director, DDE
Pondicherry University.*
2. Prof. V. Nirmala
*Programme Coordinator
Department of Economics
Pondicherry University.*

Academic Support Committee

1. Dr. A. Punitha
*Asst. Professor, DDE
Pondicherry University.*
2. Dr. V. Umasri
*Asst. Professor, DDE
Pondicherry University.*
3. Dr. Sk. Md. Nizamuddin
*Asst. Professor, DDE
Pondicherry University.*

Administrative Support Committee

1. Lt Cdr Raj Kumar
*Deputy Registrar, DDE
Pondicherry University.*
2. Dr. Arvind Gupta
*Asst. Director, DDE
Pondicherry University.*

COURSE WRITER

B.M. Aggarwal, Visiting Professor and Author of Books in the Area of Mathematics, Statistics and Operations Research.

© This book may not be duplicated in any way without the written consent of the publisher and Pondicherry University except in the form of brief excerpts or quotations for the purpose of review. The information contained herein is for the personal use of the DDE students, Pondicherry University and may not be incorporated in any commercial programs, other books, databases, or any kind of software without written consent of the publisher. Making copies of this book or any portion, for any purpose other than your own is a violation of copyright laws. The author and publisher have used their best efforts in preparing this book and believe that the content is reliable and correct to the best of their knowledge.

Printed and Published by:

Mrs. Meena Pandey

Himalaya Publishing House Pvt. Ltd.,

"Ramdoot", Dr. Bhalerao Marg, Girgaon, Mumbai - 400 004.

Phone: 022-23860170, 23863863; **Fax:** 022-23877178

E-mail: himpub@bharatmail.co.in; **Website:** www.himpub.com

For:

Pondicherry University

Directorate of Distance Education,

R.V. Nagar, Kalapet, Puducherry – 605 014.

Tel. 0413-2654 439/440; E-mail: director.dde@pondiuni.edu.in

Website: <https://dde.pondiuni.edu.in>

SYLLABUS - BOOK MAPPING TABLE

Basic Statistics

Syllabus	Mapping in Book
Unit I: Measures of Central Tendency: Meaning of average – types of average: arithmetic mean, median, mode, geometric mean, harmonic mean, quartiles, deciles and percentiles.	Unit I: Measures of Central Tendency (Pages 1 – 49)
Unit II: Measures of Dispersion: Meaning of dispersion – types of dispersion: range, quartile deviation, mean deviation, standard deviation and variance (along with absolute measure, the relative measure or coefficient of each type of dispersion) – coefficient of variation – combined standard deviation.	Unit II: Measures of Dispersion (Pages 50 – 69)
Unit III: Skewness and Kurtosis: Skewness – meaning of skewness and symmetry in a distribution – symmetrical distribution – asymmetrical or skewed distribution – negatively skewed and positively skewed – definition, types and measures of kurtosis.	Unit III: Skewness and Kurtosis (Pages 70 – 82)
Unit IV: Correlation and Regression: Concept of correlation – types of correlation – simple correlation – Karl Pearson's product moment coefficient of correlation measure – partial correlation: definition and measure – multiple correlation: definition and measure – Spearman's rank correlation coefficient (when ranks are given, when ranks are not given and when equal ranks are given) – properties and uses of correlation – Introduction to Regression.	Unit IV: Correlation and Regression (Pages 83 – 121)
Unit V: Index Numbers: Definition of index number – types of index number – price index – quantity index – value index – simple index number – weighted index number – construction of index number – problems in construction – methods in construction – simple and weighted – Laspeyre's price index (CPI in India) – Paasche's price index – Fisher's ideal index – splicing of index number – deflating (finding real wages).	Unit V: Index Numbers (Pages 122 – 164)

CONTENTS

Unit I : Measures of Central Tendency

1 – 49

- 1.1 Introduction
- 1.2 Definitions
- 1.3 Types of Averages
- 1.4 Importance (Significance) of Measures of Central Tendency
- 1.5 Arithmetic Mean (A.M.) for Ungrouped Data
- 1.6 Combined Mean of Two Groups
- 1.7 Weighted A.M.
- 1.8 Properties of Arithmetic Mean
- 1.9 Arithmetic Mean for Continuous Series or Grouped Data
- 1.10 Steps to Find the A.M. in Grouped Data
- 1.11 Median (General Introduction)
- 1.12 Median for Grouped Data
- 1.13 The Steps for Finding the Median for Grouped Data
- 1.14 Mode (Introduction)
- 1.15 Definitions of Mode
- 1.16 Mode for Ungrouped Data
- 1.17 Determination of Mode in a Continuous Series
- 1.18 Empirical Relation among Mean, Median and Mode
- 1.19 Merits, Demerits and Uses of Mean, Median and Mode
- 1.20 Computations of Quartiles, Deciles and Percentiles
- 1.21 Answers to 'Check Your Progress'
- 1.22 Summary
- 1.23 Key Terms
- 1.24 Self-Assessment Questions and Exercises
- 1.25 References

Unit II: Measures of Dispersion

50 – 69

- 2.1 Introduction
- 2.2 Definition
- 2.3 Objectives of Dispersion
- 2.4 Importance of Dispersion
- 2.5 Characteristics of a Good Measure of Dispersion
- 2.6 Merits and Demerits of Measures of Dispersion
- 2.7 The Range

- 2.8 The Interquartile Range or the Quartile Deviation
- 2.9 The Mean Deviation
- 2.10 Variance Standard Deviation
- 2.11 Coefficient of Variation
- 2.12 Answers to 'Check Your Progress'
- 2.13 Summary
- 2.14 Key Terms
- 2.15 Self-Assessment Questions and Exercises
- 2.16 References

Unit III: Skewness and Kurtosis

70 – 82

- 3.1 Introduction
- 3.2 Definitions
- 3.3 Difference between Dispersion and Skewness
- 3.4 Positively Skewed and Negatively Skewed Distributions
- 3.5 Comparison between Symmetrical and Skewed Distributions
- 3.6 Measures of Skewness (Coefficient of Skewness)
- 3.7 Kurtosis
- 3.8 Answers to 'Check Your Progress'
- 3.9 Summary
- 3.10 Key Terms
- 3.11 Self-Assessment Questions and Exercises
- 3.12 References

Unit IV: Correlation and Regression

83 – 121

- 4.1 Introduction
- 4.2 Definition of Correlation
- 4.3 Importance (or Utility) of Correlation
- 4.4 Kinds of Correlation
- 4.5 Positive and Negative Correlation
- 4.6 Linear and Non-linear Correlation
- 4.7 Correlation Based on Number of Variables
- 4.8 Some Important Points about the Study of Correlation Analysis
- 4.9 Measures of Correlation
- 4.10 Properties of Correlation Coefficient
- 4.11 Probable Error

- 4.12 Merits and Demerits of Rank Correlation Method
- 4.13 Regression Analysis
- 4.14 Regression Lines or Regression Equations
- 4.15 Properties of Regression Coefficients
- 4.16 Why There are Two Regression Lines?
- 4.17 Examples of Irreversible Relation
- 4.18 Coefficient of Determination
- 4.19 Answers to 'Check Your Progress'
- 4.20 Summary
- 4.21 Key Terms
- 4.22 Self-Assessment Questions and Exercises
- 4.23 References

Unit V: Index Numbers

122 – 164

- 5.1 Introduction
- 5.2 Definition
- 5.3 Types of Index Numbers
- 5.4 Construction of Index Numbers
- 5.5 Methods of Construction
- 5.6 Comparison between Laspeyre's Index Number and Paasche's Index Number
- 5.7 Selection of an Average for the Construction of Index Numbers
- 5.8 Problems in the Construction of Index Numbers
- 5.9 Advantages or Uses of Index Numbers
- 5.10 Limitations of Index Numbers
- 5.11 Splicing of Index Numbers
- 5.12 Answers to 'Check Your Progress'
- 5.13 Summary
- 5.14 Key Terms
- 5.15 Self-Assessment Questions and Exercises
- 5.16 References

Unit I Measures of Central Tendency

NOTES

Learning Objectives:

By the end of this unit the learners would be able to:

- To understand the importance of an Average
- Difference between AM, GM, and HM
- What are partition values i.e. Quartiles, deciles and percentiles

Structure:

- 1.1 Introduction
- 1.2 Definitions
- 1.3 Types of Averages
- 1.4 Importance (Significance) of Measures of Central Tendency
- 1.5 Arithmetic Mean (A.M.) for Ungrouped Data
- 1.6 Combined Mean of Two Groups
- 1.7 Weighted A.M.
- 1.8 Properties of Arithmetic Mean
- 1.9 Arithmetic mean for Continuous Series or Grouped Data
- 1.10 Steps to Find the A.M. in Grouped Data
- 1.11 Median (General Introduction)
- 1.12 Median for Grouped Data
- 1.13 The Steps for Finding the Median for Grouped Data
- 1.14 Mode (Introduction)
- 1.15 Definitions of Mode
- 1.16 Mode for Ungrouped Data
- 1.17 Determination of Mode in a Continuous Series
- 1.18 Empirical Relation among Mean, Median and Mode
- 1.19 Merits, Demerits and Uses of Mean, Median and Mode
- 1.20 Computations of Quartiles, Deciles and Percentiles

NOTES

- 1.21 Answers to 'Check Your Progress'
- 1.22 Summary
- 1.23 Key Terms
- 1.24 Self-Assessment Questions and Exercises
- 1.25 References

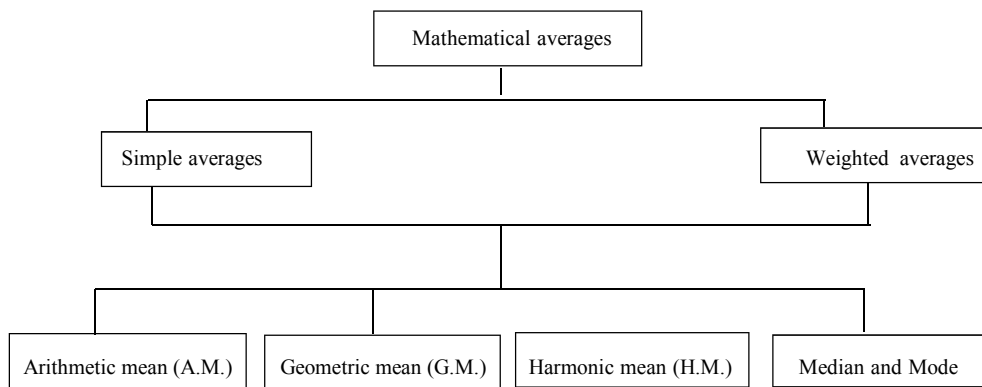
1.1 INTRODUCTION

The purpose of measures central tendency is to determine the centre of the data values or possibly the most typical value which should represent the data.

1.2 DEFINITIONS

1. **According to Simpson and Kafka**, *'A typical value that other figures tend to cluster around or that splits their number in half is a measure of central tendency. Therefore, an entire series of numbers with different magnitudes of the same variable can be described or represented as an average. That total value is ordinary. We can compare individual items within a group using the central tendency measure, and we can compare several series of data according to their central tendencies. Averages are generated numbers rather than the raw facts.'*
2. **According to Lawrence J. Kaplan**, *"Measures of location, often known as average, measures of central tendency, or centre location, are one of the most popular categories of summary statistics. A single value that is reflective of all the observations and is easy for the mind to understand is what is sought after when calculating the average value for a group of observations. The single value is the point at which the many components group together."*
3. **Crum and Smith** say, *"An average is sometimes called a 'measure of central tendency' because individual value of the variable usually cluster around it."*
4. **Central tendency** is a metric in statistics that establishes one score as being representative of the entire distribution. Finding the single score that is most typical or representative of the entire group is the aim of central tendency.
7. **According to Croxton and Cowden**, *"An average is a single value within the range of the data that is used to represent all the values in the series. Since an average is somewhere within the range of the data it is sometimes called a measure of central value."*

1.3 TYPES OF AVERAGES



NOTES

1.4 IMPORTANCE (SIGNIFICANCE) OF MEASURES OF CENTRAL TENDENCY

1. It reduces the large number of values to one figure which serves as the representative of the whole data.
2. Clark and Schkade State “An Average is an attempt to find one single figure to describe whole of the figures.”
3. Croxton and cowden say “An average value is a single value within the range of the data that is used to represent at of the values in the series.”
4. Statistics is some times called the science of averages.
5. Average can be used for comparing one set of data with others. Assume that department A's average monthly sales are compared to department B's average monthly sales.
6. We can know about the universe from a sample since the mean of a sample provides a decent indication of the population's mean
7. Per capita income is calculated with the help of central tendency i.e., Arithmetic mean.
8. The use of averages in standard-setting, estimation, planning, and other administrative choices is beneficial. For instance, the average number of passengers transported by rail on different passenger lines.

Requisites (Desiderata) of a Good Measures of Central Tendency

1. It needs to be precisely specified.
2. In light of all observations.
3. Must be simple to comprehend.
4. Compilation need to be simple.
5. Should be least impacted by sampling variability.
6. It should be able to undergo additional algebraic.

1.5 ARITHMETIC MEAN (A.M.) FOR UNGROUPED DATA

NOTES

$$\text{A.M.} = \frac{\text{Sum of the Values}}{\text{Number of Values}} \quad \dots(I)$$

\Rightarrow A.M. \times Number of values = Sum of the values.

\Rightarrow If each value in a distribution of any series is replaced by the A.M. of that series and these A.M.s, are added, the result = sum of the values.

If X_1, X_2, \dots, X_n are N values of a variable, then their arithmetic mean (generally denoted by \bar{X}) is given by

$$\text{A. M.} = \frac{X_1 + X_2 + X_3 + \dots + X_N}{N} = \frac{\sum_{i=1}^N X_i}{N} \quad (\text{Individual Series}) \quad \dots(II)$$

If $f_1, f_2, f_3, \dots, f_N$ are the respective frequencies of X_1, X_2, \dots, X_N , then

$$\text{A.M.} = \frac{f_1 X_1 + f_2 X_2 + f_3 X_3 + \dots + f_N X_N}{f_1 + f_2 + f_3 + \dots + f_N} = \frac{\sum_{i=1}^N f_i X_i}{\sum_{i=1}^N f_i} \quad (\text{Discrete Series}) \quad \dots(III)$$

The frequency of a value is the number of times that value appears in a distribution.

The A.M. of the values 3, 3, 3, 3, 5, 5, 7, 11, 15, 15

$$= \frac{3+3+3+3+5+5+7+11+15+15}{10} = \frac{70}{10} = 7.$$

The A.M. of the these values can also be represented as

$$= \frac{3 \times 4 + 5 \times 2 + 7 \times 1 + 11 \times 1 + 15 \times 2}{4 + 2 + 1 + 1 + 2} = 7.$$

(The frequency of 3 is 4, frequency of 5 is 2, etc.)

$$= \frac{\text{Sum of the products of the values by their respective frequencies}}{\text{Sum of the frequencies}} = \frac{\sum fx}{\sum f} = \bar{X}$$

$\Rightarrow \sum fX = \text{Sum of the value and } \sum f = \text{Number of values.}$

Hence (I), (II) and (III) give the same definition to the A.M.

A.M. of N values $X_1, X_2, X_3, \dots, X_N$ is denoted by \bar{X}

$$\therefore \text{From (I) } \bar{X} = \frac{\text{Sum of the values}}{\text{Number of values}}$$

$$\Rightarrow (\text{A.M.}) \times \text{Number of values} = \text{Sum of the values i.e., } N\bar{X} = \sum_{i=1}^N X_i \quad \text{or} \quad \bar{X} \sum_{i=1}^N f_i$$

$$= \sum_{i=1}^N f_i X_i$$

\Rightarrow The aggregate of these replacements will be equal to the sum of the individual values if we replace each value in a series by its mean.

⇒ A.M. is a figure of equal division, *i.e.*, if we want to divide profit or loss equally among different companies, we total the profit or loss of all the companies and divide by the number of companies. In general whenever we want equal division we always take A.M.

Note: Σ is a symbol for summation.

For example,
$$\sum_{i=1}^5 X_i = X_1 + X_2 + X_3 + X_4 + X_5$$

and
$$\sum_{i=1}^7 X_i = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 \text{ etc.}$$

NOTES

1.6 COMBINED MEAN* OF TWO GROUPS

If there are two groups G_1, G_2 , such the

	Group I (G_1)	Group II (G_2)
Arithmetic mean	\bar{X}_1	\bar{X}_2
Number of values (or observations)	N_1	N_2

Combined mean of group I and group II (\bar{X}_{12})

$$\begin{aligned} & \frac{\text{Sum of the values in Group I}}{\text{Number of values in Group I}} + \frac{\text{Sum of the values in Group II}}{\text{Number of values in Group II}} \\ &= \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2} \\ &= \frac{\text{Sum of the values in both the groups**}}{\text{Number of values in both the groups}} \end{aligned}$$

If there are p groups, then their combined arithmetic mean \bar{X} is given by

$$\bar{X} = \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2 + \dots + N_p \bar{X}_p}{N_1 + N_2 + \dots + N_p}.$$

where \bar{X}_p denotes the arithmetic mean of p th group and N_p denotes number of value in the p th group.

* Mean ⇒ Arithmetic mean, the word 'arithmetic' is many times omitted.

** Here also mean = $\frac{\text{Sum of the values}}{\text{Number of values}}$

NOTES

If $N_1 = N_2 = N$, i.e., if each group has the same number of values, then Combined arithmetic of the two groups, i.e.,

$$\bar{X}_{12} = \frac{N\bar{X}_1 + N\bar{X}_2}{N + N} = \frac{\bar{X}_1 + \bar{X}_2}{2} = \text{A.M. of the A.Ms. of the two groups.}$$

This is applicable for more group also.

1.7 WEIGHTED A.M.

So far we have given equal importance to each item or value in the series but this concept may be misleading in many situations. Suppose a man purchases mangoes at ₹ 20 per kg., apples at ₹ 30 per kg., and oranges at ₹ 16 per kg. then average rate of purchase will be ₹ $\frac{20+30+16}{3} = \frac{66}{3} = ₹ 22$ per kg. provided he purchases only 1 kg. of each fruit but if he purchases different quantities of each fruit, then this result will not hold good.

In such cases, we have to use weighted arithmetic mean = $\frac{\sum wX}{\sum w}$.

In the above example, if the man purchases 5 kg. of mangoes, 2 kg. of apples and 6 kg. of oranges then the average rate of purchase per kg.

$$\begin{aligned} &= \frac{5 \times 20 + 2 \times 30 + 6 \times 16}{5 + 2 + 6} \\ &= \frac{\text{Total amount spent}}{\text{Total quantity purchased}} \\ &= \frac{100 + 60 + 96}{11} \\ &= \frac{256}{11} = ₹ 19.69 \text{ per kg.}^* \end{aligned}$$

When we are seeking for the mean of means as our average, the weighted mean is really helpful. When there are two groups, then to find the combined mean, we weigh the average of each group by the number of items in it, i.e., if we add these products and divide by the total number of items in the two groups, we may determine the products of the means by the number of things in the respective groups. Additionally, the computation of standardised mortality and birth rates as well as the building of index numbers and comparison of the outcomes of two or more institutions with different student populations all employ weighted A.M.

Steps to find the arithmetic in ungrouped data

(a) When no value is repeated (Individual series)

(i) Find the sum of all the value, i.e., find $\sum_{i=1}^N X_i$.

* Please Note: The average rate of purchase in this case, will not be = $\frac{20 + 30 + 16}{3} = ₹ 22$ per kg.

(ii) Divide the sum by the number of values, *i.e.*, find $\frac{\sum_{i=1}^N X_i}{N} = \text{A.M.}$

(b) When the values are repeated (Discrete series)*

(i) Add the results after multiplying each number by the corresponding frequency,

i.e., find $\sum_{i=1}^N f_i X_i$.

(ii) Divide $\sum_{i=1}^N f_i X_i$ by the total frequency, *i.e.*, find $\frac{\sum_{i=1}^N f_i X_i}{\sum_{i=1}^N f_i} = \text{A.M.}$

(c) When the weights are given

(i) Add the results after multiplying each value by the corresponding weight, *i.e.*,

find $\sum_{i=1}^N W_i X_i$.

(ii) Divide $\sum_{i=1}^N W_i X_i$ by the sum of the weights $\left(\sum_{i=1}^N W_i\right)$, *i.e.*, find $\frac{\sum_{i=1}^N W_i X_i}{\sum_{i=1}^N W_i} = \text{A.M.}$

Note: In normal practice $\sum_{i=1}^N X_i$ is written as $\sum X$ and $\sum_{i=1}^N f_i X_i = \sum fX$ and so on.

1.8 PROPERTIES OF ARITHMETIC MEAN

(1) *The mean is increased or decreased by the same constant amount that is added to or removed from each number in the series.*

Let the value be 3, 5, 9, 15, 16.

The A.M. of these values = $\frac{3+5+9+15+16}{5} = \frac{48}{5} = 9.6$.

If we add 2 to each value, the value become 5, 7, 11, 17, 18.

New mean = $\frac{5+7+11+17+18}{5} = \frac{58}{5} = 11.6 = 9.6 + 2$.

i.e., the mean is also added by 2.

Similarly, if the mean age of 5 boys is 10 years, then their mean age after 3 years will become 13 years because in 3 years, the age of each boy will increase by 3 years.

(2) *The mean is multiplied or divided by a constant if each value in a series is multiplied or divided by a constant.*

* In individual series values are not repeated, *i.e.*, frequency of each value = 1, but in discrete series values may be repeated.

NOTES

NOTES

The mean of the values

3, 5, 9, 15, 16 is 9.6.

If we multiply each value by 3, the values become.

9, 15, 27, 45, 48.

$$\text{The new mean} = \frac{9+15+27+45+48}{5} = \frac{144}{5} = 28.8 = 9.6 \times 3$$

i.e., the mean is also multiplied by 3.

(3) *The deviations of all the values from their mean added together algebraically equal 0.*

Let the values be 4, 6, 9, 15, 16

$$\text{Mean} = \frac{4+6+9+15+16}{5} = 10.$$

$$\text{Deviation of 4 from 10} = 4 - 10 = -6.$$

$$\text{Deviation of 6 from 10} = 6 - 10 = -4.$$

$$\text{Deviation of 9 from 10} = 9 - 10 = -1.$$

$$\text{Deviation of 15 from 10} = 15 - 10 = 5.$$

$$\text{Deviation of 16 from 10} = 16 - 10 = 6.$$

$$\begin{aligned} \text{Algebraic sum of the deviations} &= (\text{Sum of the deviations with their signs}) \\ &= (-6) + (-4) + (-1) + 5 + 6 = 0. \end{aligned}$$

If \bar{X} is the mean of the n values, $X_1, X_2, X_3, \dots, X_n$, the algebraic sum of the deviations of all the value from \bar{X} .

$$= (X_1 - \bar{X}) + (X_2 - \bar{X}) + (X_3 - \bar{X}) + \dots + (X_n - \bar{X})$$

$$= (X_1 + X_2 + X_3 + \dots + X_n) - N\bar{X} = \text{Sum of the values} - N\bar{X}$$

$$= n\bar{X} - n\bar{X} = 0 \quad [\because \text{Sum of the value} = \text{means} \times \text{number of values.}]$$

The algebraic sum of all values' departures from the mean for a frequency distribution.

$$= \sum f(X - \bar{X}) = \sum fX - \bar{X}\sum f = \bar{X}\sum f - \bar{X}\sum f = 0 \quad \left[\because \bar{X} = \frac{\sum fX}{\sum f} = \bar{X} \sum f = \sum fX \right]$$

This property suggests that the mean can be thought of as a point of equilibrium, meaning that if we consider the deviations of all the values from their mean, the sum of the positive deviations will be equal to the sum of the negative deviations.

(4) The lowest, or less than the sum of the squared deviations of the values from any other value outside mean, is the sum of the squared deviations of all the values from their mean.

Notes:

- (i) When the mean and number of values are provided but just one value in the series is missing, the missing value is = Number of values \times Mean – Sum of known values.
- (ii) If one or more values were incorrectly used while determining the mean, then
Correct mean

$$= \frac{1}{N} \left[\begin{array}{l} \text{Sum of all the values including wrong values} - \text{the value} \\ \text{(or values) taken wrongly} + \text{correct value (or values)} \end{array} \right]$$

NOTES**SOLVED EXAMPLES**

Example 1. Find the arithmetic mean of 3, 6, 24, 48.

Solution.

$$\begin{aligned} \text{A.M.} &= \frac{\text{Sum of the values}}{\text{Number of values}} \\ &= \frac{3+6+24+48}{4} = \frac{81}{4} = 20.25. \end{aligned}$$

Example 2. Calculate the A.M. of the following observations:

32, 35, 36, 37, 39, 41, 43, 47, 48

Solution.

$$\text{A.M.} = \frac{32+35+36+37+39+41+43+47+48}{9} = \frac{358}{9} = 39.77.$$

Example 3. If 5, 8, 6 and 1 occur with frequencies 3, 2, 4 and 1 respectively find A.M.

Solution.

$$\begin{aligned} \text{A.M.} &= \frac{\sum fx}{\sum f} = \frac{5 \times 3 + 8 \times 2 + 6 \times 4 + 1 \times 1}{3 + 2 + 4 + 1} \\ &= \frac{15 + 16 + 24 + 1}{10} \\ &= \frac{56}{10} = 5.6 \end{aligned}$$

Example 4. Find the mean of 97, 78, 83, 64, 53.

Solution.

$$\begin{aligned} \text{Mean} &= \frac{97 + 78 + 83 + 64 + 53}{5} \\ &= \frac{375}{5} = 75 \end{aligned}$$

NOTES

Example 5. Find the A.M. from the following data:

X	5	10	15	20	25	30	35	40
f	5	9	13	21	20	15	8	3

Solution.

Total Frequency = $\Sigma f = 5 + 9 + 13 + 21 + 20 + 15 + 8 + 3 = 94$ = Number of values.

X	f	fX
5	5	25
10	9	90
15	13	195
20	21	420
25	20	500
30	15	450
35	8	280
40	3	120
	$\Sigma f = 94$	$\Sigma fX = 2,080$

ΣfX = Sum of the products of X values with their respective frequencies.
 = Sum of the value = 2,080.

$$\begin{aligned}\text{Mean} &= \frac{\text{Sum of the values}}{\text{Number of values}} \\ &= \frac{\Sigma fX}{\Sigma f} = \frac{2,080}{94} = 22.127.\end{aligned}$$

Example 6. The A.M. of the following table is 17 years. Find the value of X .

Age (years)	8	20	26	29
No. of persons	3	2	X	1

Solution.

$$\text{A.M.} = \frac{\Sigma fX}{\Sigma f} = \frac{8 \times 3 + 20 \times 2 + 26 \times X + 29 \times 1}{3 + 2 + X + 1} = 17$$

$$\Rightarrow \frac{24 + 40 + 26X + 29}{6 + X} = 17 \Rightarrow 93 + 26X = 102 + 17X \Rightarrow 9X = 9 \Rightarrow X = 1.$$

Example 7. Numbers 3.2, 5.8, 7.9, and 4.5 have frequencies of X , $X + 2$, $X - 3$, and $X + 6$ correspondingly. Find the value of X if the arithmetic mean is 4.876.

Solution.

$$\text{A.M.} = \frac{\sum_{i=1}^n f_i X_i}{\sum_{i=1}^n f_i}$$

$$\begin{aligned}
&= \frac{X \times 3.2 + (X + 2) \times 5.8 + (X - 3) \times 7.9 + (X + 6) \times 4.5}{X + X + 2 + X - 3 + X + 6} \\
&= \frac{3.2X + 5.8X + 7.9X + 4.5X + 2 \times 5.8 - 3 \times 7.9 + 6 \times 4.5}{4X + 5} \\
&= \frac{21.4X + 14.9}{4X + 5} = 4.876. \text{ (given)} \\
\Rightarrow 21.4X + 14.9 &= 4 \times 4.876X + 5 \times 4.876 \Rightarrow 21.4X - 19.504X = 24.38 - 14.9 \\
\Rightarrow 1.896 X &= 9.48 \quad \Rightarrow X = \frac{9.48}{1.896} = 5.
\end{aligned}$$

NOTES

Example 8. Ten students averaged 35 points per grade. Later, the moderator gave 4 students each 2 grace marks and 2 students each 1 grace mark. Discover the final average score.

Solution.

$$\begin{aligned}
\text{Total marks scored by 10 students} &= 35 \times 10 = 350. \\
\text{Total grace marks} &= 4 \times 2 + 2 \times 1 = 10. \\
\text{Total marks after moderation} &= 350 + 10 = 360 \\
\text{Average marks after moderation} &= \frac{360}{10} = 36.
\end{aligned}$$

Example 9. A person bought three different kinds of pencils. The following information is pertinent:

Quality	Price Per Pencil (₹)	Money Spent (₹)
A	1	50
B	1.5	30
C	2.00	20

Calculate the average price per pencil.

Solution.

$$\text{Number of pencils purchased of quality A} = \frac{50}{1} = 50.$$

$$\text{Number of pencils purchased of quality B} = \frac{30}{1.5} = 10.$$

$$\text{Number of pencils purchased of quality C} = \frac{20}{2} = 10$$

$$\begin{aligned}
\text{Average price per pencil} &= \frac{\text{Total amount spent}}{\text{Total number of pencils purchased}} \\
&= \frac{50 + 30 + 20}{50 + 10 + 10} = \frac{100}{80} = ₹ 1.25.
\end{aligned}$$

NOTES

Example 10. A man bought 5 rolls of ribbons, each 60 metres long, at the rate of 4, 6, 10, 12 and 15 meters a rupee respectively. Find the average price of ribbons.

Solution.

Cost of the roll purchased at the rate of 4 metres a rupee.

Cost of the roll purchased at the rate of 6 metres a rupee = $\frac{60}{6} = ₹ 10$.

Cost of the roll purchased at the rate of 10 metres a rupee = $\frac{60}{10} = ₹ 6$.

Cost of the roll purchased at the rate of 12 metres a rupee = $\frac{60}{12} = ₹ 5$.

Cost of the roll purchased at the rate of 15 metres a rupee = $\frac{60}{15} = ₹ 4$.

Total length of 5 rolls = $60 \times 5 = 300$ metres = Total length purchased
Total cost of 5 rolls = $₹ (15 + 10 + 6 + 5 + 4) = ₹ 40$.

Average rate = $\frac{300}{40} = 7.5$ metres a rupee.

Example 11. The A.M. of the following distribution is 1.46.

No. of Accidents	0	1	2	3	4	5	Total
No. of Days	46	f_1	f_2	25	10	5	200

Find f_1 and f_2 .

Solution.

(X)	(f)	fX
0	46	0
1	f_1	f_1
2	f_2	$2f_2$
3	25	75
4	10	40
5	5	25
Total	$86 + f_1 + f_2$	$140 + f_1 + 2f_2$

$$\text{Total frequency} = 86 + f_1 + f_2 = 200 \quad \Rightarrow f_1 + f_2 = 114 \quad \dots(i)$$

$$\text{Mean} = \frac{140 + f_1 + 2f_2}{200} = 1.46 \quad \Rightarrow f_1 + 2f_2 = 152 \quad \dots(ii)$$

Subtracting (i) from (ii), we get $f_2 = 38$.

Putting $f_2 = 38$ in (i), we get $f_1 = 76$.

Example 12. The arithmetic average of the following distribution is 40. Find the value of k .

X	20	25	30	$k + 10.5$	60
f	2	7	15	10	6

Solution.

X	f	fX
20	2	40
25	7	175
30	15	450
$(k + 10.5)$	10	$10k + 105$
60	6	360
$\Sigma f = 40$	$\Sigma fX = 1130 + 10k$	

Example 13. A company's employees made an average of ₹ 24,000 each year. Employees' mean yearly incomes were ₹ 19,000 and ₹ 25,000 respectively for men and women. Find out the company's gender representation in the workforce.

Solution.

Let N_1, N_2 denote respectively the number of male and female employees.

$$\text{Total yearly salary paid to all the employees} = ₹ 24,000 (N_1 + N_2)$$

$$\text{Total yearly salary paid to male employees} = ₹ 25,000 (N_1)$$

$$\text{Total yearly salary paid to female employees} = ₹ 19,000 (N_2)$$

The sum of the salaries paid to all the employee is equal to the sum of the salaries paid to male and female employees.

$$\Rightarrow 24,000 (N_1 + N_2) = 25,000 (N_1) + 19,000 (N_2) \Rightarrow \frac{N_1}{N_2} = \frac{5}{1} \Rightarrow N_1 = 5N_2.$$

$$\text{Percentage of female employees} = \frac{N_2}{N_1 + N_2} \times 100 = \frac{N_2}{5N_2 + N_2} \times 100 = 16 \frac{2}{3} \%.$$

$$\therefore \text{Percentage of male employees} = 100 - 16 \frac{2}{3} = 83 \frac{1}{3} \%.$$

Example 14. The average monthly salary for 10 employees in Factory A is ₹ 4,000, whereas that for employees in Factory B is ₹ 3,700. Find the number of workers in B if the median monthly salary for all employees in companies A and B is ₹ 3,800.

NOTES

NOTES

Solution.

Let the numbers of workers in factory B be X , then

$$\frac{10 \times 4,000 + 3,700X}{10 + X} = 3,800$$

$$\Rightarrow 40,000 + 3,700X = 38,000 + 3,800X$$

$$\Rightarrow 2,000 = 100X \Rightarrow X = 20$$

Example 15. A man invested his savings as follows:

₹ 10,000 in Post Office Savings Bank at 8% p.a.; ₹ 6,000 in National Bank at 7% p.a.; ₹ 4,000 in a Private Firm at 10% p.a.

Solution.

$$\text{Interest from Post Office S/B} = \frac{10,000 \times 8}{100} = ₹ 800$$

$$\text{Interest from National Bank} = \frac{6,000 \times 7}{100} = ₹ 420$$

$$\text{Interest from Private Firm} = \frac{4,000 \times 10}{100} = ₹ 400$$

Total Interest on ₹ (10,000 + 6,000 + 4,000 = 20,000) = ₹ (800 + 420 + 400) = ₹ 1,620

$$\text{Hence, average percentage rate of interest} = \frac{\text{Total interest}}{\text{Total amount invested}} \times 100$$

$$= \frac{1,620}{20,000} \times 100 = 8.10\%$$

Example 16. 300 students received 45 in the English test on average. What are the remaining students' average marks if the first 100 students' average marks are 70 and 20, respectively?

Solution.

$$\text{Total marks of 300 students} = 300 \times 45 = 13,500.$$

$$\text{Total marks of first 100 student} = 100 \times 70 = 7,000.$$

$$\text{Total marks of last 100 student} = 100 \times 20 = 2,000.$$

$$\text{Average marks of remaining 100 students} = \frac{13,500 - 7,000 - 2,000}{100} = 45.$$

Example 17. 100 students showed up for the test. Below are the outcomes for those who failed.

Marks	5	10	15	20	25	30
No. of Students	4	6	8	7	3	2

Find out the average score of the students who passed if the overall average score for all 100 students was 68.6.

Solution.

$$\text{Number of students who passed} = 100 - (4 + 6 + 8 + 7 + 3 + 2) = 70.$$

$$\text{Total marks obtained by 100 students} = 100 \times 68.6 = 6,860.$$

Total marks obtained by the students who failed

$$= 4 \times 5 + 6 \times 10 + 8 \times 15 + 7 \times 20 + 3 \times 25 + 2 \times 30 = 475.$$

$$\text{Total marks obtained by students who passed} = 6,860 - 475 = 6,385.$$

$$\text{Average marks of students who passed} = \frac{6,385}{70} = 91.2.$$

Example 18. The average pay for a company with 1,000 employees was found to be ₹ 541.20. Later, after salaries had been paid out, it was revealed that two employees' salaries had been entered incorrectly as " ₹ 891" and " ₹ 495." It was ₹ 591 and ₹ 555 that were the proper salaries. Identify the proper arithmetic mean.

Solution.

Total salary calculated wrongly, of 1,000 employees

$$= 541.20 \times 1,000 = ₹ 5,41,200.$$

Correct total salary of 1,000 employees

$$= 5,41,200 - (891 + 495) + (591 + 555) \\ = ₹ 5,40,960.$$

$$\text{Correct arithmetic mean} = \frac{5,40,960}{1,000} = ₹ 540.96.$$

Example 19. The average weights of two groups of students were 162 kg and 148 kg, respectively. When will both groups' combined mean weight be 155 kgs?

Solution.

Let N_1 be the number of student in the first group and N_2 the number of students in the second group, then

	Group I	Group II
Number of Students	N_1	N_2
Mean	162 kgs.	148 kgs.

$$\text{Combined mean} = 155 \text{ kgs.}$$

$$\therefore 155 = \frac{162N_1 + 148N_2}{N_1 + N_2}$$

$$\Rightarrow 155N_1 + 155N_2 = 162N_1 + 148N_2 \Rightarrow 7N_2 = 7N_1 \Rightarrow N_1 = N_2.$$

\Rightarrow The number of items in the two group should be same.

NOTES

NOTES

Second Method:

If $N_1 = N_2$, then the combined A.M. of two groups is

$$= \frac{\text{Mean of 1}^{\text{st}} \text{ group} + \text{Mean of 2}^{\text{nd}} \text{ group}}{2}.$$

Since this conditions is satisfied in this question, therefore $N_1 = N_2$.

Example 20. 15 wrestlers, on average, weigh 110 kg. Five of them have an average weight of 100 kg. while the remaining 5 are 125 kg. What does the remainder's intended weight mean?

Solution.

Mean weight of 15 wrestlers = 110 kgs.

Total weight of 15 wrestlers = $110 \times 15 = 1650$ kgs.

Total weight of 5 wrestlers whose mean weight is 100 kgs. = $100 \times 5 = 500$ kgs.

Total weight of 5 wrestlers whose mean weight is 125 kgs. = $125 \times 5 = 625$ kgs.

Total weight of remaining 5 wrestlers = $1,650 - 500 - 625 = 525$ kgs.

Mean weight of these 5 wrestlers = $\frac{525}{5} = 105$ kgs.

Example 21. The average of 10 numbers is 12.5. The first six's mean is 15 while the latter five's mean is 10. Identify the sixth number.

Solution.

Sum of ten numbers = $12.5 \times 10 = 125$

Sum of first six numbers = $6 \times 15 = 90$

Sum of last five numbers = $10 \times 5 = 50$

Sum of first six numbers + Sum of last five numbers

= Sum of first six numbers + Sixth number + Sum of last four numbers

= Sum of first six numbers + Sum of last four numbers + 6th number

= Sum of ten numbers + 6th number

$\Rightarrow 125 + 6\text{th number} = 90 + 50 \Rightarrow \text{Sixth number} = 140 - 125 = 15.$

Second Method:

Let \bar{X} be the average of last four numbers, then

Sum of ten numbers = Sum of first six numbers + Sum of last four numbers

$$= 6 \times 15 + \bar{X} \times 4 = 125$$

$$\Rightarrow \bar{X} = \frac{125-90}{4} = \frac{35}{4}$$

$$\text{Sum of last four numbers} = \frac{35}{4} \times 4 = 35$$

$$\therefore 90 + \text{Sixth number} + 35 = 90 + 50 = 140$$

[\therefore Sum of the last five numbers is 50]

$$\Rightarrow \text{Sixth number} = 140 - 90 - 35 = 15.$$

NOTES

1.9 ARITHMETIC MEAN FOR CONTINUOUS SERIES OR GROUPED DATA

To find the mean of a large number of values, it may be quite lengthy to add up all the value and divide by the number of values. In these situations, we create groups with a range of values in each, and we count the number of items that fall in each category. For instance, if a class has 75 students, then to find the average marks of students in any subject, we have to add the marks obtained by all the 75 students and divide by 75 which may become lengthy. In such cases, we divide the marks in small groups. If the total marks are 100, then we can form groups. Let each group contain 10 marks. Then the groups will be 0-10, 10-20, 20-30, 30-40, 40-50, 50-60, 60-70, 70-80, 80-90, 90-100. The frequency of each group is then determined by counting the number of pupils who belong to each group.

In grouped data, the middle value of each group serves as the group representative. i.e., Each item in a group of grouped data is taken to be equal to the middle value of that group. **(Please note)**

\therefore Sum of the value in each group

$$= \text{Middle value of the group} \times \text{Number of values in that group.}$$

Further the procedure is the same as for finding the combined mean of two or more groups. The method will be made clear in the example that follows.

SOLVED EXAMPLES

Example 22. Calculate the frequency distribution's arithmetic mean as follows:

Class Interval	Frequency
10-20	4
20-40	10
40-70	26
70-120	8
120-200	2
Total	50

Solution.

$$\text{Mid value of first class (or group)} = \frac{10+20}{2} = 15.$$

$$\text{Sum of the values in first class} = 15 \times 4 = 60.$$

NOTES

$$\text{Mid value of second class} = \frac{20+40}{2} = 30.$$

$$\text{Sum of the values in the second class} = 30 \times 10 = 300.$$

$$\text{Mid value of the third class} = \frac{40+70}{2} = 55.$$

$$\text{Sum of the value in third class} = 55 \times 26 = 1,430.$$

$$\text{Mid value of the fourth class} = \frac{70+120}{2} = 95.$$

$$\text{Sum of the values in the fourth class} = 95 \times 8 = 760.$$

$$\text{Mid value of the fifth class} = \frac{120+200}{2} = 160.$$

$$\text{Sum of the value in the fifth class} = 160 \times 2 = 320.$$

$$\text{Mean} = \frac{\text{Sum of the values in all the classes}}{\text{No. of values in all the classes}}$$

$$= \frac{60+300+1,430+760+320}{4+10+26+8+2}$$

$$= \frac{2,870}{50} = 57.4.$$

The above steps can be written in the tabular form as:

Class-Interval	Mid-Values (X)	Freq. (f)	fX
10-20	15	4	60
20-40	30	10	300
40-70	55	26	1430
70-120	95	8	760
120-200	160	2	320
70-120	95	$\Sigma f = 50$	$\Sigma fX = 2870$

$$\text{A.M.} = \frac{\Sigma fX}{\Sigma f} = \frac{2,870}{50} = 57.4.$$

1.10 STEPS TO FIND THE A.M. IN GROUPED DATA

- (i) Find the middle value of each group by using the formula

$$\text{Middle Value} = \frac{\text{lower limit} + \text{upper limit}}{2}$$

- (ii) Multiply the middle value of each group by its respective frequency to find the sum of the values in that group.
- (iii) Find the sum of the values in all the group by adding all the products obtained in Step II.

$$\text{A.M} = \frac{\text{Sum of the values in all the groups, i.e., sum of the products of the middle values of all the groups by their respective frequencies}}{\text{Sum of the frequencies of all the groups.}}$$

Example 23. Find the missing frequency from the data below if the arithmetic mean is 33.

Loss per Shop	No. of Shops
0-10	10
10-20	15
20-30	30
30-40	—
40-50	25
50-60	20

Solution.

Let the missing frequency be: f

Loss per Shop	Mid Value (X)	No. of Shops	fX
0-10	5	10	50
10-20	15	15	225
20-30	25	30	750
30-40	35	f	$35f$
40-50	45	25	1125
50-60	55	20	1100
		$\Sigma f = 100 + f$	$\Sigma fX = 3250 + 35f$

$$\begin{aligned} \text{Mean} &= \frac{\Sigma fX}{\Sigma f} \\ &= \frac{3,250 + 35f}{100 + f} = 33. \end{aligned}$$

$$\Rightarrow 3,250 + 35f = 3,300 + 33f$$

$$\Rightarrow 2f = 50 \quad \Rightarrow f = 25.$$

Example 24. The frequency of the class interval (30-40) in the subsequent frequency distribution is unknown. Determine whether the distribution's arithmetic mean is 28.

Profit per Shop (₹)	0-10	10-20	20-30	30-40	40-50	50-60
No. of Workers	12	18	27	(x)	17	6

NOTES

NOTES

Solution.

Profit per Shop	Mid Points (X) ₹	Frequency f	fX
0-10	5	12	60
10-20	15	18	270
20-30	25	27	675
30-40	35	x	35x
40-50	45	17	765
50-60	55	6	330
		N = 80 + x	$\Sigma fX = 2,100 + 35x$

$$\bar{X} = \frac{\Sigma fX}{N} \Rightarrow 28 = \frac{2,100 + 35x}{80 + x}$$

$$\Rightarrow 28(80 + x) = 2,100 + 35x \Rightarrow 2,240 + 28x = 2,100 + 35x$$

$$\Rightarrow 28x - 35x = 2,100 - 2,240 \Rightarrow -7x = -140$$

$$\Rightarrow 7x = 140 \Rightarrow x = 20.$$

Hence, 20 is the missing frequency.

1.11 MEDIAN (GENERAL INTRODUCTION)

As opposed to the arithmetic mean, which is determined using the values of each item in the series, the median is referred to as a position average, according to Simpson and Kafka. The position of a value in a series is referred to as its “position.” The median is positioned in a series so that the number of items below it in magnitude equals the number of things above it in magnitude. Therefore, the median is a value in a series that is exceeded by an equal number of other values. The centre value is hence called median.

According to **Connor**, “The median is that value of the variable which divides the group into two equal parts. one part comprising all values greater, and the other, all values less than the median.”

According to **A.L. Bowley**, “If the numbers of the group are ranked in order; according to the measurement under consideration then the measurement of the number most nearly one half is the medium.”

Ungrouped data are organised either in ascending or descending order of magnitude in order to calculate the median. The middle value after the data has been arranged in one of two ways yields the median (Preferably arrange the data in ascending order).

If the number of items is odd, then $\left(\frac{N+1}{2}\right)$ th value gives the median and if the number of items is even, then the arithmetic mean of $\frac{N}{2}$ th and $\left(\frac{N}{2}+1\right)$ th value gives the median value. For example, consider the following series:

2, 10, 5, 7, 15, 17, 21, 13, 4
and 2, 10, 5, 7, 15, 17, 21, 13, 4, 22

After arranging the above series in ascending order, we get

2, 4, 5, 7, 10, 13, 15, 17, 21, 'A'
and 2, 4, 5, 7, 10, 13, 15, 17, 21, 22, 'B'

Series 'A' contains 9 items, therefore, $\frac{9+1}{2} = 5$ th value, i.e., 10 is the median.

Series 'B' contains 10 values, therefore, A.M. of $\frac{N}{2} = \frac{10}{2} = 5$ th value and

$\left(\frac{N}{2} + 1\right) = 6$ th value gives the median. In this case $\frac{10+13}{2} = 11.5$ is the median.

Which is a value half way between the two central values.

When a median value is surrounded by values that are of comparable magnitude, the definitions of the median given above do not apply. For instance, there is no value in a series of values such as 12, 13, 14, 15, 16, 17 and 18 that is situated such that three values are smaller and three greater than it. The median number is, however, 15, thus Croxton and Cowdon have provided a revised definition of median as taken into consideration such circumstances. The median is the value at which half or more of the items in a series are equal to, less than, or equal to it, and half or more of the things are equal to, equal to, or higher than it.

The description above makes it clear that finding the median requires the data to be sorted in either ascending or descending order, but finding the mean does not. The size of the items does not affect the median; rather, the location of the items in the array does.

1.12 MEDIAN FOR GROUPED DATA

The separate elements in grouped data become unrecognisable, and counting cannot reveal the intermediate item. In order to obtain the value that divides the total number of objects in half, one must enter a class. When we divide the number of frequencies (N) in half, we discover that the centre item belongs to a class. In order to identify this class, we add the frequencies together until we locate the lowest class

*where the total frequency is higher than $\frac{N}{2}$. This class is called the **median class**.*

Median class is the lowest class whose cumulative frequency* contain the value $\frac{N}{2}$ when the class intervals are arranged in ascending order. It is obvious that the median value is more than the median class's lower limit and less than or equal to the median class's higher limit. assuming that all items fall into this median class at an equal rate.

* $l_2 - l_1 = \text{Class interval} = i$.

NOTES

NOTES

When we reach the frequency equivalent to $\frac{N}{2}$, we cease moving toward the upper limit of this class. We arrive at a value within the median class—which is assumed to have $\frac{N}{2}$ items on either side—by doing this method.

As mentioned earlier. The median value cannot exceed the upper limit of the median class but must be at least as high as the lower limit.

$$\text{Median} = \text{Lower limit of median class} + \frac{\frac{\text{Total frequency}}{2} - \text{Cumulative frequency of the class preceding the median class}}{\text{Frequency of the median class}} \times \text{Class interval of the median class}$$

1.13 THE STEPS FOR FINDING THE MEDIAN FOR GROUPED DATA

1. Divide the number of items in the distribution by 2, that is, compute the value $\frac{N}{2}$.
2. Accumulate, *i.e.*, cumulate the frequencies.
3. Find the class whose cumulative frequency is the first to exceed $\frac{N}{2}$. This is the median class.
4. Find the lower limit of the median class.
5. Then carry out the subsequent actions: The frequencies we have accumulated before to joining the median class are subtracted from $\frac{N}{2}$. Subtract this difference from the median class's frequency. Add the size of the median class's class interval to the quotient you just obtained.
6. The lower limit of the median class is increased by the outcome of the operation in Step 5. This total provides the median.

The formula for this procedure is

$$\text{Median} = l_1 + \left(\frac{\frac{N}{2} - Cf_{-1}}{f_{med}} \right) \times i \quad \dots (i)$$

where l_1 = the lower limit of the median class.

N = the total frequency.

Cf_{-1} = cumulative frequency of the class preceding the median class,

f_{med} = the frequencies of the median class, and

i = the size of the class interval of the median class.

For finding the median for grouped data also, the class intervals should be arranged in ascending order of magnitude.

SOLVED EXAMPLES

Example 25. Calculate the median from the following values

18, 16, 14, 11, 13, 10, 9, 20

Solution.

Arranging the value in ascending order, we get

9, 10, 11, 13, 14, 16, 18, 20

Number of items = 8.

Since the number of items is even, the A.M. of $\frac{N}{2}$ th and $\left(\frac{N}{2}+1\right)$ th value is the median, *i.e.*, the A.M. of 4th and 5th value will be median,

4th value = 13,

5th value = 14

$$\text{A.M. of 13 and 14} = \frac{13+14}{2} = 13.5^*$$

$$\therefore \text{Median} = 13.5.$$

Example 26. Find the median of the observations

4, 12, 7, 9, 14, 16, 21, 3, 17

Solution.

Arranging the data in ascending order, we get

3, 4, 7, 9, 12, 14, 16, 17, 21

Since the number of values is odd, *i.e.*, 9, $\frac{N+1}{2}$ th value, *i.e.*, $\frac{9+1}{2} = 5$ th value is the median.

5th value = 14.

\therefore Median = 14.

Example 27. Find the median of the numbers

25, 24, 23, 32, 40, 27, 30, 25, 20, 10, 15, 45

Solution.

Arranging the data in ascending order, we have

10, 15, 20, 23, 24, 25, 25, 27, 30, 32, 40, 45

The number of items in the series is 12 which is an even number.

NOTES

* If we arrange the data in descending order, the values become 20, 18, 16, 14, 13, 11, 10, 9. Now 4th value = 14, 5th value = 13, A.M. of 14 and 13 = 13.5 = Median.

NOTES

\therefore The A.M. of 6th and 7th item is the median

$$6\text{th value} = 25$$

$$7\text{th item} = 25.$$

Since 6th and 7th items are equal, therefore, the median is 25.

Example 28. (a) The median of the following observations arranged in ascending order is 42. Find x .

$$22, 24, 33, 37, x + 1, x + 3, 44, 47, 51, 58.$$

Solution. The number of observation is 10 which is even. Hence mean of 5th and 6th value is median.

Mean of 5th and 6th value

$$= \frac{x+1+x+3}{2} = x + 2 = 42 \text{ (given)} \Rightarrow x = 40.$$

Example 29. (b) The median of eight observation 31, 48, 37, 34, 45, 36, 41 and x is 38 where $37 < x < 41$. Find the value of x .

Solution.

Arranging the data in ascending order, we get

$$31, 34, 36, 37, x, 41, 45, 48$$

Median is the A.M. of 4th and 5th value, *i.e.*,

$$\text{Median} = \frac{37+x}{2} = 38 \quad \Rightarrow x = 39$$

Example 30. Calculate the median from the following data.

$$(a) \quad 15, 21, 60, 65, 70, 45, 54, 50, 40, 30, 26$$

$$(b) \quad 21, 13, 17, 11, 19, 9, 16, 23, 14$$

$$(c) \quad 15, 20, 25, 28, 16, 17, 9, 11$$

Ans. (a) 45, (b) 16, (c) 16.5

Solution.

(a) Arranging the data in ascending order we get

$$15, 21, 26, 30, 40, 45, 50, 54, 60, 65, 70$$

Total number of values = 11 is odd number.

$$\text{Median is } = \frac{11+1}{2} = 6\text{th value}$$

$$\therefore \text{Median} = 45$$

- (b) Arranging the data in ascending order we get

9, 11, 13, 14, 16, 17, 19, 21, 23

Total number of values = 9

$$\therefore \frac{9+1}{2} = 5 \text{ i.e. 5th value is median}$$

Hence value 16 is median

- (c) Arranging the values in ascending order we get

9, 11, 15, 16, 17, 20, 25, 28

Number of values = 8 i.e. Even

So A.M. of $\frac{n}{2}$ th value and $\left(\frac{n}{2}+1\right)$ th value is median

$$\frac{n}{2} \text{ i.e. 4th value} = 16$$

$$\left(\frac{n}{2}+1\right) \text{ i.e. 5th value} = 17$$

A.M. of 16 and 17 = 16.5 = median

Example 31. The marks out of 50 obtained by 100 students in the test are given below. Find the median marks.

Marks	20	25	28	29	33	38	42	45
No. of Students	6	20	24	28	15	4	2	1

Solution.

Cumulative frequency table after arranging the marks in ascending order is as under:

Marks Obtained	Frequency (No. of Student)	Cumulatively Frequency
20	6	6
25	20	26
28	24	50
29	28	78
33	15	93
38	4	97
42	2	99
45	1	100

Total frequency = $\Sigma f = N = 100$ which is even. Hence, Median value will be the average of $\frac{N}{2}$ th value, i.e., $\frac{100}{2}$ th = 50th value and $\left(\frac{N}{2}+1\right)$ th = 51st value.

50th value = 28.

51st value = 29.

NOTES

NOTES

\therefore A.M. of 28 and 29 is $\frac{28+29}{2} = 28.5$.

\Rightarrow Above 50% of student obtained marks 28.5 or less and the remaining 50% obtained marks 28.5 or more.

Example 32. Find the median of the data:

X	160	150	152	161	156
f	5	8	6	3	7

Solution.

Arrange the data in ascending order, we have

X	f	Cumulative Frequency
150	8	8
152	6	14
156	7	21
160	5	26
161	3	29
Total	$N = \sum f = 29$	

Total number of observations $N (= 29)$ is odd.

$$\therefore \left(\frac{N+1}{2} \right) \text{th item} = \left(\frac{29+1}{2} \right) \text{th item} = 15 \text{th item.}$$

Now, 15th item lies in the row with 21 as cumulative frequency, since $14 < 15 < 21$.

\therefore Median = size of 15th item = 156.

Example 33. Find the median of the following data:

Income (₹)	100	150	80	200	250	180
No. of Persons	24	26	16	20	6	30

Solution.

Arranging the data in ascending order, we have

Income (in ₹)	No. of Persons	Cumulative Frequency
80	16	16
100	24	40
150	26	66
180	30	96
200	20	116
250	6	122
Total	$N = \sum f = 122$	

Hence $N (= 122)$ is even and $\frac{N}{2} = 61$.

\therefore Median is the mean of value of $\left(\frac{N}{2}\right)$ th and $\left(\frac{N}{2}+1\right)$ th items,

i.e., Median = Mean of 61st and 62nd value

$$= ₹ \left(\frac{150+150}{2} \right) = ₹ 150.$$

Example 34. Find the missing frequencies f_1 and f_2 if the median of the following frequency distribution is 46:

Variable	Frequency
10–20	12
20–30	30
30–40	f_1
40–50	65
50–60	f_2
60–70	25
70–80	18
Total	229

Solution.

We have the following table:

Variable	Frequency (f)	Cumulative Frequency ($c.f.$)
10–20	12	12
20–30	30	42
30–40	f_1	$42 + f_1$
40–50	65	$107 + f_1$
50–60	f_2	$107 + f_1 + f_2$
60–70	25	$132 + f_1 + f_2$
70–80	18	$150 + f_1 + f_2$
Total	$N = \Sigma f = 150 + f_1 + f_2$	

It is given that $N = 229 \Rightarrow \frac{N}{2} = 114.5$.

$$\therefore 150 + f_1 + f_2 = 229 \Rightarrow f_1 + f_2 = 79 \quad \dots(i)$$

Now, since the median is given to be 46, hence median class is 40 – 50.

$$\therefore l_1 = 40, f = 65, Cf_{-1} = 42 + f_1, i = 10.$$

$$\therefore \text{Median} = l_1 + \frac{\frac{N}{2} - Cf_{-1}}{f} \times i$$

NOTES

NOTES

$$= 40 + \frac{114.5 - (42 + f_1)}{65} \times 10 = 40 + \frac{(72.5 - f_1) \times 2}{13}$$

$$\Rightarrow 46 = 40 + \frac{145 - f_1}{13} \quad [\text{Median} = 46 \text{ (given)}]$$

$$\Rightarrow 46 - 40 = \frac{145 - 2f_1}{13} \quad 6 \times 13 = 145 - 2f_1$$

$$\Rightarrow 2f_1 = 145 - 78 \quad \Rightarrow 2f_1 = 67 \quad \Rightarrow f_1 = 33.5 = 34^*$$

Hence from (i), $f_2 = 79 - 34 = 45$.

Hence, the missing frequencies are $f_1 = 34$ and $f_2 = 45$.

Example 35. Determine the median graphically from the data given below:

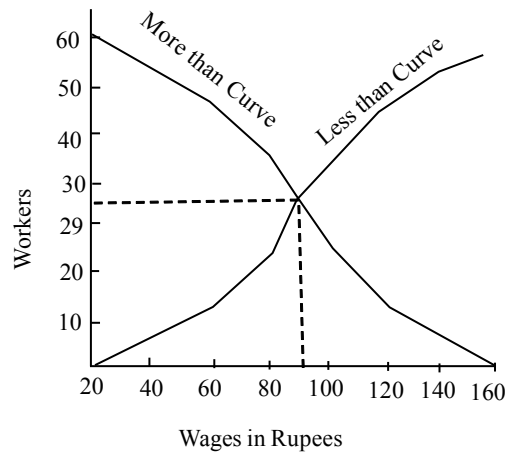
Wages (₹)	No. of Workers	Wages (₹)	No. of Workers
20-40	4	40-60	6
60-80	10	80-100	16
100-120	12	120-140	7
140-160	3		58

Solution.

We shall use the second method of locating the Median by drawing two ogives (one 'less than' and the other 'more than'). In the 'less than' ogive, the upper limit of the first class interval (20-40) would be the starting point and in the 'more than' ogive, the lower limit of this class-interval or 20 would be the first value of the variable. From the point of intersection of the two ogives we will draw a perpendicular on the x-axis and the point where it touches the x-axis would be the value of the Median.

Wages Less than (₹)	No. of Workers	Wages More than (₹)	No. of Workers
40	4	20	58
60	10	40	54
80	20	60	48
100	36	80	38
120	48	100	22
140	55	120	10
160	58	140	3

The value of the Median comes to 91.25.



Locating median Graphically

NOTES

1.14 MODE (INTRODUCTION)

The value with the highest frequency in the series, or the most frequent value, is the mode. For instance:

- (i) The modal price is the retail price that the majority of customers pay for a good.
- (ii) In a factory, the modal salary is the pay rate given to the majority of employees.
- (iii) The modal class is the one with the highest frequency in a grouped frequency distribution.

1.15 DEFINITIONS OF MODE

1. The value that appears the most frequently in a collection of objects and around which the other items are clustered most densely is known as the mode, according to Zizek.
2. In words of **Croxtan and Cowdon**, “*The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. It may be regarded as a most typical of a series of values*”.
3. According to **A.M. Tuttle**, “*Mode is the value which has the greatest frequency or density in its immediate neighbourhood*”.

In case of dispute between the owner of a factory and Trade Union leader, the Trade Union leader will always fight the case on the basis of modal wage whereas the owner will plead his case on the basis of average wage, because due to higher wages of Top Officials, the average wage will always be more than the modal wage.

NOTES

1.16 MODE FOR UNGROUPED DATA

In case of ungrouped data, the value occurring most frequently is the mode, for example, in the series 2, 4, 5, 5, 5, 8, 9, 10. The mode is 5. Here also it is preferable to arrange the data as discrete frequency distribution. Since the value 5 is repeated maximum number of times, 5 is the mode.

In the following series, modal value has been circled:

- 41, 42, 45, 44, 45, 48, 50, 45, 47, 51, 56.

Arranging the distribution as frequency distribution, we get

<i>Value</i>	<i>Frequency</i>
41	1
42	1
45	3
44	1
48	1
50	1
47	1
51	1
56	1

- Consider the series 27, 28, 26, 35, 40, 37, 38, 28, 27, 25, 30, 28, 40

Arranging as frequency distribution, we get

<i>Value</i>	<i>Frequency</i>
41	—
25	1
26	1
27	2
28	3
30	1
35	1
37	1
38	1
40	2

In the case of a discrete frequency distribution (ungrouped data) mode can be located simply by inspection. Here the value having maximum frequency will represent the mode.

Example 36. Determine mode from the following distribution:

X	10	12	14	16	18	20	22
<i>f</i>	4	6	10	11	21	10	5

Solution.

In the above discrete frequency distribution the variable value 18 has the maximum frequency 21. Therefore, 18 is the mode of the given discrete frequency distribution.

Example 37. Find the mode from the following data.

12, 14, 16, 18, 26, 16, 20, 16, 11, 12, 16, 15, 20, 24

Solution. Value 16 occurs four times i.e. frequency of 16 is 4 which is highest in the data.

Therefore mode is 16

2nd Method

Arranging the data in ascending order, we get

11, 12, 12, 14, 15, 16, 16, 16, 16, 18, 20, 20, 24, 26

16 has maximum frequency

\therefore Mode = 16

Example 38. Calculate the value of mode from the following data.

Marks: 16, 18, 22, 26, 15, 14, 14, 10, 11, 14

Here in the data 14 has maximum frequency i.e., 3

\therefore Mode = 14

2nd Method

Arranging the data in ascending order we get 10, 11, 14, 14, 14, 15, 16, 16, 18, 22

Here 14 is occurring largest number of times.

\therefore Mode = 14

Mode by Grouping Method

In a unimodal distribution where the distribution has only one mode and the highest concentration of values is around one value only, we do not face any difficulty in locating the modal value. However, when two or more adjacent values show a nearly similar frequency concentration, difficulties are encountered. In these situations, an effort is made to use the Grouping Method to determine the value of concentration. The values are initially organised in ascending order and their frequencies are noted when using the grouping method. The grouping table typically includes the following six columns.

Column 1: The maximum frequency is indicated by placing a mark or a circle, and the original frequencies are listed in ascending order.

Column 2: The frequencies are arranged in groups of two, with the totals shown next to each pair and the highest total denoted.

Column 3: After excluding the initial frequency, the remaining frequencies are paired together, and the frequency with the highest sum is noted.

NOTES

NOTES

The frequencies are organised into threes in column

Column 4. The highest sum is highlighted.

Column 5: The remaining frequencies are grouped in threes after the initial frequency, and the frequency with the highest sum is noted.

Column 6: The remaining frequencies are grouped in threes after the first two are eliminated, and the frequency with the highest sum is noted.

After the grouping table is complete, an analysis table is created to identify the value or observation that is repeated the most.

Example 39. The measurements of 230 university students' collar sizes are shown in the accompanying table. Identify the collar's model size.

<i>Collar Size (cm)</i>	32	33	34	35	36	37	38	39	40	41
<i>No. of Students</i>	7	14	30	28	35	34	16	14	36	16

Solution.

Here, we observe nearly equal concentration of frequencies for the collar sizes 36, 37 and 40. Thus, grouping method is used in locating the modal class. Following the steps listed above in grouping method, we get the following table.

Determining Mode by Grouping Method

<i>Collar Size</i>	<i>Frequency</i>					
	<i>Col. 1</i>	<i>Col. 2</i>	<i>Col. 3</i>	<i>Col. 4</i>	<i>Col. 5</i>	<i>Col. 6</i>
32	7	} 21	} 44	} 51	} 72	} (93)
33	14					
34	30	} 58	} (63)	} (97)	} (85)	} 64
35	28					
36	35	} (69)	} 50	} 66	} (66)	
37	34					
38	16	} 30	} 50			
39	14					
40	(36)	} 52				
41	16					

After forming the grouping table, values having highest frequency are counted with the help of an analysis table as shown below:

Analysis Table

Coloum Number	Collae Size Values Contributing to the Highest Frequency									
	32	33	34	35	36	37	38	39	40	41
1										
2					√	√				
3				√	√					
4				√	√	√				
5					√	√	√			
6			√	√	√					
No. of Times	–	–	1	3	5	3	1	–	1	–

From analysis table, we observe that size 36 is repeated the maximum number of times. Therefore, modal size is 36 cm.

NOTES

1.17 DETERMINATION OF MODE IN A CONTINUOUS SERIES

In a continuous frequency distribution, frequencies are given for various classes and the Modal Class is the one with the highest frequency. The following interpolation formula is used to acquire the precise value of the mode after finding the modal class:

$$\begin{aligned}\text{Mode} &= l_1 + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times i = l_1 + \left[\frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \right] \times i \\ &= l_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i\end{aligned}$$

where l_1 = the lower limit of the modal class.

f_m = the frequency of the the modal class.

f_{m-1} = the frequency of the class preceding the modal class.

f_{m+1} = the frequency of the class succeeding the modal class.

i = size of the modal class, i.e., class interval of the modal class.

Example 40. The mode of the following series is 36. Find the missng frequency in it.

Class Interval	0–10	10–20	20–30	30–40	40–50	50–60	60–70
Frequency	8	10	–	16	12	6	7

NOTES

Solution.

Since the mode of ten given series is 36 and maximum frequency 16 lies in the class 30-40, so the modal class is 30-40.

Let x be the missing frequency.

$$l_1 \Rightarrow \text{lower limit of the modal class} = 30$$

$$i \Rightarrow \text{class interval of the modal class} = 10$$

$$f_m \Rightarrow \text{frequency of the modal class} = 16$$

$$f_{m-1} \Rightarrow \text{frequency of the class preceding the modal class} = x$$

$$f_{m+1} \Rightarrow \text{frequency of the class succeeding the modal class} = 12$$

$$\begin{aligned} \text{Mode} &= l_1 + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times i \\ &= 30 + \left\{ 10 \times \frac{(16-x)}{(32-x-12)} \right\} = 36 \end{aligned}$$

$$\Rightarrow \frac{10 \times (16-x)}{(20-x)} = 6$$

$$\Rightarrow 160 - 10x = 120 - 6x$$

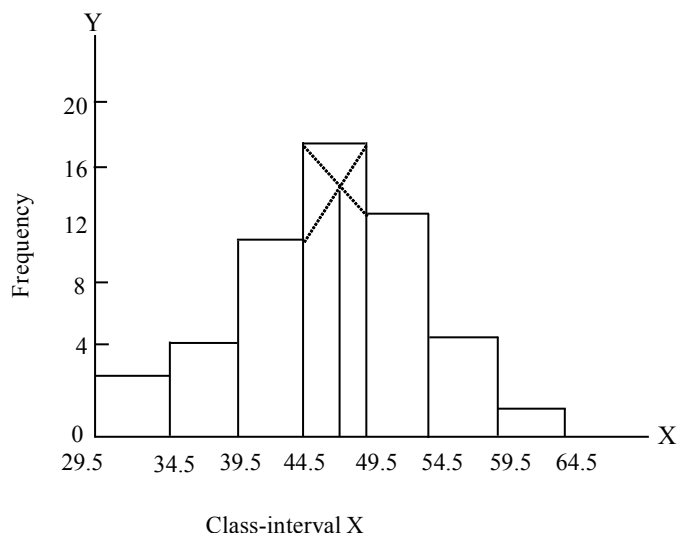
$$\Rightarrow 4x = 40 \Rightarrow x = 10$$

Hence, the missing frequency is 10.

Example 41. The Following table give the weight (in Kg.) of 60 students. Find the value of Mode graphically

Weight (in Kg.)	No. of Students
29.5–34.5	3
34.5–39.5	5
39.5–44.5	12
44.5–49.5	18
49.5–54.5	14
54.5–59.5	6
59.5–64.5	2

Solution.



The modal line touches the X-axis at 47.5. Hence, the value of mode is 47.5 kg. If we calculate mode directly by the formula, it would be:

$$Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} (l_2 - l_1) \text{ Modal class is } 44.5 - 49.5$$

$$Z = 44.5 + \frac{18 - 12}{36 - 12 - 14} (5) = 44.5 + \left(\frac{6}{10} \times 5 \right) = 47.5$$

Example 42. The frequency distribution of marks obtained by 60 students of a class in college is given below:

Marks	30-34	35-39	40-44	45-49	50-54	55-64
No. of Students	3	5	12	18	14	8

Draw Histogram for the distribution and find the modal value.

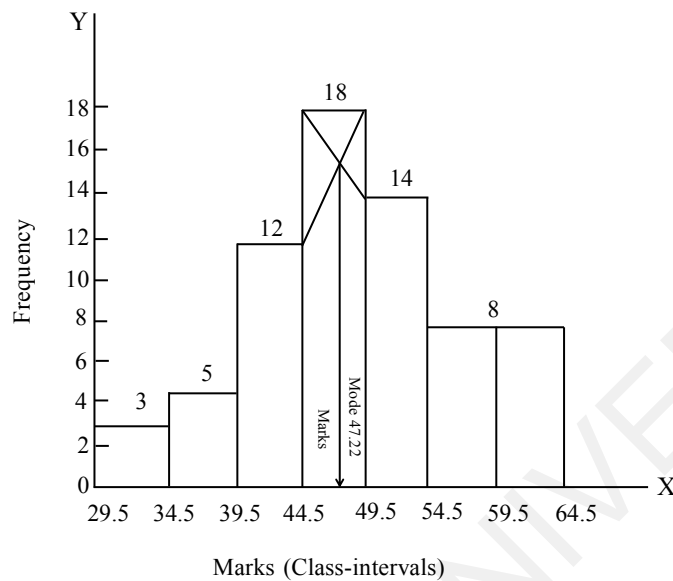
Solution.

To draw the histogram, we first convert the distribution into continuous class intervals as given in the following table:

Marks	No. of Students
29.5-34.5	3
34.5-39.5	5
39.5-44.5	12
44.5-49.5	18
49.5-54.5	14
54.5-59.5	4
59.5-64.5	4

NOTES

NOTES



Mathematical Computation:

Mode lies in the class 44.5 – 49.5

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 - \Delta_2} \times i$$

$$L = 44.5, \Delta_1 = |18 - 12| = 6, \Delta_2 = |18 - 14| = 4, i = 5$$

$$\text{Mode} = 44.5 + \frac{6}{6 + 4} \times 5 = 47.22 \text{ Marks}$$

1.18 EMPIRICAL RELATION AMONG MEAN, MEDIAN AND MODE

For a symmetrical distribution

Mean = Median = Mode.

For a moderately asymmetrical (Skewed) distribution.

$$3 (\text{Mean} - \text{Median}) = \text{Mean} - \text{Mode}$$

$$\Rightarrow \text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

$$\text{Mode} = 3 \text{ median} - 2 \text{ mean}$$

$$\Rightarrow 3 \text{ median} = \text{mode} + 2 \text{ mean}$$

$$\begin{aligned} \Rightarrow 3 \text{ median} &= 3 \text{ mode} + 2 \text{ mean} - 2 \text{ mode} \\ &= 3 (\text{mode}) + 2 (\text{mean} - \text{mode}) \end{aligned}$$

$$\Rightarrow \text{Median} = \text{mode} + \frac{2}{3} (\text{median} - \text{mode})$$

$$\text{Mode} = 3 \text{ median} - 2 \text{ mean}$$

$$\Rightarrow 2 \text{ mean} = 3 \text{ median} - \text{mode}$$

$$\begin{aligned}
 &= 3 \text{ median} - 3 \text{ mode} + 2 \text{ mode} \\
 &= 2 \text{ mode} + 3 (\text{median} - \text{mode})
 \end{aligned}$$

$$\Rightarrow \text{Mean} = \text{mode} + \frac{3}{2} (\text{median} - \text{mode})$$

Note: If in a distribution one value occurs more frequently than any other value, the distribution is called *unimodal*. If in a distribution two different values have equal and maximum frequency associated with them the distribution is known as *bimodal*. When all the values of the observations are non-repetitive or unique, no mode exists.

Example 43. State why the average should be applied in the following circumstance:

- (i) To ascertain the typical shoe size offered in a store.
- (ii) Measuring agricultural holdings' sizes.
- (iii) To ascertain the typical salaries at a certain industrial company.
- (iv) To determine the per capita income in several cities.
- (v) To identify the typical beauty in a class of female students.
- (vi) Average size of readymade garments.
- (vii) Average height of a male in a nation.
- (viii) Average expenditure of a student in a college.

Solution.

(i) Mode, (ii) Median, (iii) Median (sometimes mode is also used), (iv) Mean, (v) Median, (vi) Mode, (vii) Mode, (viii) Mode.

1.19 MERITS, DEMERITS AND USES OF MEAN, MEDIAN AND MODE

Arithmetic Mean

Merits

1. It is the most popular and widely used Central Tendency metric.
2. It is rigidly defined. This signifies that whatever method of computation may be adopted, the result will remain the same.
3. All of the observations are the basis for it.
4. It can be treated algebraically, allowing us to determine the combined mean of two or more groups. In statistical analysis, it is often employed.
5. It is least impacted by sampling-related variations.
6. It is very simple to understand.
7. It can be used for comparison purpose.
8. The algebraic sum of all values' departures from their A.M. equals zero. This serves as the foundation for the quick method for finding the mean.

NOTES

NOTES

Demerits

1. It is highly affected by the extreme values. For example, in an industry the workers get less pay as compared with top officials, But if we calculate the average wage paid so all employees, it will not represent the workers because the high salary of top officials will increase the average wage much more than what a normal worker gets.
2. It's possible that the calculated average value won't match any of the provided values; for instance, the mean of the numbers 5, 9, and 10 is 8, which isn't a number in the series.
3. It cannot be determined graphically or through visual examination.
4. It provides both large and small observations, which skew the average, equal weight.
5. If even one observation in the series is missing, it is impossible to calculate the arithmetic mean. Also, in distribution with open-ended classes the mean cannot be determined without making certain assumptions regarding the class-end values.
6. Two presumptions are made in order to calculate the mean in continuous series:
 - (i) The frequencies are evenly distributed.
 - (ii) The middle value of each class serves as the class representative.

This might not always be the case.

7. On occasion, it comes to irrational conclusions. As an illustration, if there are three classes with 9, 11, and 15 students each, the average number of pupils in

$$\text{each class} = \frac{9 + 11 + 15}{3} = \frac{35}{3} = 11.67 \text{ which is absurd.}$$

Uses

Arithmetic mean is very common average and is widely used. It is generally used in all the subjects of studies may be Social and Economic studies. It is frequently used in Business and Commerce, for example, 'Average cost of production', 'Average price' and 'Average yield per acre', etc. It is used to compute additional statistical measures, such as the standard deviation and coefficient of variation, in addition to comparing two or more series.

Median**Merits**

1. If the data are qualitative, the only average that may be utilised is the median. For instance, we can determine the median for traits like honesty, cleverness, and beauty.
2. Extreme values have no impact on it. It is frequently employed as a measure of Central tendency in asymmetrical distributions because of this property.
3. If the class intervals are open-ended (*i.e.*, undetermined). Median can be calculated unless it falls in the open-ended class interval.

4. The total absolute deviations of all values from the median are at their lowest.
5. It can be located graphically.
6. *Even if the data is not complete, i.e., some values are missing, the median can be calculated if the number of items and the middle item is known.*
7. It can occasionally be calculated from visual inspection.
8. It is rigorously defined, meaning that no matter what computation method is used, the value of the function remains constant.
9. The median value is unaffected if the class intervals are different or unequal.

NOTES**Demerits**

1. It cannot be treated algebraically, hence it is impossible to calculate the combined median of two or more groups.
2. It is affected by sampling fluctuations more than the mean.
3. It is the average of the two middle values if there are an equal number of observations. Thus, it cannot be described precisely under such circumstances.
4. It fails to represent a satisfactory average when there is a great variation among the items of the series.
5. It gives equal weight to all observations but if extreme observations are given more weight, then median as an average becomes inappropriate.
6. $\text{Median} \times \text{Number of observations} = \text{sum of the observations.}$

Uses

It proves very useful when extreme values of the series are either not available or are abnormal.

Mode**Merits**

1. Mode is the value which occurs most frequently in the series, *i.e.*, the value having maximum frequency. Hence it is one of the values in the series and not an isolated value.
2. It can be found graphically.
3. It is not affected by extreme value.
4. It is of great use to the manufacturers of garments, shoes, etc., because most common size, *i.e.*, the model size helps them to decide the quantity to be manufactured for each size.
5. It can be found by inspection.
6. It is the most used average in certain situations, *e.g.*, the average number of pupils in a section, the average size of shoes, the average marks for the class, etc.

NOTES

Demerits

1. It is not rigidly defined. If we calculate mode by different methods, results may be different.
2. Mode can be called the representative average only if the number of values is too large.
3. Series may be bimodal or multimodal, *i.e.*, the mode is not unique.
4. It is most affected by fluctuations of sampling.
5. It is not capable of further algebraic treatment. We cannot find the combined mode.
6. Mode is based only on concentrated values.
7. It does not yield the sum of the value as the mean does when multiplied by the number of observations.

1.20 COMPUTATIONS OF QUARTILES, DECILES AND PERCENTILES

These partition values are calculated in the exact same way as the Median is calculated. The values of the lower (Q_1) and upper (Q_3) quartiles would be the values of and items, respectively, in a series of discrete series and individual observations. The deciles' value in such a case would be as follows:

$$D_1 = \text{value of } \left(\frac{N+1}{4} \right)^{\text{th}} \text{ item}$$

$$D_2 = \text{value of } \frac{2(N+1)}{10}^{\text{th}} \text{ item}$$

$$D_3 = \text{value of } \frac{3(N+1)}{10}^{\text{th}} \text{ item and so on.}$$

The value of the percentiles would be:

$$P_1 = \text{value of } \left(\frac{N+1}{100} \right)^{\text{th}} \text{ item}$$

$$P_2 = \text{value of } \frac{2(N+1)}{100}^{\text{th}} \text{ item}$$

.....

$$P_{99} = \text{value of } \frac{99(N+1)}{100}^{\text{th}} \text{ item}$$

In continuous series in the calculation of quartiles, deciles and percentiles $\frac{N+1}{4}$,

$\frac{N+1}{10}$ and $\frac{N+1}{100}$ would be replaced by $\frac{N}{4}$, $\frac{N}{10}$ and $\frac{N}{100}$ respectively. The values would have to be interpolated here as was done in case of the computation of Median.

The following examples would illustrate the above points.

Example 44. The following table gives the marks obtained by a batch of 30 B. Com. Students in a class-test in statistics.

Roll No.	Marks Obtained	Roll No.	Marks Obtained
1	33	16	24
2	32	17	33
3	55	18	42
4	47	19	38
5	21	20	45
6	50	21	26
7	27	22	33
8	12	23	44
9	68	24	48
10	49	25	52
11	40	26	30
12	17	27	58
13	44	28	37
14	48	29	38
15	62	30	35

Find the value of Q_1 , Q_2 , D_6 and P_{70}

Solution.

Obtained by 30 students arranged in order of magnitude.

Serial No.	Marks	Serial No.	Marks	Serial No.	Marks
1	12	11	33	21	47
2	17	12	35	22	48
3	21	13	37	23	48
4	24	14	38	24	49

NOTES

NOTES

5	26	15	38	25	50
6	27	16	40	26	52
7	30	17	42	27	55
8	32	18	44	28	58
9	33	19	44	29	62
10	33	20	45	30	68

Q_1 = The value of $\left(\frac{N+1}{4}\right)^{\text{th}}$ or $\left(\frac{30+1}{4}\right)^{\text{th}}$ or 7.75th item

= Size of 7th item + 0.75 (size of 8th item minus size of 7th item)

$$= 30 + 0.75 (32 - 30) = 31.5$$

Q_3 = The value of $\frac{3(N+1)}{4}$ or $\frac{3(30+1)}{4}$ or 23.25th item

$$= 48 + 0.25 (49 - 48) = 48.25$$

D_6 = The value of $\frac{6(N+1)}{10}$ item for $\frac{6(30+1)}{10}$ or 18.6th item

$$= 44 + 0.6 (44 - 44) = 44$$

P_{70} = The value of $\frac{70(N+1)}{100}$ item or $\frac{70(30+1)}{100}$ item or the value of 21.7th item.

$$= 47 + 0.7(48 - 47) = 47.7.$$

LIST OF FORMULAE

Abbreviations

\bar{X} = Arithmetic Mean

X = The variable

ΣX = Sum of all the values of the variable X .

N = Number of values (or observations) or (Σf)

f = Frequency

ΣfX = Sum of the product of variable (X) and its frequency (f)

m = Mid-value of the class interval

Σfm = Sum of the products of mid-points and the frequencies

A = Assumed mean.

$X - \bar{X}$ = x , i.e., deviations of X variable from the mean.

\bar{X}_{12} = Combined mean of two groups.

\bar{X}_1	= Arithmetic of the first group
\bar{X}_2	= Arithmetic group of second group.
N_1	= Number of observations in the first group
N_2	= Number of observations in the second group.
d	= $X - A$, i.e., deviation of X variable from an assumed mean A .
Σd	= Sum of the deviations of X variable taken from an assumed mean.
c	= Common factor (in most cases common factor is taken as class interval).
d'	= $\frac{X - A}{c}$, i.e., step deviations of x -variable from assumed mean a and divided by common factor.
$\Sigma d'$	= Sum of step deviations
Σfd	= Sum of the product of frequencies and their respective deviations.
$\Sigma fd'$	= Sum of the product of deviations and their respective step deviations.
$(X - \bar{X})^2$	= x^2 , i.e., square of the deviations of X variable from mean.
\bar{X}_w	= Weighted arithmetic Mean
W	= Weights
ΣWX	= Sum of the product of variable X and weight W .

Arithmetic Mean, Properties and Weighted Mean

Type of Series Method	Direct Method	Short Cut Method	Step Deviation
1. Individual Observations $\bar{X} = A + \frac{\Sigma d'}{N} \times c$ (Ungrouped data)	$\bar{X} = \frac{\Sigma X}{N}$	$\bar{X} = A + \frac{\Sigma d}{N}$	
2. Discrete Series Frequency $\bar{X} = A + \frac{\Sigma d'}{N} \times c$ Distributions	$\bar{X} = \frac{\Sigma fx}{N}$	$\bar{X} = A + \frac{\Sigma fd}{N}$	
3. Continuous Series $\bar{X} = A + \frac{\Sigma d'}{N} \times c$ (Grouped data)	$\bar{X} = \frac{\Sigma fm}{N}$	$\bar{X} = A + \frac{\Sigma fd}{N}$	

m in the mean or mid-value of the group; c = Common factor.

NOTES

NOTES

Check Your Progress

I. Multiple Choice Questions

- The algebraic sum of the deviations of 10 observations measured from 15 is 7. Thus, the mean is
 (a) 105 (b) 70
 (c) 15.7 (d) None of these.
- The algebraic sum of the deviations from mean is _____.
 (a) maximum (b) least
 (c) zero (d) None of these.
- The sum of squares of deviations from mean is _____.
 (a) least (b) zero
 (c) maximum (d) None of these.
- An additional observation 15 is included in a series of 11 observations and its mean remains unaffected. The mean of the series was _____.
 (a) 11 (b) 15
 (c) 165 (d) 4.
- The median of a series is 10. Two additional observations, 7 and 20 are added to the series. The median of the new series will be
 (a) 9 (b) 20
 (c) 7 (d) 10.

II. State whether the following Statements are True or False

- Arithmetic mean is always the best measure of central tendency.
- The median is a computational measure of central tendency.
- The median of distribution can be determined graphically.
- The value of mean is least affected by sampling fluctuations.
- The sum of the absolute deviations from arithmetic mean is the least.

III. Fill in the Blanks

- Measures of Central tendency are called averages of _____. (first order, third order)
- If a constant amount is added to each value in a distribution then the arithmetic mean _____ (remains unchanged, is also added by the same constant amount)
- If the arithmetic mean of the ages of a group of boys is 20 years, then the arithmetic mean of their ages after 2 years will be _____. (22 years, the same)
- For calculating mode, class intervals have _____ (to be equal, no restriction)
- For finding quartiles, the data should be arranged in _____. (ascending order, descending order)

1.21 ANSWERS TO 'CHECK YOUR PROGRESS'

I. Multiple Choice Questions

1. (c)
2. (c)
3. (a)
4. (b)
5. (d)

II. State Whether the following Statements are True or False

1. False (In many situations like qualitative data, open-ended class intervals, size of garments, etc., mean cannot be used.)
2. False (Median is a positional measure.)
3. True
4. True
5. False

III. Fill in the Blanks

1. first order,
2. is also added by the same constant amount;
3. 22 years;
4. equal;
5. ascending order;

1.22 SUMMARY

- **Meaning and Definition of Central Value or Average:** It is one single value that represents the whole series.
- **Objects and Functions of Averages:** A tool to represent the salient features of a mass of complex data, useful for comparison, to know about universe from sample, helpful for making planning decisions in various fields, to establish mathematical relationship.

Mode. Mode is the value occurring most frequently in a series (or group) of items and around which the other items are distributed most densely.

- **Choice of a Suitable Average**

Arithmetic Mean (Mean): If all values are given equal importance, e.g., average marks, average salary, average profit, A.M. is the most preferred average.

Median: In case of qualitative data which cannot be measured quantitatively, e.g., average intelligence, beauty, honesty.

NOTES

NOTES

Mode: Used in business to determine most stylish or most frequently occurring item, e.g., modal stylish shirt, modal accidental spot, most fashionable garment, modal garment, modal pay of workers in a factory, modal size of shoes.

- **Mathematical Properties of Median**

- (i) It is a positional average, therefore influenced by the position of items in the arrangement and not by the size of items.
- (ii) The sum of the deviations of items about median, ignoring \pm signs, will be less than the sum of deviations about any other point.

- **Relationship between Mean, Median and Mode**

- (i) For symmetrical distribution (bell-shaped curve)

Mean = Median = Mode.

- (ii) For positively skewed (asymmetric) distribution

Mean > Median > Mode.

- (iii) For negatively skewed (asymmetric) distribution Mean < Median < Mode.

- (iv) For moderately asymmetric distribution mean – mode = 3(mean – median)

1.23 KEY TERMS

- Arithmetic mean is the average of first order.
- Median is the middle most value.
- Mode is the most frequently occurring value.
- Quartiles divide the distribution in to four equal parts.
- Deciles divides the distribution in to 10 equal parts and percentiles in to 100 equal parts.

1.24 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. Why are measures of central value calculated?
2. Define an average. Why do we call an average to be a measure of central tendency?
3. What are the different types of averages? Give name of each.
4. Give four characteristics of a good average.
5. Define arithmetic mean. In one sentence, differentiate a simple arithmetic mean from a weighted arithmetic mean.

Long Answer Questions

Questions about Calculating Arithmetic Mean

1. Calculate the A.M. of from the following data.

Marks	4	8	12	16	20
Number of Students	6	12	18	15	9

[Ans. 12.6]

2. The mean marks of 100 students were found to be 40. Later on it was discovered that a score of 53 was misread as 83. Find the corrected mean corresponding to the corrected score.

[Ans. 39.7 marks]

3. The mean age of a combined group of men and women is 30 years. If the mean age of the group of men is 32 years and that for the group of women is 27 years, find the percentage of men and women in the group.

[Ans. 60% men and 40% women]

4. The mean weight of 150 students in a class is 60 kg. The mean weight of boys in the class is 70 kg and that of the girls is 55 kgs. Find the number of girls in the class.

[Ans. 100 girls]

5. In a certain examination, the “average grade of all students in Class ‘A’ is 68.4 and of students in class ‘B’ is 71.2. If the average of both classes combined is 70, find the ratio of the number of students in class ‘A’ to the number of students in class ‘B’.

[Ans. 3/4]

6. The mean weight of 15 students is 110 kg. The mean weight of 5 of them is 100 kg and another 5 is 125 kg. What is the mean weight of the remainder?

[Ans. 105 kg]

7. Fifty students took up a test. The result of those who passed the test is given below:

Marks	4	5	6	7	8	9
No. of Students	10	10	9	6	4	3

If the average of all the 50 students was 5.16 marks, find the average of those who failed.

[Ans. 1.625]

8. The mean marks in statistics of 100 students of a class was 72. The mean of marks of boys are 75, while their number was 70. Find out the mean marks of girls in the class.

[Ans. 65]

9. The mean age of a combined group of men and women is 25 years. The mean age of the group of men is 26 years and that of the group of women is 21 years. Find out the percentage of men and women in the group.

[Ans. 80]

NOTES

NOTES

10. Calculate the mean from the following data:

x	10	12	18	16	15	19	18	17
f	8	12	6	3	19	13	17	14

[Ans. 15.826]

11. Find the average number of children per family for the two sub-groups separately as well as for the distribution as a whole:

Group 1		Group 2	
No. of Children	No. of Families	No. of Children	No. of Families
0	10	4-5	20
1	50	6-7	12
2	60	8-9	4
3	40	10-11	4
	160		40

[Ans. 6.1, 2.67]

12. Find the A.M. of first natural numbers. [Ans. $x + 1/3$]
 13. Find the A.M. of first 100 natural numbers. [Ans. 50.5]
 14. A.M. of the following frequency distribution is 67.45. Find f_3 :

Height (in inches)	61	64	67	70	73
Frequency	15	54	f_3	81	24

[Ans. 113]

15. The arithmetic mean of 10 values is 12.5. The A.M of first 6 values is 15 and that of last 5 values is 10. Find the 6th value. [Ans. 15]

Questions for Calculating Median and Mode

1. The marks obtained by 15 students in a class test are given below: 6, 9, 10, 12, 18, 19, 23, 23, 24, 28, 37, 48, 49, 53 and 60

Find (i) Mean, (ii) Median [Ans. $\bar{X} = 27.93$, Median = 23]

2. Calculate mean, median and mode from the following data of the heights in inches of a group of students: 61, 62, 63, 61, 63, 64, 64, 60, 65, 63, 64, 65, 66, 64

Now suppose that a group of students whose heights are 60, 66, 59, 68, 67 and 70 inches is added to the original group. Find mean, median and mode of the combined group.

[Ans. $\bar{X} = 63.2$, $M = 63.5$; $Mo = 64$ for the original group; $\bar{X} = 63$, 42 Median = 64, Mode = 64 for the combined.]

3. The numbers of telephone calls received in 245 successive one minute intervals at an exchange are shown in the following frequency distribution:

Number of Calls	0	1	2	3	4	5	6	7
Frequency	14	21	25	43	51	40	39	12

Calculate mean, median and mode.

[Ans. Mean = 3.763, Median = 4, Mode = 4]

4. The numbers of fully formed tomatoes on 100 plants were counted with the following results:

Plants	Tomatoes
2	0
5	1
7	2
11	3
18	4
24	5
12	6
8	7
6	8
4	9
3	10

- (i) How many tomatoes were there in all?
 (ii) What was the average number of tomatoes per plant?
 (iii) What was the mode or modal number of tomatoes?

[Ans. (i) 486, (ii) $\bar{X} = 4.86$, (iii) Mode = 5]

NOTES

1.25 REFERENCES

1. D.N. Elhance, Veena Elhance and B.M. Aggarwal, 2007, Fundamentals of Statistics. Kitab Mahal, New Delhi.
2. B.M. Aggarwal, 2012, Business Statistics (With Lab Work), Himalaya Publishing House, Mumbai.
3. B.M. Aggarwal, Dr. Puja A. Gulati, Neha Aggarwal, 2022, Statistics for Business and Economics. Kitab Mahal, New Delhi.

NOTES

Unit II Measures of Dispersion

Learning Objectives:

By the end of this unit the learners would be able to:

- To understand the importance of dispersion
- Uses of S.D in difference situations
- To understand coefficient of variation

Structure:

- 2.1 Introduction
- 2.2 Definition
- 2.3 Objectives of Dispersion
- 2.4 Importance of Dispersion
- 2.5 Characteristics of a Good Measure of Dispersion
- 2.6 Merits and Demerits of Measures of Dispersion
- 2.7 The Range
- 2.8 The Interquartile Range or the Quartile Deviation
- 2.9 The Mean Deviation
- 2.10 Variance Standard Deviation
- 2.11 Coefficient of Variation
- 2.12 Answers to 'Check Your Progress'
- 2.13 Summary
- 2.14 Key Terms
- 2.15 Self-Assessment Questions and Exercises
- 2.16 References

2.1 INTRODUCTION

Averages of the first order are the name given to measures of central tendency. However, these averages are not sensitive to the data's variation. The mean, median, and mode of two distributions may be the same, but the data variability in the two situations may be very different. Take a look at groups A and B.

2.2 DEFINITION

The spread or scatter of numbers from a measure of central tendency is known as dispersion.

A measure of dispersion is intended to quantify how far each observation (or item) deviates from the mean. Here, we will merely take into account the quantity of variance rather than its direction.

D.C. Brooks and **W.F.L. Dick** define dispersion as “*Dispersion or spread is the degree of the scatter or variation of variable about a central value*”.

A measure of variation or dispersion, which typically takes the form of an average departure from a central value, reflects the level of scatter indicated by observations.

In the words of Fritz Kafka and George Simpson, “An average does not give the whole tale. It is hardly a complete representation of a mass unless we are aware of how the different components disperse around it. If we are to determine how representative the value is, a more thorough description of the series is required.”

According to **Spiegel**, “*The degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data*”.

Because they are based on the variances of the various values from the mean or other measures of central tendency, which are known as averages of the first order, measures of dispersion are referred to as averages of the second order.

The difference between an average and the actual value is called **deviation** from the average value and measures the variability, spread or scatter of the value from that average. Deviations can be taken from an arbitrary value.

NOTES

2.3 OBJECTIVES OF DISPERSION

Main objectives of measuring dispersion are:

1. One of the main goals of evaluating dispersion is to understand how often various values of an item deviate from the average of a series.
2. To understand the distribution of values on each side of the central tendency or the make-up of a series.
3. Understanding the difference between the highest and lowest value (i.e., the range of values).
4. To assess the degree of variation by comparing the discrepancy between two or more series expressed in various units.
5. To determine whether or not the central tendency accurately captures the series. The metrics of central tendency do not accurately depict the series if the dispersion is higher.

NOTES

2.4 IMPORTANCE OF DISPERSION

1. A conclusion formed just from the central tendency is meaningless without understanding the variance among the series' distinct elements with respect to the average.
2. Dispersion can be used to quantify wealth and income distribution disparities.
3. Dispersion is used to compare and gauge the level of monopolistic and economic power in a nation.
4. Price and output control both use dispersion.

2.5 CHARACTERISTICS OF A GOOD MEASURE OF DISPERSION

1. It should be straightforward to comprehend and compute.
2. It should have strict definitions.
3. It must be based on every component of the series.
4. Extreme goods shouldn't have a significant negative impact.
5. The effects of sampling variations should be minimised.
6. It ought to be open to more algebraic analysis.

2.6 MERITS AND DEMERITS OF MEASURES OF DISPERSION

As was already said, there are various metrics of dispersion, each with proportional benefits and drawbacks. However, the benefits and drawbacks that apply to all categorical measures of dispersion are listed below.

Merits

1. They reveal a statistical series' dispersal nature.
2. They discuss a series' average value's dependability or reliability.
3. They allow statisticians to compare two or more statistical series based on characteristics like their uniformity, consistency, or equity.
4. They make it possible for one to control the variability of an event that is under his control.
5. They make it easier to conduct additional statistical analysis of the series using tools such as variance analysis, coefficient of skewness, coefficient of kurtosis, coefficient of correlation, etc.
6. They add to the measures of central tendency by providing ever-more details about the characteristics of a series.

NOTES

Demerits

1. They are susceptible to misunderstandings and incorrect generalisations by a statistician with a biased viewpoint.
2. There are numerous ways to calculate the dispersion, making them susceptible to producing inaccurate findings.
3. With the exception of one or two, the majority of dispersion methods entail challenging computational procedures.
4. They are unable to provide any information about the symmetrical or skewed nature of a sequence on their own.
5. Similar to measurements of central tendency, most measures of dispersion do not provide a layperson with a clear understanding of a series.

2.7 THE RANGE

The range, which is the difference between a data set's maximum value and minimum value, is the most basic indicator of dispersion.

Example 1. Find the range for the following three sets of data:

Set 1	15	15	15	15	15	15	15	15	15	5
Set 2	8	7	15	11	12	5	13	11	15	9
Set 3	5	5	5	5	5	15	15	15	15	15

Solution.

The greatest and lowest numbers in each of these three sets are 15 and 5, respectively.

The range is 10 in each instance since it is the difference between the data's maximum value and least value. The range, however, does not provide any insight into the spread or dispersion of the series between the greatest and lowest number. The information mentioned above makes this clear.

The difference between the upper limit of the highest class and the lower limit of the lowest class is used to compute the range in a frequency distribution.

Example 2. Find the range for the following frequency distribution:

<i>Size of Item</i>	<i>Frequency</i>
20 - 40	7
40 - 60	11
60 - 80	30
80 - 100	17
100 - 120	5
Total	70

NOTES

Solution.

In this case, the lowest class has a lower limit of 20 while the highest class has an upper limit of 120. The range is therefore $120 - 20 = 100$. Keep in mind that the frequencies have no effect on the range. The formula $L-S$, where L is the biggest value and S is the smallest value in a distribution, is used to compute the range symbolically. With the following formula, the coefficient of range is determined:

$$\frac{L-S}{L+S} \text{ This is the relative measure}$$

The following example will show that the coefficient of range is more suitable for comparison-related purposes.

Example 3. Calculate the coefficient of range separately for the two sets of data given below:

Set 1	8	10	20	9	15	10	13	28
Set 2	30	35	42	50	32	49	39	31

Solution.

It can be seen that the range in both the sets of data is the same:

$$\text{Set 1} \quad 28 - 8 = 20$$

$$\text{Set 2} \quad 50 - 30 = 20$$

$$\text{Coefficient of range in set 1 is: } \frac{28-8}{28+8} = \frac{20}{36} = 0.55$$

$$\text{Coefficient of range in set 2 is: } \frac{50-30}{50+30} = \frac{20}{80} = 0.25$$

Limitations

There are several significant range restrictions:

1. It does not include all of the things in a distribution and is solely based on two items.
2. It varies greatly from sample to sample when based on the same population.
3. It fails to provide any insight into the distribution pattern. The information provided in instances 1 and 3 made this clear.
4. Lastly, it is impossible to compute the range for open-ended distributions. Despite these range restrictions, there are certain benefits as well.

Advantages

1. It is primarily utilised when one needs to quickly get a sense of how variable a piece of data is.
2. The range is thought to be an appropriate indicator of variability when the sample size is very small. As a result, it is frequently employed in quality control when it is necessary to continuously monitor the variability of raw materials or final goods.

- When predicting the weather, the range is a useful metric. The meteorological division uses the range by providing the highest and lowest temperatures. The average person can use this information to determine the magnitude of potential temperature variation on a given day.

2.8 THE INTERQUARTILE RANGE OR THE QUARTILE DEVIATION

NOTES

The quartile deviation, also known as the interquartile range, is a more accurate indicator of variation within a distribution than the range. By omitting the 25% of the distribution at either end, the middle 50% of the distribution is used in this instance. The interquartile range, then, represents the variation between the third quartile and the first quartile.

Symbolically, interquartile rang = $Q_3 - Q_1$

As illustrated below, the interquartile range is frequently lowered to the semi-interquartile range or quartile deviation:

$$\text{Semi-interquartile range or Quartile deviation} = \frac{Q_3 - Q_1}{2}$$

The middle 50% of the items have little variance when the quartile deviation is low. In contrast, a large quartile deviation indicates that the centre 50% of the items vary widely. It should be noticed that the two quartiles, Q_3 and Q_1 , are equally spaced from the median in a symmetrical distribution. Symbolically

$$M - Q_1 = Q_3 - M$$

The majority of business and economic statistics are asymmetrical, thus this is rarely the case. However, we can suppose that the interquartile range contains roughly 50% of the observations.

It should be highlighted that the quartile deviation or the interquartile range is an exact indicator of dispersion. It can be transformed into the following relative measure of dispersion:

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

The calculation of the upper and lower quartiles is all that is necessary to determine a quartile deviation. It is not attempted here to compute the two quartiles because that has already been covered in the chapter before.

Merits of Quartile Deviation (QD)

- It is regarded as a better measure of dispersion than range.
- It works well in the situation of open-ended distribution.
- It is especially useful in highly skewed or unpredictable distributions since it is unaffected by the extreme values of a distribution.

NOTES

Limitations of Quartile Deviation

1. It doesn't encompass every item in a distribution, just like the range.
2. It cannot be manipulated mathematically.
3. Based on the same population, it differs significantly from sample to sample.
4. Because it is a positional average, it is not taken into account as a dispersion measure.

It does not depict a distribution around an average, only a distance on a scale.

The interquartile range or quartile deviation has a limited practical utility due to the aforementioned drawbacks.

2.9 THE MEAN DEVIATION

The average deviation is another name for the mean deviation. The absolute amounts by which the various items depart from the mean are averaged, as the name suggests. We ignore positive and negative signs when computing the mean deviation since positive and negative deviations from the mean are equivalent. Symbolically

$$MD = \frac{\sum |x|}{n}$$

where, MD = mean deviations

$|x|$ = deviation of an item from the mean,* ignoring positive and negative signs

n = the total number of observations.

Let us take an example

Example 4.

<i>Size of Item</i>	<i>Frequency</i>
2-4	20
4-6	40
6-8	30
8-10	10

Solution.

We set up worksheet for calculating the mean deviation

<i>Size of Items</i>	<i>Mid-points (m)</i>	<i>Frequency (f)</i>	<i>mf</i>	<i>d from</i>	<i>f d </i>
2-4	3	20	60	-2.6	52
4-6	5	40	200	-0.6	24
6-8	7	30	210	1.4	42
8-10	9	10	90	3.4	34
Total		100	560		152

* Occasionally, deviations are taken from the median.

$$x = \frac{\sum fm}{n} = \frac{560}{100} = 5.6$$

$$MD(\bar{x}) = \frac{\sum f |d|}{n} = \frac{152}{100} = 1.52$$

NOTES**Merits of Mean Deviation**

1. The fact that mean deviation is straightforward to comprehend and straightforward to compute is a big benefit.
2. It considers each and every component of the distribution. As a result, a change in any item's value will have an impact on how much the mean deviation is off.
3. The mean deviation's value is less affected by the values of extreme items.
4. Since deviations from a central value are taken into account, it is feasible to compare the ways in which various distributions are formed in meaningful ways.

Limitations of Mean Deviation

1. It cannot be subjected to more algebraic analysis.
2. It might occasionally produce inaccurate results. When deviations are taken from the median rather than the mean, the mean deviation produces the best results. However, median is not a good indicator in a series with wide variability in the items.
3. The method is incorrect just mathematically since it ignores the algebraic signs when deviating from the mean.

Example 5. Calculate Semi-interquartile range and the Coefficient of Q.D. from the following

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of Student	11	18	25	28	30	33	22	15	22

Solution.

Calculation of Quartile Deviation

Marks (X)	Frequency (f)	Cumulative Frequency (c.f.)
0-10	11	11
10-20	18	29
20-30	25	54
30-40	28	82
40-50	30	112
50-60	33	145
60-70	22	167
70-80	15	182
80-90	22	204

NOTES

First Quartile = Q_1 = the value of $\left(\frac{204}{4}\right)^{\text{th}}$ or 51st item which is in 20-30 group.

Third Quartile = Q_3 = the value of 3 $\left(\frac{204}{4}\right)^{\text{th}}$ or 153rd item which is in 60-70 group.

$$\text{The value of } Q_1 = l_1 + \frac{l_2 - l_1}{f_1} \left(\frac{N}{4} - cf_{-1} \right) = 20 + \frac{10}{25} (51 - 29) = 28.8$$

$$\text{The value of } Q_3 = l_1 + \frac{l_2 - l_1}{f_1} \left(\frac{3N}{4} - cf_{-1} \right) = 60 + \frac{10}{22} (153 - 145) = 63.64$$

$$Q.D. = \frac{Q_3 - Q_1}{2} = \frac{63.64 - 28.8}{2} = 17.42 \text{ marks}$$

$$\text{Coefficient of } Q.D. = \frac{Q_3 + Q_1}{Q_3 - Q_1} = \frac{63.64 + 28.8}{63.64 - 28.8} = 0.37$$

Example 6. Calculate Quartile Deviation and its relative measure:

Variable	Frequency	Variable	Frequency
20-29	306	50-59	96
30-39	182	60-69	42
40-49	144	70-79	34

Solution.

Calculation of Quartile Deviation*

Variable (X)	Frequency (f)	Cumulative Frequency (c.f.)
19.5-29.5	306	306
29.5-39.5	182	488
39.5-49.5	144	632
49.5-59.5	96	728
59.5-69.5	42	770
69.5-79.5	34	804

Q_1 = The value of $\left(\frac{804}{4}\right)^{\text{th}}$ of 201st item

which lies in 20-29 group [which in exclusive series] = (19.5 - 29.5)*

$$Q_1 = l_1 + \frac{l_2 - l_1}{f_1} \left(\frac{N}{4} - cf_{-1} \right) = 19.5 + \frac{10}{306} (201 - 0) = 19.5$$

$$+ \left(\frac{10}{306} \times 201 \right) = 26.07$$

Q_3 = The value of $3\left(\frac{804}{4}\right)^{\text{th}}$ or 603rd item which lies in the 40-49 group.

[which in exclusive series = (39.5 - 49.5)]

$$Q_3 = 39.5 + \frac{10}{144} (603 - 488) = 39.5 + \left(\frac{10}{144} \times 115 \right) = 47.49$$

$$(i) \text{ Quartile Deviation} = \frac{Q_3 - Q_1}{2} = \frac{47.49 - 26.07}{2} = 10.71$$

$$(ii) \text{ Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{47.49 - 26.07}{47.49 + 26.07} = 0.287$$

Example 7. Estimate an appropriate measure of dispersion for the following data:

Income (₹)	No. of Persons
Less than 50	54
50-70	100
70-90	140
90-100	300
110-130	230
130-150	126
Above 150	51
	1000

* There is no class preceding the class 19.5–29.5.

NOTES

NOTES

Solution.

Since the data has open ends. Quartile Deviation would be a suitable measure:

Calculation of Quartile Deviation

Income (₹) (X)	No. of Persons (f)	Cumulative Frequency (c.f)
Less than 50	54	54
50-70	100	154
70-90	140	294
90-110	300	594
110-130	230	824
130-150	125	949
Above 150	51	1000

Q_1 is the value of $\frac{1000}{4}$ or 250th item which is in 70-90 group.

Q_3 is the value of $3\left(\frac{1000}{4}\right)$ or 750th item which is in 110-130 group.

$$Q_1 = 70 + \frac{20}{140}(250 - 154) = 70 + \left(\frac{20}{140} \times 96\right) = 70 + 13.7 = 83.7$$

$$Q_3 = 110 + \frac{20}{230}(750 - 594) = 110 + \left(\frac{20}{230} \times 156\right) = 110 + 13.7 = 123.5$$

$$\text{Quartile Deviation or } Q.D. = \frac{Q_3 - Q_1}{2} = \frac{123.5 - 83.7}{2} = 19.9$$

2.10 VARIANCE STANDARD DEVIATION

The most significant, trustworthy, and often used measure of dispersion is the standard deviation. This measure of variation is referred to as "standard" for the following reasons, most likely:

- Of all the variations measures, it is the most widely used and flexible in terms of the range of applications.
- Any symmetrical curve, such as a normal curve, has a constant area within a fixed range of standard deviations from the mean on either side of it. For example, any normal curve's area within a range of 1 standard deviation from the mean is always 68.27 percent of the total area, and its area within a range of 2 standard deviations is 95.45 percent of the total area.

- (iii) Root mean square deviation from the mean is the least because the sum of squares of deviations from the mean is the lowest compared to the sum of squares of deviations from the median or mode.

Because it is employed in so many other statistical procedures, such as sampling techniques, correlation and regression analysis, determining the coefficient of variation, skewness, kurtosis, etc., the standard deviation is the most crucial of all the measures of dispersion. Other names for standard deviation include "Mean Error," "Mean Square Error," and "Root-Mean Square Deviation." The standard deviation is always computed around the mean; in contrast to the mean deviation, which may be computed around any average, the mean deviation is **the positive squareroot of the arithmetic mean of the squared deviations of all the values from their mean***. Positive SD is always true. It is obvious that a measurement of the spread in a group of data is the standard deviation. Each deviation from the mean would be 0 if every observation's values were the same. The minimal value of the standard deviation, 0, would exist in such a perfectly uniform distribution. On the other hand, the standard deviation increases as the number of items that deviate from the mean increases.

The standard deviation is the square root of the arithmetic mean of the squares of all deviations, where deviations are measured from the arithmetic mean of the observations, according to Yule and Kennard. Because the sum of squared deviations from the mean equals the smallest, the standard deviation is always calculated from the mean.

Second Formula for S.D.

$$\begin{aligned}
 (\text{SD})^2 &= \frac{\sum (X - \bar{X})^2}{N} = \frac{\sum X^2 + \sum \bar{X}^2 - 2\sum X\bar{X}}{N} \\
 &= \frac{\sum X^2}{N} + \frac{N\bar{X}^2}{N} - 2\bar{X} \frac{\sum X}{N} \\
 &= \frac{\sum X^2}{N} + \bar{X}^2 - 2\bar{X}^2 = \frac{\sum X^2}{N} - \bar{X}^2 \\
 &= \frac{\sum X^2}{N} - \left(\frac{\sum X}{N} \right)^2.
 \end{aligned}$$

$$\begin{aligned}
 \therefore \sigma &= \text{S.D.} = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N} \right)^2} \\
 \text{and variance} &= \frac{\sum X^2}{N} - \left(\frac{\sum X}{N} \right)^2 = (\text{SD})^2
 \end{aligned}$$

In case frequencies of different values (or observations) are given, then

$$\text{S.D.} = \sqrt{\frac{\sum fX^2}{\sum f} - \left(\frac{\sum fX}{\sum f} \right)^2}$$

NOTES

* SD is always positive.

NOTES

$$\text{and variance} = \sigma^2 = \frac{\sum fX^2}{\sum f} - \left(\frac{\sum fX}{\sum f} \right)^2 = (\text{SD})^2$$

Calculations of S.D. for Continuous Series

In continuous series X values are the middle points (Middle values) of the class intervals. Other calculations remain the same as described earlier for discrete series.

The Main Characteristics of Standard Deviation

The major traits of the standard deviation are:

1. It has a strict definition.
2. All observations are the basis for it.
3. Because the deviations are minimal for values that are near to the mean, the variance and standard deviation are likewise modest. When all values are equal, variance and standard deviation will both be zero.
4. The variance and standard deviations must remain unchanged if the same amount is added to or subtracted from each number, i.e., S.D. is independent of change in origin.
5. If all values are multiplied by the same number, say k, whether or not the units are the same, the standard deviation of the modified values equals k. (standard deviation of the original values). In other words, the S.D. depends on the scale change.
6. If multiple samples are taken from the same population, it can be seen that the standard deviation is less influenced from sample to sample than measurements of dispersions.
7. Coefficient of S.D. = $\frac{\text{S.D.}}{\text{Mean}}$

Properties (or Mathematical Properties) of Standard Deviation

1. Combined S.D. of two or more groups can be calculated (Refer Formula 5.16, 3.17, 3.20).

2. S.D. of first N natural numbers = $\frac{1}{12}\sqrt{N^2 - 1}$

$$\text{e.g., S.D. of first four natural numbers} = \frac{1}{12}\sqrt{(4)^2 - 1} = \frac{\sqrt{15}}{12} = \frac{\sqrt{5}}{4}$$

$$\text{S.D. of first two natural numbers} = \sqrt{\frac{1}{12}[(2)^2 - 1]} = \sqrt{\frac{3}{12}} = \frac{1}{2}$$

3. The sum of the squares of the deviations of all the values from the arithmetic mean is minimum.

$$\text{i.e., } S(X - \bar{X})^2 < S(X - A)^2 \text{ where } A \text{ may be any constant.}$$

4. In a normal distribution, the following relationships hold good.

Mean \pm S.D. covers 68.27% of the items.

Mean \pm 2 S.D. covers 95.45% of the items.

Mean \pm 3 S.D. covers 99.73% of the items.

NOTES

2.11 COEFFICIENT OF VARIATION

$$\text{Karl Pearson's Coefficient of Variation} = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100\%$$

The relative measure of dispersion known as the coefficient of variation has no units. It is typically stated as a percentage. The variability of two or more distributions is compared using this method. The distribution is stated to be less stable, uniform, consistent, homogeneous, and equitable but more variable if the coefficient of variations is higher.

The main advantage of the coefficient of variation is that the uniformity, consistency, equitability and homogeneity of entirely different distributions, measured in different units can be compared. The comparison is very helpful for actuaries.

Example 8. Given the following numbers: 1, 2, 3, 4, 5.

Calculate: (i) Coefficient of variation.

Solution.

$$(i) \text{ A.M.} = \frac{1}{5}(1 + 2 + 3 + 4 + 5) = \frac{1}{5} \times 15 = 3.$$

X	$X - 3 = d$	d^2
1	-2	4
2	-1	1
3	0	0
4	1	1
5	2	4
		$\Sigma d^2 = 10$

$$s = \sqrt{\frac{\Sigma d^2}{n}} = \sqrt{\frac{10}{5}} = \sqrt{2} = 1.4142$$

$$\begin{aligned} \text{C.V.} &= \frac{\text{S.D.}}{\text{Mean}} \times 100\% = \frac{\sigma}{\text{A.M.}} \times 100\% \\ &= \frac{1.4142}{3} \times 100\% = 47.14\% \end{aligned}$$

NOTES

Direct Method

X^2 : 1, 4, 9, 16, 25

$$\begin{aligned} \text{SD} &= \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} \\ &= \sqrt{\frac{1+4+9+16+25}{5} - 9} \\ &= \sqrt{\frac{55}{5} - 9} = \sqrt{2} = 1.4142 \end{aligned}$$

Example 9. If the S.D. of a set of observations is zero then all observations are equal. Comment.

Solution. Yes.

$$(\text{S.D.})^2 = \frac{\sum (X - \bar{X})^2}{N}.$$

$$\text{If S.D.} = 0 \Rightarrow (\text{S.D.})^2 = 0 \Rightarrow \frac{\sum (X - \bar{X})^2}{N} = 0 \Rightarrow \sum (X - \bar{X})^2 = 0,$$

which is possible only if each value is equal to \bar{X} .

Example 10. Given X : 9, 7, 5, 11, 1, 5, 7, 3.

(a) Calculate (i) standard deviation. Also calculate the measure with $-x$ and compare them with the corresponding values of X .

Solution.

$$(i) \sum X^2 = (1)^2 + (3)^2 + 2 \times (5)^2 + 2 \times (7)^2 + (9)^2 + (11)^2 = 360.$$

$$\text{S.D.} = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\frac{360}{8} - (6)^2} = \sqrt{45 - 36} = 3.$$

If X is negative,

$$\text{S.D.} = 3 \quad \left[\because \text{If } X \text{ is -ve } \sum X^2 \text{ and } \left(\frac{\sum X}{N}\right)^2 \text{ will remain unchanged.} \right]$$

\therefore S.D. remains unchanged with negative X .

LIST OF FORMULAE

$$1. \text{ Range} = L - S$$

where, L = value of the largest item and S = value of the smallest item.

$$2. \text{ Coefficient of range} = \frac{L - S}{L + S}$$

3. Interquartile range = $Q_3 - Q_1$, where, Q_3 and Q_1 are upper and lower quartiles, respectively.

4. Semi-interquartile range or Quartile deviation = $\frac{Q_3 - Q_1}{2}$

5. Coefficient of semi-interquartile range or Quartile deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

6. Mean deviation = $\frac{\sum |x|}{N}$

where $|x|$ stands for deviations from the mean ignoring plus and minus signs.

7. Mean deviation in a grouped frequency = $\frac{\sum f|d|}{N}$

where $|d|$ stands for deviations from the mean ignoring plus and minus signs.

8. Variance of $x = \frac{\sum (x - \bar{x})^2}{N}$

9. Standard deviation σ of ungrouped data

$$= \sqrt{\frac{\sum (x - \bar{x})^2}{N}} = \sqrt{\frac{\sum d^2}{N}}$$

where d = deviation from the mean

10. Standard deviation in a grouped series

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

Where $d = x - \mu$

11. Standard deviation using the arbitrary mean

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

where $d = x - A$ (arbitrary mean)

12. Standard deviation by the step-deviation method

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

where d' stands for deviations divided by C , the class interval. C is used to simplify the calculations.

NOTES

NOTES

13. Combined standard deviation of two series

$$\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

where, σ_1 = standard deviation of first group; σ_2 = standard deviation of second group; n_1 and n_2 being the number of observations in group 1 and group 2 respectively; $d_1 = (\bar{x}_1 - \bar{x}_{12})$ and $d_2 = (\bar{x}_2 - \bar{x}_{12})$.

14. Coefficient of variation

$$CV = \frac{\sigma}{\bar{x}}(100)\%$$

A relative measure of dispersion, which is free from unit of measurement. It enables us to compare two or more distributions having different units of measurement.

Check Your Progress

I. Multiple Choice Questions

- S.D. is independent of:
 - Change of origin
 - Change of Scale
 - Both
 - None
- Which of the following measures is most affected by extreme values
 - S.D.
 - Q.D.
 - M.D.
 - Range
- Which of the following is a unit free number:
 - S.D.
 - Variance
 - M.D.
 - C.V.
- If every item in a set of data is increased by 10, then which of the following measures will remain unchanged?
 - Mean
 - Median
 - Mode
 - Variance
 - None of the above
- The standard deviation of 30 items in a data set is $5/3$. If every item in the data set is multiplied by 3, then the variance of the group will be:
 - 5
 - 8
 - 15
 - 25
 - 41

NOTES

II. State whether the following Statements are True or False

1. Mean deviation is better than S.D.
2. The standard deviation is free from all those defects with which the other measures suffer.
3. The variance and the coefficient of variation are the same.
4. Root mean square deviation about arithmetic mean is the least.
5. Reducing each and every item by 5 will reduce standard deviation also by 5.
6. In an ordinary data, standard deviation is less than quartile deviation.
7. Relative measure of dispersion is a unit free pure number.
8. Standard deviation and mean deviation are relative measures.
9. Absolute measures of variation are used for comparing variability in distributions.
10. Standard deviation can be calculated from any average.
11. Measures of dispersion are the averages of second order.
12. The variance is equal to square of S.D.
13. Standard deviation and coefficient of variation are always expressed in the same unit.

III. Fill in the Blanks

1. To find the range, we consider only _____.
2. If in a series coefficient of variation is 64% and mean is 10, the standard deviation shall be _____.
3. Variance is always _____.
4. If each of the 10 values of a set are equal to 5, the standard deviation will be equal to _____.

2.12 ANSWERS TO 'CHECK YOUR PROGRESS'**I. Multiple Choice Questions**

1. Change of origin
2. Range
3. C.V.
4. None of the above
5. 25

II. State whether the following Statements are True or False

1. False
2. True
3. False

NOTES

4. True
5. False (S.D. Will remain unchanged)
6. False
7. True
8. False
9. False
10. False
11. True
12. True
13. False

III. Fill in the Blanks

1. Extreme values
2. 6.4
3. Positive
4. Zero

2.13 SUMMARY

We should understand S.D. and C.V., We should understand range mean deviation, Difference between M.D. and S.D., Meaning of Dispersion, Definition of Dispersion, Range, Q.D. Mean Deviation their properties, S.D. meaning and uses, Coefficient of variation.

2.14 KEY TERMS

- **Coefficient of variation:** A measure of relative variability, the coefficient of variation expresses the standard deviation as a percentage of the mean.
- **Dispersion:** A set of data's dispersion or variability.
- **Interquartile range:** The discrepancy between the first and third quartiles' values is known as the interquartile range.
- **Mean deviation:** A measure of dispersion that provides the average absolute differences between each item and the mean (i.e., disregarding plus and minus signs).
- **Measures of dispersion:** Measures that reveal the spread of a distribution are known as dispersion metrics.
- **Quartile deviation or semi-interquartile rang:** The difference between the upper and lower quartiles is divided by two to produce the quartile deviation, also known as semi-interquartile range, which is a measure of dispersion.

- **Range:** The difference between a data set's biggest and smallest values.
- **Standard deviation:** The square root of a series' variation is the standard deviation. It displays the distribution of the data.
- **Standard score:** The result of transforming an observation by dividing by the standard deviation after removing the mean. In order to express an observation above or below the mean, it is done so in units of standard deviation.
- **Standardized variable:** A variable that reflects an interest value's x in terms of how far above or below (or by how many standard deviations) the mean it is from. The term "standardised normal variable" also applies to it.
- **Statistic:** A statistic is a synthesis metric computed from sample data.
- **Variance:** The average squared deviation of each series' individual items from the mean.

NOTES

2.15 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. Why is it preferred to mean deviation?
2. What are the advantages of range?
3. In what situation is it used as a unit of measurement?
4. What is the importance of dispersion?
5. Explain the units for measuring standard deviation.
6. Write any two characteristics of standard deviation.

Long Answer Questions

1. Prove that for the positive values S.D. = what their difference
2. Calculate mean deviation (from A.M) for the following values 4800, 4600, 4400, 4200, 4000 by 240 coefficient of M.D = 0.054
3. Why the word 'standard' is used in standard deviation?
4. Why measures of dispersion are called average of second order?

2.16 REFERENCES

1. D.N. Elhance, Veena Elhance and B.M. Aggarwal, 2007, Fundamentals of Statistics. Kitab Mahal, New Delhi.
2. B.M. Aggarwal, 2012, Business Statistics (With Lab Work), Himalaya Publishing House, Mumbai.
3. B.M. Aggarwal, Dr. Puja A. Gulati, Neha Aggarwal, 2022, Statistics for Business and Economics. Kitab Mahal, New Delhi.

NOTES

Unit III Skewness and Kurtosis

Learning Objectives:

By the end of this unit the learners would be able to:

- To understand the symmetry of the Distribution
- To understand the peak of the Distribution
- Bowley's coefficient of Skewness
- Platykurtic, Mesokurtic and leptokurtic curves

Structure:

- 3.1 Introduction
- 3.2 Definitions
- 3.3 Difference between Dispersion and Skewness
- 3.4 Positively Skewed and Negatively Skewed Distributions
- 3.5 Comparison between Symmetrical and Skewed Distributions
- 3.6 Measures of Skewness (Coefficient of Skewness)
- 3.7 Kurtosis
- 3.8 Answers to 'Check Your Progress'
- 3.9 Summary
- 3.10 Key Terms
- 3.11 Self-Assessment Questions and Exercises
- 3.12 References

3.1 INTRODUCTION

Skewness refers to a distribution's asymmetry and indicates how the distribution is shaped. Even while two distributions may have the same mean and standard deviation, they may have very distinct forms, for example, one distribution may be symmetrical and the other asymmetrical.

3.2 DEFINITIONS

According to Morris Hamburg, “Skewness refers to the symmetry or lack of symmetry in the shape of a frequency distribution. This characteristic is of particular importance in connection with judging the typicality on certain measures of central tendency.”

According to Padan and Linguist, “A distribution is said to be skewed if it is lacking symmetry that its measures tend to pile up at one end or the other.”

According to Wessel, Willett and Simone, “Skewness is lack of symmetry. Any measure of skewness indicates the difference between the manner in which items are distributed in a particular distribution compared with a normal distribution.”

The mean, median, and mode all coincide and are located in the centre of a symmetrical distribution*. But in an asymmetric distribution, these three values are pulled apart.

NOTES

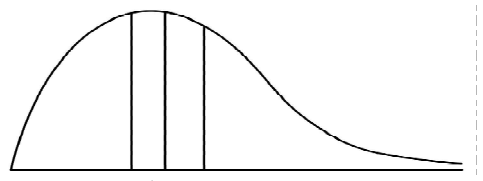
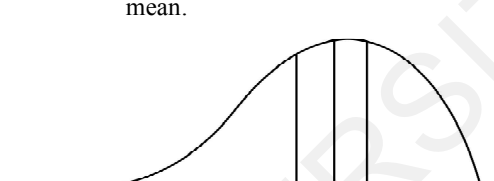
3.3 DIFFERENCE BETWEEN DISPERSION AND SKEWNESS

1. Skewness deals with the symmetry of the distribution around the centre value, whereas dispersion deals with the spread of values about the central value.
2. The terms "dispersion" and "skewness" refer to the amount of variation, the direction of the variation, and the degree to which the central tendency deviates from symmetry.
3. Dispersion enables us to determine how well a centre value represents the entire distribution. The nature of deviations on either side of a centre value, however, is the subject of skewness.
4. While skewness reveals whether the concentration is more in greater values or smaller values, dispersion is beneficial in explaining the degree of variability in the data.

3.4 POSITIVELY SKEWED AND NEGATIVELY SKEWED DISTRIBUTIONS

- (a) In a *positively skewed* distribution:
 - (i) more than half the area under the curve is on the right side of the mode,
 - (ii) $\text{mean} > \text{median} > \text{mode}$, and
- (b) In a *negatively skewed* distribution:
 - (i) more than half the area under the curve is on the left side of the mode,
 - (ii) $\text{mean} < \text{median} < \text{mode}$, and

* A distribution is said to be symmetrical if the values equidistant from the mean have the same frequency,

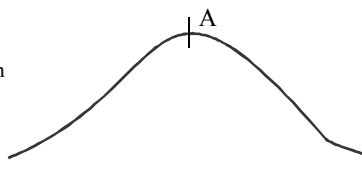
<p><i>Positively Skewed Distribution</i> <i>(Skewed to the right)</i></p> <p>Mean > Median > Mode Values above the mean occur less frequently than values below the mean.</p>  <p>Mode Median Mean Mean > Median > Mode (Mean is largest)</p>	<p><i>Negatively Skewed Distribution</i> <i>(Skewed to the left)</i></p> <p>Mean < Median < Mode Values above the mean occur more frequently than values below the mean.</p>  <p>Mean < Median < Mode (Mode is largest)</p>
---	---

3.5 COMPARISON BETWEEN SYMMETRICAL AND SKEWED DISTRIBUTIONS

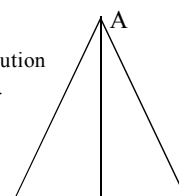
<i>Symmetrical Distribution</i>	<i>Skewed (Asymmetrical) Distribution</i>
1. Mean = Median = Mode	1. Mean, median and mode do not coincide, <i>i.e.</i> , Mean \neq Median \neq Mode.
2. The sum of the positive deviations from the median or mode is equal to the sum of the negative deviations.	2. The sum of the positive deviations from the median or mode is not equal to the sum of the negative deviations. The amount of difference between the sums of positive and negative.

* A distribution is said to be symmetrical (*i.e.*, zero skewness) about the mean if the values equidistant from the mean have equal frequencies.

Normal Distribution
(Symmetrical and bell-shaped)



Symmetrical Distribution
but not normal.



3.6 MEASURES OF SKEWNESS (COEFFICIENT OF SKEWNESS)

Karl Pearson's measure (coefficient) of skewness [based on mean, median, mode and standard deviation].

$$\text{Karl Pearson's first measure of skewness, } Sk_{(p)} = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}} = \frac{\bar{X} - Z}{\sigma}$$

Karl Pearson's second measure of skewness = $\frac{3(\text{Mean} - \text{Median})}{\text{S.D.}}$

Prof. Pearson claims that the results of his coefficient of skewness shall lie in the interval ± 3 but they are rarely achieved in actual practice.

SOLVED EXAMPLES

Example 1. Distinguish between a symmetrical and skewed distribution.

Solution.

In a symmetrical distribution, Mean = Median = Mode, i.e., if we cut the figure into two halves along the axis of symmetry, the two halves exactly overlap each other. In a skewed or asymmetric distribution.

$$\text{Mean} \neq \text{Median} \neq \text{Mode}$$

and we use the empirical formula

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median}).$$

[This formula is used mainly when the distribution is moderately asymmetrical]

Example 2. In a moderately skewed distribution if mean = 24.6 and median = 25.1, find mode

Solution.

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median})$$

$$24.6 - \text{Mode} = 3(24.6 - 25.1)$$

$$= -1.5$$

$$\Rightarrow \text{Mode} = 24.6 + 1.5$$

$$= 26.1$$

Example 3. Mean, Median and coefficient of variation of 100 items are found to be 90, 84 and 60 respectively. Find the coefficient of skewness.

Solution.

$$\frac{\text{SD}}{\text{Mean}} \times 100 = \text{C.V.}$$

$$\text{SD} = \frac{\text{Mean} \times \text{C.V.}}{100}$$

$$= \frac{90 \times 60}{100} = 54$$

$$\text{Coefficient of skewness} = \frac{3(\text{Mean} - \text{Median})}{\text{SD}}$$

$$= \frac{3(90 - 84)}{54} = \frac{1}{3} = 0.33$$

* This formula is based on the empirical relation between mean, median and mode, i.e., Mean - Mode = 3(Mean - Median). Z represents mode.

NOTES

NOTES

Example 4. Distribution of wage rates in the cotton textiles industry, when compared with wage rates in all manufacturing industries in 1999 shows the following results.

Industry Wage (in ₹)	Mean Monthly Wages (in ₹)	Median Monthly Wage (in ₹)	S.D.
Cotton textiles	57.00	55.00	1.2
All manufacturing	54.00	53.00	1.2

Give your comments on the structure of distribution.

Solution.

Pearson's coefficient of skewness for cotton industries

$$= \frac{3(\text{Mean} - \text{Median})}{\text{S.D.}}$$

$$= \frac{3(57 - 55)}{1.2} = 5$$

$$\text{Coefficient of skewness for all manufacturers} = \frac{3(54 - 53)}{1.2} = 2.5$$

Since the coefficient of skewness is more for cotton industries, the wage rates in cotton industries are more asymmetrical (variable) as compared to wage rates for all industries.

Example 5. Finding the Karl Pearsonian coefficient of skewness of the distribution from a moderately skewed distribution of retail pricing for men's shirts reveals that the mean price is ₹ 200 and the median price is ₹ 170. If the coefficient of variation is 20 percent.

Solution.

$$\bar{X} = ₹ 200, \text{Median} = ₹ 170 \text{ and C.V.} = 20\%.$$

$$\text{Coefficient of variation} = \frac{\sigma}{\bar{X}} = 0.20 \Rightarrow \sigma = 0.20 \times 200 \Rightarrow \sigma = 40$$

$$[\because \bar{X} = 200]$$

$$\text{Pearsonian coefficient} = \frac{3(\text{Mean} - \text{Median})}{\sigma} = \frac{3(200 - 170)}{40} = 2.25$$

Example 6. Given: Mean = 30, S.D. = 8, Karl Pearson's coefficient of skewness = 0.4. Find the median and mode.

Solution.

$$0.4 = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}} = \frac{30 - \text{Mode}}{8}$$

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

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

$$\Rightarrow \text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median})$$

$$\Rightarrow 30 - 26.8 = 3(30 - \text{Median})$$

$$\Rightarrow \frac{3.2}{3} = 30 - \text{Median} \Rightarrow \text{Median} = 28.93$$

$$\therefore \text{Mode} = 26.89, \text{Median} = 28.93.$$

NOTES

Example 7. Pearson's coefficient of skewness for a distribution is 0.4 and coefficient of variation is 30%. Its mode is 88. Find the mean and the median.

Solution.

$$\text{Pearson's coefficient of skewness} = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}}$$

$$\frac{1 - \frac{\text{Mode}}{\text{Mean}}}{\frac{\text{S.D.}}{\text{Mean}}} = \frac{1 - \frac{88}{\text{Mean}}}{0.3} = 0.4 \Rightarrow 1 - \frac{88}{\text{Mean}} = 0.4 \times 0.3 = 0.12$$

$$\Rightarrow \frac{88}{\text{Mean}} = 1 - 0.12 = 0.88$$

$$\Rightarrow 0.88 \text{ Mean} = 88 \Rightarrow \text{Mean} = \frac{88}{0.88} = 100$$

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median})$$

$$100 - 88 = 3(100 - \text{Median}) \Rightarrow \text{Median} = 96.$$

Example 8. The coefficient of skewness based on quartiles in the frequency distribution is 0.5. If the median is 11 and the total of the upper and lower quartiles is 28. Find the top and lower quartile values.

Solution.

Coefficient of skewness based on quartiles is Bowley's coefficient of skewness

$$= \frac{Q_1 + Q_2 = 2Q_3}{Q_3 - Q_1}$$

where, Q_1 = First Quartile, Q_2 = second Quartile = Median, Q_3 = Third Quartile

$$\text{Coefficient of skewness} = \frac{28 - 2 \times 11}{Q_3 - Q_1} = 0.5 \Rightarrow 6 - 2 = Q_3 - Q_1$$

[Substituting given values]

$$\Rightarrow Q_3 - Q_1 = 12 \quad \dots(1)$$

$$\text{Also, } Q_3 + Q_1 = 28 \quad \dots(2)$$

Adding (1) and (2), we get

$$2Q_1 = 40, Q_1 = 20, Q_3 - Q_1 = 12 \Rightarrow Q_3 = Q_1 + 12 = 20 + 12 = 32$$

$$Q_1 = 20, Q_3 = 32$$

NOTES

Example 9. The arithmetic mean and coefficient of variation for a distribution that is highly skewed are 100 and 35, respectively. The individual has a 0.2 coefficient of skewness. Identify the median and mode.

D.U., B.A. (Hons.) Business Economics, 2013

Solution.

$$\text{Karl Pearson's coefficient of skewness} = \frac{\text{mean} - \text{mode}}{\text{S.D.}}$$

$$\Rightarrow \frac{100 - \text{mode}}{35} = 0.2 \quad [\text{C.V.} = \frac{\text{S.D.}}{\text{mean}} \times 100 = 35 \Rightarrow \frac{\text{S.D.}}{100} \times 100 = 35 \Rightarrow \text{S.D.} = 35]$$

$$\Rightarrow \text{mode} = 100 - 7 = 93$$

$$\text{mean} - \text{mode} = 3(\text{mean} - \text{median}) \quad (\text{Empirical formula})$$

$$100 - 93 = 3(100 - \text{median}) \Rightarrow \frac{70}{3} = 100 - \text{median}$$

$$\Rightarrow \text{median} = 100 - \frac{70}{3} = \frac{230}{3} = 76.66$$

$$\Rightarrow \text{mean} > \text{median} > \text{mode} \Rightarrow \text{Positive Skewness}$$

Example 10. Distinguish between a symmetrical and skewed distribution.

Solution.

In a symmetrical distribution. Mean = Median = Mode and if we cut the figure into two halves along the axis of symmetry, the two halves exactly overlap each other.

In a skewed (asymmetric) distribution

$$\text{Mean} \neq \text{Median} \neq \text{Mode}$$

and we use the empirical formula

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median})$$

[This formula is used mainly when the distribution is **moderately** asymmetrical.]

Example 11. The sum of 20 observations is 300, its sum of squares is 5,000 and median is 15. Calculate the variations and skewness coefficients.

Solution.

$$n = 20, \Sigma X = 300, \Sigma X^2 = 5000, \text{Median} = 15$$

$$\text{S.D.} = \sqrt{\frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2} = \sqrt{\frac{5000}{20} - \left(\frac{300}{20}\right)^2} = \sqrt{250 - 225} = \sqrt{25} = 5$$

$$\text{Mean} = \frac{\Sigma X}{n} = \frac{300}{20} = 15$$

Since mean and median are equal, Karl Pearson's coefficient of skewness = 0

$$\text{C.V.} = \frac{\text{S.D.}}{\text{Mean}} \times 100 = \frac{5}{15} \times 100 = 33\frac{1}{3}\%$$

NOTES

Example 12. You are given that mean = 50, coefficient of variation = 40% and coefficient of skewness = -0.4. You are required to find out standard deviation, mode and median.

Solution.

Given $\bar{X} = 50$. Coefficient of variation (C.V.) = 40%, Coefficient of Skewness = -0.4, SD = ?, Mode = ?, Median = ?

$$(a) \quad CV = \frac{\text{S.D.}}{\bar{X}} \times 100 \Rightarrow 40 = \frac{\text{S.D.}}{50} \times 100 \Rightarrow \text{S.D.} = \frac{40 \times 50}{100} = 20$$

$$(b) \quad \text{Coefficient of Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{SD}} \Rightarrow -0.4 = \frac{50 - \text{Mode}}{20}$$

$$\Rightarrow -8 = 50 - \text{Mode} \Rightarrow \text{Mode} = 58$$

$$\text{Mode} = 3 \text{ Median} - 2 \bar{X}^* \Rightarrow 58 = 3 \text{ Median} - 2 \times 50 \Rightarrow 3 \text{ Median} = 158$$

$$\Rightarrow \text{Median} = \frac{158}{3} = 52.7 \text{ approx.}$$

Example 13. From the information below, calculate the Coefficient of Skewness. Mode = 11, Sum of two quartiles = 22, Difference of two quartiles = 8, Mean = 8

Solution.

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median}) \Rightarrow 8 - 11 = 3(8 - \text{Median}) \Rightarrow \text{Median} = 9$$

$$\text{Bowley's coefficient of skewness} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} = \frac{22 - 2 \times 9}{8} = \frac{1}{2} = 0.5$$

3.7 KURTOSIS

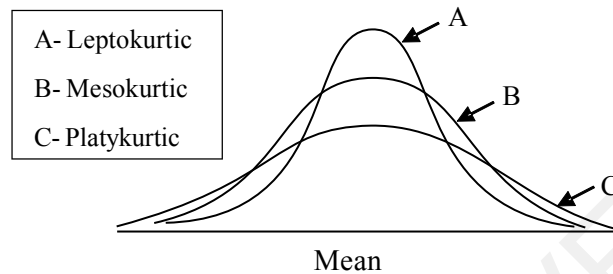
According to *Simpson and Kafka*, "Kurtosis refers to degree of peakedness or flatness in the region about the mode in frequency curve. The degree of Kurtosis of a distribution is measured relative to the **peakedness of a normal curve.**"

According to *Croxtan and Cowden*, "measure of kurtosis indicates the degree to which a curve of the frequency distribution is peaked or flat-topped."

* Mean - Mode = 3(mean - median) = 3 mean - 3 median \Rightarrow mode = 3 median - 2 mean.

NOTES

Different symmetrical curves, even having the same range may have different shapes of their peak. By measuring kurtosis, we may get a sense of how a frequency distribution's middle hump looks and functions. The Kurtosis is illustrated in the following figure.



The shape of its hump is recognised as a conventional one for type B curves, which are neither flat nor peaked and are referred to as normal curves. Mesokurtic curves are those that have humps that resemble a normal curve and are said to have a normal kurtosis. Kurtosis is reported to be absent from or negative for the type A leptokurtic curves, which are more pronounced than the normal curve. Contrarily, curves of type C, which are flatter than the normal curve, are known as platykurtic curves and are considered to have excess kurtosis or positive kurtosis.

As a measure of kurtosis, Karl Pearson gave the coefficient **Beta two** (β_2) and its derivatives **Gamma two** (γ_2) defined as follows:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4} \quad [\because \mu_2 = \sigma^2 = (\text{S.D.})^2]$$

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\sigma^4} - 3 = \frac{\mu_4 - 3\sigma^4}{\sigma^4}$$

For a normal or mesokurtic curve (Type B), $\beta_2 = 3$ or $\gamma_2 = 0$, for a leptokurtic curve (Type A), $\beta_2 > 3$ or $\gamma_2 > 0$ and for a platykurtic curve (Type C), $\beta_2 < 3$ or $\gamma_2 < 0$.

The following passage by British statistician W.S. Gosset, who wrote under the pen name Student, is notable because it humorously illustrates how the terms platykurtic and leptokurtic are used. “platykurtic curves are like platypus, squat with short tails; leptokurtic curves are high with long tails like the kangaroos noted for leaping.”

Summarising the results, we have:

- (i) Initial thought of the origin = Mean
- (ii) Variance is the second moment about the mean
- (iii) Third point regarding skewness measured by the mean
- (iv) Fourth point regarding the average measures kurtosis

LIST OF FORMULAE

1. Skewness = Mean – Mode
2. Coefficient of skewness (Karl Pearson's formula)

$$(i) \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

$$(ii) \frac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}}$$

The second formula is to be used when mode is ill-defined.

3. Skewness (Bowley's measure)

$$Sk_B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 + Q_1 - 2Q_2} \quad \text{where } Q_1 = \text{Lower quartile} \\ Q_3 = \text{Upper quartile}$$

4. Coefficient of skewness

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

5. Kelly's coefficient of skewness

$$SK_K = \frac{P_{90} + 2P_{50} + P_{10}}{P_{90} - P_{10}} \quad (\text{Based on percentiles})$$

Where P_{90} = Value of 90th percentile
 P_{10} = Value of 10th percentile
 P_{50} = Value of the median

6. Bowley's another measure of coefficient of skewness based on moments.

$$\text{Coefficient of skewness } \beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

NOTES

Check Your Progress

I. Multiple Choice Questions

1. For a distribution of data. If the arithmetic mean > median > mode, which of the following observations is true?
 - (a) The distribution is symmetrical.
 - (b) The distribution is positively skewed.
 - (c) The distribution is negatively skewed.
 - (d) Most of the values are concentrated at the higher end of the distribution.
2. In a negatively skewed distribution
 - (a) the mean is larger than the median.
 - (b) the mean is smaller than the median.

NOTES

- (c) the mean is the same as the median.
 (d) the mean is the same as the median and the mode.
3. The shape of a probability distribution is such that it tails off to right end when the distribution is:
 (a) symmetrical. (b) skewed left.
 (c) positively skewed. (d) none of these.
4. Which of the following observations is true with regard to a positively skewed distribution?
 (a) Mode > Median > Mean (b) Median > Mean > Mode
 (c) Mean > Median > Mode (d) Mode > Mean > Median
5. Which of the following indicates that the distribution curve of the data is negatively skewed?
 (a) Arithmetic Mean < Median < Mode.
 (b) Median < Mode < Arithmetic Mean.
 (c) Mode < Arithmetic Mean < Median.
 (d) Mode < Median < Arithmetic Mean.
6. When the coefficient of skewness = 0, the distribution is:
 (a) J-shaped (b) U-shaped
 (c) L-shaped (d) Bell-shaped.

II. Fill in the Blanks

1. In positively skewed distribution the relation mean, median and mode is _____.
2. In a negatively skewed distribution the relation between mean, median and mode is _____.
3. In a symmetric distribution the relation between mean, median and mode is _____.
4. In a moderately asymmetrical distribution the relation between mean, median and mode is _____.
5. The limits of Bowley's coefficient of skewness are _____.
6. If mean and mode are equal then coefficient of skewness is _____.

3.8 ANSWERS TO 'CHECK YOUR PROGRESS'**I. Multiple Choice Questions**

1. (b)
 2. (b)
 3. (c)
 4. (c)
 5. (a)
 6. (d)

II. Fill in the Blanks

1. Mean > Median > Mode,
2. Mean < Median < Mode,
3. Mean = Median = Mode,
4. Mean – Median = 3(Mean – Mode),
5. ± 1 ,
6. Zero.

NOTES

3.9 SUMMARY

- Dispersion, Skewness, Difference between dispersion and skewness, coefficient of skewness (Karl Pearson's as well as Bowley's coefficient).
- **Kurtosis:** Platykurtic, Leptokurtic, Mesokurtic

3.10 KEY TERMS

- A skewness index calculated from quartile values. It ranges from 1.
- The standard deviation of a particular data set divided by the difference between the mena and the mode.
- A skewness metric based on percentiles.
- The frequency polygon's level of "peakedness" or "flatness."
- A distribution where the majority of the observations are concentrated in the tails and close to the mode.
- A less-peaked distribution than a leptokurtic curve.
- A notion that represents various facets of a given distribution. Moments allow us to quantify a series' central tendency, dispersion or unpredictability, skewness, and pickeness of the curve.
- The longer tail of the distribution extends to the left when more observations are located to the right of the mean.
- A "flat" distribution that resembles a table or plateau.
- The longer tail of the distribution extends to the right when more observations are to the left of the mean.

3.11 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. "In a negatively skewed distribution, mean, median and mode as calculated are respectively mean = 25, median = 28, and mode = 22."
Do you agree? Write your comments.

NOTES

2. Show by calculation whether a group is negatively skewed or positively skewed if median = 32 and mode = 30.

Long Answer Questions

1. What you have learned about summary by distribution.
2. Difference between Karl Pearson's coefficient of skewness and Bowley's coefficient of skewness.
3. Difference between dispersion and Skewness.

Practice Questions

1. Karl Pearson's coefficient of skewness is 0.32. Its S.D. is 6.5 and mean 29.6. Find the value of mode and median. [Ans. Mode = 27.52, Median = 28.91]
2. Show by calculation whether the group is negatively skewed or positively skewed if median = 24 and mean = 26. [Ans. Positively skewed]

[Hint: Karl Pearson's coefficient of skewness = $\frac{3(\text{Mean} - \text{Median})}{\text{S.D.}}$]

Since the S.D. is always positive, the sign of (Mean – Median) will determine the nature of skewness. Since (Mean – Median) = 26 – 24 = 2 is positive, skewness is positive.]

3. The sum of a set of 100 numbers is 4,000, the sum of their squares is 1,62,500 and median is 41. Find the coefficient of skewness. [Ans. –0.6]
4. You are given the information: Coefficient of skewness = 0.8, Mean = 40, and Mode = 36. Find the value of S.D. [Ans. 5]
5. The following information was obtained from the records of a factory relating to wages. Arithmetic mean = ₹ 56.8, Median = ₹ 59.5, Standard deviation = ₹ 12.4. Give as much information as you can about the distribution of wages. [Ans. Distribution is negatively skewed, Mode = 64.9]
6. For a moderately skewed data, the arithmetic mean is 100, the coefficient of variation is 35 and Karl Pearson's coefficient of skewness is 0.2. Find the mode and the median. [Ans. Median = 97.67, Mode = 93]
7. For a moderately skewed distribution, mean = 150, mode = 140 and S.D. = 45. Find (i) Pearson's coefficient of skewness, and (ii) median.

[Ans. (i) 0.22 (ii) 146.67]

3.12 REFERENCES

1. D.N. Elhance, Veena Elhance and B.M. Aggarwal, 2007, Fundamentals of Statistics. Kitab Mahal, New Delhi.
2. B.M. Aggarwal, 2012, Business Statistics (With Lab Work), Himalaya Publishing House, Mumbai.
3. B.M. Aggarwal, Dr. Puja A. Gulati, Neha Aggarwal, 2022, Statistics for Business and Economics. Kitab Mahal, New Delhi.

Unit IV Correlation and Regression

Learning Objectives:

By the end of this unit the learners would be able to:

- Understand the importance and limitations of correlation analysis.
- Distinguish between: (a) linear and non-linear correlation, (b) positive and negative correlation and (c) simple, partial and multiple correlation.
- Recognise when a scatter diagram suggests a relationship between two variables.
- Calculate and interpret coefficient of correlation for individual observations as well as for bivariate grouped data.

Structure:

- 4.1 Introduction
- 4.2 Definition of Correlation
- 4.3 Importance (or Utility) of Correlation
- 4.4 Kinds of Correlation
- 4.5 Positive and Negative Correlation
- 4.6 Linear and Non-linear Correlation
- 4.7 Correlation Based on Number of Variables
- 4.8 Some Important Points about the Study of Correlation Analysis
- 4.9 Measures of Correlation
- 4.10 Properties of Correlation Coefficient
- 4.11 Probable Error
- 4.12 Merits and Demerits of Rank Correlation Method
- 4.13 Regression Analysis
- 4.14 Regression Lines or Regression Equations
- 4.15 Properties of Regression Coefficients
- 4.16 Why There are Two Regression Lines?
- 4.17 Examples of Irreversible Relation
- 4.18 Coefficient of Determination
- 4.19 Answers to 'Check Your Progress'
- 4.20 Summary
- 4.21 Key Terms
- 4.22 Self-Assessment Questions and Exercises
- 4.23 References

NOTES

4.1 INTRODUCTION

So far we have been dealing with the analysis of one variable only. For example, calculations of mean, median, mode, standard deviation, etc., involved only one variable. The degree (strength) of the statistical association between two or more variables, however, is the subject of the study of correlation, which examines the correspondence of movement (going jointly) between two variables or a series of paired items. For instance:

1. Demand declines if prices rise.
2. The supply grows if the price rises.
3. A family's expenses will rise as their income does.
4. Sale of woollen garments increases as winter approaches.
5. When the number of students in a college is large, the sale in the canteen is also large, but during vacations when students do not come to college, the sale in the canteen declines.
6. Smoking more may lead to an increase in lung cancer cases.

The two variables move together in the six cases above, either moving in the same direction or the opposite direction. However, there are instances where the two variables behave independently of one another and have no propensity to "go together." We do not measure variation in one series but rather compare variation in two or more series when dealing with correlation. Instead of dealing with one series, we deal with the association or relationship between two series. The two series could vary simultaneously in the same direction, simultaneously in the opposite direction, or not at all.

We need a significant number of items in the series in order to measure the association of the series through correlation. We cannot make generalisations about how the variables fluctuate together if we just know two or three pairs of values. * Furthermore, there must be matching throughout; there cannot be a blank in one series where there is a value in the other series. This type of analysis in which change in the value of one variable causes change in the value of the other variable is called **Bivariate Analysis**.

4.2 DEFINITION OF CORRELATION

There are numerous definitions of correlation. The following are some of the key definitions:

1. Correlation assesses how closely two variables are related, specifically how closely they are related linearly.
2. According to Croxton and Cowden, "*When the relationship is of quantitative nature the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation.*"

* If we have only a few pairs of values, then the results of correlation analysis may not be valid.

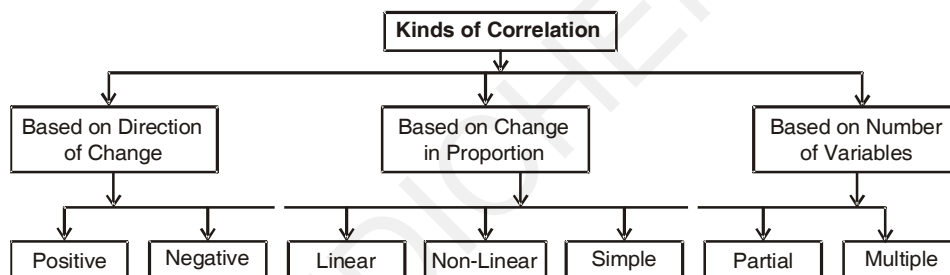
3. In the words of Boddington, “*whenever some definite connection exists between the two or more groups, classes or series of data, there is said to be correlation.*”
4. According to L.R. Connor, “two or more values are said to be correlated if they vary in sympathy, so that changes in one tend to be accompanied by comparable changes in the other(s).”

NOTES

4.3 IMPORTANCE (OR UTILITY) OF CORRELATION

1. Using just one figure, the correlation coefficient aids in determining the strength of the association between two variables.
2. The existence of a relationship between two or more variables makes it possible for us to forecast future events. For instance, if wheat output has increased while all other parameters remain the same, we may anticipate a decrease in the price of wheat.
3. If two variables are highly correlated, we can estimate one variable's value using the value of the other variable. Regression equations are employed in order to do this.
4. In corporate organisations, correlation makes decision-making easier. Correlation analysis is also used to predict how specific variables would behave.

4.4 KINDS OF CORRELATION

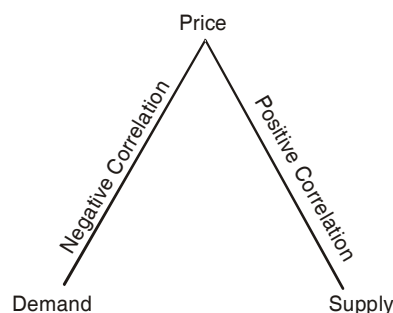


4.5 POSITIVE AND NEGATIVE CORRELATION

The correlation between two variables is said to be positive if they move in the same direction. The correlation between two variables is considered to be negative if they move in opposition to one another. There is no correlation between them if they don't move together at all and the variables are said to be independent.

Example:

- (i) Price and demand have a negative correlation since they move in the opposing directions.



NOTES

- (ii) Price and supply are positively correlated because they move in the same direction.

4.6 LINEAR AND NON-LINEAR CORRELATION

It is not necessary for there to be a proportionate change in order to measure the degree of correlation between the movements within two or more series. For instance, the law of demand states that when all other parameters are held constant, an increase in a commodity's price is followed by a drop in its demand, but we are unable to identify any proportionality between price and demand.

The correlation is said to be linear if the values of the two variables change proportionately, *i.e.*, if x and y are two variables, then in linear correlation is $\frac{\Delta y}{\Delta x}$ always constant, where Δx is a change in the value of x corresponding to a change Δy in the value of y . In linear correlation, the graph drawn with the pair of values of x and y is always a straight line. The correlation is referred to as non-linear or curvilinear if the change in the variables is not proportional. In this book, we will deal only with linear correlation.

Illustrations

(i)	x	1	2	3	4	5
	y	2	4	6	8	10

is an example of linear correlation.

(ii)	x	1	2	3	4
	y	3	5	8	15

is an example of non-linear correlation.

4.7 CORRELATION BASED ON NUMBER OF VARIABLES

- (i) **Simple Correlation:** This type of correlation involves just two variables, and it examines how they relate to one another.
- (ii) **Partial Correlation:** When there are more than two variables present, the link between them is only explored while holding the other variables **constant**. For example, production of wheat depends on many factors like rainfall, quality of seed, manure, etc., and if we study the relation between production of wheat per hectare and quality of seed keeping rainfall and manure constant, then the correlation is called partial correlation.
- (iii) **Multiple Correlation:** It is the simultaneous study of the link between two or more variables, *e.g.*, in the above example given in partial correlation if we study the relationship between production and all other factors simultaneously, the relationship will be called **multiple correlation**.

Note: Mathematical treatment of Partial and Multiple Correlation is beyond the scope of this book.

There are so many other cases in which the two variables move together. This correspondence of movement (*i.e.*, going togetherness) of the series of paired items is called *concomitant variation* or *covariation*. The study of covariation leads to correlation analysis.

4.8 SOME IMPORTANT POINTS ABOUT THE STUDY OF CORRELATION ANALYSIS

1. The series should contain a sufficient number of items. We are unable to generalise a value pair's "going togetherness" based on just two or three pairs of values.
2. In correlation analysis, we focus on the association or relationship between two or more series rather than just one series.
3. Rather than measuring variation in just one series, we compare variation across two or more series to determine whether the series fluctuate simultaneously in the same direction, simultaneously in the opposite direction, or not at all.
4. Proportionate change, or a constant change in the other variable for every unit change in one variable, is not required to assess correlation, which is the degree of association between the movement of two or more series. It is enough for the sequence of paired items to, on average, show concomitance of movement.
5. We study only **Linear Correlation**.
6. Cause and effect relationships are not always present in correlation. Any one or more of the following can cause a correlation between two variables:
 - (i) **Cause and Effect:** Two variables have a cause and effect relationship. For instance, since heat impacts temperature, the two may be connected. Similarly, increase in the cost of advertisement affects sales, increase in radius affects the measurement of circumference of a circle. However, social sciences rarely uncover these kinds of connections. Numerous other factors have an impact on a variable at the same time. For instance, a change in demand, inflation, export policies, and a plethora of other reasons may result in a price increase. Finding the cause of the price increase is essentially impossible. Thus, correlation rarely establishes a cause-and-effect link in the social sciences.
 - (ii) **Effect of Third Variable:** Sometimes a third variable or multiple variables interact to affect both of the associated variables. For instance, there may be a strong link between wheat yield and rice yield. In truth, it's possible that both of these commodities' yields have been impacted by rainfall and other factors like fertilisers, quality of seeds, irrigation facilities, etc. There might not be a connection between wheat and rice yields. In such cases, correlation can give misleading conclusions.
 - (iii) **One Cause affecting Two Variables:** When we find that there is increase or decrease together in two variables, it is related to common cause. For example, when the rainfall is scanty (very little) the yield of rice as well as wheat is low and viceversa. Similarly, when petrol prices are increased, the airfares and taxi fares are also increased.
 - (iv) **Chance:** It is entirely possible for there to be a high degree of association via coincidence and in such cases, these variables are not in any way related to each other and therefore, in fact, do not have any relationship.

NOTES

NOTES

For example, the population of country is increasing and production of sugar is also increasing — it is purely a matter of chance. Similarly, number of divorces per year and the export of television sets, production of shoes with the agriculture production, relationship between cars produced and children born in a country. This kind of correlation is called '*chance or spurious or nonsense*' correlation. It is important to realise that true correlation only exists between related variables.

The aforementioned arguments demonstrate that correlation is just a statistical association and does not necessarily imply a causal connection between the variables.

4.9 MEASURES OF CORRELATION

In case of ungrouped data, the following methods of measuring correlation are normally used:

1. Scatter diagram method.
2. Karl Pearson's coefficient for measuring linear correlation.
3. Method of Rank differences (Spearman's Rank Correlation Coefficient).

Scatter Diagram

The graphical representation of pairs of numerical values for the two variables is called a scatter diagram or dot diagram. On the graph, a dot represents each pair of values. The kind and strength of the correlation between two variables can be determined from the scatter of points and the scatter diagram's direction. The correlation is said to be perfect positive if every point is on a straight line with a positive slope (i.e., rising line).

The correlation coefficient in this instance is $r = +1$. The correlation is considered to be perfect negative if every point is on a straight line with a downward slope (falling line). In this case the coefficient of correlation $r = -1$.*

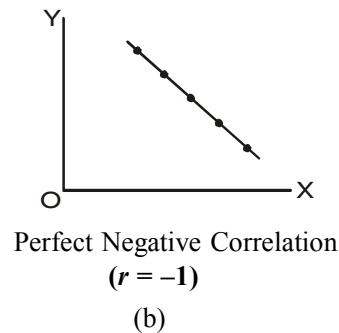
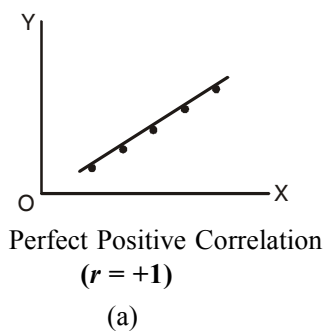
In a nutshell, the correlation is considered to be perfect if all of the points (i.e., dots) form a line.

In general, a relationship is direct and is referred to as positive if low values of one variable correlate with low values of the other variable and high values correlate with high values, and if these points generally trace a line from the lower left corner to the upper right corner. On the other hand, a connection is said to be negative if high values of one variable are paired with low values of the other, forming a path that roughly extends from the upper left corner to the lower right corner.

For instance, low sales typically correspond to a small labour force, whereas high sales correspond to a large labour force. Here the correlation is positive. In certain types jobs high productivity efficiency is generally achieved with low age and low productivity efficiency with high age. Negative relationships are those that have an inverse relationship.

* The correlation coefficient varies between $+1$ and -1 .

The condition is known as no correlation and the correlation coefficient $r = 0$ if the independent variable has no effect on the dependent variable, that is, if one variable does not help to determine the value of the other variable (or if the dots do not follow any regular pattern).



Limited Degree of Correlation: Although a correlation between two variables may exist in social science, an increase in one variable need not always be followed by a corresponding or equal increase (or reduction) in the other. When there are unequal changes in the two variables in the same direction, correlation is said to be restricted positive; when there are unequal changes in the opposite direction, correlation is said to be limited negative. The restricted degree of correlation can range from high (0.75–1) to moderate (0.25–0.75) to low (0.25–0.75), depending on the situation.

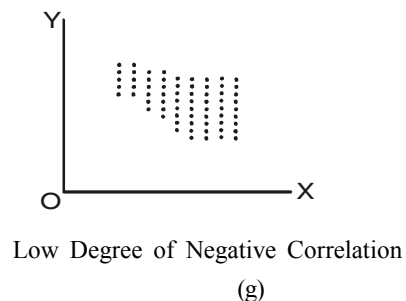
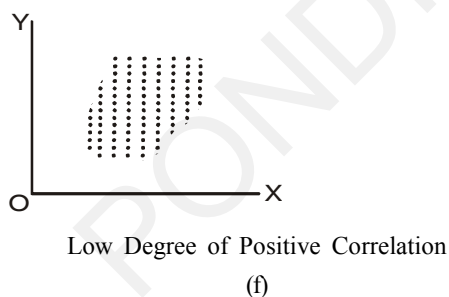
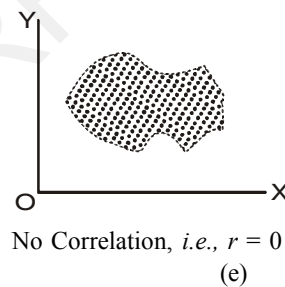
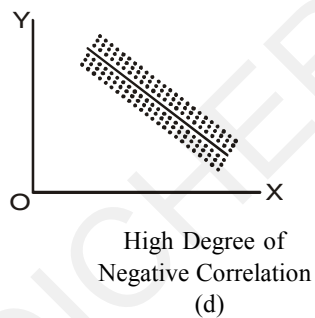
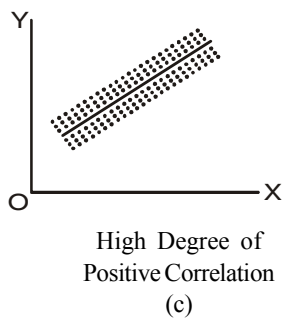


Fig. 4.1

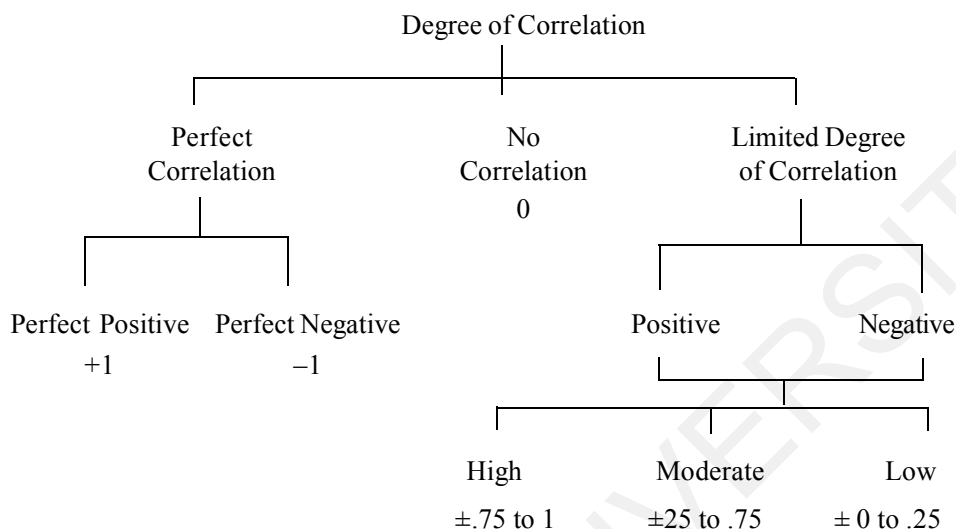
There need not be an origin on the Y-axis when constructing scatter diagrams. In order to handle the lowest values in the series, both axes can start at 0.

The value of the correlation coefficient, or r , is always between -1 and $+1$.

Karl Pearson's formula is used to calculate the approximate degree of correlation shown in the diagram below.

NOTES

NOTES



Merits and Limitations of the Scatter Diagram Method

Merits

1. It is a quick and simple approach to determine the type of correlation between two variables without using mathematics.
2. The method can be used to estimate the missing value of the dependent variable for a given value of the independent variable by drawing a line of best fit by hand between the shown dots.
3. The scatter diagram's shape indicates whether the correlation is linear or nonlinear, allowing us to determine the type of relationship that exists between the two variables. The scatter plot reveals if there is a positive or negative association.
4. The approach is unaffected by the values of the extreme observations.

Demerits

It simply provides a broad notion of the relationship between the two variables. The approach provides a general notion of the correlation's direction as well as its strength. However, this approach does not provide a numerical assessment of the strength or amount of association.

Note: If $r = -1$ or $+1$ the correlation is said to be perfectly negative or perfectly positive. The degree of the linear relationship between X and Y is shown by an intermediate value of r between -1 and $+1$, while the direction of the relationship is indicated by its sign. No linear relationship exists between the two variables if $r = 0$. The correlation coefficient is a simple integer that is unaffected by the measurement units. Its value is unaffected by changes to origin and scale, i.e., it doesn't change if each value of X or Y (or both) is added, subtracted, multiplied, divided, or multiplied and divided by a constant that is the same or different for X and Y. But if one variable is multiplied by a positive and other by a negative constant, the sign of the correlation coefficient will change.

Calculation of Correlation Coefficient (Karl Pearson's Method)*

Karl Pearson's coefficient of correlation for ungrouped data can be calculated using any one of the three approaches listed below:

- A. Actual Mean Method
- B. Direct Method
- C. Short Cut Method

A. Actual Mean Method: In this method, we make use of the following formula for computing correlation, *i.e.*,

$$\begin{aligned}
 r &= \frac{\text{Covariance}(X, Y)}{(\text{S.D. of } X)(\text{S.D. of } Y)} \\
 &= \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N\sigma_x \sigma_y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\left[\sqrt{\sum (X - \bar{X})^2} \sqrt{\sum (Y - \bar{Y})^2} \right]} \\
 &= \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}} \text{ where } x = X - \bar{X}, y = Y - \bar{Y}
 \end{aligned}$$

where symbols have their usual meanings. *This approach is appropriate when X and Y's means are both whole numbers and not fractions. (Please note).*

Steps:

1. Determine the mathematical means of the X and Y series.
2. Determine the X series' deviations from the X mean and represent these deviations by the letter x.
3. Square these deviations and obtain the total, *i.e.*, $\sum x^2$.
4. Discover the Y series' deviations from the Y mean and identify these deviations by y.
5. Square these deviations and obtain the total, *i.e.*, $\sum y^2$.
6. To determine the total, multiply these X and Y series determined deviations. *i.e.*, $\sum xy$.

B. Direct Method: The following shortened version of the formula may be used to determine the value of r when the mean values of the two series in a bivariate data set are fractional values and the number of observations and their volume in the two series is not particularly big.

$$\begin{aligned}
 r &= \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N\sigma_x \sigma_y} \\
 &= \frac{\frac{\sum XY}{N} - \frac{\sum X}{N} \frac{\sum Y}{N}}{\sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} \sqrt{\frac{\sum Y^2}{N} - \left(\frac{\sum Y}{N}\right)^2}}
 \end{aligned}$$

* Karl Pearson's coefficient of correlation is also called product-moment coefficient of correlation.

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

NOTES

For this formula, we need

N = the number of paired observations.

ΣX = the sum of observations on the X variable.

ΣY = the sum of observations on the Y variable.

ΣX^2 = the sum of squares of observations on X variable.

ΣY^2 = the sum of squares of observations on Y variable.

ΣXY = the sum of products of observations on X and Y variables.

Assumptions of the Karl Pearson's Coefficient of Correlation

The following presumptions form the foundation of Karl Pearson's coefficient of correlation:

- (i) **Linear Relationship:** In this method, it is assumed that the two variables have a linear relationship. In this scenario, a straight line forms a cluster of the paired observations on the two variables represented on a scatter diagram.
- (ii) **Causal Relationship:** When investigating correlation, we anticipate that the forces influencing the values in the two series will have a cause and effect relationship.

Merits and Demerits of Karl Pearson's Coefficient of Correlation***Merits***

1. It is a crucial and well-liked technique for determining how two variables relate to one another. It provides an exact and numerical value representing the strength of the link between the two variables. It is simple to interpret r 's value.
2. It assesses the strength of the correlation (direction of the relationship) between the two variables.

Demerits

1. Extreme elements have an impact on the coefficient's value.
2. In comparison to other methods, its computational process is challenging and time-consuming.
3. It presumes that the two variables have a linear connection.
4. There is a correlation between ± 1 . This yardstick needs to be used very carefully; otherwise, it could be understood incorrectly.

4.10 PROPERTIES OF CORRELATION COEFFICIENT

Three crucial characteristics of the coefficient of correlation are as follows:

Property I: The range of the correlation coefficient, r , is -1 to $+1$, i.e., $-1 \leq r \leq 1$.

Property II: The correlation coefficient is unaffected by changes to the scale and origin.

Property III: The correlation coefficient is a pure number without any units because it is the ratio of two quantities with the same units.

Remark 1. Adding or subtracting a fixed amount from the observations of the variables X and Y constitutes a change in origin, while doubling or dividing the values of X and Y constitutes a change in scale.

For example,

The correlation between X -series, and Y -series will be same as between $\frac{X-3}{2}$, and $\frac{Y-5}{3}$ series or $X-3$ and $Y-5$ series or $2X$ and $3Y$ series. But the correlation between $\frac{X-3}{2}$ and $\frac{5-Y}{3}$ will be opposite in sign.

Remark 2. Property II is very useful for reducing computational work involved in the calculation of coefficient of correlation and forms the basis for short cut method.

SOLVED EXAMPLES

Example 1. Determine the correlation coefficient between the father's height and the son's height using the given information:

Height of Father (in inches)	64	65	66	67	68	69	70
Height of Son (in inches)	66	67	65	68	70	68	72

Solution.

In this example, X and Y are not fractional values, so Actual Mean Method will be most suitable.

Computation of r (Actual Mean Method)

Height of Father (X)	Height of Son (Y)	$(X - \bar{X}) =$ $(X - 67) = x$	$(Y - \bar{Y}) =$ $(Y - 68) = y$	$(X - \bar{X})^2$ $= x^2$	$(Y - \bar{Y})^2$ $= y^2$	$(X - \bar{X}) \times (Y - \bar{Y})$ $= xy$
64	66	-3	-2	9	4	6
65	67	-2	-1	4	1	2
66	65	-1	-3	1	9	3
67	68	0	0	0	0	0
68	70	1	2	1	4	2

NOTES

NOTES

69	68	2	0	4	0	0
70	72	3	4	9	16	12
$\Sigma X = 469$	$\Sigma Y = 476$			$\Sigma x^2 = 28$	$\Sigma y^2 = 34$	Σxy

$$\bar{X} = \frac{\Sigma X}{N} = \frac{469}{7} = 67; \bar{Y} = \frac{\Sigma Y}{N} = \frac{476}{7} = 68$$

Since the actual means of X and Y are whole numbers, we can use actual mean method for computing r , i.e.,

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{[\Sigma(X - \bar{X})^2 \cdot \Sigma(Y - \bar{Y})^2]}}$$

$$= \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{25}{\sqrt{928 \times 34}} = \frac{25}{\sqrt{[952]}} = \frac{25}{30.85} = 0.81$$

Example 2. Calculate the correlation coefficient from the following data.

X	12	9	8	10	11	13	7
Y	14	8	6	9	11	12	3

Let's multiply each value of X by 2 and add 6 to that result. Similar to this, 2 is deducted from each value of Y before being multiplied by 3. What is the expected association between the new X and Y series.

Solution.

X	Y	$x - 10 = x$	$Y - 9 = y$	x^2	y^2	xy
12	14	2	5	4	25	10
9	8	-1	-1	1	1	1
8	6	-2	-3	4	9	6
10	9	0	0	0	0	0
11	11	1	2	1	4	2
13	12	3	3	9	9	9
7	3	-3	-6	$\frac{9}{28}$	$\frac{36}{84}$	$\frac{18}{46}$

$$\Sigma X = 70, \bar{X} = \frac{70}{7} = 10, \Sigma Y = 63, \bar{Y} = \frac{63}{7} = 9.$$

$$r = \frac{46}{\sqrt{28 \times 84}}$$

$2X + 6 = X'$	$3Y - 2 = Y'$	$X' - 26 = d_x$	$Y' - 25 = d_y$	d_x^2	d_y^2	$d_x d_y$
30	40	4	15	16	225	60
24	22	-2	-3	4	9	6
22	16	-4	-9	16	81	36

NOTES

26	25	0	0	0	0	0
28	31	2	6	4	36	12
32	34	6	9	36	81	54
20	7	-6	-18	36	324	108
182	175			118	756	276

$r = \frac{276}{\sqrt{112} \times \sqrt{756}} = \frac{46}{\sqrt{28 \times 84}}$ which is same as r calculated earlier which indicates that coefficient of correlation is independent of change of origin and scale.

$$\bar{X}' = \frac{182}{7} = 26$$

$$\bar{Y}' = \frac{175}{7} = 25$$

Example 3. Find Karl Pearson's Coefficient between X and Y .

X	20	30	40	50	60	70	80
Y	14	25	30	32	40	45	65

Solution.

X	Y	X^2	Y^2	XY
20	14	400	196	280
30	25	900	625	750
40	30	1600	900	1200
50	32	2500	1024	1600
60	40	3600	1600	2400
70	45	4900	2025	3150
80	65	6400	4225	5200
$\Sigma X = 350$	$\Sigma Y = 251$	$\Sigma X^2 = 20300$	$\Sigma Y^2 = 10595$	$\Sigma XY = 14580$

$$\begin{aligned}
 r &= \frac{N \Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N \Sigma X^2 - (\Sigma X)^2} \sqrt{N \Sigma Y^2 - (\Sigma Y)^2}} \\
 &= \frac{7 \times 14580 - 350 \times 251}{\sqrt{7 \times 20300 - (350)^2} \sqrt{7 \times 10595 - (251)^2}} \\
 &= \frac{102060 - 87850}{\sqrt{142100 - 122500} \sqrt{74165 - 63001}} \\
 &= \frac{14210}{140 \times 105.65} = 0.960
 \end{aligned}$$

NOTES

Example 4. Calculate Karl Pearson's coefficient of correlation from the following data.

$$N = 10, \Sigma X^2 = 290, \Sigma X = 50, \Sigma Y = -30, \Sigma Y^2 = 300, \Sigma XY = -115$$

Solution.

Karl Pearson's coefficient of correlation

$$\begin{aligned} &= \frac{N \Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N \Sigma X^2 - (\Sigma X)^2} \sqrt{N \Sigma Y^2 - (\Sigma Y)^2}} \\ &= \frac{10(-115) - 50 \times (-30)}{\sqrt{10 \times 290 - (50)^2} \sqrt{10 \times 300 - (-30)^2}} \\ &= \frac{-1150 + 1500}{\sqrt{400} \sqrt{2100}} \\ &= \frac{350}{20 \times 45.825} = 0.38 \end{aligned}$$

Example 5. Calculate the coefficient of correlation between X and Y .

$X :$	1	2	3	4	5
$Y :$	3	4	6	7	12

Solution.

X	Y	XY	X^2	Y^2	
1	3	3	1	9	$\Sigma X, Y = 117$
2	4	8	4	16	$\Sigma X^2 = 55$
3	6	18	9	36	$\Sigma Y^2 = 254$
4	7	28	16	49	$\Sigma X = 15$
5	12	60	25	144	$\Sigma Y = 32$
$\Sigma X = 15$	$\Sigma Y = 32$	$\Sigma XY = 117$	$\Sigma X^2 = 55$	$\Sigma Y^2 = 254$	

$$\begin{aligned} r &= \frac{N \Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N \Sigma X^2 - (\Sigma X)^2} \sqrt{N \Sigma Y^2 - (\Sigma Y)^2}} \\ &= \frac{5 \times 117 - 15 \times 32}{\sqrt{5 \times 55 - (15)^2} \sqrt{5 \times 254 - (32)^2}} \\ &= \frac{585 - 480}{\sqrt{275 - 225} \sqrt{1270 - 1024}} \\ &= \frac{105}{7.07 \times 15.6844} \\ &= 0.9468 \end{aligned}$$

4.11 PROBABLE ERROR*

To ascertain the significance or reliability of the value of Pearsonian coefficient of correlation, probable error is used. A probable error is a number that, when added to or subtracted from the correlation coefficient, creates a range in which the correlation coefficients of other groups randomly chosen from the same series will fall with a probability of 50%.

The likely error of a correlation coefficient, according to Horace Secrist, is a quantity that, when added to or removed from the average correlation coefficient, yields values within which there is even a chance that a correlation coefficient from a series chosen at random will fall.

Wheldon states that probable error specifies the upper and lower bounds above and below the magnitude of the coefficient computed within which there is equal probability that coefficient of correlation similarly derived from additional samples will fall.

The following formula is used to determine the probability error:

$$\text{Probable Error of 'r'} = 0.6745 \frac{1-r^2}{\sqrt{N}}$$

where, 0.6745 is a constant number,

r is Pearsonian coefficient of correlation

N is number of items.

If the coefficient of correlation of is 0.802 or 0.8. Its probable error will be

$$\begin{aligned} & 0.6745 \frac{1-r^2}{\sqrt{N}} \quad \text{or} \quad 0.6745 \frac{1-(0.8)^2}{\sqrt{N}} \\ &= 0.6745 \frac{1-64}{10} = 0.6745 \times \frac{0.36}{10} \\ &= 0.024282 \text{ or } 0.0243 \text{ when } N = 100 \end{aligned}$$

The coefficient to correlation is written as: $r = +0.8 \pm 0.0243$.

The limits of the above coefficient of correlation would be $0.8 + 0.0243 = 0.8243$ and $0.8 - 0.0243 = 0.7757$. If another sample of 100 value is chosen at random from the same universe from which this sample is taken the value of coefficient of correlation will be between $+0.8243$ and $+0.7757$.

* Probable error of a statistic may be defined to be a quantity such that exactly half of the possible values of the statistic differ from its expected value by more than that quantity. If r is a statistics and θ its expected value, then the probability that the difference of r and θ exceeds the probable error

(P.E.) is exactly $\frac{1}{2}$ i.e., $P(|r - \theta| > PE) = \frac{1}{2}$

NOTES

Conditions for the Use of Probable Error

Only under the following circumstances is the estimate of potential mistake useful:

1. The data closely resemble a normal frequency distribution, to start with (bell shaped curve).
2. A sample must have been used for the statistical measure for which the probability error was determined.
3. The sample's items had to be chosen using a random sample approach, in which there is a probability that each item will be chosen for the sample.

Functions of Probable Error of r

The probable error of the correlation coefficient has two functions:

1. **Establishing Limits:** The probable error of the correlation coefficient establishes the two limits ($r \pm P.E.$) within which there is a 50% chance that the correlation coefficients of randomly chosen samples from the same universe will lie.
2. **Meaning of " r ":** Karl Pearson's coefficient of correlation is thought to have a significant level of probable error. Following are some guidelines for interpreting the significance of the correlation coefficient depending on the likely error:
 - (i) There is no indication of a connection between two variables if the " r " (coefficient of correlation) value is smaller than the probable error ($r \pm P.E.$).
 - (ii) Correlation between the two variables is significant and certain to exist if " r " is greater than six times the probable error ($r > 6P.E.$). The correlation coefficient will be more significant if its value is greater than six times the probability of error. It is not important if " r " is less than six times the probable error ($r > 6P.E.$).
 - (iii) When ' r ' is less than 0.3, correlation should not be taken into account at all if the probability error is reasonably small.
 - (iv) If the probable error is small, correlation is definitely existing where ' r ' is above 0.5.

Properties of Probable Error

1. The maximum coefficient of correlation in the universe should be $r \pm P.E.$
2. There is no proof of correlation if the value of r is less than six times the probable error.
3. A substantial connection exists if the value of r is greater than six times the probable error.
4. Only when there are a lot of observational pairs should the probable error as a measure for assessing the coefficient of correlation be utilised. If n is small, the likelihood of mistake could produce false findings.

Conditions for the Use of Probable Error

1. The data must approximate a normal distribution.
2. Sampling has been done in a random manner.
3. A sample must have been used to construct the statistical measure for which the probability error is determined.

SOLVED EXAMPLES

Example 6. Find probable error when coefficient of correlation is 0.8 and the number of items is 64 and interpret the value of r .

Solution.

$$\begin{aligned} \text{P.E.} &= 0.6745 \frac{1-r^2}{\sqrt{N}} \\ &= 0.6745 \frac{1-(0.8)^2}{\sqrt{64}} = 0.0303 \\ \frac{r}{\text{PE}} &= \frac{0.8}{0.0303} = 26.6 \Rightarrow r > 6 \text{ PE} \end{aligned}$$

\therefore Value of r is highly significant.

Example 7. For what value of N , the coefficient of correlation equals to 0.4 will be significant

Solution.

$$\begin{aligned} \text{P.E.} &= 0.6745 \left(\frac{1-r^2}{N} \right) \\ r \text{ will be significant if} \\ r &> 6 \text{ PE} \\ \Rightarrow r &> 6 \times 0.6745 \left(\frac{1-(0.4)^2}{\sqrt{N}} \right) \\ \Rightarrow 0.4 &> 6 \times 0.6745 \left[\frac{1-0.16}{\sqrt{N}} \right] \\ \Rightarrow 0.4 \sqrt{N} &> 4.047 \times 0.84 \\ \Rightarrow \sqrt{N} &= \frac{4.047 \times 0.84}{0.4} = N = 72 \end{aligned}$$

Example 8. The value of coefficient of correlation of a sample is 0.6 where as the number of values within the sample is 36. For a different sample inside the same universe, determine the bounds of r in which r lies.

NOTES

Solution.

$$PE = 0.6745 \left(\frac{1 - r^2}{N} \right) = 0.6745 \left[\frac{1 - (0.6)^2}{\sqrt{36}} \right] = 0.072$$

Limits of r are $0.6 \pm 0.072 = 0.672$ and 0.528 .

Spearman's Rank Correlation Coefficient

Correlation by Rank When two sets of qualitative observations are ranked, such as when two independent observers rate the productivity of a group of workers as bad, fair, good, or very good, we can correlate the results using a coefficient. This will also reveal whether the preferences of the two observers for a certain quality or characteristic are similar or dissimilar. One person can assign ranks to two attributes, or two people can assign ranks to one characteristic, such as beauty, honesty, or intelligence.

The steps we use to determine the rank correlation coefficient are as follows:

1. First, rank the two series, Xs and Ys, similarly among themselves, with rank 1 going to the greatest (or smallest) value, rank 2 going to the second largest (or second smallest), and so on.
2. Determine the disparities (D) between the corresponding X and Y ranks.
3. Square these disparities and calculate the sum of their squares. *i.e.*, $\sum D^2$
4. Use the following formula to determine the rank correlation coefficient:

$$R = 1 - \frac{6 \sum D^2}{N(N^2 - 1)} \quad \dots(i)$$

Where N stands for the quantity of value pairs.

When no value in any of the two series is repeated, the aforementioned formula is applicable. (Repeated values are ranked equally and are referred to as tied values). The event that ties, we then give the mean of the ranks they both share to each of the linked observations. For instance, if the third and fourth biggest values of a variable are tied,

each of them given the rank $= \frac{3+4}{2} = 3.5$ and if the fifth, sixth and seventh largest

values of a variable are the same, we assign to each the rank $= \frac{5+6+7}{3} = 6$.

The formula below is used to determine the rank correlation coefficient when part of the values are repeated and average rankings are assigned:

$$R = 1 - \frac{6 \left[\sum D^2 + \sum \left(\frac{m^3 - m}{12} \right) \right]}{N(N^2 - 1)} \quad \dots(ii)$$

$$1 - \frac{6 \left[\sum D^2 + \frac{1}{1x} \left((m_1^3 - m_1) + (m_1^3 - m_2) + \dots \right) \right]}{N(N^2 - 1)}$$

where m is the quantity of times a given value is repeated. Values might be repeated in one series or in both series. Repetition might occur in a single value or in multiple values.

If all the values occur once, i.e., if no value is repeated, $m = 1$, and the formula (2) reduces to formula (1)

- Notes:**
1. Ranks can be allotted either in ascending order (i.e., rank 1 to the smallest value, rank 2 to the next larger value, i.e., higher than the smallest and so on) or in descending order (i.e., rank 1 to the largest value, rank 2 to the next smaller value, i.e., lower than the largest and so on) but whichever method is selected must be used for both the variables, i.e., both the series.
 2. If two or more data points have the same value (sometimes referred to as tied values), the ranks that would have been distributed independently must be averaged, and this average rank must be assigned to each point.
 3. The number of values in a series is the same as the highest rank in that series. The greatest rank in each series is equal to the number of pairs of values if two series have an equal number of values.

4.12 MERITS AND DEMERITS OF RANK CORRELATION METHOD

Merits

1. When compared to Karl Pearson's method, this one is more straightforward and straightforward.
2. This approach is particularly helpful when exact measurements of the variables under inquiry are unavailable or impossible to obtain. In other words, when the factors under study are qualitative in nature.
3. Both qualitative and quantitative data can be used with this strategy.
4. This technique can be used with erratic data as well.

Demerits

1. This approach is only applicable to ungrouped data.
2. Because the effect of extreme values is essentially ignored by the ranking system, the findings are simply approximations of the true magnitude of the data.
3. As the number of paired observations rises, the calculating process gets more challenging.
4. Further algebraic treatment is not possible.
5. Rank correlation do not give linear correlation.

Coefficient of correlation can be measured from regression coefficient also which is being dealt under regression.

NOTES

NOTES

Note: (i) If $\Sigma D = 0$, then $R = 1$

Performance	Judge I	Judge II
1	3	3
2	5	5
3	1	1
4	6	6
5	2	2
6	7	7
7	4	4

Then $\Sigma D = 0$ and $D^2 = 0$ and R will be equal to 1.

(ii) If the ranks assigned by the two judges are in opposite order as shown in the following table, then the rank correlation coefficient will be equal to -1 .

Performance	Judge I	Judge II	D	D^2
1	1	7	-6	36
2	2	6	-4	16
3	3	5	-2	4
4	4	4	0	0
5	5	3	2	4
6	6	2	4	16
7	7	1	6	36
			$\Sigma D = 0$	$\Sigma D^2 = 112$

$$R = 1 - \frac{6 \Sigma D^2}{N(N^2 - 1)} = 1 - \frac{6 \times 112}{7 \times 48} = 1 - 2 = -1 \text{ (Please Note)}$$

Example 9. Find out coefficient of rank correlation from the following data:

X	10	12	8	15	20	25	40
Y	15	10	6	25	16	12	18

Solution.

X	Rank (R_1)	Y	Rank R_2	$D = (R_1 - R_2)$	D^2
10	6	15	4	2	4
12	5	10	6	-1	1
8	7	6	7	0	0
15	4	25	1	3	9
20	3	16	3	0	0
25	2	12	5	-3	9
40	1	18	2	-1	1
$N = 7$				$\Sigma D^2 = 24$	

$$R = 1 - \frac{6\sum D^2}{N(N^2 - 1)} = 1 - \frac{6(24)}{7(7^2 - 1)} = 1 - \frac{144}{336} = 0.57$$

Example 10. From the following data find out coefficient of rank correlation between price and supply.

Price	4	6	8	10	12	14	16	18
Supply	10	15	20	25	30	35	40	45

Solution.

X	Rank (R_1)	Supply	Rank (R_2)	$D = (R_1 - R_2)$	D^2
4	8	10	8	0	0
6	7	15	7	0	0
8	6	20	6	0	0
10	5	25	5	0	0
12	3	30	3	0	0
14	4	35	4	0	0
16	2	40	2	0	0
18	1	45	1	0	0

To correlation between price and supply is perfect positive.

$$R = 1 - \frac{6\sum D^2}{N(N^2 - 1)} = 1 - \frac{0}{8(8^2 - 1)} = 1$$

Example 11. The results of 10 students' grades in mathematics and statistics were determined to have a coefficient of rank correlation of 0.5. The difference between one student's scores in the two subjects was then discovered to be 3 instead of 7, which was incorrect. What is the appropriate rank correlation coefficient?

Solution.

Edward Spearman's formula for the coefficient or rank correlation is:

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

Here, $R = 0.5$, $N = 10$, $\sum D^2 = ?$

$$\begin{aligned} 0.5 &= 1 - \frac{6\sum D^2}{10(10^2 - 1)} = 1 - \frac{6\sum D^2}{10 \times 99} = 1 - \frac{\sum D^2}{165} \text{ or } \frac{\sum D^2}{165} \\ &= 1 - 0.5 \text{ or } \sum D^2 = 82.5 \end{aligned}$$

Now, Correct value of $\sum D^2 = \text{Incorrect } \sum D^2 - (\text{Incorrect difference in ranks})^2 + (\text{Correct difference in ranks})^2$

$$= 82.5 - (3)^2 + (7)^2 = 122.5$$

NOTES

NOTES

Hence, the correct rank correlation coefficient

$$= 1 - \frac{6 \times 122.5}{10(10^2 - 1)} = 1 - \frac{122.5}{165} = \frac{42.5}{165} = 0.2576$$

Example 12. If the rankings of these colour shades by two typical consumer judges are highly associated, a company that is unsure of how customers will react to their product in ten various colour shades decides to create it in those colour shades.

Colour no.	1	2	3	4	5	6	7	8	9	10
Ranking by Judge I	6	4	3	1	2	7	9	8	10	5
Ranking by Judge II	4	1	6	7	5	8	10	9	3	2

You are to decide if the firm should introduce the new product in the market.

Solution. Is there a consensus between the two judges approving the launch of the product by the company?

D is	6-4	4-1	3-6	1-7	2-5	7-8	9-10	8-9	10-3	5-2
or $ D $ is	2	3	3	6	3	1	1	1	7	3
D^2 is	4	9	9	36	9	1	1	1	49	9
$\Sigma D^2 = 128$										

$$\begin{aligned} \text{Rank correlation} &= 1 - \frac{6 \Sigma D^2}{N(N^2 - 1)} = 1 - \frac{6 \times 128}{10(100 - 1)} \\ &= 1 - 0.78 = 0.22. \end{aligned}$$

Because there is little agreement between the two judges given the poor rank correlation, the company shouldn't develop the product to put it on the market.

Example 13. Three judges ranked the contestants in a beauty pageant in the following order:

First Judge	1	6	5	10	3	2	4	9	7	8
Second Judge	3	5	8	4	7	10	2	1	6	9
Third Judge	6	4	1	8	1	2	3	10	5	7

Choose the judges who have the closest affinity for aesthetic preferences using the rank correlation method.

Solution.

In order to determine which pair of judges is closest to sharing a similar aesthetic sense, we shall evaluate the Rank Correlation between the assessments of:

(i) 1st and 2nd judge, (ii) 1st and 3rd judge and (iii) 2nd and 3rd judge.

The Ranks given by the three judges would be denoted by R_1 , R_2 and R_3

$$N = 10$$

Rank Correlation Coefficient

$$R = 1 - \frac{6 \Sigma D^2}{N^3 - N} = 1 - \frac{6 \Sigma D^2}{N(N^2 - 1)}$$

∴ Correlation Coefficient between first and second judge

$$= 1 - \frac{6 \times 200}{10(10^2 - 1)} = 1 - \frac{6 \times 20}{99} = 1 - \frac{40}{33} = -0.212$$

Calculation of Rank Coefficient of Correlation

R_1	R_2	R_3	Rank Difference in Pairs			D_{12}^2	D_{13}^2	D_{23}^2
			$R_1 - R_2$ $= D_{12}$	$R_1 - R_3$ $= D_{13}$	$R_2 - R_3$ $= D_{23}$			
1	3	6	-2	-5	-3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	-3	-4	-1	9	16	1
10	4	8	6	2	-4	36	4	16
3	7	1	-4	2	6	16	4	36
2	10	2	-8	0	8	64	0	64
4	2	3	2	1	-1	4	1	1
9	1	10	8	-1	9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	-1	1	2	1	1	4
						200	60	214

Correlation Coefficient between first and third judge

$$= 1 - \frac{6 \times 60}{10 \times 99} = 1 - \frac{4}{11} = 0.636$$

Correlation Coefficient between second and third judge

$$= 1 - \frac{6 \times 214}{10 \times 99} = 1 - \frac{214}{165} = -0.297$$

As a result, there is a negative connection between the first and second judges, meaning that they have different perspectives about what is beautiful. The views of the second and third judges are also at odds with one another. However, the first judge and the third judge have comparable viewpoints (their correlation is positive), meaning that they share many of the same preferences. Consequently, the first and third judges share the same (common) aesthetic preferences.

Example 14. Find out coefficient of rank correlation from the following:

X	50	33	40	10	15	15	65	24	15	57
Y	12	12	24	6	15	5	20	9	6	19

NOTES

NOTES

Solution.

X	Y	$R_1(X)$	$R_2(Y)$	$R_1 - R_2 = D$	D^2
50	12	3	5.5	-2.5	6.25
33	12	5	5.5	-0.5	0.25
40	24	4	1	+3.0	9.00
10	6	10	8.5	+1.5	2.25
15	15	8	4	+4.0	16.00
15	5	8	10	-2.0	4.00
65	20	1	2	-1.0	1.00
24	9	6	7	-1.0	1.00
15	6	8	8.5	-.05	0.25
57	19	2	3	-1.0	1.00
$\Sigma D^2 = 41$					

Here in the first series, *i.e.*, X series value 15* is repeated 3 times, in the Y series, the value 12 and 6 are each repeated twice.

\therefore Rank correlation coefficient

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

$$= 1 - \frac{6 \left[\Sigma D^2 + \frac{m_1(m_1^2 - 1)}{12} + \frac{m_2(m_2^2 - 1)}{12} + \frac{m_3(m_3^2 - 1)}{12} \right]}{N(N^2 - 1)}$$

$$= 1 - \frac{6 \left[41 + \frac{3(9 - 1)}{12} + \frac{2(4 - 1)}{12} + \frac{2(4 - 1)}{12} \right]}{10 \times (100 - 1)} = 1 - \frac{6 \left[41 + \frac{24 + 6 + 6}{12} \right]}{990} = 1 - \frac{44 \times 6}{990}$$

$$= 0.733$$

4.13 REGRESSION ANALYSIS

A regression equation is a mathematical formula that predicts the values of one dependent variable from the known values of one or more independent variables (explanatory variables). This phrase comes from Sir Francis Galton's (1822-1911) pioneering investigations on heredity, in which he compared the heights of sons and fathers. Galton demonstrated how the heights of sons of tall fathers over time declined

toward the population mean height. To put it another way, sons of unusually tall fathers typically are shorter than their fathers, but sons of unusually short fathers typically are taller.

Today, the term "regression" is used to describe all different kinds of prediction issues; nevertheless, it does not always indicate a regression towards the population mean.

For Example:

1. A production manager might be required to estimate the amount of raw materials that should be purchased during a particular year to meet the production targets and to compensate for the wastage of materials that occurs during the course of production and storage etc.
2. A sales manager might submit the estimates of sales potential in different regions for the budgeting purposes.
3. Expected sales from each of the groups of new recruits might have to be estimated for their placement in different sectors of the market.

Morris Hamburg claims, "The term regression analysis refers to the methods by which estimates are made of the values of a variable from a knowledge of the values of one or more other variables and to the measurement of the errors involved in this estimation process."

M.M. Blair says, "Regression is the measure of the average relationship between two or more variables in terms of the original units of the data".

According to Taro Yamane, "Regression analysis is one of the methods most frequently used in economics and business research to discover a relationship between two or more variables that are causally connected."

The average association between two or more variables in terms of the original units of data is thus measured mathematically using regression analysis.

4.14 REGRESSION LINES OR REGRESSION EQUATIONS

J.R. Stockton says, "The device used for estimating the value of one variable from the value of the other consists of a line through the points drawn in such a manner as to represent the average relationship between the two variables. Such a line is called the **line of regression**."

As per the method of least squares, the two regression lines are:

$$Y - \bar{Y} = b_{YX}(X - \bar{X}) \quad [\text{Regression Equation of } Y \text{ on } X]$$

$$\text{and } X - \bar{X} = b_{XY}(Y - \bar{Y}) \quad [\text{Regression Equation of } X \text{ on } Y]$$

where, \bar{X} = mean of X values : \bar{Y} = mean of Y value

NOTES

NOTES

$$b_{YX} = \text{Regression coefficient of } Y \text{ on } X = r \frac{\sigma_y}{\sigma_x}$$

$$b_{XY} = \text{Regression coefficient of } X \text{ on } Y = r \frac{\sigma_x}{\sigma_y}$$

$$b_{YX} = r \frac{\sigma_y}{\sigma_x} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{N\sigma_x\sigma_y} \times \frac{\sigma_y}{\sigma_x} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{N\sigma_x^2} = \frac{\frac{\Sigma xy}{N}}{\frac{\Sigma x^2}{N}} = \frac{\Sigma xy}{\Sigma x^2}$$

$$b_{XY} = r \frac{\sigma_x}{\sigma_y} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{N\sigma_x\sigma_y} \left(\frac{\sigma_x}{\sigma_y} \right) = \frac{\Sigma xy}{\Sigma y^2}$$

where, $x = X - \bar{X}$, $y = Y - \bar{Y}$, i.e., the deviations are taken from the actual mean of X and Y .

$$b_{YX} = \frac{\frac{\Sigma XY}{N} - \left(\frac{\Sigma X}{N} \right) \left(\frac{\Sigma Y}{N} \right)}{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X^2}{N} \right)} \quad \text{if no deviations are taken.}$$

$$b_{XY} = \frac{\frac{\Sigma XY}{N} - \left(\frac{\Sigma X}{N} \right) \left(\frac{\Sigma Y}{N} \right)}{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X^2}{N} \right)} \quad \text{if no deviations are taken.}$$

If the deviations are taken from the assumed mean of X say A and assumed mean of Y say B then,

$$b_{YX} = \frac{\frac{\Sigma d_x d_y}{N} - \left[\frac{\Sigma d_x}{N} \right] \left[\frac{\Sigma d_y}{N} \right]}{\frac{\Sigma d_x^2}{N} - \left(\frac{\Sigma d_x}{N} \right)^2}$$

$$b_{XY} = \frac{\frac{\Sigma d_x d_y}{N} - \left(\frac{\Sigma d_x}{N} \right) \left(\frac{\Sigma d_y}{N} \right)}{\frac{\Sigma d_y^2}{N} - \left(\frac{\Sigma d_y}{N} \right)^2} \quad \text{where } d_x = X - A, d_y = Y - B$$

4.15 PROPERTIES OF REGRESSION COEFFICIENTS

1. Both the Regression Coefficient should be of the same sign:

$$(b_{yx})(b_{xy}) = \left[r \frac{\sigma_y}{\sigma_x} \right] \left[r \frac{\sigma_x}{\sigma_y} \right] = r^2 = \text{Coefficient of determination}^*$$

\Rightarrow Between the two regression coefficients, r represents the geometric mean. Since the square of a number is always positive, the product of two regression coefficients, or r^2 , is also always positive. However, this is only possible if the two regression coefficients have the same sign, that is, they must be both positive or both negative.

2. Determining the Correlation Coefficient's Sign from the Regression Coefficients:

The correlation coefficient is negative if both the regression coefficients are positive. [Please note].

3. r will have the same sign as that of b_{yx} and b_{xy} .

4. Change of Origin and Change of Scale:

Although neither regression coefficient is affected by the change in origin, if the scale changes for X and Y are not the same, they depend on the change of scale. If the change of scale in X and Y is identical, the regression coefficients are independent of the change of scale also.

5. Both the Regression Coefficients cannot be more than 1:

Since, the product of the two regression coefficients is equal to r^2 which is less than or equal to one, both the regression coefficients cannot be more than 1. Sometimes it may happen that one of the regression coefficients is less than 1 but very close to 1 and the other is more than 1 in such a way that the product of two regression coefficients is more than 1. Such cases are also not possible. **The product of two regression coefficients should not be more than 1.**

6. If $b_{yx} \times b_{xy} = 1$, i.e., $r^2 = 1$ or $r = \pm 1$, i.e., when there is perfect correlation between the two variables, then $b_{yx} = \frac{1}{b_{xy}}$ and $b_{xy} = \frac{1}{b_{yx}} = \frac{1}{b_{YX}}$ i.e., the two regression coefficients are reciprocal of each other.

$$7. \frac{b_{yx}}{r} = \frac{\sigma_y}{\sigma_x}, \frac{b_{xy}}{r} = \frac{\sigma_x}{\sigma_y}$$

8. The correlation coefficient will be bigger than the arithmetic mean of the two regression coefficients.

NOTES

* Coefficient of determination = $r^2 = (\text{Karl Pearson's linear coeeration coefficient})^2$

\therefore Coefficient of determination is always positive whether r is positive or negative.

NOTES

4.16 WHY THERE ARE TWO REGRESSION LINES?

Two regression lines as shown earlier.

Equation I. $Y - \bar{Y} = b_{YX} (X - \bar{X})$ is used to predict the values of Y from the known values of X .

Equation II. *i.e.*

$X - \bar{X} = b_{XY} (Y - \bar{Y})$ is used to predict the values of X from known values of Y .

For the reasons listed below, a single regression equation is insufficient to predict the values of both X and Y :

1. Since the foundation and underlying assumptions used to derive these equations are very different, the two regression lines are not reversible. While the regression equation of X on Y is achieved by minimising the sum of the squares of the errors parallel to the X -axis, the regression equation of Y on X is obtained by doing the opposite. *i.e.*, the horizontal and vertical deviations both have to be minimised.
2. There will only be one regression line if the correlation is perfect, or $r = 1$. In this case, the two regression lines meet.

Regression Relation is Irreversible

The regression relationship is irreversible because we require two regression lines to forecast the values of X and Y . It is not possible to predict the values of X from the given value of Y using the same regression equation that is used to predict the value of Y from a given value of X . Regression is an average, irreversible, and useful relationship.

4.17 EXAMPLES OF IRREVERSIBLE RELATION

1. If a family's income rises, their expenses will too, but there is no assurance that their income will follow suit if their expenses rise.
2. A good crop will result if the rainfall is timely and beneficial, but a good crop does not guarantee that the rainfall was timely and beneficial.

4.18 COEFFICIENT OF DETERMINATION

Coefficient of determination (r^2): The linear correlation's square. It gauges how much of the variation in Y values can be accounted for by the linear connection with X . From zero (none of the variations are explained) to one (all of the variances are explained), the scale runs the gamut.

HINTS FOR QUICK RECALL

- **Correlation:** When the values of two variables change as a result of a change in one of the variables, this relationship is known as correlation (or covariation),

i.e., cost of advertisement affects sales, increase in radius affects the circumference of a circle. Price of a commodity affects demand.

- **Positive Correlation:** The relationship between two variables is referred to as positive when both move in the same direction, that is, when both increase or decrease together. or straight, *e.g.*, increase in heat increases temperature, increase in height increases weight, increase in price of commodity increases amount of supply.
- **Negative Correlation:** The correlation is negative or inverse when the change is in the opposite direction, that is, when one variable is rising and the other is falling. For instance, a commodity's demand may decline as a result of rising pricing. Sale of woollen garments decreases when day temperature increases, decrease in yield of crops affects the increase in price.

- **Degree of Correlation:**

Positive: Perfect (+1), High (+0.75 to 1), Moderate (+0.25 to 0.75), Low (0 to +0.25)

Negative: Perfect (−1), High (−0.75 to −1), Moderate (−0.25 to −0.75)

No Correlation = 0.

- **Methods of Studying Correlation:**

Scatter Diagram: It is a simple visual graphic method. When the trend of plotted points is upward the correlation is positive, when it is downward the correlation is negative.

Karl Pearson's Coefficient of Correlation: It deals only with linear correlation.

Karl Pearson's coefficient of correlation is represented by r .

The limits of r are -1 to $+1$, *i.e.*, $-1 \leq r \leq 1$.

NOTES

LIST OF FORMULAE

1. Karl Pearson's Coefficient of Correlation

Actual Mean Method

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{N \sigma_x \times \sigma_y} = \frac{\Sigma xy}{N \cdot \sigma_x \sigma_y} = \frac{\Sigma xy}{N} \times \frac{1}{\sigma_x} \times \frac{1}{\sigma_y} \text{ where } x = X - \bar{X}, y = Y - \bar{Y}$$

$$= \frac{\Sigma xy}{N \sqrt{\frac{\Sigma x^2}{N}} \sqrt{\frac{\Sigma y^2}{N}}} = \frac{\Sigma xy}{\sqrt{\Sigma x^2} \sqrt{\Sigma y^2}}$$

Direct Method

$$\frac{\frac{\Sigma XY}{N} - \frac{\Sigma X}{N} \cdot \frac{\Sigma Y}{N}}{\sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2} \sqrt{\frac{\Sigma Y^2}{N} - \left(\frac{\Sigma Y}{N}\right)^2}}$$

NOTES

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

2. Coefficient of correlation with original data

$$r = \frac{\Sigma XY - \frac{\Sigma X \times \Sigma Y}{N}}{\sqrt{\left(\Sigma X^2 - \frac{(\Sigma X)^2}{N} \right) \left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{N} \right)}}$$

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

3. Coefficient of correlation with original data when deviations are taken from means

$$r = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma (X - \bar{X})^2 \Sigma (Y - \bar{Y})^2}}$$

4. Coefficient of correlation by direct method

$$r = \frac{\Sigma xy}{N \sigma_x \sigma_y} = \frac{\Sigma xy}{\sqrt{\Sigma x^2} \times \sqrt{\Sigma y^2}}$$

where, r = coefficient of correlation, Σxy = sum of the product of deviations in X and Y series from their arithmetic means, σ_x = Standard deviation of the series X , σ_y = standard deviation of the series Y and N = total number of observations.

5. Coefficient of correlation by the short-cut method

$$r = \frac{\Sigma dx dy - \frac{\Sigma dx \times \Sigma dy}{N}}{\sqrt{\left(\Sigma dx^2 - \frac{(\Sigma dx)^2}{N} \right) \left(\Sigma dy^2 - \frac{(\Sigma dy)^2}{N} \right)}}$$

where, dx and dy are the deviations of individual observation in X and Y series from their respective assumed means, Σdx and Σdy are the sums of the deviations in X and Y series respectively.

6. A simplified form of the above formula

$$r = \frac{N \Sigma f dx dy - (\Sigma f dx)(\Sigma f dy)}{\sqrt{[N \Sigma f dx^2 - (\Sigma f dx)^2][N \Sigma f dy^2 - (\Sigma f dy)^2]}}$$

7. Coefficient of rank correlation between two ranked variables

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

where, r_s stands for Spearman's rank correlation, d for the difference between the ranks of the two variables and N for number of paired observations.

Explanation of Symbols

- r = Karl Pearson's coefficient of correlation.
- x = $(X - \bar{X})$, deviations taken from actual mean of X series.
- y = $(Y - \bar{Y})$, deviations taken from actual mean of Y series.
- σ_x = Standard deviation of X series.
- σ_y = Standard deviation of Y series.
- ΣX = Sum of the values of X series.
- ΣY = Sum of the values of Y series.
- ΣX^2 = Sum of square of the values of X series.
- ΣY^2 = Sum of square of the values of Y series.
- ΣXY = Sum of the products of X, Y values.
- N = No. of pairs of observations.

Check Your Progress

I. Multiple Choice Questions

- Spurious correlation is _____.
 - Bad relation between two variables
 - Very low correlation between two variables
 - Correlation between two variables having no causal relation
 - Negative correlation
- For a group of eight students, the sum of squares of differences in ranks for commerce and economics marks was 50. The value of the rank correlations coefficient will be _____.
 - 0.5
 - 0.4
 - 0.6
 - None of these
- The coefficient of correlation _____.
 - has no limit
 - is less than 1
 - is more than -1
 - lies between -1 and +1
- If the sum of the product of deviations of X and Y series from their means is zero, the correlation coefficient will be _____.
 - 0
 - 1
 - 1
 - None

NOTES

NOTES

5. If the coefficient of correlation between two data sets is -0.85 and the variances of the two sets are 16 and 9 respectively, then the covariance between the data sets is _____.

- (a) -10.20 (b) -4.20
 (c) 3.20 (d) 7.20
 (e) 10.20

6. The correlation coefficient of a variable with itself is

- (a) 1 (b) -1
 (c) 0 (d) (a) or (b) above
 (e) None of the above

7. Which of the following statements is true?

- (a) r^2 is always less than or equal to the correlation coefficient (r) in magnitude.
 (b) r is always less than or equal to r^2 in magnitude.
 (c) $r^2 > 0$ implies $r > 0$.
 (d) r^2 is always less than or equal to r .
 (e) r is always less than or equal to r^2 .

II. State whether the following Statements are True or False

1. Correlation coefficient establishes cause and effect relationship between two variables.
2. Coefficient of correlation is a relative measure of association between two variables.
3. Correlation coefficient is an absolute measure of relationship between two variables.
4. r is expressed in the units of original data.
5. Correlation coefficient is always positive.
6. $|r|$ is $0 \leq |r| \leq 1$.
7. Scatter diagram helps in determining the degree of correlation.
8. If $r = +1$ or -1 , the correlation will be perfect.
9. Correlation coefficient measures only the degree of linear relationship.
10. Negative correlation means that both the variables have decreasing values.
11. Positive correlation means that both variables have increasing values.
12. If the direction of change in the two variables is the same, the two variables will be positively correlated.
13. If the change in the two variables is in opposite direction, the correlation will be negative.
14. If $X = 2Y$, X and Y will be perfectly correlated.
15. If $X = 20 - Y$, there will be a perfect negative correlation between X and Y .
16. If $X = Y + 2$, $r = +1$.
17. $r = 0$ means absence of correlation.
18. Coefficient of correlation between height and weight is 1.4.
19. Correlation coefficient is independent of the change of origin and scale.
20. The correlation coefficient increases with the number of paired observations.
21. The coefficient of correlation can be calculated only if both the variables are in the same unit.

22. The correlation between the age of applicants for life insurance and premium of insurance is positive.

III. Fill in the Blanks

- Karl Pearson's coefficient of correlation measures _____.
(linear correlation, non-linear correlation)
- Maximum positive value of coefficient of correlation is _____. (+1, +1.5)
- Correlation between price and demand is _____. (positive, negative)
- Correlation between price and supply is _____. (positive, negative)
- If the correlation is perfect positive, its value is _____. (+1, 0)
- Covariance measures _____ variation of two variables.
- Covariance may be _____.
- If $r = 0.5$, $\Sigma xy = 120$, $\Sigma y = 8$, $\Sigma x^2 = 90$, then $x =$ _____.
- For finding the correlation between two attributes we consider _____.
- If the sum of the products of deviations of x and y series from their means is zero, then the coefficient of correlation will be _____.
- Two judges in a beauty contest rank the 12 entries as follows:

Judge A	1	2	3	4	5	6	7	8	9	10	11	12
Judge B	12	11	10	9	8	7	6	5	4	3	2	1

Then rank correlation coefficient between them will be _____.

NOTES

4.19 ANSWERS TO 'CHECK YOUR PROGRESS'

I. Multiple Choice Questions

- (c)
- (b)
- (d)
- (a)
- (a)
- (a)
- (a)

II. State whether the following Statements are True or False

- False. (Not necessary). Suppose there is a danger of famine, the price will increase and the demand will also increase. But we cannot say that one of the price or demand, is the cause and the other is effect, because the cause is famine.
- True. Because it gives the relation between two series.
- False.
- False. It is free from units.

NOTES

5. False. It can be positive or negative.
6. True. $-1 \leq r \leq 1$ but $0 \leq |r| \leq 1$.
7. False It only gives a rough idea about the correlation between two variables.
8. True.
9. True. Refer to Karl Pearson's coefficient of correlation.
10. False. The correlation is negative when both the variables increase or decrease in opposite sense.
11. False. The correlation is positive, when both the variables either increase together or decrease together.
12. True.
13. True.
14. True. In Q. 14, the line has positive slope and in this case, the relationship is linear and the scatter diagram will be a straight line.
15. True. In Q. 15, the line has negative slope. In this case, the relationship is linear and the scatter diagram will be a straight line.
16. True. The scatter diagram will be a straight line with positive slope therefore $r = 1$
17. True.
18. False. There is no fixed correlation between height and weight.
19. True.
20. False.
21. False. Correlation between price and demand is negative but they are not measured in the same units.
22. True. As the age increases, the premium of insurance for fresh policies also increases.

III. Fill in the Blanks

1. Linear correlation,
2. +1,
3. negative,
4. positive
5. +1,
6. Joint
7. Positive,
8. negative or zero,
9. rank correlation coefficient
10. Zero
11. -1, 12. 0.87.

4.20 SUMMARY

- Introduction
- Definitions
- Importance of correlation
- Kinds of co-relation
- Linear correlation and rank co-relation

4.21 KEY TERMS

- **Correlation analysis:** A statistical data analysis focused on determining if two variables are related to one another.
- **Correlation coefficient (r):** A metric that shows how closely two variables are related. It ranges between -1 and $+1$.
- **Cov (X, Y):** This term refers to the covariance between X and Y , which is calculated as the average of the products of the variances from the means for n pairs of X and Y series. A relationship between two variables that can be illustrated by a curved line is called a curvilinear relationship.
- **Direct relationship:** When two variables move in the same direction, there is a direct relationship between them.
- **Inverse relationship:** When two variables are related, it may be shown that they are moving in the opposite directions.
- **Linear relationship:** When the points in a scatter plot congregate around a straight line with a nonzero slope, there is a certain form of association between the two variables known as a linear relationship.
- **Karl Pearson's:** Karl Pearson's measure of correlation, the Pearson's correlation coefficient. The product moment correlation coefficient is another name for it.

The most common correlation coefficient is this one.

- **Rank correlation:** A technique for figuring out correlation when the data aren't available in numerical form and the ranking technique is employed instead.
- **Rank-correlation coefficient:** A measurement of how closely two variables are related that is based on the ranks of observations rather than their numerical values.
- **Scatter diagram:** A depiction of the paired X and Y data that reveals the general pattern of relationships between the two variables.

NOTES

4.22 SELF-ASSESSMENT QUESTIONS AND EXERCISES**Short Answer Questions**

1. Distinguish between univariate and bivariate analysis with examples.
2. Define correlation. Give examples from practical life.
3. What are the different types of correlation? Name them.
4. Briefly explain the meaning of four types of correlation.
5. What is a scatter diagram? What does it show?
6. How is the Karl Pearson's coefficient of correlation obtained?
7. Mention the three basic assumptions of Karl Pearson's coefficient of correlation.
8. Mention two merits and two demerits of Karl Pearson's coefficient of correlation.
9. Positive correlation is found between number of children born and export over last decade. Yes or No.
10. If the coefficient of correlation between X and Y is zero, does it mean that there is the absence of any type of relation between them?
11. State the nature of the following correlations (positive, negative or no correlation)
12. Explain the Algebraic Properties of correlation coefficient.
13. What will you interpret if (i) $r = 0$, (ii) $r = +1$, (iii) $r = -1$, (iv) $r = +0.8$ and (v) $r = -0.8$?

Long Answer Questions**Karl Pearson's Correlation**

1. Using Karl Pearson's approach and the actual mean value of the following series, determine the coefficient of correlation between the ages of the husband and wife:

Age of Husband (X)	20	23	27	31	35	38	40	42
Age of Wife (Y)	18	20	24	30	32	34	36	38

[Ans. $r = +0.994$]

2. Information relating to price and quantity of a commodity for five months is given below:

Months	1	2	3	4	5
Price (X)	10	10	11	12	12
Quantity (Y)	5	6	4	3	2

Determine the price-to-quantity correlation coefficient and make a note of its size and sign.

[Ans. -0.949]

3. A work study practitioner made specific observations regarding a worker's rate of work completion. In terms of how he rated the employee:

90	95	115	95	85	110	90	95	95	95
----	----	-----	----	----	-----	----	----	----	----

The worker actually performed at the ratings:

95	90	110	100	85	105	95	100	105	95
----	----	-----	-----	----	-----	----	-----	-----	----

respectively. How would you place the work study practitioner insofar as his ability in rating is concerned?

[Ans. $r = 0.79$. Since there is a high degree of positive correlation between the observation of the practitioner and work actually done by the worker, the ability of the work study practitioner in rating is quite good]

4. (i) Compute the coefficient of correlation between the corresponding values of X and Y in the following table:

X	2	4	5	6	8	11
Y	18	12	10	8	7	5

- (ii) Add 6 and multiply each X -value in the table by 2. Add 15 and multiply each value of Y in the table by 3. Discover the correlation factor between the two fresh sets of data. Why do you or do not you achieve the same outcome as in (i)?

[Ans. -0.92]

5. Find out coefficient of correlation between X and Y .

X	10	12	8	15	20	25	40
Y	15	10	6	25	16	12	8

[Ans. -0.19]

6. The total of the multiplication of deviation of X and $Y = 3,044$.

No. of pairs of the observation is 10.

Total deviations of $X = -170$, total of deviations of $Y = -20$.

Total of the squares of deviations of $X = 8,238$. Total of the squares of deviations of $Y = 2,264$.

Calculate the correlation coefficient when X and Y have arbitrary means of 82 and 68, respectively.

[Ans. 0.7812]

7. The following data should be used to calculate the coefficient of correlation. Comment on the outcome.

Experience X	16	12	18	4	3	10	5	12
Performance Y	20	22	24	17	19	20	18	20

[Ans. 0.8247]

8. The correlation coefficient between two series is found to be 0.751. Find the probable error and indicate whether correlation is significant.

[Ans. Significant]

NOTES

NOTES

9. Calculate the coefficient of correlation from the following information:

- (i) Number of pairs of x and y series = 16
 (ii) SD of x -series = 3.21
 (iii) SD of y -series = 3.42

Summation of products of deviations of x and y series from their mean = 142

[Ans. 0.808]

10. By comparing deviations from each series' respective means, the following results are found between the two series. Calculate the correlation coefficient.

	x -series	y -series
Number of items	7	7
Sum of the squares of deviations from their arithmetic mean	28	76
Sum of products of deviations of x and y series from their respective means	46	

[Ans. 0.997]

11. How would you interpret if the correlation coefficient r is equal to.

(1) 0 (2) -1 (3) 1 (4) 0.2 (5) 0.9

- [Ans. (1) No linear correlation
 (2) Perfect negative correlation
 (3) Perfect positive correlation
 (4) Low degree of positive correlation
 (5) High degree of positive correlation.]

12. Calculate the coefficient of correlation from the following data and comment on the result.

Experience X	16	12	18	4	3	10	5	12
Performance Y	20	22	14	17	19	20	18	20

[Ans. 0.8247]

Rank Correlation

1. Find rank correlation coefficient from the following data:

X	8	36	98	25	75	82	92	62	65	35
Y	84	51	91	60	68	62	86	58	35	49

[Ans. 0.4545]

2. Calculate the rank order correlation between price and supply:

Price	4	6	8	10	12	14	16	18
Supply	10	15	20	25	30	35	40	45

[Ans. 0.4545]

3. Determine the Spearman's coefficient of correlation between the grades that judges X and Y issued to 10 students in a particular competitive test as given below:

S. No.	1	2	3	4	5	6	7	8	9	10
Marks by X Judge	52	53	42	60	45	41	37	38	25	27
Marks by Y Judge	65	68	43	38	77	48	35	30	25	50

4. Three judges rank ten contestants in a beauty pageant in the following order. Use Spearman's rank correlation coefficient to identify the judges who have the closest affinities for aesthetics.

I Judge	1	6	5	10	3	2	4	9	7	8
II Judge	3	5	8	4	7	10	2	1	6	9
III judge	6	4	9	8	1	2	3	10	5	7

[Hint: Find r_{12} , r_{23} and r_{13} by rank correlation method][Ans. 1st Judge and 3rd Judge]

5. It was discovered that the rank correlation coefficient of the grades received by 10 students in mathematics and statistics was 0.5. The difference between a student's scores in the two subjects was then discovered to have been mistakenly calculated as 3, 6. Which rank correlation coefficient is appropriate?

[Ans. 6]

A debate competition with 10 competitors had a rank correlation coefficient of 0.6, according to the calculations. Later on, it was found that the difference in some participants' ranks was actually 8 instead of 3. Calculate the appropriate rank correlation coefficient.

6. The following eight students scored in accountancy and economics. Determine the rank correlation coefficient.

Accountancy (X)	25	30	38	22	50	70	30	90
Economics (Y)	50	40	60	40	30	20	40	70

[Ans. 0.02]

4.23 REFERENCES

1. D.N. Elhance, Veena Elhance and B.M. Aggarwal, 2007, Fundamentals of Statistics. Kitab Mahal, New Delhi.
2. B.M. Aggarwal, 2012, Business Statistics (With Lab Work), Himalaya Publishing House, Mumbai.
3. B.M. Aggarwal, Dr. Puja A. Gulati, Neha Aggarwal, 2022, Statistics for Business and Economics. Kitab Mahal, New Delhi.

NOTES

Unit V Index Numbers

Learning Objectives:

By the end of this unit the learners would be able to:

- Understand Index numbers
- Their definition
- Types of Index numbers
- Comparison between Laspeyre's and Paasche's Index number
- Selection of an average
- Advantages or Uses of Index numbers
- Limitations of Index numbers
- Splicing of Index numbers
- Deflating of Index numbers

Structure:

- 5.1 Introduction
- 5.2 Definition
- 5.3 Types of Index Numbers
- 5.4 Construction of Index Numbers
- 5.5 Methods of Construction
- 5.6 Comparison between Laspeyre's Index Number and Paasche's Index Number
- 5.7 Selection of an Average for the Construction of Index Numbers
- 5.8 Problems in the Construction of Index Numbers
- 5.9 Advantages or Uses of Index Numbers
- 5.10 Limitations of Index Numbers
- 5.11 Splicing of Index Numbers
- 5.12 Answers to 'Check Your Progress'
- 5.13 Summary
- 5.14 Key Terms
- 5.15 Self-Assessment Questions and Exercises
- 5.16 References

5.1 INTRODUCTION

One of the most often used statistical tools is the index number. Index figures reveal whether prices are rising or falling, industrial production is increasing or decreasing, and whether sales are up or down from a prior period. They are used to gauge economic health and act as warning signs of inflationary or deflationary trends.

Index numbers are basically used in describing relative changes generally expressed in percentage form (without using the word ‘percentage’) in economic variables like production, wages, price, employment etc., over time. Index numbers can also be used to study the relative changes on the basis of geographic locations or some other characteristics, e.g., incidence of crimes, number of road accidents, etc. Even the number of rooms in the dwelling of a family may be used as an index of the family’s economic status. In light of this, index numbers can be thought of as instruments that gauge changes in a phenomenon's relative intensity through time, space, or another factor. The phenomenon can be the costs of goods, the actual output of things, or both. Marketed or consumed or such concepts as intelligence, beauty or efficiency.

In general, index number are understood to express relative changes in price of a single good but in common practice it is used to represent the variations in a group of related items or series. The series to be grouped may relate to production, prices, unemployment or cost of living index. Such index numbers are described as barometers of economic activity.

5.2 DEFINITION

As a result, an **index number** may be thought of as a tool that enables us to aggregate variations in many series in order to produce a number that accurately captures the overall effects of a change in the constituent elements.

An index number is a statistical tool that quantifies how closely connected variables have changed in magnitude over time or space.

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, geographical location or some other characteristic.*”

According to **Wheldon**, “*A index number is a statistical device for indicating the relative movements of the data where measurement of actual movements is difficult or incapable of being made.*”

“Index Numbers are devices for measuring differences in the magnitude of a group of related variables.”

– **Croxtton and Cowdon**

“Index Numbers are used to measure the changes in some quantity which we cannot observe directly.”

– **Dr. A.L. Broley**

NOTES

NOTES

“An index number is a quantity which by reference to a base period shows by its variation, the changes in the magnitudes over a period of time.”

– John. I. Griffen

For calculating index numbers we consider two time periods:

- (i) **Base Period** [Reference Period]
- (ii) **Current Period.**
 - (i) Base Period is the point of time with which all later changes are compared, *i.e.*, the period with respect to which changes are measured.
 - (ii) Current Period is the period of time in which change is measured considering the prices (or other figures) in the base period as standard.

$$\begin{aligned}\text{Simple Price Index number or price relative} &= \frac{P_1}{P_0} = 100 = I_{01} \\ &= \frac{\text{Current year's price per unit}}{\text{Base year's price per unit}} \times 100\end{aligned}$$

Quantity Index number =

$$\frac{\text{Current year's quantity produced or consumed}}{\text{Base year's quantity produced or consumed}} \times 100$$

Notes:

- Variables (*i.e.*, price, production, consumption, etc.) are supposed to remain constant over the base period.
- Base period and Current period are normally taken as one year, so we call them as Base year and Current year.
- When the comparison is made in respect of prices the index number is called *Index number of prices of price index number or price relative or price index*.
- If $p_1', p_1'', p_1''' \dots$ are the prices per unit of different commodities in the current year and

$p_0', p_0'', p_0''' \dots$ are the prices of the same commodities in the base year and

$q', q'', q''' \dots$ are the respective quantities purchased, then

$$\text{Price Index} = \frac{p_1' q' + p_1'' q'' + p_1''' q''' + \dots}{p_0' q' + p_0'' q'' + p_0''' q''' + \dots} \times 100^*$$

i.e., in price index numbers, quantities are used as weights.

If we can consider only the quantities produced or consumed in the Current year and Base year, the index number is called *quantity index number*. ** However, **True value** (or simply value) for any period is the product of the quantities purchased multiplied by the prices paid per unit during that period. Quantity weights and price weights are the two different categories of weights. The amount of a good created, distributed, or consumed over time is denoted by the quantity weight, or q . The price is per unit. Price and the amount produced, distributed, or consumed are combined to create value weight. Value is represented by the symbols $p q$ = Price per unit Quantity.

* Quantities for Base year and Current year should not change.

** Price per unit is taken as weight in quantity index numbers.

5.3 TYPES OF INDEX NUMBERS

- (i) **Price index numbers:** Whole sale price index number
- (ii) **Quantity index numbers:** Index of industrial production
- (iii) **Value index numbers:** Index of departmental store sales
- (iv) **Special purpose index numbers:** Index of business activity
- (v) **Consumer price index number:** Cost of living index number

(i) Price Index Numbers

(ii) Quantity Index Numbers:

The physical volume of things produced, distributed, or consumed is measured by quantity index numbers, which also allow for comparison. Production (Quantity) index numbers are extremely important as indicators of the level of output in the economy or in certain sectors of it, even though the Price index numbers are more commonly employed.

How much is created and how current level of production compares to earlier levels indicate the economy's direction of movement. Given that it mitigates the effects of inflation and price swings brought on by a competitive market, a quantity index is helpful in production decision-making. Quantity index numbers are often employed for commodities that experience significant price fluctuations. Quantity index numbers are crucial for assessing the standard of living in terms of physical production.

$$\text{Quantity index number} = \frac{\text{Quantity used in the current year}}{\text{Quantity used in the base year}} \times 100 = \frac{q_1}{q_0} \times 100$$

The issues that the statistician faces when creating quantity index numbers are similar to those that are involved with price index numbers. We weigh things and use prices or values as weights when measuring changes in amounts. Industrial production index.

(iii) Value Index Numbers:

Value indices contrast total value (price plus quantity) in a particular year with total value in the base year.

$$\text{Value index number} = \frac{\text{Sum of the current year values}}{\text{Sum of the base year values}} \times 100 = \frac{\sum p_1 q_1}{\sum p_0 q_0} \times 100$$

These index numbers are not very popular.

(iv) Special Purpose Index Numbers:

Index of production-workers employment in manufacturing industries, measurement of commercial and industrial efficiency. Special purpose index numbers include and index number of sales opportunities, which are typically calculated by geographic areas by market research organisations.

(v) Consumer Price Index Numbers

NOTES

5.4 CONSTRUCTION OF INDEX NUMBERS

The many approaches to creating index numbers of prices for various commodities are as follows:

NOTES

- (i) Simple aggregate of prices
- (ii) Weighted aggregate of prices
- (iii) Simple arithmetic mean of price relatives
- (iv) Simple geometric mean of price relatives
- (v) Weighted arithmetic mean of price relatives
- (vi) Weighted geometric mean of price relatives
- (vii) Family Budget method.

(i) Simple Aggregate of Prices

In this method, the price for standard unit for each of the items included in the group is aggregated for the base year as well as for the current year separately and then the total price for the current year is expressed as a percentage of total price for the base year.

If $p_1', p_1'', p_1''', \dots, p_1^{(n)}$ be the price per standard unit of different commodities in the current year and $p_0', p_0'', p_0''', \dots, p_0^{(n)}$ be the corresponding prices in the base year, then price relative is given by

$$I_{01} = \frac{p_1' + p_1'' + p_1''' + \dots + p_1^{(n)}}{p_0' + p_0'' + p_0''' + \dots + p_0^{(n)}} \times 100 = \frac{\sum p_1}{\sum p_0} \times 100$$

$$= \frac{\text{Sum of the unit prices of all the commodities in the current year}}{\text{Sum of the unit prices of all the commodities in the base year}} \times 100$$

(ii) Weighted Aggregate of Prices

In this method, the prices in the current year and the base year are both weighted by *same* quantities. The ratio of the aggregate of the weighted prices for the current year to the aggregate of the weighted prices for the base year expressed as a percentage gives the required index number. **Please note that same quantities should be used for base year as well as current year.**

If $w', w'', \dots, w^{(n)}$ are the respective weights, then the *weighted aggregative* price index is given by

$$I_{01} = \frac{p_1'w' + p_1''w'' + \dots + p_1^{(n)}w^{(n)}}{p_0'w' + p_0''w'' + \dots + p_0^{(n)}w^{(n)}} \times 100 = \frac{\sum p_1w}{\sum p_0w} \times 100$$

$$= \frac{\text{Sum of the values of all the commodities in the current year}}{\text{Sum of the values of all the commodities in the base year}} \times 100$$

where $p_1', p_1'', p_1''', \dots$ and $p_0', p_0'', p_0''', \dots$ have the same meaning as in simple aggregate of prices.

- Notes:**
1. We may use q instead of w .
 2. Same weights should be used in Current Year as well as Base Year.

Index of simple aggregate of prices suffers from the limitations that:

1. It is highly influenced by the high priced items. The utility of an index number derived using the simple aggregate approach of real prices is greatly constrained by this restriction, as one kg of tea may cost many times more than the cost of a kilogramme of potatoes.
2. There is another limitations also that the unit of measurement should be same for all commodities and prices per unit should be for the same unit for all commodities, *e.g.*, if one person uses the unit quintal for one commodity, kg for the other commodity and litre for the third commodity and the other person uses the unit kg for all commodities, then the index number will differ in both the cases.
3. The weighted aggregative index has a drawback in that it requires a constant pattern of consumption, *i.e.*, that the quantities purchased in the base year and non-base year (current year) are the same. It may not pose a very serious problem if both the periods are close together but if the time interval is more, this assumption may be unrealistic and the index may lose much of its usefulness.

But still if choice is to be made between weighted and unweighted price index, weighted price index is always preferred because the use of weights assures that relative importance of the items in the index have been taken into consideration.

(iii) Simple Arithmetic Mean of Price Relatives

$$\begin{aligned}
 I_{01} &= \frac{1}{n} \left[\frac{p_1'}{p_0} \times 100 + \frac{p_1''}{p_0} \times 100 + \frac{p_1'''}{p_0} \times 100 + \dots + \frac{p_1^{(n)}}{p_0} \times 100 \right] \\
 &= \frac{100}{n} \left[\frac{p_1'}{p_0} + \frac{p_1''}{p_0} + \frac{p_1'''}{p_0} + \dots + \frac{p_1^{(n)}}{p_0} \right] \\
 &= \frac{100}{n} \sum \left(\frac{p_1'}{p_0} \right) = \frac{1}{n} \left[\sum \left(\frac{p_1'}{p_0} \right) \times 100 \right] \\
 &= \frac{\text{Sum of the price relatives of all the commodities}}{\text{Total no. of commodities}}
 \end{aligned}$$

$$\text{where } \frac{p_1'}{p_0} \times 100, \frac{p_1''}{p_0} \times 100, \frac{p_1'''}{p_0} \times 100, \dots, \frac{p_1^{(n)}}{p_0} \times 100$$

are the price relatives of items 1, 2, 3, ... respectively and n the number of items or commodities included in the group.

Since we are comparing price per kg with price per kg and price per tonne with price per tonne, this difficulty does not arise when constructing index numbers by aggregate of prices, which was mentioned that the price of all quantities must be expressed in the same units. However, in this method, as well, we discover a hidden assumption that in the base year, wholesalers sell an equal number of rupees' worth of each commodity. This presumption is inconsistent with reality.

NOTES

NOTES

Suppose in 1995 a wholesaler sells three commodities A , B and C worth ₹ 100 each at the rates of ₹ 2, 4 and 5 per unit of A , B and C respectively, then number of units of A , B and C respectively are 50, 25 and 20. Suppose the price per units of A , B and C in 1998 becomes ₹ 2.5, 5 and 10 respectively, then the price index for these three items will be $\frac{2.5}{2} \times 100$, $\frac{5}{4} \times 100$, $\frac{10}{5} \times 100$, i.e., 112.5, 125 and 200. By selling the same number of units in the current year as in the base year, the shopkeeper will get 50×2.5 , 25×5 and 20×10 , i.e., ₹ 112.5, 125 and ₹ 200 which imply that we have embedded the quantities 50, 25 and 20 in the index. In other words, we have unintentionally weighted the data.

(iv) Simple Geometric Mean of Price Relatives

The index number computed by this method is given by

$$I_{01} = \left[\left\{ \frac{p_1'}{p_0'} \times 100 \right\} \left\{ \frac{p_1''}{p_0''} \times 100 \right\} \left\{ \frac{p_1'''}{p_0'''} \times 100 \right\} \dots \left\{ \frac{p_1^{(n)}}{p_0^{(n)}} \times 100 \right\} \right]^{\frac{1}{n}}$$

$$= 100 \sqrt[n]{\frac{p_1'}{p_0'} \times \frac{p_1''}{p_0''} \times \frac{p_1'''}{p_0'''} \times \dots \times \frac{p_1^{(n)}}{p_0^{(n)}}}$$

$$\therefore \log I_{01} = \log 100 + \frac{1}{n} \sum \log \frac{p_1}{p_0}$$

where $\frac{p_1'}{p_0'} \times 100$, $\frac{p_1''}{p_0''} \times 100$, $\frac{p_1'''}{p_0'''} \times 100$, ..., $\frac{p_1^{(n)}}{p_0^{(n)}} \times 100$, are the price relatives of items 1, 2, 3, ..., n respectively.

(v) Weighted Arithmetic Mean of Price Relatives

When the weights w_1, w_2, \dots, w_n are attached to the respective commodities, the index number, calculated as the weighted arithmetic mean of price relatives, is given by

$$I_{01} = \frac{100}{\sum w} \left(\frac{p_1'}{p_0'} w_1 + \frac{p_1''}{p_0''} w_2 + \dots + \frac{p_1^{(n)}}{p_0^{(n)}} w_n \right)$$

$$\text{i.e., } \frac{100}{\sum w} \sum \frac{p_1}{p_0} w = \frac{1}{\sum w} \sum \left(\frac{p_1}{p_0} \times 100 \right) w$$

where $\left(\frac{p_1'}{p_0'} \times 100 \right) w_1$, $\left(\frac{p_1''}{p_0''} \times 100 \right) w_2$, $\left(\frac{p_1'''}{p_0'''} \times 100 \right) w_3$, ..., $\left(\frac{p_1^{(n)}}{p_0^{(n)}} \times 100 \right) w_n$ are the weighted price relatives of items 1, 2, 3, ..., n respectively.

(vi) Weighted Geometric Mean of Price Relatives

The index number calculated by this method is not very much in use. It is given by

$$I_{01} = 100 \left\{ \left(\frac{p_1'}{p_0'} \right)^{w_1} \times \left(\frac{p_1''}{p_0''} \right)^{w_2} \times \dots \times \left(\frac{p_1^{(n)}}{p_0^{(n)}} \right)^{w_n} \right\}^{\frac{1}{\sum w}}$$

where w_1, w_2, \dots, w_n are the weights attached to the respective commodities and $\Sigma W = w_1 + w_2 + w_3 + \dots + w_n$

[If $I_1, I_2, I_3, \dots, I_n$ are the price relatives with $w_1, w_2, w_3, \dots, w_n$ as respective weights, then

$$I_{01} = \left[(I_1)^{w_1} (I_2)^{w_2} \times (I_3)^{w_3} \dots (I_n)^{w_n} \right]^{\frac{1}{\Sigma w}}$$

$$\log I_{01} = \frac{1}{\Sigma w} [w_1 \log I_1 + w_2 \log I_2 + w_3 \log I_3 + \dots + w_n \log I_n]$$

$$= \frac{1}{\Sigma w} [\Sigma w \log I]$$

where $I_I = \frac{P_1'}{P_0'} \times 100, I_2 = \frac{P_2''}{P_0''} \times 100 \dots]$

(vii) Family Budget Method of Constructing Consumer Price Index Numbers

Current year's index number or consumer price index number or cost of living index number

$$= \frac{\text{Sum of the products of price relatives and base year values}}{\text{Sum of base year values}} = \frac{\Sigma IV}{\Sigma V}$$

where $I = \frac{\text{Current year price}}{\text{Base year price}} \times \frac{p_1}{p_0} \times 100$ and $V = p_0 q_0$

I and V are calculated for different commodities.

5.5 METHODS OF CONSTRUCTION*

A weighted aggregate price index number called Laspeyre's index number uses base year values as weights.

If $p_1', p_1'', p_1''' \dots p_1^{(n)}$ denote current year prices per unit for different commodities and $p_0', p_0'', p_0''', \dots p_0^{(n)}$ denote corresponding per unit base year prices and $q_0', q_0'', q_0''', \dots, q_0^{(n)}$ be respective quantities for base year, then Laspeyre's price index number

* Quantity index numbers are obtained by interchanging p and q .

If the quantity index number is prepared by using Laspeyre's method, the base year prices will be used as weights.

$$Q_{01} = \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times 100$$

In Paasche's method current year prices are used as weights

$$Q_{01} = \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1} \times 100$$

Fisher's ideal quantity index number

$$= Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}} \times 100$$

NOTES

NOTES

$$(I_{01}) = \frac{p_1' q_0' + p_1'' q_0'' + p_1''' q_0''' + \dots + p_1^{(n)} q_0^{(n)}}{p_0' q_0' + p_0'' q_0'' + p_0''' q_0''' + \dots + p_0^{(n)} q_0^{(n)}} \times 100 = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

$$= \frac{\text{Sum of the amounts paid for base year quantities at current year prices}}{\text{Sum of the amounts paid for base year quantities at base year prices}} \times 100$$

If the quantity index number is prepared by using Laspeyre's method, the base year prices will be used as weights

$$Q_{01} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$$

Paasche's Price Index Number

A weighted aggregate price index number, Paasche's index number uses current year quantities as weights.

If $p_1', p_1'', p_1''', \dots, p_1^{(n)}$ denote current year prices per unit for different commodities and $p_0', p_0'', p_0''' \dots p_0^{(n)}$ denote corresponding per unit base year, prices and $q_1', q_1'', q_1''', \dots, q_1^{(n)}$ be the respective quantities for current year then Paasche's price index number

$$(I_{01}) = \frac{p_1' q_1' + p_1'' q_1'' + p_1''' q_1''' + \dots + p_1^{(n)} q_1^{(n)}}{p_0' q_1' + p_0'' q_1'' + p_0''' q_1''' + \dots + p_0^{(n)} q_1^{(n)}} \times 100 = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

$$= \frac{\text{Sum of the amounts paid for Current Year quantities at Current Year Prices}}{\text{Sum of the amounts paid for Current Year quantities at Base Year Prices}} \times 100$$

In quantity index numbers by Paasche's method current year prices are used as weights.

$$Q_{01} = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100$$

Fisher's Ideal Index Number

The Fisher's Ideal Index number, named after Irving Fisher, is the geometric mean of Laspeyre's Index number and Paasche's Index number.

Fisher's Ideal Index Number (I_{01})

$$= \sqrt{\text{Laspeyre's Index number} \times \text{Paasche's Index number}}$$

$$= 100 \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1}}$$

Fisher's ideal quantity index number

$$= Q_1' = 100 \times \sqrt{\frac{q_1 p_0}{q_0 p_0} \times \frac{q_1 p_1}{q_0 p_1}}$$

5.6 COMPARISON BETWEEN LASPEYRE'S INDEX NUMBER AND PAASCHE'S INDEX NUMBER

Laspeyre's index number employs base year quantities as weights which do not change from year to year. But with the passage of time conditions may change and the relative importance of the different items in the index numbers may also change. Keeping in view this aspect it would be advisable to use Paasche's index number where we use current year quantities as weights. But due to changing conditions weights may change in every period and to obtain new quantities (weights) for each period becomes quite expensive and difficult. In light of this, Paasche's index number is frequently preferred over Laspeyre's index number from a practical standpoint. Because Paasche's price index number employs base year values as weights, Laspeyre's price index number typically outperforms it and in times of rising prices, consumption is shifted to cheaper substitutes but in Laspeyre's index numbers weights of expensive as well as cheaper items are not changed so that expensive items retain their higher weights and cheaper items retain their lower weights resulting in an *upward bias*. On the other hand, the contrary of this happens in case of Paasche's method and the index has a *downward bias*. This however does not mean that Laspeyre's index number would always be more than the Paasche's index number.

The Laspeyre's price index overestimates the damage done to the consumer by the price changes, *since it neglects the consumer's ability to react to the changes by altering the amounts of the products he buys*. By a similar argument, the Paasche's index underestimates the value of price index.

However if the change from the base period to the current period is not substantial both the indices give satisfactory results. Under such circumstances, Laspeyre's method becomes more popular as in this index weights are constant and therefore the construction of the index is easier. It should also be noted that these two index values will yield the same result if the prices of all commodities vary in the same proportion, in which case the weighing system is meaningless. Additionally, as the two weighting systems are equal, both indices would provide the same value of the index if the quantities of all commodities changed in the same proportion.

Fisher's index number is called ideal – Why?

Fisher's index number is called ideal because of the following reasons:

- (a) Geometric mean is useful in averaging percentages and ratios. Index numbers indicate percentage changes and the Geometric mean of Laspeyre's and Paasche's index numbers is known as Fisher's ideal index number.*
- (b) It takes into account both Current Year as well as Base Year quantities.

* Theoretically, Geometrical mean is considered to be the best average in the construction of index numbers but due to certain problems in calculations, Arithmetic mean is considered for all practical purposes.

NOTES

NOTES

- (c) It satisfies certain mathematical tests such as time reversal and factor reversal tests.
- (d) It is free from bias upwards as well as downwards (Please note that Paasche's index numbers have a downward bias, whereas Laspeyre's index numbers have an upward tilt.)

5.7 SELECTION OF AN AVERAGE FOR THE CONSTRUCTION OF INDEX NUMBERS

The optimal average for the production of index numbers is G.M., according to theory. But due to certain difficulties in calculating G.M., practically A.M. is used for constructing index numbers.

5.8 PROBLEMS IN THE CONSTRUCTION OF INDEX NUMBERS

The following problems are mainly faced in the construction of index numbers:

- (i) Definition of the Purpose,
- (ii) Selection of the Base Period,
- (iii) Selection of Items,
- (iv) Selection of Sources of Data and Collection of Data,
- (v) Selection of Average,
- (vi) System of Weighting.

(i) Definition of the Purpose

There are no all purpose index numbers. Therefore, before constructing an index number the specific purpose, *i.e.*, objective for which it is designed must be clearly and rigorously defined. Haberler has rightly said. "Different index number are constructed to fulfill different objectives and before setting to construct a particular index number one must clearly define one's objective of study because it is on the objective of the study that the nature and format of the index number depends".

(ii) Selection of the Base Period

In the creation of index numbers, choosing the appropriate Base Period is crucial. Base Period is a reference point with which changes in other periods are measured. About the selection of Base Period, following observations may be noted.

Morris Hamburg observers, "*It is desirable that the Base Period be not too far away in times from the present. The further away we move from the Base Period, the dimmer are our recollections of Economic conditions prevailing at that time. Consequently comparisons with these remote periods tend to Lose significance and to become rather tenuous in meaning.*"

According to **George Simpson** and **Fritz Kafka**, “Since practical decisions are made in terms of index numbers, and economic practices so often are a matter of the short run, we wish to make comparisons between a base which lies in the same general economic framework as the years of immediate interest. Therefore, we choose a base relatively close to the years being studied.” The base year shouldn't be older than ten years in the modern, rapidly changing world. There is a psychological reason also for taking a recent period as base. **T.W. Lewis** and **R.A. Fox** write, *A fairly recent base is advantageous simply because it keeps the values of the index fairly close to the 100 marks and thereby helps comprehension of changes in the index. If, for example, the base year is too far in the past, current values might be around the 400, 500 or 600 level. If two consecutive values of the index are, say, 573 and 604, further calculation is necessary to appreciate the size of the increase; furthermore there is a psychological effect. If a price index stands at 604 this fact tends to give people the impression that prices are high. A fairly recent base also ensures that comparisons are based on relatively homogeneous quantities ...*

NOTES

(iii) Selection of Items

Selection of items or ‘Regimen’ or ‘Basket’. * It is neither possible nor required to include all the goods or commodities in any index number. Each index number tries to measure changes pertaining to a particular group.

Selection of items also depends on the purpose of index number. Moreover, items selected should be such as are widely consumed. The things chosen for an index number should be accurate, similar, relevant, and representational. The likelihood of an average inaccuracy will often decrease as the number of items increases.

Since we are limited in the quantity of items we may include. Always strike a balance between the quantity of things and the acceptable level of accuracy. However, we must aim for a manageable amount of things and a decent quality of accuracy.

(iv) Selection of Sources of Data and Collection of Data

For sources of data and collection of data, we are mainly concerned with the prices. Use of wholesale prices or retail prices depends on the objective of study. Price quotations should be obtained from important markets. In order to ensure better results, it is advisable to take a standard price which implies representative price of a commodity for whole interval under consideration.

-
- * (a) Geometric mean is the best measure for measuring the relative changes.
- (b) Geometric mean gives more importance to small items and less importance to bigger items. Thus, it helps in the evaluation of the real position.
- (c) Good index numbers possess the characteristic of reversibility and that can be had if geometric mean is used.

NOTES

(v) Selection of Average

Since index numbers measure the relative changes, Geometric mean* should be the best average but due to certain difficulties in calculations with G.M., for all practical purposes Arithmetic mean is used.

(vi) System of Weighting

According to **John I. Giffin**, To put it simply, weighting is intended to give component series value that is appropriate for their actual significance. It is recommended to apply a proper weighting scheme so that each commodity can have a reasonable impact on the index. When prices are indexed without weight, each commodity is assigned the same weight. However, in actuality, various commodities require varying levels of priority. In an unweighted index number, if price of wheat is doubled and that of tobacco is halved, there may or may not be any change in the index number.

But increase in the price of wheat will affect adversely than the benefit which will accrue to the consumer by the reduction in tobacco price. That is the main reason for assigning weights to the various items while constructing index numbers.

The weights may be according to:

1. The value or quantity produced.
2. The value or quantity consumed.
3. The value or quality sold.

For general index number the weights may be according to production and for cost of living index number weights may be according to consumptions. As per the need weights may be selected by Paasche's method, Laspeyre's method or Fisher's method.

5.9 ADVANTAGES OR USES OF INDEX NUMBERS

Some of the main advantages of index numbers are:

1. **Measuring Changes in Price Level or Money Worth:** One of the most significant uses of index numbers is to track changes in the value of money over time. Index numbers can be used to determine how changing money values affect various societal segments, which aids in the development of corrective actions for inflationary or deflationary conditions.
2. **Understanding the Change in Standard of Living:** Index figures are useful in identifying changes in people's living standards. Even though people's salaries in money may rise, if index numbers indicate a reduction in the value of money, living standards may fall despite higher incomes in money. Consequently, index numbers reflect changes in real income.
3. **Adjustments in Salaries and Allowances:** The government and private businesses can use the cost of living index as a useful tool to determine what changes should be made to employee salaries and benefits in order to help them maintain their quality of living. The cost of living index rising means that employee compensation and benefits would also rise. If the salaries and

allowances are increased in equal proportion to the cost of living index, the workers can maintain their original standard of living.

4. **Useful to Business Community:** When analysing the state of the economy, the business community can use price index figures as a useful tool and take useful decisions, with regard to their production keeping in view the trend in prices.
5. **Information regarding Production:** Index numbers of production show whether the level of agricultural and industrial production in the economy is increasing or decreasing.
6. **Information regarding Foreign Trade:** Information on international trade is usefully provided by the exports and imports index. Accordingly, export-import policies are formulated.
7. **Useful to Government:** The government formulates its monetary and fiscal policies and implements specific measures for the nation's economic development with the use of index numbers. In other words, with the help of index numbers Government formulates appropriate policies and other programs to increase investment, output, income, employment, trade, price level, consumption, etc.
8. **Useful to Politicians:** The analysis of index numbers helps the politicians to know about the actual economic condition of the country. Accordingly, they offer constructive criticism to the Government and make useful suggestions for improvement.
9. **Useful in Contract Escalation:** The study of index numbers (*i.e.*, past behaviour or trends) help in adjusting prices in contracts extending over long periods of time.
10. The balance between overall supply and demand at the main market level is represented by index numbers.
11. Index figures are used to determine production costs, plan plant building and production schedules, buy raw materials, assess inventories, and plan general investment programmes.
12. Based on previous price developments, index numbers are used to predict future pricing, marketing, and sales policies.
13. Index numbers are also used to analyse the changing interrelationships among individual commodities.

5.10 LIMITATIONS OF INDEX NUMBERS

With regard to the limitations of index numbers, **Coulbourn** has rightly said, *“In this changing world it is difficult to escape from the theoretical defects and in future, as far as we can see, it will not be possible, from theoretical point of view, to make use of the best method of constructing the index number.”*

NOTES

NOTES

Main limitations are:

1. **Do not Cover all Commodities:** Index numbers track variations in the price of specified, price-defined limited commodities. It does not represent an all-encompassing indicator of the "general price level" or the "purchasing power of money." Not all commodity classes that influence the overall price level are included.
2. **Retail Transactions not Covered:** Because the majority of index figures are built using wholesale prices, they do not measure price changes in retail transactions. Only retail prices are used in real life. Therefore, the index numbers are generally misleading.
3. **International Comparison is not Possible:** There is no internationally accepted base to be used by all countries for preparing index numbers. Therefore, different countries prepare index numbers on different base periods and the comparison at the international level is not possible.
4. **Effect of Time:** With the change in time, general habits of the people, tastes, purchasing capacities, etc. may change. Therefore, the index numbers constructed on old consumption pattern may not give true picture at all times.
5. **Limited Use:** There are always preset index numbers for a certain goal. It's possible that the index numbers created for one purpose won't work for another. Index numbers prepared to know about the economic condition of one class of people may not represent another class of people.
6. **Unscientific Weightage:** There is no scientific method of according weights to most of the index numbers. Consequently, index numbers do not offer precise conclusions.

5.11 SPLICING OF INDEX NUMBERS

Index number splicing is the process of combining two or more series of overlapping index numbers to produce a single index number on a shared base. Splicing, then, is a statistical technique for making a series continuous by joining an old index number to a revised series. Only when the index numbers are made up of the same components and have the same year can they be joined together.

When an old index number with an old base is being terminated and a new index with a new base is being begun, splicing is typically done.

Spliced index number =

$$\frac{\text{Current year's new index number} \times \text{New base year's old index number}}{100}$$

SOLVED EXAMPLES

Example 1. A company has three major items that it produces, say items *A*, *B* and *C*. The table below gives prices and units sold in each of three years, *i.e.*, 1991, 1992, 1993.

Item	1991		1992		1993	
	Price	Qty.	Price	Qty.	Price	Qty.
A	0.50	30	0.75	30	1.00	35
B	2.00	25	4.00	20	3.00	25
C	3.00	20	3.50	25	3.50	20

Use 1991 as the base and answer the following:

- The unweighted aggregate price index for 1993.
- The unweighted aggregate quantity index for 1993.
- The weighted aggregate index for 1993, using Laspeyre's Method to calculate weights.
- The weighted aggregate index for 1993, using Paasche's Method to calculate weights.
- The weighted average of relatives price index for 1993, using 1991 values as weights.
- The unweighted average for relatives quantity index for 1993.

Solution.

Item	1991		1993		P_0Q_0	P_1Q_0	P_1Q_1	P_0Q_1
	Price	Qty.	Price	Qty.				
	P_0	Q_0	P_1	Q_1				
A	0.50	30	1.00	35	15	30	35	17.50
B	2.00	25	3.00	25	50	75	75.00	50.00
C	3.00	20	3.50	20	60	70	80.00	60.00
	$\Sigma P_0 = 5.50 \quad \Sigma Q_0 = 75$		$\Sigma P_1 = 7.50 \quad \Sigma Q_1 = 80$		$\Sigma P_0Q_0 = 125$	$\Sigma P_1Q_0 = 175$	$\Sigma P_1Q_1 = 180$	$\Sigma P_0Q_1 = 127.50$

- Unweighted aggregate price index for 1993 = $\frac{\Sigma P_1}{\Sigma P_0} \times 100 = \frac{7.50}{5.50} \times 100 = 136$
- Unweighted aggregate quantity index for 1993 = $\frac{\Sigma Q_1}{\Sigma Q_0} \times 100 = \frac{80}{75} \times 100 = 107$
- Weighted aggregate index for 1993 using Laspeyre's Method

$$= \frac{\Sigma P_1Q_0}{\Sigma P_0Q_0} \times 100 = \frac{175}{125} \times 100 = 140$$

NOTES

NOTES

(d) Weighted aggregate index for 1993 using Paasche's Method

$$= \frac{\sum P_1 Q_1}{\sum P_0 Q_1} \times 100 = \frac{180}{127.50} \times 100 = 141$$

(e) Weighted average of relative price indexes for 1993 using 1991 values as weights

$$= \frac{\left[\sum \frac{P_1}{P_0} \times P_0 Q_0 \right] \times 100}{\sum P_0 Q_0} = \frac{\left[\frac{1.00}{0.50} \times 15 + \frac{3.00}{2.00} \times 50 + \frac{3.50}{3.00} \times 60 \right] \times 100}{(15 + 50 + 60)}$$

$$= \frac{17,500}{125} \times 100 = 140$$

(f) Unweighted average for relative quantity indexes for 1993 = $\frac{\sum \frac{Q_1}{Q_0}}{3} \times 100$.

$$= \frac{\left[\frac{35}{30} + \frac{25}{25} + \frac{20}{20} \right]}{3} \times 100 = \frac{317}{3} = 106$$

Example 2. The cost of living index uses the following weights:

Food: 40, Rent: 15 Clothing: 20, Fuel: 10, Miscellaneous: 15.

During the period 1991-2001, the cost of living index raised from 100 to 205.72. Rent increased by 60% within the same time period, clothing increased by 180%, fuel increased by 75%, and other items increased by 165%. How much did the price of food change in percentage terms?

Solution.

Let the percentage change in food price = x .

Average percentage price change for the whole groups

$$= \frac{\sum IW}{\sum W} - 100$$

$$= \frac{40x + 60 \times 15 + 180 \times 20 + 75 \times 10 + 165 \times 15}{40 + 15 + 20 + 10 + 15} - 100 = 205.72 - 100.$$

(Please note)

$$\Rightarrow \frac{40x + 7725}{100} = 105.72 \Rightarrow x = 71.18.$$

Example 3. The cost of living index rises from 110 to 200 over a specific time period, while the worker's pay increases from ₹ 3,250 to ₹ 5,000. Does the employee actually benefit, and if yes, how much in actual dollars?

Solution.

The compensate the worker fully, his salary after increase in price index should be

$$= 3,250 \times \frac{200}{110} = ₹ 5,909$$

∴ If the salary is raised to ₹ 5,000, the worker will not gain in real terms but will rather suffer a loss and will not be fully compensated. To compensate the worker fully his pay should be raised to 5909.

Example 4. The index number for various groups and their corresponding weights are shown in the following table for 2000 (base year = 1995):

Group	Group Index Number	Group Weight
Food	130	60
Clothing	280	5
Lighting and fuel	190	7
Rent	300	9
Miscellaneous	200	19

- Determine the total cost of living index value for 2000.
- Assuming a person made ₹ 1,500 per month in 1995, what should that person make in 2000 in order for his level of living to be the same as it was in 1995?

Solution.

- Overall cost of living index in 2000

$$= \frac{130 \times 60 + 280 \times 5 + 190 \times 7 + 300 \times 9 + 200 \times 19}{60 + 5 + 7 + 9 + 19}$$

$$= \frac{5,800 = 1,400 + 1,330 = 2,700 + 3,800}{100} = \frac{17,030}{100} = 170.3$$

- The salary of the man in 2000 should be = $\frac{170.3}{100} \times 1500 = 2554.5$

Example 5. The following details on employment in an industrial town are provided:

Item of Consumption	Consumer Price Index 2005 (2000 = 100)	Proportion of Expenditure on Item
Food	132	60%
Clothing	154	12%
Fuel and lighting	147	16%
Housing	178	8%
Miscellaneous	158	4%

In 2000, the average monthly salary was \$2,000. What percentage of wages should be designated for dearness allowance? What should the town's average monthly wage be for each worker be in order to prevent the worker's standard of life from dropping below that of 2000?

Solution.

Weighted average of price relatives

$$= \frac{132 \times 60 + 154 \times 12 + 147 \times 16 + 178 \times 8 + 158 \times 4}{100}$$

NOTES

NOTES

$$= \frac{7920 + 1848 + 2352 + 1424 + 632}{100} = 141.76$$

The average wage in 2005 should be $2000 \times \frac{141.76}{100} = 2835.2$

$$\text{Dearness allowance} = \frac{2835.2 - 2000}{2000} \times 100 = 41.76\%$$

Example 6. An investigation into the spending habits of middle-class families in a particular city found that, on average, the percentage costs for the various categories were as follows: food — 45; rent — 15; clothing — 12; fuel and light — 8; and miscellaneous — 20.

In comparison to a fixed base period, the group index numbers for the current year were 410, 150, 343, 248 and 285 respectively. Determine the current year's cost of living index number.

In the base year, Mr. X received ₹ 2,400, and in the current year, he received ₹ 4,300. Indicate how much more he should have gotten as a bonus allowance to maintain his previous standard of living.

Hint: Cost of living index for the current year = $\frac{\sum IW}{\sum W}$

$$= \frac{410 \times 45 + 150 \times 15 + 343 \times 12 + 248 \times 8 + 285 \times 20}{45 + 15 + 12 + 8 + 20}$$

$$= \frac{18,450 + 2,250 + 4,116 + 1,984 + 5,700}{100} = 325$$

i.e., if Mr. X is to maintain his base year standard of living his salary in the current year should be

$$= \frac{325}{100} \times 2400 = ₹ 7,800$$

Extra allowance required to be paid = ₹ (7,800 – 4,300) = ₹ 3,500.

Example 7. Below is incomplete information on the cost of living index that was gleaned from a partially damaged record:

Group	Group Index	% of Total Expenditure
Food	134	60
Clothing	140	Not available
Housing	105	20
Fuel and electricity	120	5
Miscellaneous	130	Not available

The cost of living index was determined to be 127.9 with % of total expenditure as weight. Calculate the weights that are utilised for clothing and other items.

Solution.

Let the % total expenditure for clothing = x

Let the % total expenditure for Misc. = $y = (100 - 60 - 20 - 5 - x) = 15 - x$

$$\Rightarrow 127.9 = \frac{134 \times 60 + 140 \times x + 105 \times 20 + 120 \times 5 + 130 \times (15 - x)}{100}$$

$$12,790 = 8,040 + 2,100 + 600 + 1,950 + 10x = 12,690 + 10x$$

$$10x = 100 \Rightarrow x = 10$$

$$\therefore y = 15 - x = 15 - 10 = 5.$$

Example 8. The following table provides the cost of living index for various commodity groups for the year 1998, using 1986 as a baseline: Food, clothing, fuel and light, rent, and miscellaneous, which are each listed as 440, 500, 350, 400, and 250, respectively, with their weights arranged in the ratio of 15: 1: 2: 3: 4.

Get the overall cost of living index number and enter ₹ 4,000 as the person's income in 1986.

What should be his salary in 1998 to maintain the standard of living as in 1986?

Solution.

Let the weights of the groups, Food, Clothing, Fuel and Light, Rent and Miscellaneous be $15a, a, 2a, 3a$ and $4a$ ($a > 0$) respectively.

Overall index for 1998 with base as 1986.

$$\frac{[400 \times 15 + 500 \times 1 + 350 \times 2 + 400 \times 3 + 250 \times 4]}{[15 + 1 + 2 + 3 + 4]} = \frac{66,000 + 500 + 700 + 1,200}{25} = 400$$

$$\text{Salary in 1998 should be } 4,000 \times \frac{400}{100} = ₹ 16,000.$$

Example 9. When the price of cigarettes rose by 50%, a smoker who continued to smoke at the same rate claimed that the price increase had raised his expenses by 5%. How much of his cost of living was attributable to purchasing cigarettes before the price change?

Solution.

Suppose the smoker spends $x\%$ on cigarette before the change of price, then his expenditure on cigarette after change of price $= \frac{3}{2} x\%$

$$\text{Change in expenditure} = \frac{3}{2}x - x = \frac{1}{2}x = 5 \Rightarrow x = 10$$

i.e., 10% of his cost of living was due to buying cigarettes before the change in price.

NOTES

NOTES

This question can also be set in the following form:

When an item's price rose by 50%, a guy who maintained his previous level of consumption said that the increase in the item's price had raised his cost of living by 5%. How much of his cost of living was attributable to purchasing the item before the price increase?

Example 10. In 1990, the average employee's net weekly salary was 800. In 1998, the consumer price index stood at 160. 1998 had a 200 increase. Determine the additional D.A. (dearness allowance) that will be given to the worker in the event that he needs to be fairly rewarded.

Solution.

$$\frac{\text{Price in 1998}}{\text{Price in 1990}} = \frac{\text{Price index in 1998}}{\text{Price index in 1990}} = \frac{200}{160} = \frac{5}{4}$$

$$\text{i.e., Price in 1998} = \frac{5}{4} \times \text{Price in 1990}$$

The employee will be rightly compensated if his pay is also increased in the same ratio in which the price has increased.

$$\text{The weekly pay of the employee in 1998 should be } 800 \times \frac{5}{4} = ₹ 1,000$$

$$\text{Additional dearness allow per week required to be paid} = 1,000 - 800 = ₹ 200.$$

Example 11. The following table lists a group of commodities' weights and price relationships.

Commodity	A	B	C	D
Price relative	125	120	127	119
Weight	W_1	$2W_1$	W_2	$W_2 + 3$

Find the numerical values of W_1 and W_2 if the set's index is 122 and the sum of the weights is 40.

Solution.

$$\text{Index number} = \frac{\sum \left(\frac{P_1}{P_0} \times 100 \right) W}{\sum W}$$

$$122 = \frac{125 \times W_1 + 120 \times 2W_1 + 127 \times W_2 + 119 \times (W_2 + 3)}{W_1 + 2W_1 + W_2 + W_2 + 3} *$$

$$\Rightarrow 366W_1 + 244W_2 + 366 = 365W_1 + 246W_2 + 357 \quad \dots (I)$$

* 1. Can also be written as $122 = \frac{125w_1 + 120 \times 2w_1 + 127w_2 + 119(w_2 + 3)}{40}$

$$W_1 = 2W_2 - 9 \quad \dots (II)$$

$$W_1 + 2W_1 + W_2 + W_2 + 3 = 40 \quad [\because \text{Sum of the weights} = 40]$$

$$\Rightarrow 3W_1 + 2W_2 + 3 = 40$$

$$2W_2 = 37 - 3W_1 \quad \dots (III)$$

Substituting (II) in (III), we get

$$W_1 = 37 - 3W_1 - 9 \Rightarrow 28 - 4W_1 \Rightarrow W_1 = 7$$

$$2W_2 = 37 - 3 \times 7 = 16 \Rightarrow W_2 = 8$$

$$W_1 = 7, W_2 = 8.$$

Example 12. Find the missing number in the following table if the ratio between Laspeyre's index number and Paasche's index number is 28: 27:

Commodity	Base Year		Current Year	
	Price	Quantity	Price	Quantity
X	1	10	2	5
Y	1	5	—	2

Solution.

$$\text{Laspeyre's index number} = \frac{\sum p_1 q_0}{\sum p_0 q_0}$$

$$\text{Paasche's index number} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

Let the price of commodity Y in the current year be ₹ x.

Commodity	Base year		Current year		$p_0 q_0$	$p_0 q_1$	$p_1 q_0$	$p_1 q_1$
	Price	Qty.	Price	Qty.				
X	1	10	2	5	20	10	10	5
Y	1	5	x	2	5x	5	2x	2

$$\sum p_1 q_0 = 20 + 5x, \sum p_0 q_0 = 25, \sum p_1 q_1 = 10 + 2x, \sum p_0 q_1 = 7$$

As per the question,

$$\frac{20 + 5x}{25} \times \frac{7}{10 + 2x} = \frac{28}{27} \Rightarrow \frac{4x}{10 + 2x} = \frac{4}{9}$$

$$\Rightarrow 36 + 9x = 40 + 8x \Rightarrow x = 4$$

Example 13. Utilizing the following information, calculate the consumer price index number for 2000 using the data for 1999 using:

- The family budget approach
- The method of aggregating expenditures

NOTES

NOTES

Commodity	Rice	Wheat	Pulses	Ghee	Oil
Weights	40	20	15	20	5
Price (per unit) 1999 (₹)	16.00	40.00	0.50	5.12	2.00
Price (per unit) 2000 (₹)	20.00	60.00	0.50	6.25	1.50

Solution.

Constructing Consumer Price Index Number:

(i) Family Budget Method

Commodity	Weights (W)	Price (₹) (per unit) 1999 (p_0)	Price (₹) (per unit) 2000 (p_1)	Price Relative $\frac{p_1}{p_0} \times 100$ (I)	Weighted Relatives (IW)
Rice	40	16.00	20.00	125	5,000
Wheat	20	40.00	60.00	150	3,000
Pulses	15	0.50	0.50	100	1,500
Ghee	20	5.12	6.25	122	2,440
Oil	5	2.00	1.50	75	375
	$\Sigma W = 100$				$\Sigma IW = 12,315$

$$\text{Cost of Living Index for 2000} = \frac{\Sigma IW}{\Sigma W} = \frac{12,315}{100} = 123.15$$

Thus, there is an increase of 23.15% in prices of 2000 with that of 1999.

(ii) Aggregative Expenditure Method

Commodity	Weights	Price (₹) 1999	Price (₹) 2000	$p_0 q_0$	$p_1 q_0$
	q_0	p_0	p_1		
Rice	40	16.00	20.00	640.00	800.00
Wheat	20	40.00	60.00	800.00	1200.00
Pulses	15	0.50	0.50	7.50	7.50
Ghee	20	5.12	6.25	102.40	125.00
Oil	5	2.00	1.50	10.00	7.50
				$\Sigma p_0 q_0$ = 1,559.90	$\Sigma p_1 q_0$ = 2,140.00

$$\text{Consumer Price Index for 2000} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

$$= \frac{2140}{1559.9} \times 100 = 137.188$$

Thus, there is increase of 37.188% in prices of 2000 with that of 1999.

Example 14. Determine Fisher's price index using the following information and determine whether the time-reverse test is met:

Commodity	Base Year (1990)		Current Year (1995)	
	Price	Quantity	Price	Quantity
	(₹)	(Kg.)	(₹)	(Kg.)
A	32	50	30	50
B	30	35	25	40
C	16	55	18	50

Solution.

Let p_0 and q_0 stand for the price and quantity for the base year 1990 and p_1 and q_1 for the current year 1995, respectively.

Commodity	Base Year				Current Year			
	p_0	q_0	p_1	q_1	p_0q_0	p_1q_0	p_1q_1	p_0q_1
A	32	50	30	50	1,600	1,500	1,500	1,600
B	30	35	25	40	1,050	875	1,000	1,200
C	16	55	18	50	880	990	900	800
					Σp_0q_0 = 3,530	Σp_1q_0 = 3,365	Σp_1q_1 = 3,400	Σp_0q_1 = 3,600

$$\text{Fisher's ideal index } (I_{0,1}) = \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \times 100$$

$$= \sqrt{\frac{3,365}{3,530} \times \frac{3,400}{3,600}} \times 100 = 90$$

$$\text{By omitting the factor 100, } (I_{0,1}) = \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} = 0.9$$

and

$$I_{1,0} = \sqrt{\frac{\Sigma p_0q_1}{\Sigma p_1q_1} \times \frac{\Sigma p_0q_0}{\Sigma p_1q_0}} = \sqrt{\frac{3600 \times 3530}{3400 \times 3365}}$$

$$(I_{0,1}) \times (I_{1,0}) = 1$$

which verifies time reversal test.

Example 15. Using the information provided about the following four commodities, calculate the Laspeyre's and Paasche's price index numbers for 1997:

NOTES

NOTES

		Commodity			
		A	B	C	D
Quantity (kg.)	in 1992	8	10	15	20
	in 1997	6	5	10	15
Price per kg.	in 1992	20	50	40	20
	in 1997	40	60	50	20

Solution.

Commodity	1992				1997			
	Price	Qty.	Price	Qty.	p_0q_0	p_1q_0	p_1q_1	p_0q_1
	p_0	q_0	p_1	q_1				
A	20	8	40	6	160	320	240	120
B	50	10	60	5	500	600	300	250
C	40	15	50	10	600	750	500	400
D	20	20	20	15	400	400	300	300
					Σp_0q_0 =1,660	Σp_1q_0 =2,070	Σp_1q_1 =1,340	Σp_0q_1 =1,070

$$\text{Laspeyre's index number} = \frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times 100 = \frac{2,070}{1,660} \times 100 = 124.69$$

$$\text{Paasche's index number} = \frac{\Sigma p_1q_1}{\Sigma p_0q_1} \times 100 = \frac{1,340}{1,070} \times 100 = 125.23$$

Example 16. “For constructing index numbers, the best method on theoretical ground is not best method from practical point of view.” Discuss.

Solution.

The optimum approach for creating index numbers is the geometric mean, according to theory, because:

- It does not give larger objects more weight.
- Index numbers quantify relative changes and for relative change G.M. is preferred.

But due to computational difficulties, Arithmetic mean is preferred in practical calculation.

Example 17. Use the following information to get the index number for 1998 using 1990 as the base year and the weighted average of price relative method:

Commodity	Price		
	Weight	1990	1998
A	2	12	24
B	8	8	12
C	4	15	27
D	5	6	18
E	1	10	12

NOTES

Solution.

Commodity	Weight w	1990 Price p_0	1998 Price p_1	$p = \frac{p_1}{p_0} \times 100$	P_w
A	2	12	24	$\frac{24}{12} \times 100 = 200$	400
B	8	8	12	$\frac{12}{8} \times 100 = 150$	1200
C	4	15	27	$\frac{27}{15} \times 100 = 180$	720
D	5	6	18	$\frac{18}{6} \times 100 = 300$	1500
E	1	10	12	$\frac{12}{10} \times 100 = 120$	120
	$\Sigma w = 20$				$\Sigma p_w = 3940$

$$P_{01} = \frac{\Sigma P_w}{\Sigma w} = \frac{3940}{20} = 197$$

Example 18. In 1998, a corporation spent ₹ 50, ₹ 48, ₹ 18, and ₹ 42. In 1999, the corporation boosted its spending on four commodities to ₹ 100, ₹ 98, ₹ 60, and ₹ 102, respectively. Calculate the price index for 1999 using the most appropriate technique in the case that the units of four commodities purchased in 1998 and 1999 are the same, i.e., 5, 2, 6 and 17.

Solution.

Most suitable method here will be the method of weighted aggregate of prices.

∴ Price index for 1999 with 1998 as base

$$\begin{aligned}
 &= \frac{\text{Weighted sum of the prices in 1999}}{\text{Weighted sum of the prices in 1998}} \times 100 = \frac{100 + 98 + 60 + 102}{50 + 48 + 18 + 42} \times 100 \\
 &= \frac{360}{158} \times 100 = 227.84
 \end{aligned}$$

NOTES

Example 19. The following weights were applied in order to determine a certain cost of living index number: 15 for food, 3 for clothing, 4 for rent, 2 for fuel and light, and 1 for miscellaneous. Determine the index for a point in time when the average annual percentage increases in prices for the four groupings of goods over the base period were 32, 54, 48, 78, and 58, respectively.

What should a company executive make now if his level of living is to remain the same as it was in the base period, which was \$20,500?

Solution.

Calculation of Cost of Living Index

Group	Average % Increase in Price	Group Index I^*	Weight W	IW
Food	32	132	15	1980
Clothing	54	154	3	462
Rent	47	147	4	588
Fuel and Light	78	178	2	356
Miscellaneous	58	158	1	158
			$\Sigma W = 25$	$\Sigma IW = 3544$

$$\text{Cost of living index} = \frac{\Sigma IW}{\Sigma W} = \frac{3544}{25} = 141.76$$

For maintaining the same standard, the business executive should get

$$\frac{20,500 \times 141.76}{100} = ₹ 29060.8$$

Example 20. Calculate price index numbers from the following data by:

- (i) Laspeyre's Method
- (ii) Paasche's Method
- (iii) Fisher's Method.

Commodity	Price	Base Year	Current Year	
		Quantity	Expenditure	Quantity
A	6	50	560	56
B	2	100	240	120
C	4	60	360	60
D	10	30	288	24
E	8	40	432	36

Solution.

Commodity	Base Year			Current Year				
	Price P_0	Quantity Q_0	Price P_1	Quantity Q_1	P_1Q_0	P_0Q_0	P_1Q_1	P_0Q_1
A	6	50	$\frac{560}{56} = 10$	56	500	300	560	336
B	2	100	$\frac{240}{120} = 2$	120	200	200	240	240
C	4	60	$\frac{360}{60} = 6$	60	360	240	360	240
D	10	30	$\frac{288}{24} = 12$	24	360	300	288	240
E	8	40	$\frac{432}{36} = 12$	36	1440	320	432	288
					2860	1360	1880	1344

$$\text{Laspeyre's index number} = \frac{\sum P_1Q_0}{\sum P_0Q_0} \times 100 = \frac{2860}{1360} \times 100 = 210.3$$

$$\text{Paasche's index number} = \frac{\sum P_1Q_1}{\sum P_0Q_1} \times 100 = \frac{1880}{1344} \times 100 = 139.9$$

$$\text{Fisher's index number} = \sqrt{210.3 \times 139.9} = 171.5$$

Example 21. Calculating price index number of the year 1996 with 1986 as base year from the following data using:

- Laspeyre's
- Paasche's
- Fisher's formulae

Commodity	Unit	1986		1996	
		Price (₹)	Value (₹)	Quantity Consumed	Value (₹)
A	Kg.	10	1,500	160	1,760
B	Kg.	12	1,080	100	1,300
C	Metre	15	900	60	960
D	Packets	9	450	40	480

NOTES

NOTES

Solution.

Table for calculation of Index number:

Commodity	p_0	q_0	p_0q_0	p_1	q_1	p_1q_1	p_0q_1	$p_1q_0^*$
A	10	150	1,500	11	160	1,760	1,600	1,650
B	12	90	1,080	13	10	1,300	1,200	1,170
C	15	60	900	16	60	960	900	960
D	9	50	450	12	40	480	360	600
Total			3,930			4,500	4,060	4,380

$$(i) \text{ Laspeyre's Index Number} = \frac{\sum p_1q_0}{\sum p_0q_0} \times 100 = \frac{4,380}{3,930} \times 100 = 111.45$$

$$(ii) \text{ Paasche's Index Number} = \frac{\sum p_1q_1}{\sum p_0q_1} \times 100 = \frac{4,500}{4,060} \times 100 = 110.84$$

$$(iii) \text{ Fisher's Index Number} = \sqrt{111.45 \times 110.84} = \sqrt{12,353.188} = 111.14$$

Example 22. Find Laspeyre's and Fisher's price index numbers for the following data for the year 1990 with base year 1985:

Commodity (₹)	1985		1996	
	Price Kg.	Quantity (₹)	Price Kg.	Quantity
Wheat	4.0	20	5.5	20
Sugar	5.0	10	7.0	9
Edible oil	10.5	5	14.5	4
Vegetables	4.5	4	8.0	4

Solution.

$$\begin{aligned} \text{Laspeyre's index number} &= \frac{\sum p_1q_0}{\sum p_0q_0} \times 100 \\ &= \frac{5.5 \times 20 + 7 \times 10 + 14.5 + 8.0 \times 4}{20 \times 4 + 10 \times 5 + 5 \times 10.5 + 4 \times 4.5} \times 100 \\ &= \frac{110 + 70 + 72.5 + 32}{80 + 50 + 52.5 + 18} \times 100 \\ &= \frac{284.5}{200.5} \times 100 = 141.89 = 141.9 \end{aligned}$$

$$\begin{aligned} \text{Paasche's index number} &= \frac{\sum p_1q_1}{\sum p_0q_1} \times 100 \\ &= \frac{20 \times 5.5 + 9 \times 7 + 4 \times 14.5 + 4 \times 8.0}{4 \times 20 + 5 \times 9 + 10.5 \times 4 + 4.5 \times 4} \times 100 \\ &= \frac{110 + 63 + 58 + 32}{80 + 45 + 42 + 18} \times 100 = \frac{263}{185} \times 100 = 142.16 \end{aligned}$$

$$\text{Fisher's Index Number} = \sqrt{\frac{284.5}{200.5} \times \frac{263}{185}} \times 100 = 142.029.$$

Example 23. From the information given below find Laspeyre's and Paasche's price index number for year II with year I as base:

Commodity	Price/Unit (in ₹)		Commodity (in kg.)	Total Value (in ₹)
	Year I	Year II		
A	24.09	12.30	5	184.50
B	20.66	10.17	8	122.04
C	18.60	16.42	16	295.56
D	9.45	10.21	25	189.00

Verify whether the above index numbers satisfy the time reversal test.

Solution.

$$\text{Quantity in kg of commodity A in year II} = \frac{184.50}{12.30} = 15$$

$$\text{Quantity in kg of commodity B in year II} = \frac{122.04}{10.17} = 12$$

$$\text{Quantity in kg of commodity C in year II} = \frac{295.56}{16.42} = 18$$

$$\text{Quantity in kg of commodity D in year II} = \frac{189.00}{10.21} = 18.5$$

$$\begin{aligned} \text{Laspeyre's index number} &= \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 \\ &= \frac{12.30 \times 5 + 10.17 \times 8 + 16.42 \times 16 + 10.21 \times 25}{24.09 \times 5 + 20.66 \times 8 + 18.60 \times 16 + 9.45 \times 25} \times 100 \\ &= \frac{61.5 + 81.36 + 262.72 + 255.5}{120.45 + 165.28 + 297.6 + 236.25} \times 100 \\ &= \frac{661.08}{819.58} \times 100 = 80.66 \end{aligned}$$

$$\begin{aligned} \text{Paasche's index number} &= \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 \\ &= \frac{184.50 + 122.04 + 295.56 + 189.00}{24.09 \times 15 + 20.66 \times 12 + 18.60 \times 18 + 9.45 \times 18.5} \times 100 \\ &= \frac{791}{361.35 + 247.92 + 334.8 + 174.92} \\ &= \frac{791}{1119} \times 100 = 70.68 \end{aligned}$$

Further students may check that time reversal test is not verified.

NOTES

NOTES

Example 24. An index is at 100 in 1991. It rises 4% in 1992, falls 4% in 1994 and rises 3% in 1995. Calculate the index number for 5 years with 1993 as the base.

Solution: First we will have to calculate the index at the old base of 1991 and then construct the new series with 1993 as the base. This is done in the following table.

Index Number with the New Base of 1993

Year	Old Index (1991 = 100)	New Index (1993 = 100)
1991	100	$\frac{100}{97.76} \times 100 = 102.32$
1992	$100 + 4 = 104$	$\frac{100}{97.76} \times 104 = 106.40$
1993	$104 \times \frac{94}{100} = 97.76$	$\frac{100}{97.76} \times 97.76 = 100.00$
1994	$97.76 \times \frac{96}{100} = 93.85$	$\frac{100}{97.76} \times 93.85 = 93.00$
1995	$93.85 \times \frac{103}{100} = 96.66$	$\frac{100}{97.76} \times 96.66 = 98.88$

Example 25. Cosmetics had a price index of 110 in 1996 with 1990 as the basis and 120 in 1997 with 1996. It then climbed by 30% more in 1998 compared to the 1997 price index before falling by 10% in 1999 from its level in 1998. Using 1990 as the baseline, find the index for 1999.

Solution.

Price index for 1996 with 1990 as base

$$I_{90, 96} = \frac{P_{96}}{P_{90}} \times 100 = 110 \Rightarrow \frac{P_{96}}{P_{90}} = 1.1$$

Similarly, $\frac{P_{97}}{P_{96}} = 1.2, \frac{P_{98}}{P_{97}} = 1.3, \frac{P_{99}}{P_{98}} = 0.9$

\therefore Price index for 1999 with 1990 as base

$$\begin{aligned}
 &= \frac{P_{1999}}{P_{1990}} \times 100 \\
 &= \frac{P_{96}}{P_{90}} \times \frac{P_{97}}{P_{96}} \times \frac{P_{98}}{P_{97}} \times \frac{P_{99}}{P_{98}} \times 100 \\
 &= 1.1 \times 1.2 \times 1.3 \times 0.9 \times 100 = 154.5
 \end{aligned}$$

Example 26. A commodity's price rose by 60% between 1980 and 1990, while its production fell by 30%. How much did the value index of production of a commodity change from its value in 1980 to 1990?

Solution.

$$\text{Value index for one commodity} = \frac{P_1 Q_1}{P_0 Q_0} \times 100$$

$$= \left(\frac{P_1}{P_0} \right) \left(\frac{Q_1}{Q_0} \right) \times 100 = 1.6 \times 0.7 \times 100^* = 112$$

\Rightarrow Value index increased by 12%.

Example 27. An investigation of the spending plans of middle-class households in a city revealed the following:

Item	% Expenditure	Prices (2004)	Prices (2005)
Food	35	140	170
Rent	15	80	90
Clothing	20	110	130
Fuel	10	60	80
Misc.	20	90	110

What differences do you observe in the cost of living in 2005 and 2004?

Solution.

Computation of Cost of Living Index

Items of Expenditure	2004 P_0	2005 P_1	W	$\frac{P_1}{P_0} \times 100$ P	PW
Food	140	170	35	$\frac{170}{140} \times 100 = 121.43$	4250.05
Rent	80	90	15	$\frac{90}{80} \times 100 = 112.50$	1687.5
Clothing	110	130	20	$\frac{130}{110} \times 100 = 118.18$	2363.6
Fuel	60	80	10	$\frac{80}{60} \times 100 = 133.33$	1333.3
Misc.	90	110	20	$\frac{110}{90} \times 100 = 122.22$	2444.4
			$\Sigma W = 100$		$\Sigma PW = 12078.85$

$$\text{Cost of living index} = \frac{\Sigma PW}{\Sigma W} = \frac{12078.85}{100} = 120.79$$

Example 28. In 2005, based on base 2004 data, a commodity's price and quantity indices were 120 and 110, respectively. Find the 2005 value index number using the 2004 base.

Solution.

$$\frac{P_1}{P_0} \times 100 = 120 \Rightarrow \frac{P_1}{P_0} = 1.2 \Rightarrow P_1 = 1.2P_0$$

NOTES

NOTES

$$\frac{q_1}{q_0} \times 100 = 110 \Rightarrow \frac{q_1}{q_0} = 1.1 \Rightarrow q_1 = 1.1q_0$$

$$\therefore \text{Value index} = \frac{p_1 q_1}{p_0 q_0} \times 100 = 1.2 \times 1.1 \Rightarrow \frac{p_1 q_1}{p_0 q_0} \times 100 = 132.$$

Example 29. Below is incomplete information on cost of living analysis gleaned from partially damaged records:

Group	Group Index	Per cent (%) of Total Expenditure
Food	268	60
Clothing	280	Not available
Housing	210	20
Fuel and Electricity	240	5
Miscellaneous	260	Not available

The cost of living index, which weighs expenses as a percentage of total income, was discovered to be 255.8. Assess the weights that are missing.

Solution.

Let the weight for clothing be x_1 and for miscellaneous be x_2 .

$$\text{Then } 60 + x_1 + 20 + 5 + x_2 = 100$$

$$x_1 + x_2 = 15$$

$$255.8 = \frac{268 \times 60 + 280x_1 + 210 \times 20 + 240 \times 5 + 260x_2}{100}$$

$$\Rightarrow 25580 = 16080 + 280x_1 + 4200 + 1200 + 260x_2$$

$$= 21480 + 280x_1 + 260x_2$$

$$\Rightarrow 4100 = 260(x_1 + x_2) + 20x_1$$

$$= 260 \times 15 + 20x_1 = 3900 + 20x_1$$

$$\Rightarrow 200 = 20x_1 \Rightarrow x_1 = 10$$

$$\therefore x_2 = 15 - 10 = 5$$

$$x_1 = 10, x_2 = 5$$

Example 30. Expenditure of a family on 3 times are in the ratio of 2 : 5 : 3. The prices of these commodities rise by 30 per cent, 20 per cent and 40 per cent respectively. By what per cent has total expenditure increased?

Solution.

The price of a commodity rise by 30 per cent means that with suitable time as base, index number of a commodity in current year is $100 + 30$, i.e., 130.

Item	Index No. (I)	Weight (W)	IW
A	130	2	260
B	120	5	600
C	140	3	420
		10	1280

$$\text{Price Index No.} = \frac{\Sigma IW}{\Sigma W} = \frac{1280}{10} = 128$$

The total expenditure has increased by $(128 - 100) = 28\%$.

Example 31. Given below are two piece index series. Splice them on the base 1994 = 100. By what per cent was the increase of steel use between 1990 and 1995?

Year	Old price index for Steel: (Base 1985 = 100)	New price index for Steel: (Base 1994 = 100)
1990	141.5	
1991	163.7	
1992	158.2	
1993	156.8	99.8
1994	157.1	100.0
1995		102.3

Solution.

Splicing of Old Price Index to New Price Index

Year	Old price index for Steel: (Base 1985 = 100)	New price index for Steel: (Base 1994 = 100)
1990	141.5	$\frac{100}{157.1} \times 141.5 = 90.67$
1991	163.7	$\frac{100}{157.1} \times 163.7 = 104.20$
1992	158.2	$\frac{100}{157.1} \times 158.2 = 99.81$
1993	156.8	$\frac{100}{157.1} \times 156.8 = 99.81$
1994	157.1	100.0
1995		102.3

Hence, the percentage increase in the price of steel between 1990 and 1995 is

$$\frac{102.30 - 90.07}{90.07} \times 100 = 13.58$$

Hence required increase is 13.58%.

Note: When the old index is spliced to the new index (Base 1994), the index number for 1994, i.e., 157.1 becomes 100.

Therefore, the multiplying factor for splicing is $\frac{100}{157.1} = 0.6365$.

NOTES

NOTES

Example 32. From the following three series of index numbers, create a spliced series of index numbers with 1995 = 100.

Year	1990	1991	1992	1993	1994	1995	1996
Index A	100	120	135				
Index B			100	115	125	145	
Index C						100	110

Solution.

Index numbers for 1990 to 1995 with 1992 as base.

1990	1991	1992	1993	1994	1995
$100 \times \frac{100}{135}$	$120 \times \frac{100}{135}$	$135 \times \frac{100}{135}$	115	125	145
= 74.07	= 88.88	= 100	115	125	145

Index numbers for complete series with 1995 as base.

1990	1991	1992	1993	1994	1995	1996
$74.07 \times \frac{100}{145}$	$88.88 \times \frac{100}{145}$	$100 \times \frac{100}{145}$	$115 \times \frac{100}{145}$	$125 \times \frac{100}{145}$	$145 \times \frac{100}{145}$	110
= 51.08	= 61.29	= 68.96	= 79.3	= 86.2	= 100	110

Miscellaneous Concepts and Examples**Deflating of Index Numbers**

Deflating is the process of modifying or correcting an inflated value. It accounts for the impact of price fluctuations. The purchasing power of money decreases as prices rise. The purchasing power of money income is halved if prices double while people's incomes stay the same. If, for example, the price of sugar in 2000 was ₹ 50 per kg and the price in 2010 become ₹ 75 per kg, then the person who could buy one kg of sugar in

2000 will be able to buy $\frac{50}{75} = \frac{2}{3} = 0.66$ kg of sugar for ₹ 50 in 2010. Thus, the value of money has gone down. The purchasing power of money is thus related to price changes. From the above example, it is clear that if the price rises by 50%, i.e., in place of 1, it becomes $1.5 \left(\frac{75}{50} = 1.5 \right)$, the purchasing power of money would be $\frac{1}{1.5} = 0.66$ which means that the price of one rupee has come down to 66 paise only. Similarly, if the price rises from ₹ 1 to ₹ 1.25 per kg, the purchasing power of money is $\frac{1}{1.25} = 0.8$ which the value of a rupee has gone down to 80 paise only.

$$\therefore \text{Purchasing power of money} = \frac{1}{\text{Consumer Price Index}}$$

To determine the real salary in periods of rising prices, the money wage should be deflated by the price index. We could determine whether a wage earner is better off or

worse off as a result of price change based just on their real wages. Similarly, national income or per capita income should be deflated for price changes to arrive at figures of real income.

$$\text{Real wage} = \frac{\text{Money wage index}}{\text{Consumer price index}} \times 100$$

$$\begin{aligned} \text{Real wage index} &= \frac{\text{Money wage index}}{\text{Consumer price index}} \\ &= \frac{\text{Real wage of the current year}}{\text{Real wage of the base year}} \end{aligned}$$

NOTES

Business Application of Deflating Technique

Example 33. (A) A management analyst is studying the inventory position of a firm, because there has been steadily rise of inventories over a period of 6 years. Find out index number of physical volume of inventory at constant prices from the following information given to you:

Year	2003	2004	2005	2006	2007	2008
Inventory	500.0	520.2	548.6	573.0	589.5	620.2
Price Index (1993 = 100)	128.2	131.4	136.2	142.1	144.5	147.2

Solution.

Year	Inventory (000 ₹)	Price Index (1993 = 100)	Deflated Inventory	Physical Value Index
2003	500.0	128.2	$\frac{500}{128.2} \times 100 = 390.0$	100
2004	520.2	131.4	$\frac{520.2}{131.4} \times 100 = 395.9$	$\frac{395.9}{390} \times 100 = 101.5$
2005	548.6	136.2	$\frac{548.6}{136.2} \times 100 = 402.8$	$\frac{402.8}{390} \times 100 = 103.3$
2006	573.0	142.1	$\frac{573}{142.1} \times 100 = 403.2$	$\frac{403.2}{390} \times 100 = 103.4$
2007	589.5	144.5	$\frac{589.5}{144.5} \times 100 = 408.0$	$\frac{408}{390} \times 100 = 104.6$
2008	620.2	147.2	$\frac{620.2}{147.2} \times 100 = 421.3$	$\frac{421.3}{390} \times 100 = 108.0$

NOTES

(B) A group of employees present the following data:

Year	2003	2004	2005	2006	2007	2008
Monthly Wages (₹)	109.50	112.20	116.40	125.08	135.40	138.10
Price Index	112.80	118.20	127.40	138.20	143.50	149.80
Real Wages (₹)	97.07	94.92	91.37	90.51	94.36	92.18

Claim that their real wages have gone down in 2008 as compared to 2003. What percentage increase in the present wages is required (if any) to provide the same buying power as they had in 2003?

Solution.

$$\frac{\text{Wage in 2003}}{\text{Price index in 2003}} = \frac{X (\text{Present needed wage})}{\text{Index No. 2008}}$$

$$\frac{109.50}{112.8} = \frac{X}{149.8} \cdot \text{By calculating this, the value of } X = 145.42.$$

$$\text{Increase needed} = \frac{X - 138.10}{138.10} \text{ or } \frac{145.42 - 138.10}{138.10} = 5.3\%$$

LIST OF FORMULAE

$$1. \text{ Price Relative} = P_{01} = \frac{p_1}{p_0} \times 100$$

$$\text{Quantity Relative} = Q_{01} = \frac{q_1}{q_0} \times 100$$

2. Unweighted Index Numbers:

$$(a) \text{ Simple Aggregate Price Index} = P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$$

$$(b) \text{ Simple Average of Price Relatives} = P_{01} = \frac{\sum \frac{p_1}{p_0} \times 100}{N}$$

3. Weighted Aggregate Method:

$$(a) \text{ Simple Aggregative with fixed weight} = P_{01} = \frac{\sum p_1 w}{\sum p_0 w} \times 100$$

$$\text{Laspeyre's Price Index} = P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

$$\text{Paasche's Index number} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

$$\text{Marshall-Edgeworth's Index} = \frac{\Sigma p_1(q_0 + q_1)}{\Sigma p_0(q_0 + q_1)} \times 100$$

$$\text{Fisher's Ideal Index} = P_{01} = \sqrt{\left[\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \right]} \times 100$$

(b) Weighted Average of Price Relatives:

$$\text{Weighted Average of Price Relatives} = P_{01} = \frac{\Sigma \left(\frac{p_1}{p_0} \times 100 \right) W}{\Sigma W}$$

Weighted Average of Price Relative (base year values as weights)

$$= P_{01} = \frac{\Sigma \left(\frac{p_1}{p_0} \times 100 \right) p_0 q_0}{\Sigma p_0 q_0}$$

Weighted Average of Price Relatives (current year values as weights)

$$= P_{01} = \frac{\Sigma \left(\frac{p_1}{p_0} \times 100 \right) p_1 q_1}{\Sigma p_1 q_1}$$

Check Your Progress

I. Multiple Choice Questions

- For comparing yearly changes in price level, the suitable index to be used is:
 - FBI with average price as base
 - CBI
 - None of these
 - (a) and (b) both
- The weights used in Laspeyre's price index are denoted as:
 - q_0
 - q_1
 - P_0
 - p_1
- The weights used in Laspeyre's quantity index are denoted as:
 - q_0
 - q_1
 - p_0
 - p_1
- The weights used in Paasche's price index are denoted as:
 - q_0
 - q_1
 - p_0
 - p_1
- The weights used in Paasche's quantity index are denoted as:
 - q_0
 - q_1
 - p_0
 - p_1
- Weighted average of relatives if base year value is taken as weights gives:
 - Fisher's index
 - Laspeyre's index
 - Paasche's index
 - Bowley's index

NOTES

NOTES

7. The formula for simple average of price relative is:

- (a) $\frac{1}{n} \sum \frac{p_0}{p_1} \times 100$ (b) $\sum \frac{p_2}{p_0} \times 100$
 (c) $\frac{1}{n} \sum \frac{p_1}{p_0} \times 100$ (d) None

II. State Whether the Following Statements are True or False

- Index numbers are specialised averages.
- An index number is a statistical indicator of a variable's or a group of related variables' changes.
- Index numbers are not proportionally expressed.
- Index numbers track changes that cannot be observed immediately.
- Index figures serve as economic indicators.
- It is not possible to forecast using index numbers.
- Index numbers are used in computing the real income or wages.
- The ideal average used in the construction of index numbers is the arithmetic mean.
- Base year used in index numbers should be a normal year.
- Weighted index numbers are better and more representative than simple index numbers.
- When constructing index numbers, the geometric mean is the most appropriate average.
- Price relatives are formed to study price changes over time.
- Quantity relatives are used to track variations in the quantity or volume of a commodity's production or consumption.
- Laspeyre's formula is a weighted aggregate index with base year quantity weights.
- Paasche's approach uses base year quantity weights to create a weighted aggregate index.
- Fisher's index formula is the arithmetic mean of Laspeyre's and Paasche's formula.

III. Fill in the Blanks

- Consumer price index number _____.
(measures changes in the retail prices, measures changes in wholesale prices)
- If the price index increases by 20% the cost of Campa Cola, which at present is ₹ 10, will _____.
(increase by ₹ 2, increase by ₹ 12)
- Index numbers are _____.
(simple averages, specialised averages)
- Index of industrial production is a _____.
(price index, quantity index)
- Index numbers indicate _____.
(relative changes, absolute changes)

5.12 ANSWERS TO 'CHECK YOUR PROGRESS'

I. Multiple Choice Questions

- (b)
- (a)
- (c)
- (b)
- (d)
- (b)
- (c)

II. State whether the following Statements are True or False

1. True
2. True
3. False
4. True
5. True
6. False
7. True
8. True
9. True
10. True
11. True
12. True
13. True
14. True
15. False
16. False

NOTES**III. Fill in the Blanks**

1. measures changes in retail price
2. increase by ₹ 2
3. specialised averages
4. quantity index
5. relative changes

5.13 SUMMARY

- Index numbers
- Introduction
- Definition
- Types of Index Numbers
- Construction of Index Numbers
- Methods of Construction
- Comparison between Laspeyre's Index Number and Paasche's Index
- Number
- Selection of an Average for the Construction of Index Numbers
- Problems in the Construction of Index Numbers
- Advantages or Uses of Index Numbers
- Limitations of Index Numbers
- Splicing of Index Numbers

NOTES

5.14 KEY TERMS

$$1. \text{ Price Relative} = P_{01} = \frac{p_1}{p_0} \times 100$$

$$\text{Quantity Relative} = Q_{01} = \frac{q_1}{q_0} \times 100$$

2. **Unweighted Index Numbers:**

$$(a) \text{ Simple Aggregate Price Index} = P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$$

$$(b) \text{ Simple Average of Price Relatives} = P_{01} = \frac{\sum \frac{p_1}{p_0} \times 100}{N}$$

3. **Weighted Aggregate Method:**

$$(a) \text{ Simple Aggregate with fixed weight} = P_{01} = \frac{\sum p_1 w}{\sum p_0 w} \times 100$$

$$\text{Laspeyre's Price Index} = P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

$$\text{Paasche's Index Number} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

$$\text{Marshall-Edgeworth's Index} = \frac{\sum p_1 (q_0 + q_1)}{\sum p_0 (q_0 + q_1)} \times 100$$

$$\text{Fisher's Ideal Index} = P_{01} = \sqrt{\left[\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1} \right]} \times 100$$

(b) **Weighted Average of Price Relatives:**

$$\text{Weighted Average of Price Relatives} = P_{01} = \frac{\sum \left(\frac{p_1}{p_0} \times 100 \right) W}{\sum W}$$

Weighted Average of Price Relative (base year values as weights)

$$= P_{01} = \frac{\sum \left(\frac{p_1}{p_0} \times 100 \right) p_0 q_0}{\sum p_0 q_0}$$

Weighted Average of Price Relatives (current year values as weights)

$$= P_{01} = \frac{\sum \left(\frac{p_1}{p_0} \times 100 \right) p_1 q_1}{\sum p_1 q_1}$$

5.15 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. Define index numbers.
2. Mention three important characteristics of index numbers.
3. Mention two important uses of index numbers.
4. Given two important limitations of index numbers.
5. Point out the steps and problems involved in the construction of index numbers.
6. What care should be taken in the selection of commodities for the index numbers?
7. What do you mean by weight in the context of index numbers?
8. Define Laspeyre's index numbers.
9. Define Paasche's index numbers.
10. What is Consumer Price Index Number?

NOTES

Long Answer Questions

1. A worker earned ₹ 9,000 per month in 1995. The cost of living index increased by 60% between 1995 and 1999.

How much extra income should the worker have earned in 1999 so that he could buy same quantities as in 1995? [Ans. ₹ 5,400]

2. A worker earned ₹ 8,000 per month in 1992. The cost of living index increased by 70% between 1992 and 1995.

How much extra income should the worker have earned in 1995 so that he could buy the same quantities as in 1992? [Ans. ₹ 5,600]

3. From the following data calculate price index numbers for 2000 with 1990 as base by: (i) Laspeyre's method, (ii) Paasche's method, and (iii) Fisher's ideal method.

Commodities		1990		2000
	Price	Quantity	Price	Quantity
A	20	8	40	6
B	50	10	60	5
C	40	15	50	15
D	20	20	20	25

[Ans. (i) 124.69, (ii) 121.76, (iii) 123.21]

4. The following table gives the cost of living index numbers for different commodity groups together with respective weights for 2000 (Base = 1980):

Group	Food	Clothing	Fuel and Lighting	Rent	Misc.
Group index	425	475	300	400	250
Group weight	62	4	6	12	16

NOTES

Obtain the overall cost of living index number. Suppose a person was earning ₹ 6,000 in 1980. What should be his salary in 2000 if his standard of living in that year is to be the same as in 1980? [Ans. ₹ 388.50, ₹ 23.310]

5. The price relative and weight of a set of commodities are given in the following table:

Commodity	C_1	C_2	C_3	C_4
Price Relative	120	127	125	119
Weight	$2w_1$	w_2	w_1	w_{2+3}

If the index for the set is 122 and the sum of the weights is 40, find w_1 and w_2 .

[Ans. $w_1 = 7$, $w_2 = 8$]

6. The following table gives the cost of living index numbers for different groups with the respective weights for the year 2000 (base year 1985).

Group	Cost of Living Index	Weight
Food	430	60
Clothing	480	5
Lighting and Fuel	320	7
Rent	390	10
Miscellaneous	300	18

- (i) Calculate the overall cost of living index number.
 (ii) Suppose, a person was earning ₹ 7,000 p.m. in 1985. What should be his salary in 2000, if his standard of living in 2000 is to be the same as in 1985?

[Ans. ₹ 397.4, ₹ 27,818]

7. Compute Fisher's ideal index number from the data given below and check whether the time reversal test and factor reversal test is satisfied.

		Base Year	Current Year	
Commodity	Price	Quantity	Price	Quantity
A	2	7	6	6
B	3	6	2	3
C	4	5	8	5
D	5	4	2	4

[Ans. 145]

5.16 REFERENCES

1. D.N. Elhance, Veena Elhance and B.M. Aggarwal, 2007, Fundamentals of Statistics. Kitab Mahal, New Delhi.
2. B.M. Aggarwal, 2012, Business Statistics (With Lab Work), Himalaya Publishing House, Mumbai.
3. B.M. Aggarwal, Dr. Puja A. Gulati, Neha Aggarwal, 2022, Statistics for Business and Economics. Kitab Mahal, New Delhi.