ਜਗਤ ਗੁਰੂ ਨਾਨਕ ਦੇਵ
ਪੰਜਾਬ ਸਟੇਟ ਓਪਨ ਯੂਨੀਵਰਸਿਟੀ
ਪਟਿਆਲਾ

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

JAGAT GURU NANAK DEV
PUNJAB STATE OPEN UNIVERSITY, PATIALA

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

M.A (ECONOMICS)

SEMESTER- II

MAEC24204T: QUANTITATIVE METHODS II

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

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

COURSE COORDINATOR AND EDITOR:

Dr. Pinky Sra

Assistant Professor in Economics
School of Social Sciences and Liberal Arts
Jagat Guru Nanak Dev Punjab State Open University, Patiala

JAGAT GURU NANAK DEV PUNJAB STATE OPEN UNIVERSITY,PATIALA

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

PREFACE

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

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

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

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

Prof. G S Batra
Dean Academic Affairs

ਜਗਤ ਗੁਰੂ ਨਾਨਕ ਦੇਵ
ਪੰਜਾਬ ਸਟੇਟ ਓਪਨ ਯੂਨੀਵਰਸਿਟੀ
ਪਟਿਆਲਾ

M.A (ECONOMICS)

SEMESTER-II

(MAEC24204T) QUANTITATIVE METHODS II

MAX. MARKS:100

EXTERNAL:70

INTERNAL:30

PASS:40%

CREDITS:6

OBJECTIVE:

This course acquaints the students with the basic principles of Microeconomics and economic activities. It will help the students to understand the subject by applying it to their day to day experiences.

INSTRUCTIONS FOR THE PAPER SETTER/EXAMINER:

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

INSTRUCTIONS FOR THE CANDIDATES:

Candidates are required to attempt any two questions each from the sections A, and B of the question paper, and any ten short answer questions from Section C. They have to attempt questions only at one place and only once. Second or subsequent attempts, unless the earlier ones have been crossed out, shall not be evaluated.

SECTION - A

Unit 1: Measures of Central Tendency: Mean, Median, Mode

Unit 2: Dispersion - Objectives and significance of Good Measures, Measures of Dispersion - Range, Quartile Deviation, Mean Deviation and Standard Deviation (ungrouped data), Co-efficient of variation (CV), Lorenz Curve

Unit 3: Correlation Analysis: Karl Pearson's (excluding grouped data) and Spearman's rank formula

Unit 4: Simple Regression Analysis: regression meaning, properties, X on Y and Yon X

SECTION – B

Unit 5: Meaning of Hypothesis, Characteristics of Hypothesis, Basic Concepts, Hypothesis Testing Procedures (Steps), Introduction to parametric and non-parametric tests.

Unit 6: Sampling distributions of a Statistics- Small Sample test or student-t test and its applications: t-test for single mean, difference of means, Paired t-test

Unit 7: Large Sample test: Introduction, Sampling of Attributes- test for Single Proportion, test for difference in proportion and F-test

Unit 8: Interpolation and Extrapolation.

Suggested Readings:

- A.M Goon, M.K Gupta and B. Dasgupta, fundamental of statistics Vol-I, World Press Calcutta
- Anderson, D.R.; Sweeney, D.J. and Williams, T.A., "Statistics for Business and 30
- Economics", 2nd edition (2011), Thompson, New Delhi.
- Gupta SC: Fundamental of statistics, S. Chand & Company. New Delhi • Gupta, SP: Statistical Methods, S. Chand & Company. New Delhi
- Lehmann, E.L. (1986): Testing Statistical hypotheses (Student Edition).
- Monga, GS: Mathematics and Statistics for Economics, Vikas Publishing House, New Delhi.
- Zacks, S. (1971): Theory of Statistical Inference, John Wiley and Sons. New York.

JAGAT GURU NANAK DEV PUNJAB STATE OPEN UNIVERSITY, PATIALA

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

M.A (ECONOMICS)

SEMESTER-II

(MAEC24204T) QUANTITATIVE METHODS II

COURSE COORDINATOR AND EDITOR: DR. PINKY SRA

SECTION A

UNIT NO:	UNIT NAME
Unit 1	Measures of Central Tendency: Mean, Median, Mode
Unit 2	Dispersion
Unit 3	Correlation Analysis: Karl Pearson’s (excluding grouped data) and Spearman’s rank formula
Unit 4	Simple Regression Analysis: regression meaning, properties, X on Y and Yon X

SECTION B

UNIT NO:	UNIT NAME
Unit 5	Meaning of Hypothesis, Characteristics of Hypothesis, Basic Concepts, Hypothesis Testing Procedures (Steps), Introduction to parametric and non-parametric tests.
Unit 6	Sampling distributions of a Statistics- Small Sample test or student-t test and its applications: t-test for single mean, difference of means, Paired t-test
Unit 7	Large Sample test: Introduction, Sampling of Attributes- test for Single Proportion, test for difference in proportion and F-test
Unit 8	Interpolation and Extrapolation

M.A (ECONOMICS)

SEMESTER -II

QUANTITATIVE METHODS II

Unit 1: Measures of Central Tendency: Mean, Median, Mode

STRUCTURE

1.0 LEARNING OBJECTIVES

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

- Meaning of Average
- Features of a good measure of Average
- Find different types of averages for various types of data
- Understand the relation that exists between different types of Averages
- Procedure to find out mean of different series
- Merits and limitations of each type of average
- Define the meaning of median
- Calculate medium for different series
- Define the meaning of mode
- Know merits and limitations of Mean and Median
- Relationship between Mean, Mode and Median
- Meaning of Quartile, Decile and Percentile

1.1 INTRODUCTION

We can say that the modern age is the age of Statistics. There is no field in modern life in which statistics is not used. Whether it is Business, Economics, Education. Government Planning or any other field of our life, statistics is used everywhere. Business manager use statistics for business decision making, Economists use statistics for economic planning, Investors use statistics for future forecasting and so on. There are many techniques in statistics that helps us in all these purposes. Average or Central Tendency is one such technique that is widely used in statistics. This technique is used almost in every walk of the life.

1.2 MEANING OF AVERAGE OR CENTRAL TENDENCY

Average or Central tendency is the most used tool of statistics. This is the tool without which statistics is incomplete. In simple words we can say that the Average is the single value which is capable of representing its series. It is the value around which other values in the series move. We can define Average as the single typical value of the series which represents the whole series data. According to **Croxton and Cowden** "An average is a single value within the range of data that is used to represent all values in the series. Since an average is somewhere within the range of the data, it is also called a measure of Central Value".

1.3 OBJECTIVES AND FUNCTIONS OF AVERAGE

- Single Value Representing Entire Data: In statistics, data is often represented using tables and diagrams. However, when the dataset is too large, presenting it comprehensively becomes challenging. An average serves to summarize such extensive data into a single value. For instance, while India's national income data is vast, calculating the per capita income provides a concise representation.
- Facilitating Comparison: Comparing two different data series can be complex due to differences in the number of items or other factors. Averages simplify this process by providing a standard measure. For example, per capita income—an average—can be used as a basis for comparison to compare the income levels of people in India and Pakistan.
- Drawing Conclusions About a Population from a Sample: Averages play a critical role in inferring conclusions about a population based on a sample. For example, the mean of a sample often serves as a representative measure for its corresponding population.
- Foundation for Other Statistical Methods: Many statistical techniques rely on the concept of averages. Without understanding averages, applying methods like dispersion, skewness, or index numbers becomes difficult. These techniques are built on the foundation provided by averages.
- Basis for Decision-Making: Averages are essential for making informed decisions. By analyzing averages, we gain insights into data trends that guide decision-making. For instance, a company may base its sales strategies on the average yearly sales of previous years.
- Establishing Precise Relationships: Averages help identify and quantify relationships between variables or items, eliminating subjective bias in analysis. For example, instead of stating subjectively that Rajesh is more intelligent than Ravi, comparing their average marks provides a factual basis for this conclusion.
- Support in Policy Formulation: Averages assist governments in developing policies by providing key indicators such as per capita income and average growth rates. These measures are fundamental in designing effective economic and social strategies.

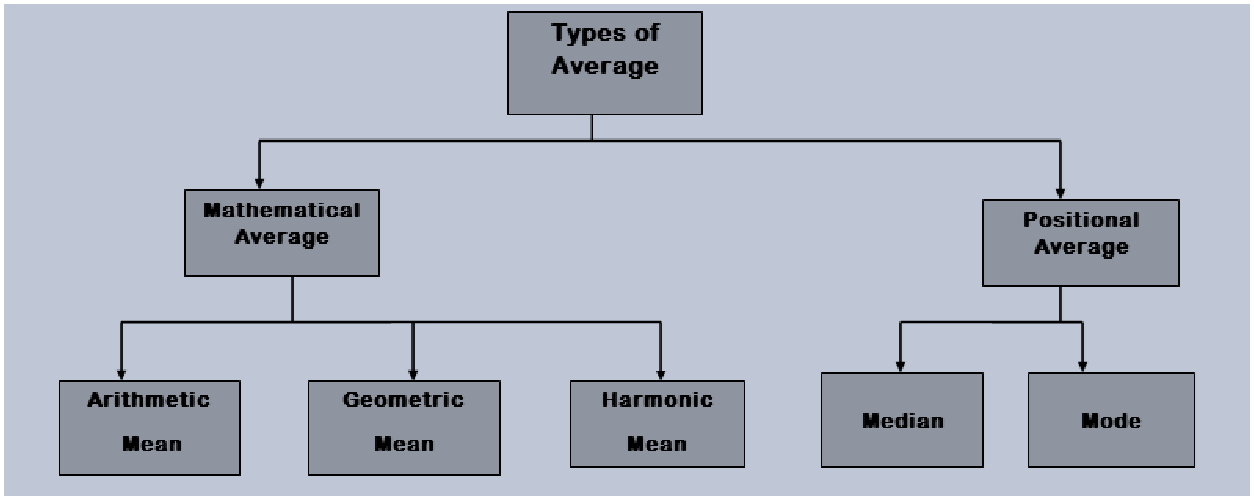
1.4 FEATURES OF GOOD AVERAGE

- **Rigidly Defined:** A good measure of average is one which is having a clear-cut definition and there is no confusion in the mind of person who is calculating the average. In case person applies his discretion while calculating the average, we cannot say that average is a good measure. Good average must have fix algebraic formula, so that whenever average of same data is calculated by two different persons, result is always same.

- **Easy to Compute**: Good average is one which does not involve much calculation and are easy to compute. A good average is one which can be calculated even by a person having less knowledge of Statistics. If it is very difficult to calculate the average, we cannot regard it as a good measure.
- **Based on all Observations:** Good average must consider all the values or data that is available in the series. If average is based on only few observations of the series, we cannot say that it is a good measure of average.
- **Not affected by Extreme Values**: A good measure of average is not affected by the extreme values present in the Data. Sometime data contains values which are not within normal limits, these values are called extreme values. If average is affected by these extreme values, we cannot claim that average is a good measure.
- **Representative of whole Series:** A good measure of average is one which represents characteristics of whole series of the data.
- **Easy to Understand:** A good measure of average is not only easy to understand but also easy to interpret.
- **Not Affected by Fluctuations in the Sampling:** If we take one sample from the universe and calculate an average, then we draw another sample from the same universe and calculate the average again, there must not be much difference between these two averages. If average significantly change with the change in sample, we cannot treat it as a good measure of average.
- **Capable of further Algebraic Treatment:** a good average is one on which we can apply further algebraic treatment. In case further algebraic treatment is not possible, we cannot say that it is a good average. Sch further algebraic treatment may be anything like calculating combined average when average of two different series is available.
- **Located Graphically:** It will be better if we can locate average graphically also. Graphs are easy to understand and interpret, so the average that can be located graphically is a good average.

1.5 MEASURES OF CENTRAL TENDENCY

There are many methods through which we can calculate average or central tendency. We can divide these methods into two categories that are Algebraic Method and Positional Average. Algebraic methods are those in which the value of average depends upon the mathematical formula used in the average. The mathematical average can further be divided into three categories that are Arithmetic Mean, Geometric Mean and Harmonic Mean. On the other hand, positional averages are those average which are not based on the mathematical formula used in calculation of average rather these depends upon the position of the variable in the series. As these depends upon the position of the variable, these averages are not affected by the extreme values in the data. Following chart shows different types of averages.



1.6 ARITHMETIC MEAN: MEANING

It is the most popular and most common measure of average. It is so popular that for a common man the two terms Arithmetic Mean and Average are one and the same thing. However, in reality these two terms are not same and arithmetic mean is just one measure of the average. We can define the arithmetic mean as: "The value obtained by dividing sum of observations with the number of observations". So arithmetic mean is very easy to calculate, what we have to do is just add up the value of all the items given in the data and then we have to divide that total with the number of items in the data. Arithmetic mean is represented by symbol A. M. or $\overline{X}$

1.7 ARITHMETIC MEAN IN DIFFERENT SERIES

1.7.1 Arithmetic Mean in case of Individual Series

Individual series are those series in which all the items of the data are listed individually. There are two methods of finding arithmetic mean in the individual series. These two methods are Direct method and Shortcut Method.

1. **Direct Method** According to this method calculation of mean is very simple and as discussed above, we have to just add the items and then divide it by number of items. Following are the steps in calculation of mean by direct method:

    1. Suppose our various items of the data are X_1, X_2, X_3 ………………… X_n
    2. Add all the values of the series and find $\sum X$.
    3. Find out the number of items in the series denoted by n.
    4. Calculate arithmetic mean dividing sum value of observation with the number of observations using following formula:

$$\overline{\text{x}} = \frac{\mathbf{X1 + X2 + X3 + \; - - - - - - - - Xn}}{\mathbf{N}} = \frac{\mathbf{\sum X}}{\mathbf{N}}$$

Where $\overline{\text{x}}$= Mean

N = Number of items

$\sum$X = Sum of observation

Example 1. The daily income of 10 families is a as given below (in rupees) :

R. No	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
marks	**50**	**60**	**65**	**88**	**68**	**70**	**83**	**45**	**53**	**58**

Find the arithmetic mean by direct method.

Solution: Computation of Arithmetic Mean

Serial No.	Daily Income (in Rs.) X
1	50
2	60
3	65
4	88
5	68
6	70
7	83
8	45
9	53
10	58
N=10	$\sum$X=640

$$\text{A. M.,} \quad \overline{\text{x}} = \frac{\text{X1 + X2 + X3 + } - - - - - - - - \text{XN}}{\text{N}} = \frac{\sum \text{X}}{\text{N}} = \frac{640}{10} = \text{Rs. } 64$$

2. **Short Cut Method:** Normally this method is used when the value of items is very large and it is difficult to make calculations. Under this method we take one value as mean which is known as assumed mean and deviations are calculated from this as you mean. This method is also known as is assumed mean method. Following are the steps of this method:

1. Suppose our various items of the data are X_1, X_2, X_3 ………………….. X_n

2. Take any value as assumed mean represented by ‘A’. This value may be any value among data or any other value even if that is not presented in data.

3. Find out deviations of items from assumed mean. For that deduct Assumed value from each value of the data. These deviations are representing as ‘dx’

4. Find sum of the deviations represented by ∑dx.

5. Find out the number of items in the series denoted by n.

6. Calculate arithmetic mean dividing sum deviations of the observation with the number of observations using following formula:

$$\overline{\times} = \mathbf{A} + \frac{\sum \mathbf{dx}}{\mathbf{N}}$$

Where $\overline{\times}$= Mean, A = Assumed Mean, N = Number of items, ∑dx = Sum of deviations

Example 2. Calculate A. M. by short - cut method for the following data

R. No	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
marks	**50**	**60**	**65**	**88**	**68**	**70**	**83**	**45**	**53**	**58**

Solution: Let assumed Mean (A) be 60

R. No.	Marks (X)	dx=X-A
1	50	-10
2	60	0
3	65	5
4	88	28
5	68	8
6	70	10
7	83	23
8	45	-15
9	53	-7
10	58	-2
N=10		∑dx=40

As $\overline{\times} = \mathrm{A} + \frac{\sum \mathrm{dx}}{\mathrm{N}}$

⇒ $\overline{\times} = 60 + \frac{40}{10} \quad = 60 + 4$

⇒ $\overline{\times} = 64$ Marks

1.7.2 Arithmetic Mean in case of Discrete Series

In individual series if any value is repeated that is shown repeatedly in the series. It makes series lengthy and make calculation difficult. In case of discrete series, instead of repeatedly showing the items we just group those items and the number of time that item is repeated is shown as frequency. In case of discrete series, we can calculate Arithmetic mean. By using Direct Method and Shortcut Method.

1. **Direct Method:** In indirect method we multiply the value of items (X) with their respective frequency (f) to find out the the product item (fX). Then we take up sun of the product and divide it with the number of items. Following are the steps

1. Multiply the value of items (X) with their respective frequency (f) to find out the the product item (fX)
2. Add up the product so calculated to find ∑ fX.
3. Find out the number of items in the series denoted by n.
4. Calculate arithmetic mean dividing sum of the product with the number of observations using following formula:

$$\overline{x}= \; = \frac{\sum \mathbf{fX}}{\mathbf{N}}$$

Where $\overline{x}$= Mean

N = Number of items

∑fX = Sum of product of observations.

Example 3. From the following data find out the mean height of the students.

Height (in cms.)	154	155	156	157	158	159	160	161	162	163
No. of Students	1	6	10	22	21	17	14	5	3	1

Solution:

Height in cms. (x)	No. of students (f)	fx
154	1	4
155	6	5
156	10	6
157	22	7
158	21	8
159	17	9
160	14	10
161	5	11
162	3	12
163	1	13
	∑f=100	∑fx=15813

∴ Mean Height $\overline{X} = \frac{\sum fX}{\sum f} = \frac{15813}{100} = 158.13$

2. **Short Cut Method:** Under this method we take one value as mean which is known as assumed mean and deviations are calculated from this as you mean. Then average is calculated using assumed mean. Following are the steps of this method:

1. Suppose our items of the data are 'X' and its corresponding frequency is 'f'.
2. Take any value as assumed mean represented by 'A'.
3. Find out deviations of items from assumed mean. For that deduct Assumed value from each value of the data. These deviations are represented as 'dx'
4. Multiply the values of dx with corresponding frequency to find out product denoted by fdx
5. Find sum of the product so calculated represented by ∑ fdx.
6. Find out the number of items in the series denoted by n.
7. Calculate arithmetic mean dividing sum deviations of the observation with the number of observations using following formula:

$$\overline{X} = \mathbf{A} + \frac{\sum \mathbf{fdx}}{\mathbf{N}}$$

Where $\overline{X}$= Mean, A = Assumed Mean, N = Number of items, ∑fdx = Sum of product of deviation with frequency.

Example 4. From the following data find out the mean height of the students.

Height (in cms.)	154	155	156	157	158	159	160	161	162	163
No. of Students	1	6	10	22	21	17	14	5	3	1

Solution: Let the Assumed Mean (A) be 150

Height in cms.
(X)
	No. of students
	(f)
		dX=(X-A)
		A=150
			fdX

154	1	4	4
155	6	5	30
156	10	6	60
157	22	7	154
158	21	8	168
159	17	9	153
160	14	10	140
161	5	11	55
162	3	12	36
163	1	13	13

	∑f=100		∑fdX=813

Applying the formula

$$\overline{X} = \mathrm{A} + \frac{\sum \mathrm{fdX}}{\sum \mathrm{f}}$$

We get

$$\overline{X} = 150 + \frac{813}{100}$$

$$= 150 + 8.13 = 158.13$$

∴ Mean Height $= 158.13$ cm

1.7.3 Arithmetic Mean in case of Continuous Series

Continuous series is also known as Grouped Frequency Series. Under this series the values of the observation are grouped in various classes with some upper and lower limit. For example, classes like 10-20, 20-30, 30-40 and so on. In the class 10-20 lower limit is 10 and upper limit is 20. So, all the observations having values between 10 and 20 are put in this class interval. Similar procedure is adopted for all class intervals. The procedure of calculating Arithmetic Mean is continuous series is just like discrete series except that instead of taking values of observations we take mid value of the class interval. The mid value is represented by 'm' and is calculated using following formula:

$$\mathbf{m} = \frac{\textbf{Lower Limit} + \textbf{Upper Limit}}{\mathbf{2}}$$

1. **Direct Method:** In indirect method we multiply the mid values (m) with their respective frequency (f) to find out the product item (fm). Then we take up sun of the product and divide it with the number of items. Following are the steps
    1. Multiply the mid values (m) with their respective frequency (f) to find out the product item (fm)
    2. Add up the product so calculated to find ∑fm.
    3. Find out the number of items in the series denoted by n.
    4. Calculate arithmetic mean by dividing sum of the product with the number of observations using following formula:

$$\overline{X} = \ = \frac{\sum \mathbf{fm}}{\mathbf{N}}$$

Where $\overline{X}$= Mean

N = Number of items

∑fm = Sum of the product of observations of mean and frequencies.

Example 5. Calculate the arithmetic mean of the following data:

Daily Wages (Rs.)	0-100	100-200	200-300	300-400	400-500	500-600	600-700	700-800	800-900
No. of Workers	1	4	10	22	30	35	10	7	1

Solution:

Daily Wages C.I	No. of Workers (f)	Mid Value m	fdm
0-100	1	50	50
100-200	4	150	600
200-300	10	250	2500
300-400	22	350	7700
400-500	30	450	13500
500-600	35	550	19250
600-700	10	650	6500
700-800	7	750	5250
800-900	1	850	850
	∑f=120		∑fm=56200

As $\overline{X} = \frac{\sum fm}{\sum f}$

∴ $\overline{X} = \frac{56200}{120}$ = 468.3333

2. **Short Cut Method:** This method of mean is almost similar to calculation in the discrete series but here the assumed mean is selected and then the deviation is taken from mid value of the observations. Following are the steps of this method:

1. Calculate the Mid Values of the series represented by 'm'.
2. Take any value as assumed mean represented by 'A'.
3. Find out deviations of items from assumed mean. For that deduct Assumed value from mid values of the data. These deviations are representing as 'dm'
4. Multiply the values of dm with corresponding frequency to find out product by fdm
5. Find sum of the product so calculated represented by ∑ fdm.
6. Find out the number of items in the series denoted by n.
7. Calculate arithmetic mean dividing sum deviations of the observation with the number of observations using following formula:

$$\overline{X} = \mathbf{A} + \frac{\sum \mathbf{fdm}}{\mathbf{N}}$$

Where $\overline{X}$= Mean, A = Assumed Mean, N = Number of items

$\sum$ fdm = Sum of product of deviation from mid values with frequency.

Example 6. Calculate the mean from the following data

Daily Wages (Rs.)	0-100	100-200	200-300	300-400	400-500	500-600	600-700	700-800	800-900
No. of Workers	1	4	10	22	30	35	10	7	1

Solution: Let the assumed mean, $A = 150$

Daily Wages C.I	No. of Workers (f)	Mid Value m	dm = m-A (m-150)	fdm
0-100	1	50	-100	-100
100-200	4	150	0	0
200-300	10	250	100	1000
300-400	22	350	200	4400
400-500	30	450	300	9000
500-600	35	550	400	14,000
600-700	10	650	500	5000
700-800	7	750	600	4200
800-900	1	850	700	700
	∑f=120			∑fdm=38,200

As $\overline{X} = \mathrm{A} + \frac{\sum \mathrm{fdm}}{\sum \mathrm{f}}$

$= 150 + \frac{38,200}{120}$ $= 150 + 318.33 = 468.33$

⇒ $\overline{X} = 468.33$

3. **Step Deviation Method:** Step Deviation method is the most frequently used method of finding Arithmetic Mean in case of continuous series. This method is normally used when the class interval of the various classes is same. This method makes the process of calculation simple. Following are the steps of this method:

   1. Calculate the Mid Values of the series represented by ‘m’.
   2. Take any value as assumed mean represented by ‘A’.

3. Find out deviations of items from assumed mean. For that deduct Assumed value from mid values of the data. These deviations are representing as ‘dm’.

4. Find out if all the values are divisible by some common factor ‘C’ and divide all the deviations with such common factor to find out dm’ which is dm/c

5. Multiply the values of dm’ with corresponding frequency to find out product denoted by fdm’

6. Find sum of the product so calculated represented by ∑ fdm’.

7. Find out the number of items in the series denoted by n.

8. Calculate arithmetic mean dividing sum deviations of the observation with the number of observations using following formula:

$$\overline{\times}= \mathbf{A} + \frac{\sum \mathbf{fdm'}}{\sum \mathbf{f}} \times \mathbf{C}$$

Where $\overline{\times}$= Mean, A = Assumed Mean, N = Number of items, C = Common Factor

∑fdm’ = Sum of product of deviation after dividing with common factors and multiplying it with frequency.

Example 7. Use step deviation method to find $\overline{\times}$ for the data given below:

Income (Rs.)	1000-2000	2000-3000	3000-4000	4000-5000	5000-6000	6000-7000
No. of Persons	4	7	16	20	15	8

Solution: Let the assumed mean A = 4500

Income (Rs.) C. I.	No of Persons f	Mid Value m	$\mathrm{dm} = \mathrm{m} - \mathrm{A}$ $= (\mathrm{m} - 4500)$	$\mathrm{dm}' = \frac{\mathrm{dm}}{\mathrm{C}}$ $\mathrm{C} = 1000$	fdm'
1000-2000	4	1500	-3000	-3	-12
2000-3000	7	2500	-2000	-2	-14
3000-4000	16	3500	-1000	-1	-16
4000-5000	20	4500	0	0	0
5000-6000	15	5500	1000	1	15
6000-7000	8	6500	2000	2	16
	∑f=70				∑fdm'=-11

As $\overline{\times}= \mathrm{A} + \frac{\sum \mathrm{fdm}'}{\sum \mathrm{f}} \times \mathrm{C}$

∴ $\overline{\times}= 4500 + \frac{(-11)}{70} \times 1000 \quad = 4500 - \frac{1100}{7}$

$$= 4500 - 157.14 \qquad = 4342.86$$

$$\overline{x} = 4342.86$$

CHECK YOUR PROGRESS- A

2. Calculate mean for the following data using the shortcut method.

700, 650, 550, 750, 800, 850, 650, 700, 950

3. Following is the height of students of class tenth of a school. Find out the mean height of the students.

Height in Inches	64	65	66	67	68	69	70	71	72	73
No. of students	1	6	10	22	21	17	14	5	3	1

4. Calculate A.M for the following frequency distribution of Marks.

Marks	5	10	15	20	25	30	35	40
No of students	5	7	9	10	8	6	5	2

5. Calculate mean for the following data

Marks	5-15	15-25	25-35	35-45	45-55	55-65
No of Students	8	12	6	14	7	3

Answers

1) 733.30	3) 20.48
2) 68.13 inches	4) 31.8

Another Special case of Continuous Series

1.7.4 Arithmetic Mean in case of Cumulative Frequency Series:

The normal continuous series gives frequency of the particular class. However, in case of cumulative frequency series, it does not give frequency of a particular class rather it gives the total of frequency including the frequency of preceding classes. Cumulative frequency series may be of two types: ' less than' type and 'more than' type. For calculating Arithmetic mean in cumulative frequency series, we convert such series into the normal frequency series and then apply the same method as in case of normal series.

Less than Cumulative Frequency Distribution

Example 8. Find the mean for the following frequency distribution:

Marks Less Than	10	20	30	40	50	60
No. of Students	5	15	40	70	90	100

Solution: Convert the given data into exclusive series:

Marks C. I.	No. of Students f	Mid Value m	$dm = m - A$ $A = 25$	$dm' = \frac{dm}{C}$ $C = 10$	fdm'
0-10	5	5	-20	-2	-10
10-20	15-5=10	15	-10	-1	-10
20-30	40-15=25	25	0	0	0
30-40	70-40=30	35	10	1	30
40-50	90-70=20	45	20	2	40
50-60	100-90=10	55	30	3	30
	∑f=100				∑fdm'=80

As $\overline{x} = A + \frac{\sum fdm'}{\sum f} \times C$

⇒ $\overline{x} = 25 + \frac{80}{100} \times 10 = 33$

⇒ $\overline{x} = 33$

More Than Cumulative Frequency Distribution

Example 9. Find the mean for the following frequency distribution

Marks More Than	0	10	20	30	40	50	60	70	80	90
No. of Students	80	77	72	65	55	43	28	16	10	8

Solution: Convert the given data into exclusive series

Marks C. I.	No. of Students f	Mid Value	$dm = m - A$ $A = 55$	$dm' = \frac{dm}{C}$ $C = 10$	fdm'
0-10	80-77=3	5	-50	-5	-15
10-20	77-72=5	15	-40	-4	-20
20-30	72-65=7	25	-30	-3	-21
30-40	65-55=10	35	-20	-2	-20
40-50	55-43=12	45	-10	-1	-12
50-60	43-28=15	55	0	0	0
60-70	28-16=12	65	10	1	12
70-80	16-10=6	75	20	2	12
80-90	10-8=2	85	30	3	6
90-100	8	95	40	4	32
	$\sum f = 80$				$\sum fdm' = -26$

As $\overline{x} = A + \frac{\sum fdm'}{\sum f} \times C$

∴ $\overline{\text{X}} = 55 + \frac{(-26)}{80} \times 10$

$= 55 - \frac{13}{4} \quad = \frac{220-13}{4} \quad = \frac{207}{4} = 51.75$

⇒ $\overline{\text{X}} = 51.75$

6.6.5 Arithmetic Mean in case of Unequal Class Interval Series:

Sometime the class interval between two classes is not same, for example 10-20, 20-40 etc. These series are known as unequal class interval series. However, it does not affect the finding of arithmetic mean as there is not precondition of equal class interval in case of arithmetic mean. So, mean will be calculated in usual manner.

Example 10. Calculate $\overline{\text{X}}$ if the data is given below:

C.I.	4-8	8-20	20-28	28-44	44-68	68-80
f	3	8	12	21	10	6

Solution:

C. I.	f	Mid Value
		m
			dm = m − A
			A = 26
				fdm

4-8	3	6	-20	-60
8-20	8	14	-12	-96
20-28	12	24	-2	-24
28-44	21	36	+10	210
44-68	10	56	+30	300
68-80	6	74	+48	288
	$\sum f = 60$			$\sum fdm = 618$

As $\overline{\text{X}} = A + \frac{\sum fdm}{\sum f}$

⇒ $\overline{\text{X}} = 26 + \frac{618}{60} \quad = 26 + 10.3 = 36.3$

⇒ $\overline{\text{X}} = 36.3$

1.7.6 Combined Arithmetic Mean:

Sometime we have the knowledge of mean of two or more series separately but we are interested in finding the mean that will be obtained by taking all these series as one series, such mean is called combined mean. It can be calculated using the following formula.

$$\overline{\mathbf{X}_{12}} = \frac{\mathbf{N_1}\overline{\mathbf{X_1}} + \mathbf{N_2}\overline{\mathbf{X_2}}}{\mathbf{N_1} + \mathbf{N_2}}$$

Where N_1 = Number of items in first series, N_2 = Number of of items in second series

$\overline{X_1}$ = Mean of first series, and $\overline{X_2}$ = Mean of second series

Example 11. Find the combined mean for the following data

	Firm A	Firm B
No. of Wage Workers	586	648
Average Monthly Wage (Rs.)	52.5	47.5

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

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

Where N_1 = Number of workers in Firm A, N_2 = Number of workers in Firm B

$\overline{X_1}$ = Mean wage of workers in Firm A, and $\overline{X_2}$ = Mean wage of workers in Firm B

We are given that

$$N_1 = 586 \qquad N_2 = 648$$

$$\overline{X_1} = 52.5 \qquad \overline{X_2} = 47.5$$

∴ Combined Mean, $\overline{X_{12}}$

$$= \frac{(586\times52.5)+(648\times47.5)}{586+648} = \frac{61{,}545}{1234} = \text{Rs. } 49.9$$

1.7.7 Correcting Incorrect Mean

Many a time it happens that we take some wrong items in the data or overlook some item. This results in wrong calculation of Mean. Later we find the correct values and we want to find out correct mean. This can be done using the following steps:

1. Multiply the incorrect mean of the data (incorrect $\overline{X}$) with number of items to find out incorrect $\sum\overline{X}$.
2. Now subtract all the wrong observation from the above values and add the correct observation to the above value to find out correct $\sum\overline{X}$.
3. Now divide the correct $\sum\overline{X}$. with the number of observations to find correct mean.

Example12. Mean wage of 100 workers per day found to be 75. But later on, it was found that the wages of two laborer's Rs. 98 and Rs. 69 were misread as Rs. 89 and Rs. 96. Find out the correct mean wage.

Solution: We know that, Correct $\sum X$ = Incorrect $\sum X -$ (Incorrect items) + (Correct Items)

Also $\overline{\mathrm{X}} = \frac{\sum \mathrm{X}}{\mathrm{N}}$

⇒ Incorrect $\sum \mathrm{X} = 100 \times 75 = 7500$

∴ Correct $\sum \mathrm{X} = 7500 - (89 + 96) + (98 + 69) \quad = 7482$

⇒ Correct $\overline{\mathrm{X}} = \frac{\text{Correct } \sum \mathrm{X}}{\mathrm{N}} \quad = \frac{7482}{100} = 74.82$

Determination of Missing Frequency

Example 13. Find the missing frequencies of the following series, if $\overline{\mathbf{X}} = \mathbf{33}$ and $\mathbf{N} = \mathbf{100}$

X	5	15	25	35	45	55
f	5	10	?	30	?	10

Solution: Let the missing frequencies corresponding to $\mathrm{X} = 25$ and $\mathrm{X} = 45$ be ‘f_1’ and ‘f_2’ respectively.

X	f	fX
5	5	25
15	10	150
25	f_1	$25\mathrm{f}_1$
35	30	1050
45	f_2	$45\mathrm{f}_2$
55	10	550
	$\sum$f=55+f_1+f_2	$\sum$fX=1775+25f_1+45f_2

Now, $\mathrm{N} = 100$ (Given)

∴ $55 + \mathrm{f}_1 + \mathrm{f}_2 = 100$

⇒ $\mathrm{f}_1 + \mathrm{f}_2 = 45$ …(i)

Also $\overline{\mathrm{X}} = \frac{\sum \mathrm{fX}}{\mathrm{N}}$

⇒ $33 = \frac{1775 + 25\mathrm{f}_1 + 45\mathrm{f}_2}{100}$

⇒ $3300 = 1775 + 25\mathrm{f}_1 + 45\mathrm{f}_2$

⇒ $25\mathrm{f}_1 + 45\mathrm{f}_2 = 1525$ …(ii)

Solving (i) and (ii), we get

$25 \times (\mathrm{f}_1 + \mathrm{f}_2 = 45) \quad \Rightarrow 25\mathrm{f}_1 + 25\mathrm{f}_2 = 1125$

$1 \times (25\mathrm{f}_1 + 45\mathrm{f}_2 = 1525) \quad \Rightarrow \underline{25\mathrm{f}_1 + 45\mathrm{f}_2 = 1525}$

$(-) \quad (-) \quad (-)$

$$-20f_2 = -400$$

$$f_2 = \frac{400}{20} = 20$$

$\therefore \qquad f_2 = 20$

Put $\qquad f_2 = 20$ in (i)

$$f_1 + 20 = 45$$

$\Rightarrow \qquad f_1 = 45 - 20 = 25$

$\therefore \qquad f_1 = 25$

$\therefore \qquad f_1 = 25,\ f_2 = 20$

1.7.8 Properties of Arithmetic mean

1. If we take the deviations of the observations from its Arithmetic mean and then sum up such deviations, then sum of such deviations will always be zero.
2. If we take the square of the deviations of items from its Arithmetic mean and then sum up sum of squares, the value obtained will always be less than the square of deviation taken from any other values.
3. If we have separate mean of two series, we can find the combined mean of the series.
4. If the value of all items in that data is increased or decreased by some constant value say 'k', then the Arithmetic mean is also increased or decreased by same 'k'. In other words, if k is added to the items, then actual mean will be calculated by deducting that k from the mean calculated.
5. If value of all items in the series is divided or multiplied by some constant 'k' then the mean is also multiplied or divided by the same constant 'k'. In other words, if we multiply all observations by 'k' then actual mean can be calculated by dividing the mean to obtained by the constant 'k'.

1.7.9 Merits and Limitations of Arithmetic Mean

- Arithmetic mean is very simple to calculate and it is also easy to understand.
- It is most popular method of calculating the average.
- Arithmetic mean is rigidly defined means it has a particular formula for calculating the mean.
- Arithmetic mean is comparatively less affected by fluctuation in the sample.
- It is most useful average for making comparison.

- We can perform further treatment on Arithmetic mean.
- We need not to have grouping of items for calculating Arithmetic mean.
- Arithmetic mean is based on all the values of the data.

Limitations of Arithmetic Mean

- The biggest limitation of Arithmetic mean is that it is being affected by extreme values.
- If we have open end series, it is difficult to measure Arithmetic mean.
- In case of qualitative data, it is not possible to calculate Arithmetic mean.
- Sometime it gives absurd result like we say that there are 20 students in one class and 23 students in other class then average number of students in a class is 21.5, which is not possible because student cannot be in fraction.
- It gives more importance to large value items than small value items.
- Mean cannot be calculated with the help of a graph.
- It cannot be located by just inspections of the items.

CHECK YOUR PROGRESS -B

1. From the following data, find the average sale per shop.

Sales in '000 (units)	10-12	13-15	16-18	19-21	22-24	25-27	28-30
No. shops	34	50	85	60	30	15	7

2. For the following data (Cumulative Series), find the average income.

Income Below in (Rs.)	30	40	50	60	70	80	90
No. of persons	16	36	61	76	87	95	100

3. Calculate the average marks for the following cumulative frequency distribution.

Marks Above	0	10	20	30	40	50	60	70	80	90
No of students	80	77	72	65	55	43	28	16	10	8

4. For a group of 50 male workers, their average monthly wage Rs.6300 and for a group of 40 female workers this average is Rs.5400. Find the average monthly wage for the combined group of all the workers.

5. The average marks of 100 students is given to be 45. But later on, it was found that the marks of students getting 64 was misread as 46. Find the correct mean.

6. Find missing frequency when the mean is 35 and number is 68.

X:	0-10	10-20	20-30	30-40	40-50	50-60
F:	4	10	12	?	20	?

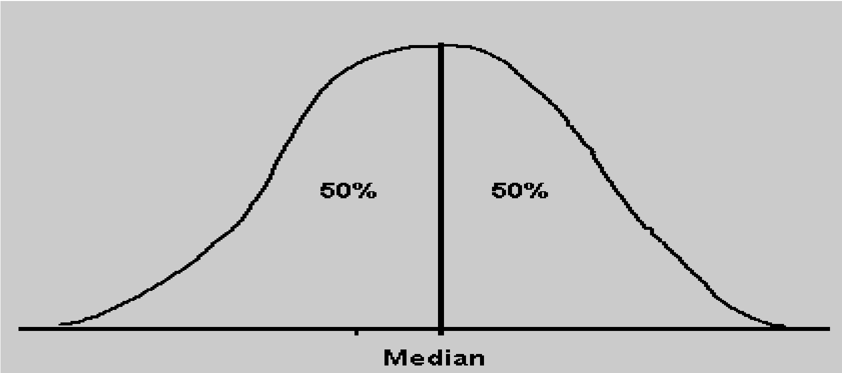
7. The mean age of a combined group of men and women is 30 years. The mean age of group of men is 32 years and women is 27 years. Find the percentage of men and women in the group

Answers

1) 17.8 (in 000 units)	4) 5900	7) Men 60%
2) 48	5) 45.18	
3) 51.75	6) 10,12	

1.8 MEDIAN: MEANING

Median is the positional measure of Central tendency. It means the median does not depend upon the value of the item under the observation, rather it depends on the position of the item in the series. Median is a value that divides the series exactly into two equal parts, it means 50% of the observations lie below the median and 50% of the observations lie above the median. However, it is important to arrange the series in ascending or descending order before calculating Median. If the series is not arranged, then Median cannot be calculated



For calculating Median

1. Series should be in ascending or descending order.
2. Series should be exclusive, not inclusive.

1.9 MEDIAN FOR DIFFERENT SERIES

1.9.1 Median in case of Individual series

For calculating the median in individual series, following are the steps:

1. Arrange the series in ascending or descending order

2. Calculate the number of observations. It is denoted by N
3. Calculate the $\left(\frac{N+1}{2}\right)^{th}$ term
4. Corresponding value to this item is the median of the data
5. In case there are even number of items in the series, this value will be in fraction. In that case take the arithmetic mean of the adjacent items in which Median is falling. For example, if it is 4.5 than take arithmetic mean of 4th item and 5th item

$$\text{Median = value of } \left(\frac{N+1}{2}\right)^{th} \text{ term}$$

When the number of observations N is odd

Example 1. Calculation median from the following observations:

15, 17, 19, 22, 18, 47, 25, 35, 21

Solution: Arranging the given items in ascending order, we get

$$15,\ 17,\ 18,\ 19,\ 21,\ 22,\ 25,\ 35,\ 47$$

Now $\quad$ $\text{Median, M} = \text{Size of } \left(\frac{N+1}{2}\right)^{th} \text{ item}$

$$M = \text{Size of } \left(\frac{9+1}{2}\right)^{th} \text{ item}$$

$$= \text{Size of } 5^{th} \text{ item}$$

$$= 21$$

$\Rightarrow \quad M = 21$

When the number of observations N is even

Example 2. Find median from the following data

28, 26, 24, 21, 23, 20, 19, 30

Solution: Arranging the given figures in ascending order, we get

$$19,\ 20,\ 21,\ 23,\ 24,\ 26,\ 28,\ 30$$

Now $\quad$ $\text{Median, M} = \text{Size of } \left(\frac{N+1}{2}\right)^{th} \text{ item}$

$$M = \text{Size of } \left(\frac{8+1}{2}\right)^{th} \text{ item}$$

$$= \text{Size of } 4.5^{th} \text{ item}$$

$$= \frac{4^{th} \text{ item} + 5^{th} \text{ item}}{2} \quad = \frac{23+24}{2} = \frac{47}{2} = 23.5$$

$\Rightarrow \quad M = 23.5$

1.9.2 Median in case of Discrete series

Following are the steps in case of discrete series:

1. Arrange the data in ascending or descending order.

2. Find the cumulative frequency of the series.
3. Find the $\left(\frac{N+1}{2}\right)^{th}$ term
4. Now look at this term in the cumulative frequency of the series.
5. Value against which such cumulative frequency falls is the median value.

$$\textbf{Median = value of } \left(\frac{\mathbf{N+1}}{\mathbf{2}}\right)^{\mathbf{th}} \textbf{ term}$$

Example 3. Calculate the value of median, if the data is as given below:

Height (in cms.)	110	125	250	200	150	180
No. of Students	8	12	3	10	13	15

Solution: Arranging the given data in ascending order, we get

Height (in cms.)	No. of Students f	Cumulative Frequency $C \cdot f$
110	8	8 (1-8)
125	12	20 (9-20)
150	13	33 (21-33)
180	15	48 (34-48)
200	10	58 (49-58)
250	3	61 (59-61)
	$\sum f = N = 61$	

Now Median, $M = \text{Size of} \left(\frac{N+1}{2}\right)^{th}$ item

$$M = \text{Size of} \left(\frac{6+1}{2}\right)^{th} \text{ item}$$
$$= \text{Size of } 31^{st} \text{ item}$$
$$= 150$$

$\Rightarrow$ Median, $M = 150$ cms.

1.9.3 Median in case of Continuous Series

Following are the steps in case of continuous series:

1. Arrange the data in ascending or descending order.
2. Find the cumulative frequency of the series.
3. Find the $\left(\frac{N}{2}\right)^{th}$ term
4. Now look at this term in the cumulative frequency of the series. The value equal to or higher than term calculated in third step is the median class.

5. Find median using following formula.

$$M = L + \frac{\frac{N}{2} - C \cdot f}{f} \times i$$

Where M = Median, L = Lower Limit of Median Class, N = Number of Observations.

c.f. = Cumulative frequency of the Median Class, f = Frequency of the class preceding Median Class, i = Class interval of Median Class

Example 4. Calculate Median

Marks	5-10	10-15	15-20	20-25	25-30	30-35
No. of Students	8	7	14	16	9	6

Solution:

C.I.	No. of Students (f)	C.f
5-10	8	8 (1-8)
10-15	7	15 (9-15)
15-20	14	29 (16-29)
20-25	16	45 (30-45)
25-30	9	54 (46-54)
30-35	6	60 (55-60)
	∑f=N=60	

$$\text{Median, } M = \text{Size of } \left(\frac{N}{2}\right)^{\text{th}} \text{ item}$$

$$M = \text{Size of } \left(\frac{60}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 30^{\text{th}} \text{ item}$$

⇒ Median lies in the class interval $20 - 25$

As $\text{Median, } M = L + \frac{\frac{N}{2} - C \cdot f}{f} \times i$

Here $L = \text{Lower limit of the median class} = 20$

$N = 60, \quad C \cdot f = 29, \quad f = 16$

$i = \text{Class – length of the median class} = 5$

∴ $M = 20 + \frac{(30-29)}{16} \times 5$

$$= 20 + \frac{5}{16} \qquad = 20 + 9.312 = 29.312$$

$\Rightarrow \qquad M = 29.312$

Inclusive Series – It must be converted to Exclusive Series before the Median is calculated.

Example 5. Find Median from the given data

X	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89
f	6	53	85	56	21	16	4	4

Solution: Converting the given data into exclusive form, we get

$$\left[\text{Correction factor} = \frac{L_2 - U_1}{2} = \frac{20-19}{2} = \frac{1}{2} = 0.5\right]$$

(0.5 is subtracted from all lower limits and added to all upper limits)

X	f	Cumulative frequency $C \cdot f$
9.5-19.5	6	6
19.5-29.5	53	59
29.5-39.5	85	144
39.5-49.5	56	200
49.5-59.5	21	221
59.5-69.5	16	237
69.5-79.5	4	241
79.5-89.5	4	245
	∑f=N=245	

Median, $M = \text{Size of} \left(\frac{N}{2}\right)^{\text{th}}$ item

$M = \text{Size of} \left(\frac{245}{2}\right)^{\text{th}}$ item

$= \text{Size of } 122.5^{\text{th}}$ item

$\therefore \qquad$ The real class limits of the median class $= (29.5 - 39.5)$

So $\qquad M = L + \frac{\left(\frac{N}{2} - C\cdot f\right)}{f} \times i$

$\Rightarrow \qquad M = 29.5 + \left(\frac{122.5 - 59}{85}\right) \times 10$

$= 29.5 + \left(\frac{63.5}{85} \times 10\right)$

$= 29.5 + \left(\frac{635}{85}\right)$

$= 29.5 + 7.47 = 36.97$

⇒ $M = 36.97$

Cumulative Series (More than and less than)

Example 6. Find median, if the data is as given below:

Marks More than	20	35	50	65	80	95
No. of Students	100	94	74	30	4	1

Solution: Converting the given data into class – interval form, we get

Marks C. I.	Frequency f	Cumulative Frequency C · f
20-35	100-94=6	6
35-50	94-74=20	26
50-65	74-30=44	70
65-80	30-4=26	96
80-95	4-1=3	99
95-110	1	100
	$\sum f = N = 100$	

Now Median, $M = \text{Size of} \left(\frac{N}{2}\right)^{\text{th}}$ item

$$M = \text{Size of} \left(\frac{100}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 50^{\text{th}} \text{ item}$$

⇒ Median lies in the class interval $= 50 - 65$

So $M = L + \frac{\left(\frac{N}{2} - C \cdot f\right)}{f} \times i$

⇒ $M = 50 + \left(\frac{50-26}{44}\right) \times 15$

$$= 50 + \left(\frac{24}{44} \times 15\right)$$

$$= 50 + 8.18 = 58.18$$

⇒ $M = 58.18$

Example 7. Find median, if the data is as given below:

Marks Less than	10	20	30	40	50	60	70	80

No. of Students	20	30	50	94	96	127	198	250

Solution: Converting the given data into class interval form, we get

Marks C. I.	No. of Students f	Cumulative Frequency C · f
0-10	20	20
10-20	30-20=10	30
20-30	50-30=20	50
30-40	94-50=44	94
40-50	96-94=2	96
50-60	127-96=31	127
60-70	198-127=71	198
70-80	250-198=52	250
	$\sum \mathrm{f} = \mathrm{N} = 250$	

Now $\quad$ Median, $\mathrm{M} = \text{Size of} \left(\frac{\mathrm{N}}{2}\right)^{\mathrm{th}}$ item

$$\mathrm{M} = \text{Size of} \left(\frac{250}{2}\right)^{\mathrm{th}} \text{ item}$$

$$= \text{Size of } 125^{\mathrm{th}} \text{ item}$$

⇒ $\quad$ Median lies are the class – interval $= 50 - 60$

So $\quad$ $\mathrm{M} = \mathrm{L} + \frac{\frac{\mathrm{N}}{2} - \mathrm{C{\cdot}f}}{\mathrm{f}} \times \mathrm{i}$

⇒ $\quad$ $\mathrm{M} = 50 + \left(\frac{125-96}{31}\right) \times 10$

$$= 50 + \left(\frac{29}{31} \times 10\right)$$

$$= 50 + \frac{290}{31} \qquad = 50 + 9.35 = 59.35$$

⇒ $\quad$ $\mathrm{M} = 59.35$

Mid – Value Series

Example 8. Find the value of median for the following data:

Mid Value	15	25	35	45	55	65	75	85	95
f	8	26	45	72	116	60	38	22	13

Solution: It is clear from the mid – value that the class size is 10. For finding the limits of different classes, apply the formula:

$\mathrm{L} = \mathrm{m} - \frac{\mathrm{i}}{2}$ and $\mathrm{U} = \mathrm{m} + \frac{\mathrm{i}}{2}$

Where, L and U denote the lower and upper limits of different classes, 'm' denotes the mid – value of the corresponding class interval and 'i' denotes the difference between mid values.

∴ Corresponding to mid – value '15', we have

$\mathrm{L} = 15 - \frac{10}{2}$ and $\mathrm{U} = 15 + \frac{10}{2}$

i. e. $\mathrm{C.I.} = 10 - 20$

Similarly other class intervals can be located

Mid Value	f	C. I.	Cumulative Frequency $\mathrm{C} \cdot \mathrm{f}$
15	8	10-20	8
25	26	20-30	34
35	45	30-40	79
45	72	40-50	151
55	116	50-60	267
65	60	60-70	327
75	38	70-80	365
85	22	80-90	387
95	13	90-100	400
	N=100		

Now Median, $\mathrm{M} = \text{Size of } \left(\frac{\mathrm{N}}{2}\right)^{\mathrm{th}}$ item

$\mathrm{M} = \text{Size of } \left(\frac{400}{2}\right)^{\mathrm{th}}$ item

$= \text{Size of } 200^{\mathrm{th}}$ item

⇒ Median lies is the class – interval $= 50 - 60$

So $\mathrm{M} = \mathrm{L} + \frac{\frac{\mathrm{N}}{2} - \mathrm{C} \cdot \mathrm{f}}{\mathrm{f}} \times \mathrm{i}$

⇒ $\mathrm{M} = 50 + \left(\frac{200-151}{116}\right) \times 10$

$= 50 + \left(\frac{49}{116} \times 10\right)$

$= 50 + \frac{490}{116}$

$= 50 + 4.224 = 54.224$

⇒ $\mathrm{M} = 54.224$

Determination of Missing Frequency

Example 9. Find the missing frequency in the following distribution if $\mathbf{N = 72, Q_1 = 25}$ and $\mathbf{Q_3 = 50}$

C.I.	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80

f	4	8	-	19	-	10	5	-

Solution: Let the missing frequencies be f_1, f_2 and f_3 respectively.

C. I.	f	Cumulative Frequency $C \cdot f$
0-10	4	4
10-20	8	12
20-30	f1	$12 + f_1$
30-40	19	$31 + f_1$
40-50	f2	$31 + f_1 + f_2$
50-60	10	$41 + f_1 + f_2$
60-70	5	$46 + f_1 + f_2$
70-80	f3	$46 + f_1 + f_2 + f_3$
	$N = 72 = \sum f$ $\sum f = 46 + f_1 + f_2 + f_3$	

Now $N = 72$

$= \sum f$

$= 46 + f_1 + f_2 + f_3$

$\Rightarrow$ $f_1 + f_2 + f_3 = 72 - 46 = 26$

$\Rightarrow$ $f_1 + f_2 + f_3 = 26$ …(i)

Also, $Q_1 = 25$ (Given)

$\Rightarrow$ Q_1 lies in the class – interval $20 - 30$

$$\Rightarrow \quad Q_1 = L + \frac{\frac{N}{4} - C \cdot f}{f} \times i$$

$$25 = 20 + \frac{\frac{72}{4} - 12}{f_1} \times 10$$

$$25 = 20 + \frac{18 - 12}{f_1} \times 10$$

$$25 - 20 = \frac{6}{f_1} \times 10$$

$$5f_1 = 60$$

$$f_1 = \frac{60}{5}$$

$\Rightarrow$ $f_1 = 12$ …(ii)

Similarly, we are given that

$Q_3 = 50$

$\Rightarrow$ Q_3 lies in the class – interval $50 - 60$

$$\Rightarrow \quad Q_3 = L + \frac{\frac{3N}{4} - C \cdot f}{f} \times i$$

$$50 = 50 + \frac{\frac{3\times72}{4}-(31+f_1+f_2)}{10} \times 10$$

$$50 = 50 + \frac{54-(31+12+f_2)}{1} \qquad (\because f_1 = 12 \text{ By (ii)})$$

$$50 - 50 = 54 - (43 + f_2)$$

$$0 = 54 - (43 + f_2)$$

$$43 + f_2 = 54$$

$$f_2 = 54 - 43$$

$$\Rightarrow \qquad f_2 = 11 \qquad \ldots\text{(iii)}$$

Putting (ii) and (iii) in (i), we get

$$f_1 + f_2 + f_3 = 26$$

$$12 + 11 + f_3 = 26$$

$$23 + f_3 = 26$$

$$f_3 = 26 - 23$$

$$\Rightarrow \qquad f_3 = 3$$

1.9.4 Merits and Limitations of Median

- Median is easy to calculate.
- It is capable of Graphic presentation.
- It is possible even in case of open-ended series.
- This is rigidly defined.
- It is not affected by extreme values.
- In case of qualitative data, it is very useful**.**

Limitations of Median

- It is not capable of further algebraic treatment.
- It is a positional average and is not based on all observations.
- It is very much affected by fluctuation in sampling.
- Median needs arrangement of data before calculation.
- In the case of continuous series, it assumes that values are equally distributed in a particular class.

TEST YOUR PROGRESS- C

1. Calculate Median

30, 45,75, 65, 50, 52, 28, 40, 49, 35, 52

2. Find Median

Wages:	100	150	80	200	250	180
No. of workers	24	26	16	20	6	30

3. Calculate Median

X;	0-5	5-10	10-15	15-20	20-25	25-30	30-35

F: 4 6 10 16 12 8 4

4. Find Median:

Income	100-200	200-400	400-700	700-1200	1200-2000
Number of firms	40	100	260	80	20

5. Find missing frequency when median is 50 and number is 100.

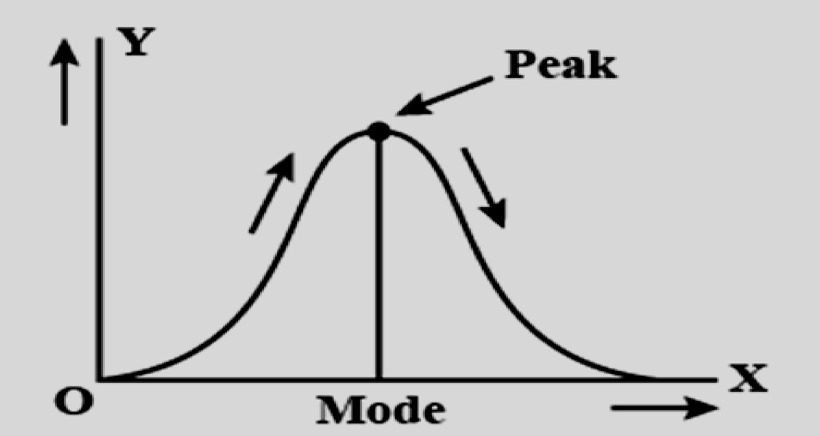
X;	0-20	20-40	40-60	60-80	80-100
F:	1 4	?	27	?	15

Answers

1) 49	4) 526.9
2) 150	5) 23,21
3) 18.125	

1.10 MODE: MEANING

Mode is another positional measure of Central Tendency. Mode is the value that is repeated most number of time in the series.   In other words, the value having highest frequency is called Mode. The term ‘Mode’ is taken from French word ‘La Mode’ which means the most fashionable item. So, Mode is the most popular item of the series.



For calculating Mode

1. Series should be in ascending or descending order.
2. Series should be exclusive, not inclusive.

Series should have equal class intervals.

1.10.1 Mode in Individual Series

In case of Individual series, following are the steps of finding the Mode.

1. Arrange the series either in ascending order or descending order.
2. Find the most repeated item.
3. This item is Mode.

Example 1. Calculate mode from the following data of marks obtained by 10 students

S. No.	1	2	3	4	5	6	7	8	9	10
Marks obtained	10	27	24	12	27	27	20	18	15	30

Solution: By Inspection

It can be observed that 27 occur most frequently i. e. 3 times. Hence, mode = 27 marks

By converting into discrete series

Marks Obtained	Frequency
10	1
12	1
15	1
18	1
20	1
24	1
27	3
30	1
	N=10

Since, the frequency of 27 is maximum i. e. 3

It implies the item 27 occurs the maximum number of times. Hence the modal marks are 27.

Mode = 27

7.3.2 Mode in discrete series:

In case of discrete series, we can find mode by two methods that are Observation Method and Grouping Method.

1. **Observation Method**: Under this method value with highest frequency is taken as mode.
2. **Grouping Method**: Following are the steps of Grouping method:
   - Prepare a table and put all the values in the table in ascending order.
   - Put all the frequencies in first column. Mark the highest frequency.
   - In second column put the total of frequencies taking two frequencies at a time like first two, then next two and so on. Mark the highest total.
   - In third column put the total of frequencies taking two frequencies at a time but leaving the first frequency like second and third, third and fourth and so on. Mark the highest total.
   - In fourth column put the total of frequencies taking three frequencies at a time like first three, then next three and so on. Mark the highest total.
   - In fifth column put the total of frequencies taking three frequencies at a time but leaving the first frequency like second, third and fourth; than fifth, sixth and seventh and so on. Mark the highest total.

- In sixth column put the total of frequencies again taking three frequencies at a time but leaving the first two frequencies. Mark the highest total.
- Value that is marked highest number of times is the mode.

Example 2. Find the modal value for the following distribution

Age (in years)	8	9	10	11	12	13	14	15
No. of Persons	5	6	8	7	9	8	9	6

Solution: Here, as maximum frequency 9 belongs to two age values 12 and 14, so its not possible to determine mode by inspection. We will have to determine the modal value through grouping and analysis table.

Grouping Table

Age (in years)	Frequency					
	G_1	G_2	G_3	G_4	G_5	G_6
8	5	11	14	19	21	
9	6					24
10	8	15				
11	7		16	24		
12	9	17			26	23
13	8		17			
14	9	15				
15	6					

Analysis Table

Group No.	8	9	10	11	12	13	14	15
G_1					×		×	
G_2					×	×		
G_3						×	×	
G_4				×	×	×		
G_5					×	×	×	
G_6			×	×	×			
Total	×	×	1	2	5	4	3	×

Since, 12 occurs maximum number of times i. e. 5 times, the modal age is 12 years

$$\text{Mode} = 12$$

1.10.3 Mode in Continuous series: In case of continuous series, we can find mode by two methods that are Observation Method and Grouping Method.

1. **Observation Method**: Under this method value with highest frequency is taken as mode class than the mode formula is applied which is given below.
2. **Grouping Method**: Following are the steps of Grouping method:

- Prepare a table and put all the classes of data in the table in ascending order.
- Put all the frequencies in first column. Mark the highest frequency.
- In second column put the total of frequencies taking two frequencies at a time like first two, then next two and so on. Mark the highest total.
- In third column put the total of frequencies taking two frequencies at a time but leaving the first frequency like second and third, third and fourth and so on. Mark the highest total.
- In fourth column put the total of frequencies taking three frequencies at a time like first three, then next three and so on. Mark the highest total.
- In fifth column put the total of frequencies taking three frequencies at a time but leaving the first frequency like second, third and fourth; than fifth, sixth and seventh and so on. Mark the highest total.
- In sixth column put the total of frequencies again taking three frequencies at a time but leaving the first two frequencies. Mark the highest total.
- Class that is marked highest number of times is the mode class.
- Apply following formula for calculating the mode:

$$\mathbf{Z = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i}$$

Where, Z = Mode, L = Lower limit of the mode class

f_m = Frequency of mode class, f_1 = Frequency of class preceding mode class
f_2 = Frequency of class succeeding mode class, i = Class interval

Example 3. Find the mode for the following frequency distribution

Age (in years)	30-35	35-40	40-45	45-50	50-55	55-60
No. of Persons	3	8	12	20	15	2

Solution: Here, the maximum frequency is corresponding to the class – interval $45 - 50$.

So, the modal class is $45 - 50$.

Now, the mode is given by the formula

$$\text{Mode, } Z = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i$$

Here L = Lower limit of modal class $= 45$

f_m = Frequency of modal class $= 20$

f_1 = Frequency of class preceeding the modal class $= 12$

f_2 = Frequency of class succeeding the modal class $= 15$

i = Class length of modal class $= 5$

$$\therefore \quad \text{Mode, } Z = 45 + \frac{20 - 12}{(2 \times 20) - 12 - 15} \times 5$$

$$= 45 + \frac{8}{40-27} \times 5$$

$$= 45 + 3.07$$

$= 48.1$ years (approx.)

⇒ $\quad Z = 48.1$ year

Example 4. Calculate mode from the following data

C.I.	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
f	2	9	10	13	11	6	13	7	4	1

Solution: Here as it is not possible to find modal class by inspection, so we have to determine it through grouping and analysis table.

Grouping Table						
C. I.	Frequency					
	G_1	G_2	G_3	G_4	G_5	G_6
0-10	2	11		21		
10-20	9		19		32	
20-30	10	23				34
30-40	13		24	30		
40-50	11	17			30	
50-60	6		19			26
60-70	13	20		24		
70-80	7		11		12	
80-90	4	5				
90-100	1					

Analysis Table										
Group No.	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
G_1				×			×			
G_2			×	×						
G_3				×	×					
G_4				×	×	×				
G_5		×	×	×						
G_6			×	×	×					
Total	×	1	3	6	3	1	1	×	×	×

Clearly the modal class is $30 - 40$

Now the mode is given by the formula

Mode, $Z = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i$

Here $L =$ Lower limit of modal class $30 - 40 = 30$

$f_m =$ Frequency corresponding to modal class $= 13$

$f_1 =$ Frequency of interval preceding modal class

$f_2 =$ Frequency of interval succeeding and

$i =$ Class length of modal class

∴ Mode, $Z = 30 + \frac{13-10}{(2\times13)-10-11} \times 10$

$= 30 + \frac{3}{26-21} \times 10$

$= 30 + \frac{30}{5}$

$= 30 + 6 \qquad = 36$

⇒ $Z = 36$

Example 5. Determine the missing frequencies when it is given that N=230, Median, M=233.5 and Mode, Z=234.

C.I	200-210	210-220	220-230	230-240	240-250	250-260	260-270
f	4	16	-	-	-	6	4

Solution: Let the missing frequencies be f_1, f_2 and f_3 respectively.

C.I	f	C · f
200-210	4	4
210-220	16	20
220-230	f_1	$20+f_1$
230-240	f_2	$20+f_1+f_2$
240-250	f_3	$20+f_1+f_2+f_3$
250-260	6	$26+f_1+f_2+f_3$
260-270	4	$30+f_1+f_2+f_3$
	$N = 230 = \sum f$ $\sum f = 30 + f_1 + f_2 + f_3$	

Now $N = 230 = \sum f$ (Given)

$= 30 + f_1 + f_2 + f_3$

⇒ $f_1 + f_2 + f_3 = 230 - 30 = 200$

⇒ $f_1 + f_2 + f_3 = 200$ …(i)

Also, Median $= 233.5$ (Given)

⇒ Median class is $230 - 240$

⇒ $M = L + \frac{\frac{N}{2} - C\cdot f}{f} \times i$

$233.5 = 230 + \frac{\frac{230}{2}-(20+f_1)}{f_2} \times 10$

$3.5 = \frac{115-20-f_1}{f_2} \times 10$

$3.5f_2 = 950 - 10f_1$

⇒ $10f_1 + 3.5f_2 = 950$ …(ii)

Now Mode $= 234$ lies in $230 - 240$

∴ $Z = L + \frac{f_2-f_1}{2f_2-f_1-f_3} \times i$

⇒ $234 = 230 + \frac{f_2-f_1}{2f_2-f_1-f_3} \times 10$

⇒ $4 = \frac{f_2-f_1}{2f_2-f_1-(200-f_1-f_2)} \times 10$ [Using (i)]

⇒ $4 = \frac{f_2-f_1}{2f_2-f_1-200-f_1-f_2} \times 10$

⇒ $4 = \frac{(f_2-f_1)\times 10}{3f_2-200}$

⇒ $12f_2 - 800 = 10f_2 - 10f_1$

⇒ $2f_2 - 800 + 10f_1 = 0$

⇒ $10f_1 + 2f_2 = 800$ …(iii)

Solving (ii) and (iii), we get

$$\begin{array}{l}
10f_1 + 3.5f_2 = 950 \\
10f_1 + 2f_2 \quad = 800 \\
(-) \quad (-) \quad (-) \\
\hline
\qquad 1.5f_2 = 150
\end{array}$$

⇒ $f_2 = \frac{150}{1.5} = 100$

$f_2 = 100$ …(iv)

Put (iv) in (iii)

$10f_1 + 2(100) = 800$

⇒ $10f_1 = 800 - 200 = 600$

⇒ $10f_1 = 600$

⇒ $f_1 = 60$ …(v)

Put (iv) and (v) in (i)

$60 + 100 + f_3 = 200$

⇒ $f_3 = 40$

∴ The missing frequencies are 60, 100 and 40.

1.10.4 Merits and Limitations of Mode

- Mode is easy to calculate.
- People can understand this in routine life.
- It is capable of Graphic presentation.
- It is possible even in case of open-end series.

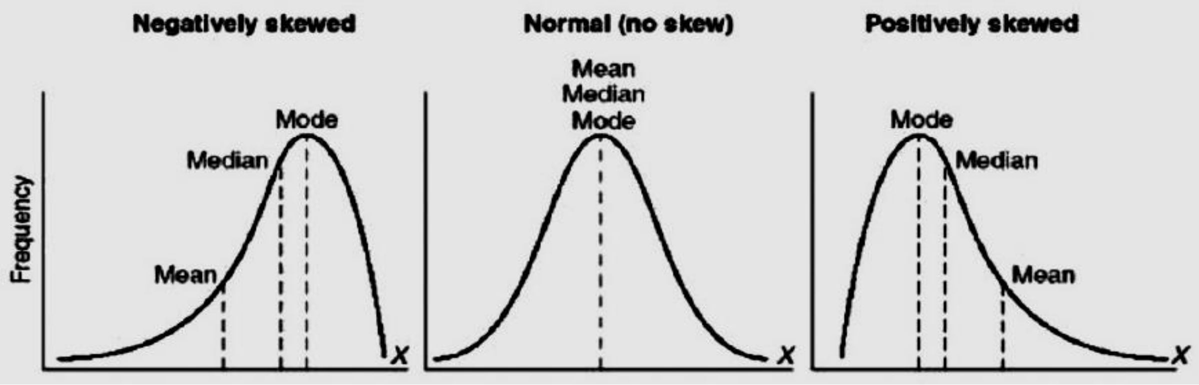
- This is rigidly defined.
- It is not affected by extreme values.
- In case of qualitative data, it is very useful**.**

Limitations of Mode

- It is not always determinable as series may be Bi-modal or Tri-modal.
- It is not capable of further algebraic treatment.
- It is positional average and is not based on all observation.
- It is very much affected by fluctuation in sampling.
- Mode needs arrangement of data before calculation.

1.11 RELATION BETWEEN MEAN, MEDIAN AND MODE

In a normal series the value of Mean, Median and Mode is always the same. However, Karl Pearson studied the empirical relation between the Mean, Median and Mode and found that in moderately skewed series the Median always lies between the Mean and the Mode. Normally it is two-thirds distance from Mode and one-third distance from Mean.



On the basis of this relation following formula emerged

Mode = 3 Median – 2 Mean
or $\boldsymbol{Z = 3M - 2\overline{X}}$

Example 6. Calculate $\mathbf{M}$ when $\overline{\mathbf{X}}$ and Z of a distribution are given to be $\mathbf{35.4}$ and $\mathbf{32.1}$ respectively.

Solution: We are given that

Mean, $\overline{\mathrm{X}} = 35.4$
Mode, $\mathrm{Z} = 32.1$

As, we know the empirical relation between Mean, Median and Mode.

i. e. $\quad$ Mode = 3 Median – 2 Mode

$\Rightarrow \quad \mathrm{Z} = 3\mathrm{M} - 2\overline{\mathrm{X}}$

⇒ $\quad M = \frac{1}{3}\left(Z + 2\overline{X}\right)$

⇒ $\quad M = \frac{1}{3}\left(32.1 + 2(35.4)\right)$

$\quad = \frac{1}{3}(32.1 + 70.8)$

$\quad = \frac{1}{3}(102.9) = 34.3$

⇒ $\quad$ Median, $M = 34.3$

CHECK YOUR PROGRESS - D

1. Find Mode:

X: 22, 24, 17, 18, 19, 18, 21, 20, 21, 20, 23, 22, 22, 22

2. Find Mode by inspection method

X	6	12	18	24	30	36	42	48
f	9	11	25	16	9	10	6	3

3. Find Mode by Grouping Method

X	21	22	25	26	27	28	29	30
F	7	10	15	18	13	7	3	2

4. Calculate mode using grouping and analysis methods.

X	100-110	110-120	120-130	130-140	140-150	150-160	160-170	170-180
f	4	6	20	32	33	17	8	2

Answers

1) 22	3) 26
2) 18	4) 56.46

1.12 SUM UP

Average is the single value that represents its series. Average is also known as Central Tendency. Five types of average Arithmetic Mean, Geometric Mean, Harmonic Mean, Median, Mode. Average or Central Tendency is one such technique that is widely used in statistics. It can be calculated in different ways likewise short cut method, direct method, and step deviation method. One can also able to determine combined average for more than one series. Incorrect mean can be corrected with the help of a correct arithmetic mean formula. Median divides the series in two equal parts. Mode is value repeated most number of time. There is an existence of relationship between mean medium and mode.

1.13 QUESTIONS FOR PRACTICE

Q1. What is central tendency? What are the uses of measuring central tendency?

Q2. What are the objectives and functions of the average?

Q3. Explain the features of good average.

Q4. What is average? Give uses and limitations of average.

Q5. What is arithmetic mean? How it is calculated.

Q6. Give properties, advantages and limitations of Arithmetic mean.

Q7. How you can calculate combined arithmetic mean.

Q8. What is median? How it is calculated?

Q9. Give merits and limitations of Median.

Q10. What is mode? How it is calculated. Give its merits and limitations.

Q11. Explain the grouping method of calculating Mode.

Q12. Give relation between Mean, Median and Mode.

Q13. What is positional average. Give various positional average.

1.14 SUGGESTED READINGS

- J. K. Sharma, Business Statistics, Pearson Education.
- S.C. Gupta, Fundamentals of Statistics, Himalaya Publishing House.
- S.P. Gupta and Archana Gupta, Elementary Statistics, Sultan Chand and Sons, New Delhi.
- Richard Levin and David S. Rubin, Statistics for Management, Prentice Hall of India, New Delhi.

M.A (ECONOMICS)

SEMESTER -II

QUANTITATIVE METHODS II

Unit 2: Dispersion

STRUCTURE

2.0 LEARNING OBJECTIVES

After studying the Unit, students will be able to:

- Explain the meaning of Dispersion
- Compare absolute and relative measures of Dispersion
- Understand features of a good measure of Dispersion

- Calculate the Range and Quartile Deviation
- Measure the Dispersion using Mean and Standard Deviation
- Compare the variation of the two series.
- concept of Coefficient of Variation (CV)
- calculation of Coefficient of Variation

2.1 MEANING OF DISPERSION

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

Series X	Series Y	Series Z
100	94	1
100	105	2
100	101	3
100	98	4
100	102	490
$\sum X = 500$	$\sum Y = 500$	$\sum Z = 500$
$\bar{X} = \frac{\sum X}{N} = \frac{500}{5} = 100$	$\bar{Y} = \frac{\sum Y}{N} = \frac{500}{5} = 100$	$\bar{Z} = \frac{\sum Z}{N} = \frac{500}{5} = 100$

We can see that in all the above series the average is 100. However, in the first series average fully represents its data as all the items in the series are 100 and average is also 100. In the second series, the items are very near to its average which is 100, so we can say that average is a good representation of the series. But in case of third series, average does not represent its data as there is a lot of difference between items and the average. In order to understand the nature of data it is very important to see the difference between items and the data. This could be done by using dispersion. Dispersion is a very important statistical tool that helps us in PROGRESS the nature of data. Dispersion shows the extent to which individual items in the data differ from its average. It is a measure of the difference between data and the individual items. It indicates how that are lacks uniformity. Following is some of the definitions of Dispersion. According to Simpson and Kafka, “The measures of the scatterness of a mass of figures in a series about an average is called a measure of variation, or dispersion”.

2.2 BENEFITS / USES OF DISPERSION

Benefits of Dispersion analysis are outlined as under:

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

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

2.3 FEATURES OF GOOD MEASURE OF DISPERSION

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

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

2.4 ABSOLUTE AND RELATIVE MEASURE OF DISPERSION

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

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

Absolute Measure of Dispersion

- Range
- Quartile Deviation
- Mean Deviation
- Standard Deviation

Relative Measure of Dispersion

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

2.5 MEASURE OF DISPERSION - RANGE

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

Range = Highest Value - Lowest Value

R = H – L

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

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

$$= \frac{\boldsymbol{H\ -\ L}}{\boldsymbol{H +\ L}}$$

A. Range in Individual Series:

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

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

Solution:

Range = Highest Value − Lowest Value

= 500 – 300 = 200

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

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

B. Range in Discrete Series:

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

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

Solution: Range = Highest Value − Lowest Value

= 500 – 300 = 200

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

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

C. Range in Continuous Series:

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

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

Solution: Range = Highest Value − Lowest Value

= 90 – 10 = 80

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

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

TEST YOUR PROGRESS (A)

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

28	110	27	77	19	94	63	25	111

2. Find Range and coefficient of Range

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

4. Calculate coefficient of Range:

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

Answers

1. 92, 0.7	3. .714
2. 35, 0.778	

2.6 MEASURE OF DISPERSION – QUARTILE DEVIATION

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

Quartile Range = $Q_3 – Q_1$

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

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

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

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

A. Quartile Deviation in Individual Series:

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

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

Solution:

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

= Value of 2nd item

= 330

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

= Value of 6th item

= 470

Quartile Range = $Q_3 - Q_1$

= 470 – 330 = 140

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

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

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

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

B. Quartile Deviation in Discrete Series:

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

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

Solution: Calculation of Quartile

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

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

= Value of 23rd item = 380

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

= Value of 69th item = 470

Quartile Range = $Q_3 - Q_1$

= 470 – 380 = 90

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

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

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

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

C. Quartile Deviation in Continuous Series:

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

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

Solution: Calculation of Quartile

Wage (Rs.) (X)	**No. of Workers (f)**	**Cumulative Frequency (cf)**
10-20	5	5
20-30	8	13
30-40	12	25
40-50	20	45

50-60	18	63
60-70	15	78
70-80	13	91
80-90	9	100

Calculation of Q_1

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

Q_1 Class = Value of 25th item

Q_1 Class = 30-40

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

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

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

Calculation of Q_3

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

Q_3 Class = Value of 75th item

Q_3 Class = 60-70

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

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

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

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

Quartile Range = $Q_3 - Q_1$

$= 68 - 40 \qquad = 28$

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

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

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

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

TEST YOUR PROGRESS (B)

1. Find Quartile deviation and coefficient of Quartile Deviation:

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

2. Find Quartile deviation and coefficient of Quartile Deviation:

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

3. Find Quartile deviation and coefficient of Quartile Deviation

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

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

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

Answers

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

2.7 MEASURE OF DISPERSION – MEAN DEVIATION

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

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

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

A. Mean Deviation in case of Individual Series:

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

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

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

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

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

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

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

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

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

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

Solution: Let us calculate Mean Median and Mode

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

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

= Value of 5th item = 9

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

Calculation of Deviations

Marks **X**	$\mathbf{D_{\bar{X}} = \|\,X - \bar{X}\,\|}$ **(Where $\bar{X}$ = 10)**	$\mathbf{D_M = \|\,X - M\,\|}$ **(Where M = 9)**	$\mathbf{D_Z = \|\,X - Z\,\|}$ **(Where Z = 8)**
5	5	4	3
7	3	2	1
8	2	1	0
8	2	1	0
9	1	0	1
11	1	2	3
13	3	4	5
14	4	5	6
15	5	6	7

	$\sum D_{\bar{X}}$ = 26	$\sum D_{M}$ = 25	$\sum D_{Z}$ = 26

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

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

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

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

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

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

B. Mean Deviation in case of Discrete Series:

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

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

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

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

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

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

Solution: Let us calculate Mean Median and Mode

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

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

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

= Value of 50.5 item = 410

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

Calculation of Deviations

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

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

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

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

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

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

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

C. Mean Deviation in case of Continuous Series:

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

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

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

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

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

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

Solution:

Wage (Rs.) (X)	**No. of Workers (f)**	**Cumulative Frequency (Cf)**	**Mid Value (m)**	**$\lvert D_M \rvert$ $\lvert m - M \rvert$**	**$\lvert f D_M \rvert$**
10-20	5	5	15	37.78	188.9
20-30	8	13	25	27.78	222.24
30-40	12	25	35	17.78	213.36
40-50	20	45	45	7.78	155.6
50-60	18	63	55	2.22	39.96
60-70	15	78	65	12.22	183.3
70-80	13	91	75	22.22	288.86
80-90	9	100	85	32.22	289.98
	N = 100				**$\sum \lvert f D_M \rvert$ = 1582.2**

Calculation of Median

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

Median Class = Value of 50th item

Median Class = 50-60

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

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

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

Calculation of Mean Deviation from Median

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

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

TEST YOUR PROGRESS (C)

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

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

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

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

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

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

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

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

5. Calculate M.D. from Mean and Median

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

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

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

Answers

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

2.8 MEASURE OF DISPERSION – STANDARD DEVIATION

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

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

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

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

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

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

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

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

A. Standard Deviation in case of Individual Series

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

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

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

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

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

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

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

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

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

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

Marks	5	7	11	16	15	12	18	12

Solution:

1. Standard Deviation using Actual Mean

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

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

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

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

2. Standard Deviation using Assumed Mean

Marks **X**	**dx = X - A** **(Where A = 11)**	$\mathbf{dx^2}$
5	-6	36
7	-4	16
11	0	00
16	5	25
15	4	16
12	1	01
18	7	49
12	1	01
∑X = 96	**∑dx = 8**	**$\sum dx^2$ = 144**

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

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

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

3. Standard Deviation by Direct Method

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

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

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

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

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

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

Solution: Mean and Standard Deviation of Player X

Score **X**	**x = X - $\bar{X}$** **(Where $\bar{X}$ = 30)**	**x^2**
26	-4	16
24	-6	36
28	-2	2
30	0	0
35	5	25
40	10	100
25	-5	25
30	0	0
45	15	225
17	-13	169

$\sum X = 300$		**$\sum x^2 = 600$**

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

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

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

Mean and Standard Deviation of Player Y

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

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

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

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

Conclusion:

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

B. Standard Deviation in case of Discrete Series

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

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

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

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

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

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

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

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

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

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

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

Solution:1. Standard Deviation using Actual Mean

Marks	f	fX	$x = X - \bar{X}$	x^2	fx^2
X			$(\bar{X} = 19)$		
5	2	10	-14	196	392
10	7	70	-9	81	567
15	11	165	-4	16	176
20	15	300	1	1	15
25	10	250	6	36	360
30	4	120	11	121	484
35	1	35	16	256	256
	N = 50	**$\sum$fX = 950**			**$\sum x^2$ = 2250**

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

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

2. Standard Deviation using Assumed Mean

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

35	1	15	225	15	225
	N = 50			**$\sum$fdx = -50**	**$\sum \mathbf{fdx^2}$ = 2300**

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

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

3. Standard Deviation using Direct Method

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

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

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

C. Standard Deviation in case of Continuous Series

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

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

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

Solution:1. Standard Deviation using Actual Mean

Marks X	m	f	fX	x = m - $\bar{X}$ ($\bar{X}$ = 21.5)	x^2	fx^2
5-10 10-15 15-20	7.5 12.5 17.5	2 7 11	15 87.5 192.5	-14 -9 -4	196 81 16	392 567 176

20-25	22.5	15	337.5	1	1	15
25-30	27.5	10	275	6	36	360
30-35	32.5	4	130	11	121	484
35-40	37.5	1	37.5	16	256	256
		N = 50	**∑fX = 1075**			**$\sum x^2$ = 2250**

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

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

2. Standard Deviation using Assumed Mean

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

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

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

3. Standard Deviation using Direct Method

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

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

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

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

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

2.8.1 Combined Standard Deviation

The main benefit of standard deviation is that if we know the mean and standard deviation of two or more series, we can calculate combined standard deviation of all the series. This feature is not available in other measures of dispersion. That’s why we assume that standard deviation is best measure of finding the dispersion. Following formula is used for this purpose:

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

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

σ_1, $\sigma_{2,}$ σ_3 = standard deviation of series 1, 2 and 3

d_1, $d_{2,}$ d_3 = difference between mean of the series and combined mean for 1, 2 and 3.

Example14. Find the combined standard deviation for the following data

	Firm A	Firm B
No. of Wage Workers	70	60
Average Daily Wage (Rs.)	40	35
S.D of wages	8	10

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

$$\overline{\text{X}_{12}}=\frac{\text{N}_1\overline{\text{X}_1}+\text{N}_2\overline{\text{X}_2}}{\text{N}_1+\text{N}_2}$$

Where N_1 = Number of workers in Firm A, N_2 = Number of workers in Firm B

$\overline{\text{X}_1}$ = Mean wage of workers in Firm A, and $\overline{\text{X}_2}$ = Mean wage of workers in Firm B

We are given that

$\text{N}_1=70$ $\text{N}_2=60$

$\overline{\text{X}_1}=40$ $\overline{\text{X}_2}=35$

∴ Combined Mean, $\overline{\text{X}_{12}}$

$$=\frac{(70\times 40)+(60\times 35)}{70+60}$$

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

$$=\text{Rs.}\,37.69$$

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

d_1 = 40 – 37.69 = 2.31

d_2 = 35 – 37.69 = -2.69

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

Example 15. Find the missing values

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

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

200 = 50 + n_2 + 90

N_2 = 60

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

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

We are given that

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

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

∴ Combined Mean, $\overline{X_{12}}$

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

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

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

$\overline{X_2} = 120$

Combined Standard Deviation =

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

d_1 = 113 – 116 = -3, d_2 = 120 – 116 = 4, d_3 = 115 – 116 = -1

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

Squaring the both sides

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

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

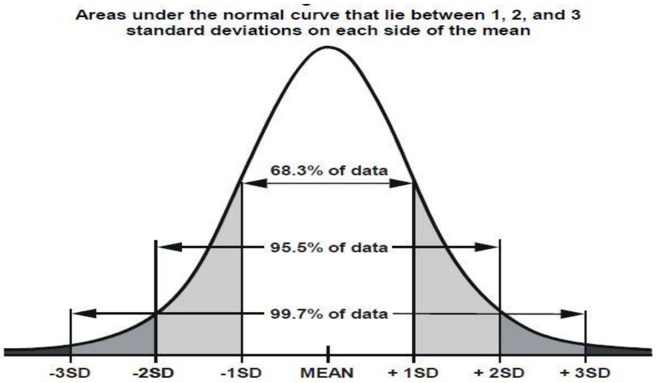
$\sigma_3 = 8$

2.8.2 Properties of Standard Deviation

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

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

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



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

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

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

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

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

7. In perfect symmetric distribution following result follows:

   Range = 6 (S.D.)

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

   Variance = $(S.D)^2$

TEST YOUR PROGRESS (D)

1. Calculate Standard Deviation and find Variance:

X:	5	7	11	16	15	12	18	12

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

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

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

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

4. Find Standard Deviation.

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

5. Find the Standard Deviation and coefficient of variation.

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

6. Find Standard Deviation.

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

7. Find combined Mean and Combined Standard Deviation:

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

8. Find missing information:

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

Answers

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

2.9 CALCULATION OF COEFFICIENT OF VARIATION

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

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

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

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

CV = (SD / μ) * 100

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

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

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

Standard Deviation (SD): The standard deviation is a measure of the dispersion or variability of a dataset.

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

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

Mean (μ): The mean is the average value of a dataset.

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

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

Coefficient of Variation (CV): The CV is the ratio of the standard deviation to the mean, expressed as a percentage.

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

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

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

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

CV = (10 / 80) * 100

CV = 12.5%

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

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

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

Numerator Component: Standard Deviation (SD)

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

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

The standard deviation is calculated using the following steps:

Compute the mean (μ) of the dataset.

Calculate the difference between each data point and the mean.

Square each difference.

Calculate the mean of the squared differences.

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

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

Denominator Component: Mean (μ)

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

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

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

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

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

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

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

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

μ = 50,333.33

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

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

Sum up the squared differences.

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

Take the square root of the result.

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

SD = √[8,666,666.67 / 12]

SD = √[722,222.22]

SD ≈ 849.84

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

CV = (849.84 / 50,333.33) * 100

CV ≈ 1.69%

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

Example 2: Consider a dataset of daily temperature readings in Celsius for a week:

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

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

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

μ = 17.86

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

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

Sum up the squared differences.

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

Take the square root of the result.

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

SD = √[0.48 / 7]

SD ≈ 0.30

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

CV = (0.30 / 17.86) * 100

CV ≈ 1.68%

The Coefficient of Variation (CV) for this dataset of daily temperature readings is approximately 1.68%. It suggests a relatively low level of variability in temperature across the week when compared to the mean. These examples illustrate the step-by-step calculation process of the Coefficient of Variation (CV) for different datasets, showcasing how the CV captures the relative variability in relation to the mean.

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

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

Low Coefficient of Variation (CV):

A low CV indicates a relatively low level of variability in the dataset compared to the mean. It suggests that the data points are clustered closely around the mean, indicating a higher level of consistency or stability. In practical terms, a low CV implies that the dataset is relatively homogeneous or has a narrow range of values. Examples where a low CV may be desirable include quality control processes, where consistency and precision are crucial.

Moderate Coefficient of Variation (CV):

A moderate CV suggests a moderate level of variability in the dataset compared to the mean. It indicates that the data points have some dispersion around the mean, but not to a significant extent.

In practical terms, a moderate CV implies that there is a certain degree of diversity or spread in the dataset, but it is not excessively high. Examples where a moderate CV may be observed include economic indicators, where some fluctuation is expected but not extreme.

High Coefficient of Variation (CV):

A high CV indicates a relatively high level of variability in the dataset compared to the mean. It suggests that the data points are spread out from the mean, indicating a higher level of dispersion or volatility. In practical terms, a high CV implies that the dataset is heterogeneous or has a wide range of values. Examples where a high CV may be observed include financial markets, where significant fluctuations and risks are present. It is important to note that the interpretation of CV values depends on the context and the nature of the dataset being analyzed. What constitutes a low, moderate, or high CV may vary across different fields, industries, or research domains. It is crucial to compare CV values within the specific context or against relevant benchmarks or standards.

Additionally, it is important to consider other factors and characteristics of the dataset alongside the CV. For example, the presence of outliers, data distribution, sample size, and specific domain knowledge may provide additional insights into the variability and its implications. Ultimately, the interpretation of CV values should be done in conjunction with the goals of the analysis, the nature of the dataset, and the specific context in which the data is being examined. It is advisable to consider the CV alongside other statistical measures and domain-specific knowledge for a comprehensive understanding of the dataset's variability.

2.11 LORENZ CURVE

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

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

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

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

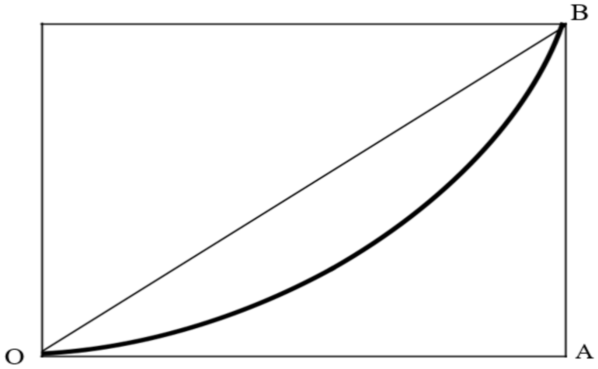
The Lorenz curve can also be summarized by a single numerical measure called the Gini coefficient. It is calculated as the area between the Lorenz curve and the line of perfect equality, divided by the total area under the line of perfect equality. The Gini coefficient ranges between 0 and 1, where 0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

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

Lorenz curve express the relation between the cumulative proportion of people with income at least equal to some specific value and the cumulative proportion of income received by these people. Lorenz curve is represented by a function L(P), which corresponds to a fraction received by the p-th lower fraction of the population when it is ordered by increasing income. The curve slope is always positive and convex, so L (0) = 0 and L (1) = 1.

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

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



2.12 LET US SUM UP

- Dispersion shows whether the average is a good representative of the series.
- High dispersion means values differ more than their average.
- There are two measures of dispersion, Absolute measure and relative measure.
- There are four methods that can be used for measuring the dispersion namely, Range, Quartile Deviation, Mean Deviation and Dispersion.
- Range is simples' method of dispersion.
- Mean deviation can be calculated from Mean, Median or Mode
- Standard Deviation is the best measure of Dispersion.
- If we know standard deviation of two series, we can calculate combined standard deviation.

2.13 QUESTIONS FOR PRACTICE

Q1. What is Dispersion? Explain its uses.

Q2. What are features of good measure of Dispersion?

Q3. What are absolute and relative measure of dispersion?

Q4. What is range? Give its merits and limitations.

Q5. What are Quartile deviations? Give its merits and limitations.

Q6. What is mean deviation. How it is calculated.

Q7. What is standard deviation? How it is calculated.

Q8. How combined standard deviation can be calculated.

Q9. Give properties of standard deviation.

Q10. How is the Coefficient of Variation (CV) calculated?

Q11. What does a high Coefficient of Variation (CV) indicate about a dataset?

Q12. When is the Coefficient of Variation (CV) particularly useful in statistical analysis?

Q13. What is the Lorenz curve and how is it used in economics?

Q14. Can you explain the concept of income inequality and its relationship to the Lorenz curve?

Q15. How is the Lorenz curve constructed and what does it represent?

2.14 FURTHER READINGS

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

M.A (ECONOMICS)

QUANTITATIVE METHODS II

SEMESTER -II

Unit 3: Correlation Analysis: Karl Pearson’s and Spearman’s rank formula

STRUCTURE

3.0 LEARNING OBJECTIVES

After studying the Unit, students will be able to:

- Define what is Correlation
- Distinguish between different types of correlation
- Understand the benefits of correlation
- Find correlation using the graphic method
- Plot a scatter diagram
- Types of Correlation
- Degrees of Correlation

- Scatter Diagram Method
- Properties of Correlation
- Different methods to calculate Coefficient of Correlation
- Spearman’s Rank Correlation

3.1 INTRODUCTION

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

3.2 MEANING OF CORRELATION

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

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

(a) X may be causing Y

(b) Y may be causing X

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

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

However, correlation has nothing to do with causation. It simply attempts to find the degree of mutual association between them. It is possible that two variables might be found highly correlated, but they are not causing the change in each other. There may be a correlation due to pure chance. For example, we may find a high degree of correlation between the number of trees in a city and number of drug addicts. However, there is no theoretical base that relates these variables together Such correlation is known as Spurious Correlation or Non-sense Correlation.

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

3.3 USES OF CORRELATIONS

1. It helps us in understanding the extent and direction of the relation between two variables. It shows, whether two variables are positively correlated or negatively correlated. It also shows whether relation between two variables is high or low.

2. Correlation also helps in the prediction of future, for example, if we know relation between monsoon and agricultural produce, we can predict that what will be the level of produce on basis of monsoon prediction. We can also predict price of Agricultural Products depending on level of produce.

3. With the help of correlation, we can find the value of one variable when the value of other variable is known. This can be done by using the statistical technique called regression analysis.

4. Correlation also helps in business and Commerce. A businessman can fix price of its product using the correlation analysis. Correlation also helps him in deciding business policy.

5. Correlation also helps government in deciding its economic policy. With the help of correlation government can study relation of various economic variables, thus government can decide their economic policies accordingly.

6. Correlation is also helpful in various statistical Analysis. Many Statistical techniques use correlation for further analysis.

3.4 TYPES OF CORRELATION

a. Positive correlation: It is a situation in which two variables move in the same direction. In this case, if the value of one variable increases the value of the other variable also increases. Similarly, if the value of one variable decrease, the value of other variables also decreases. So, when both the variables either increase or decrease, it is known as a positive correlation. For example, we can find a Positive correlation between Income and Expenditure, Population and Demand for food products, Incomes and Savings, etc. The following data shows positive correlation between two variables:

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

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

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

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

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

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

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

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

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

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

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

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

3.5 DEGREES OF CORRELATION

Here degrees of correlation shown in the following table:

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

3.6 SCATTER DIAGRAM METHOD

Scatter Diagram is one of the oldest and simple methods of measuring the correlation. This is a graphic method of measuring the correlation. This method uses diagram representation of bivariate data to find out degree and direction of correlation. Under this method, values of the data are plotted on a graph paper by taking one variable on the x-axis and other variable on the y-axis. Normally independent variable is shown on x-axis whereas the value of the dependent variable is taken on the y-axis. Once all the values are drawn on the graph paper, we can find out degree of correlation between two variables by looking at direction of dots on the graph. Scatter Diagram shows whether two variables are co-related to each other or not. It also shows the direction of correlation whether positive or negative and the shows extent of correlation whether high or low. The following situations are possible in the scatter diagram.

1. Perfect Positive Correlation: After we plot two variables on the graph, if the points of graph fall in a straight line that moves from lower left-hand side to the upper corner on the right-hand side, then it is assumed that there is perfect positive correlation between the variables.

2. Perfect Negative Correlation: After drawing the variables on the graph, if all the points fall in a straight line but direction of the points is downward from right-hand corner to left-hand side corner, then it is assumed that there is perfect negative correlation between the variable.

3. High Degree of Positive Correlation: If we draw two variables on the graph and we find that the points move in upward direction from left-hand corner to the right-hand corner but not in a straight line, rather these are in narrow band, we can assume that there is high degree of positive correlation between the variables.

4. High Degree of Negative Correlation: After plotting the dots on a graph, if we find that all the dots move downward from left-hand corner to the right-hand side corner but not in a straight line rather in a narrow band, we can say that there is high degree of negative correlation between the variables.

5. Low Degree of Positive Correlation: In case the dots drawn on a graph paper moves upward from left side to right side but the dots are widely scattered, it can be said that there is low degree of positive correlation between the variables.

6. Low Degree of Negative Correlation: In case the points drawn on a graph are in downward direction from left side to right side but the points are widely scattered, it is the situation of low degree of negative correlation between the variables.

7. Zero or No Correlation: Sometime find that the dots drawn on a graph paper do not move in any direction and are widely scattered in the graph paper, we can assume that there is no correlation between the two variables.

3.7 PROPERTIES OF CORRELATION

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

3.8 DIFFERENT METHODS TO CALCULATE COEFFICIENT OF CORRELATION

A. Direct method of calculating Correlation

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

1. Take two series X and Y.

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

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

Example 1: Find coefficient of correlation

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

Solution:

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

$N = 6$, Coefficient of Correlation, $r = \frac{N\sum XY-(\sum X)(\sum Y)}{\sqrt{N\sum X^2-(\sum X)^2}\sqrt{N\sum Y^2-(\sum Y)^2}}$

$$= \frac{6\times90-21\times24}{\sqrt{6\times91-(21)^2}\sqrt{6\times106-(24)^2}} \qquad = \frac{540-504}{\sqrt{546-441}\sqrt{636-576}}$$

$$= \frac{36}{\sqrt{105}\sqrt{60}} \qquad = \frac{36}{10.246\times7.7459} \qquad = \frac{36}{79.31} = 0.4539$$

$\Rightarrow \qquad r = 0.4539$

B. Actual Mean method of calculating Correlation

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

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

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

Example 2. Calculate Karl Pearson’s coefficient of correlation

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

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

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

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

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

X	Y	$x = (X - \overline{X})$	x^2	$y = (Y - \overline{Y})$	y^2	xy
50	11	-8	64	-3	9	24
50	13	-8	64	-1	1	8
55	14	-3	9	0	0	0
60	16	2	4	2	4	4
65	16	7	49	2	4	14
65	15	7	49	1	1	7
65	15	7	49	1	1	7

60	14	2	4	0	0	0
60	13	2	4	-1	1	-2
50	13	-8	64	-1	1	8
$\sum X=580$	$\sum Y=140$		$\sum x^2=360$		$\sum y^2=22$	$\sum xy=70$

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

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

Coefficient of Correlation, $r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}} = \frac{70}{\sqrt{360\times22}} = \frac{70}{\sqrt{7920}} = 0.7866$

$\Rightarrow \qquad r = 0.7866$

C. Assumed Mean method of calculating Correlation

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

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

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

Example 3. Compute coefficient of correlation from the following figures

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

Solution: Here, $N = 7$

Coefficient of Correlation, r is given by

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

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

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

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

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

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

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

D. Step Deviation method of calculating Correlation

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

1. Take two series X and Y.
2. Take any value as assumed mean for series X.
3. Take deviations of series X from its assumed mean and it is denoted by ‘dx’.
4. Divide the value of ‘dx’ so obtained by some common factor to get dx'
5. Find sum of deviations denoted by $\sum dx'$.
6. Take square of deviation of series X denoted by $dx'^{\ 2}$

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

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

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

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

Solution:

X	Y	dx= X-A A=15	dx'=dx/C$_1$ C$_1$=5	dy= Y-B B=30	dy'=dy/C$_2$ C$_2$=5	dx’^{2}	dy'^{2}	dx'dy'
5	40	-10	-2	10	2	4	4	-4
10	35	-5	-1	5	1	1	1	-1
15	30	0	0	0	0	0	0	0
20	25	5	1	-5	-1	1	1	-1
25	20	10	2	-10	-2	4	4	-4
			∑dx'=0		∑dy'=0	∑dx'^{2} =10	∑dy'^{2} =10	∑dx'dy'=- 10

Here, $N = 5$, Coefficient of Correlation, $r = \frac{N\sum dx'dy' - (\sum dx')(\sum dy')}{\sqrt{N\sum dx'^2 - (\sum dx')^2}\sqrt{N\sum dy'^2 - (\sum dy')^2}}$

$$= \frac{5\times(-10) - 0\times 0}{\sqrt{5\times10-0^2}\,\sqrt{5\times10-0^2}} \quad = \frac{-50}{\sqrt{50}\times\sqrt{50}} = -1$$

$\Rightarrow \qquad\qquad r = -1$

E. Calculating Correlation with help of Standard Deviations

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

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

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

Example 5. Given No. of pairs of observations = 10, $\sum xy = 625$, X Series Standard Deviation = 9, Y Series Standard Deviation =8, Find 'r'.

Solution: We are given that $N = 10$, $\sigma_X = 9$, $\sigma_Y = 8$ and $\sum xy = 625$

Now $r = \frac{\sum xy}{N\sigma_X\sigma_Y} = \frac{625}{10\times9\times8} = \frac{625}{720} = 0.868$

$\Rightarrow$ $r = +.868$

Example 6. Given No. of pairs of observations = 10

X Series Arithmetic Mean =75, Y Series Arithmetic Mean =125

X Series Assumed Mean =69, Y Series Assumed Mean =110

X Series Standard Deviation =13.07, Y Series Standard Deviation =15.85

Summation of products of corresponding deviation of X and Y series =2176, Find 'r'.

Solution: We are given that

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

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

and $\sum xy = 2176$

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

$$= \frac{2176-10(75-69)(125-110)}{10\times13.07\times15.85} \qquad = \frac{2176-900}{2071.595} \qquad = 0.6159 \approx 0.616$$

⇒ $r = +0.616$

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

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

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

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

X	Y
5	15
10	18

While the correct values were

X	Y
6	18
10	19

Find the correct value of correlation coefficient.

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

∴ Correct value of $\sum X$=50-5-10+6+10 $= 51$

Incorrect value of $\sum Y = 106$

∴ Correct value of $\sum Y = 106 - 15 - 18 + 18 + 19 \ = 110$

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

∴ Correct value of $\sum X^2 = 448 - 5^2 - (10)^2 + 6^2 + (10)^2 \ = 459$

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

∴ Correct value of $\sum Y^2 = 1896 - 15^2 - (18)^2 + (18)^2 + 19^2 = 2032$

Incorrect value of $\sum XY = 879$

∴ Correct value of $\sum XY = 879 - (5 \times 15) - (10 \times 18) + (6 \times 18) + (10 \times 19) \ = 952$

Thus, the corrected value of coefficient of correlation

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

$$= \frac{6\times952-51\times110}{\sqrt{6\times459-(51)^2}\sqrt{6\times2032-(110)^2}} \qquad = \frac{5712-5610}{\sqrt{2754-2601}\sqrt{12,192-12,100}}$$

$= \frac{102}{\sqrt{153}\,\sqrt{92}} \qquad = \frac{102}{12.369\times 9.59} \qquad = \frac{102}{118.618} = 0.8599$

⇒ $r = +0.8599$

TEST YOUR UNDERSTANDING (A)

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

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

2. Calculate Karl Pearson's coefficient of correlation

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

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

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

4. Find correlation between marks of subject A Subject B

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

5. Find correlation between Height of Mother and Daughter

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

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

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

8. Following values were obtained during calculation of correlation:

N = 30; ∑X = 120 ∑X2 = 600 ∑Y = 90 ∑Y2 = 250 ∑XY = 335

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

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

Find correct correlation.

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

	X series	Y series

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

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

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

Answers

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

3.9 SPEARMAN’S RANK CORRELATION

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

A. Spearman’s Rank Correlation when ranks are given

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

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

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

Solution:

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

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

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

$$= 1 - \frac{6\times 28}{8(8^2-1)} = 1 - \frac{168}{8(64-1)} \quad = 1 - \frac{168}{8(63)}$$

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

⇒ Rank Correlation Coefficient $= 0.67$

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

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

Solution:

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

Now $r_{12} = 1 - \frac{6\sum {D_1}^2}{N(N^2-1)} \quad = 1 - \frac{6\times 144}{10(10^2-1)} \quad = 1 - \frac{864}{10(100-1)} \quad = 1 - \frac{864}{10(99)}$

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

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

$$r_{23} = 1 - \frac{6\sum {D_2}^2}{N(N^2-1)} \qquad = 1 - \frac{6\times112}{10(10^2-1)} \qquad = 1 - \frac{672}{10(100-1)}$$

$$= 1 - \frac{672}{10(99)} \quad = 1 - \frac{672}{990} \qquad = 1 - 0.679 = 0.321$$

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

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

$$= 1 - \frac{6\times182}{10(10^2-1)} \qquad = 1 - \frac{1092}{10(100-1)} \qquad = 1 - \frac{1092}{10(99)} \quad = 1 - \frac{1092}{990}$$

$$= 1 - 1.103 = -0.103$$

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

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

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

B. Spearman's Rank Correlation when ranks are not given

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

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

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

Solution:

X	Ranks R_1	Y	Ranks R_2	Difference of Ranks $D = R_1 - R_2$	D^2
60	4	44	8	-4	16
70	2	74	3	-1	1
53	7	54	6	+1	1
59	5	64	5	0	0

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

Here N = 8, Rank Coefficient of Correlation, $r = 1 - \frac{6\sum D^2}{N(N^2-1)}$

$$= 1 - \frac{6\times 28}{8(8^2-1)} = 1 - \frac{168}{8(64-1)}$$

$$= 1 - \frac{168}{8(63)} = 1 - \frac{168}{504} = 1 - 0.33 = 0.67$$

⇒ Rank Correlation Coefficient = 0.67

C. Spearman's Rank Correlation when there is repetition in ranks

1. Assign the ranks in descending order to series X by giving first rank to highest value and second rank to value lower than higher value and so on. If two items have same value, assign the average rank to both the item. For example, two equal values have ranked at 5th place than rank to be given is 5.5 to both i.e., mean of 5th and 6th rank. ($\frac{5+6}{2}$).
2. Similarly assign the ranks to series Y.
3. Calculate the difference between ranks of both the series denoted by $\sum D$.
4. Take square of deviations and calculate the value of D^2.
5. Calculate sum of square of deviations denoted by $\sum D^2$.
6. Apply following formula $r = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}(m_1{}^3 - m_1) + \frac{1}{12}(m_2{}^3 - m_2)\right\}}{N(N^2-1)}$

Where m = no. of times a particular item is repeated.

Example 12. Find the Spearman's Correlation Coefficient for the data given below

X	110	104	107	82	93	93	115	95	93	113
Y	80	78	90	75	81	70	87	78	73	85

Solution: Here, in X series the value 93 occurs thrice $(m_1 = 3)$, i.e. at 7^{th}, 8^{th} and 9^{th} rank. So, all the three values are given the same average rank, i.e. $\frac{7+8+9}{3} = 8^{th}$ rank.

Similarly, in Y series the value 78 occurs twice $(m_2 = 2)$, i.e. at 6^{th} and 7^{th} rank. So, both the values are given the same average rank, i.e. $\frac{6+7}{2} = 6.5^{th}$ rank.

X	Ranking of X	Y	Ranking of Y	Difference of Ranks $D = R_1 - R_2$	D^2

	R_1		R_2		
110	3	80	5	-2	4
104	5	78	6.5	-1.5	2.25
107	4	90	1	+3	9
82	10	75	8	+2	4
93	8	81	4	+4	16
93	8	70	10	-2	4
115	1	87	2	-2	1
95	6	78	6.5	-0.5	0.25
93	8	73	9	-1	1
113	2	85	3	-1	1
					$\sum D2\ = 42.5$

Here $N = 10$, Spearman's Rank Correlation Coefficient, $r = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}(m_1{}^3 - m_1) + \frac{1}{12}(m_2{}^3 - m_2)\right\}}{N(N^2-1)}$

i.e. $r = 1 - \frac{6\left\{42.50 + \frac{1}{12}(3^3-3) + \frac{1}{12}(2^3-2)\right\}}{10(10^2-1)}$

$= 1 - \frac{6\left\{42.50 + \frac{24}{12} + \frac{6}{12}\right\}}{10(100-1)} \quad = 1 - \frac{6\left\{42.50 + 2 + \frac{1}{2}\right\}}{10\times 99} \quad = 1 - \frac{6\{42.5+2.5\}}{990} \quad = 1 - \frac{6\times 45}{990}$

$= 1 - 0.2727 = 0.7273$

⇒ Rank Correlation Coefficient =0.7273

Example 13. The rank correlation coefficient between the marks obtained by ten students in Mathematics and Statistics was found to be 0.5. But later on, it was found that the difference in ranks in the two subjects obtained by one student was wrongly taken as 6 instead of 9. Find the correct rank correlation.

Solution: Given $N = 10$, Incorrect $r = 0.5$

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

⇒ $0.5 = 1 - \frac{6\sum D^2}{10(10^2-1)} = 1 - \frac{6\sum D^2}{10\times 99}$

⇒ Incorrect $\sum D^2 = \frac{990}{6} \times 0.5 = 82.5$

∴ The corrected value of $\sum D^2 = 82.5 - 6^2 + 9^2 \quad = 82.5 - 36 + 81 = 127.5$

∴ Correct Rank Correlation Coefficient, $r = 1 - \frac{6\times 127.5}{10(10^2-1)}$

$= 1 - \frac{765}{10(100-99)} \quad = 1 - \frac{765}{10\times 99} \quad = 1 - \frac{765}{990} \quad = 1 - 0.7727 = 0.2273$

TEST YOUR UNDERSTANDING (B)

1. Find Rank correlation on base of following data.

X	78	36	98	25	75	82	90	62	65	39
Y	84	51	91	60	68	62	86	58	53	47

In Dance competition following ranks were given by 3 judges to participants. Determine which two judges have same preference for music:

1st Judge	1	6	5	10	3	2	4	9	7	8
2nd Judge	3	5	8	4	7	10	2	1	6	9
3rd Judge	6	4	9	8	1	2	3	10	5	7

3. Find Rank correlation on base of following data.

X	25	30	38	22	50	70	30	90
Y	50	40	60	40	30	20	40	70

4. Find Rank correlation on base of following data.

X	63	67	64	68	62	66	68	67	69	71
Y	66	68	65	69	66	65	68	69	71	70

Answers

1) .82	2) I and II -.2121, II and III -.297, I and III .6364, so judge I and III	3) 0	4) 0.81

3.10 LET US SUM UP

- Correlation shows the relation between two or more variables.
- The value of the coefficient of correlation always lies between -1 and +1.
- Correlation may be positive or negative.
- Correlation may be linear or nonlinear.
- Karl Person's coefficient of correlation is the most popular method of correlation.
- It can deal only with quantitative data.
- Spearman's Rank correlation calculated correlation on the basis of ranks given to data.
- It can deal with qualitative data also.

3.11 QUESTIONS FOR PRACTICE

Q1. What is Correlation?

Q2. What are uses of measuring correlation?

Q3. Explain the properties of Correlation Coefficient.

Q4. Give different types of correlation.

Q5. What are the various degrees of correlation coefficient?

Q6. What do you mean by scatter diagram?

Q7. Give Karl Persons method of calculating correlation.

Q8. Give Karl Pearson's coefficient of correlation in case of actual and assumed mean.

Q9. What are the merits and limitations of Karl Pearson's method?

Q10. What is Spearman's Rank correlation? How it is determined.

Q11. In case of repeated ranks how would you determine Spearman's Rank correlation?

Q12. What are the limitations of Spearman's Rank correlation

3.12 SUGGESTED READINGS

- J. K. Sharma, Business Statistics, Pearson Education.
- S.C. Gupta, Fundamentals of Statistics, Himalaya Publishing House.
- S.P. Gupta and Archana Gupta, Elementary Statistics, Sultan Chand and Sons, New Delhi.
- Richard Levin and David S. Rubin, Statistics for Management, Prentice Hall of India, New

Delhi.

- M.R. Spiegel, Theory and Problems of Statistics, Schaum's Outlines Series, McGraw Hill Publishing Co.

M.A (ECONOMICS)

QUANTITATIVE METHODS II

SEMESTER -II

Unit 4: Simple Regression Analysis: Regression Meaning, Properties, X On Y and Yon X

STRUCTURE

4.0 LEARNING OBJECTIVES

After studying the Unit, leaner will be able to:

- Describe what is regression
- Distinguish between different types of Regression
- Understand the benefits of Regression
- Show how correlation and regression are related
- Understand the properties of regression coefficients

4.1 INTRODUCTION

Statistics has many applications in our life whether it's business life or our routine life. There are many techniques in statistics that can help us in prediction. Regression analysis is a tool in statistics that can help in the prediction of one variable when the value of other variable is known if there exists any close relation between two or more variables, though such relation may be positive or negative. The technique of Regression can be widely used as a powerful tool in almost all fields whether science, social science, Business, etc. However, particularly, in the fields of business and management this technique is very useful for studying the relationship between different variables such as Price and Demand, Price and Supply, Production and Consumption, Income and Consumption, Income and Savings, etc.

When we find a regression between two or more variables, we try to understand the behavior of one variable with help movement of the other variable in a particular direction. For example, if the correlation coefficient between value of sales and amount spent on advertisement say +0.9, it means that if advertisement expenditure is increased, Sale is also likely to increase, as there is a very high positive relation between the two variables. However, correlation only tells the relation between two variables, but it does not tell the extent to which a change in one variable will affect the change in other variables. For this purpose, we have to calculate the co-efficient of Regression. The regression Coefficient is a statistical measure that tries to find out the value of one variable known as the dependent variable when the value of another variable known as an independent variable is known. Thus, in the case of two variables, like Advertisement expenditure and amount of Sales, we can estimate the likely amount of Sales if the value of Advertisement expenditure is given. Similarly, we can predict the value of Advertisement expenditure required, to achieve a particular amount of Sales. This can be done using the two regression coefficients

4.2 MEANING OF REGRESSION ANALYSIS

Many experts have defined the term Regression in their own way. Some of these definitions are given as, according to Sir Francis Galton, the term regression analysis is defined as "the law of regression that tells heavily against the full hereditary transmission of any gift, the more bountifully the parent is gifted by nature, the rarer will be his good fortune if he begets a son who is richly endowed as himself, and still more so if he has a son who is endowed yet more largely."

In the words of Ya Lun Chou, "Regression analysis attempts to establish the nature of the relationship between variables that is to study the functional relationship between the variables and thereby provide a mechanism for prediction or forecasting".

4.3 LIMITATIONS OF REGRESSION ANALYSIS

Though Regression is a wonderful statistical tool, still it suffers from some limitations. The following are the limitations of Regression analysis:

- Regression analysis assumes that there exists cause and effect relationship between the variables and such a relation is not changeable. This assumption may not always hold good and thus could give misleading results.

- Regression analysis is based on some limited data available. However, as the values are based on limited data it may give misleading results.
- Regression analysis involves very lengthy and complicated steps of calculations and analysis. A layman may not be using this technique.
- Regression analysis can be used only with quantitative data. It cannot be used with data of a qualitative nature, such as hard work, beauty, etc.

4.4 DIFFERENT TYPES OF REGRESSION ANALYSIS

1. Simple and Multiple Regression

- Simple Regression: When only two variables are under study, it is known as a simple regression. For example, we are studying the relationship between Sales and Advertising expenditure. Let's consider sales as Variable X and advertising as variable Y. The X = a+b Y is the regression equation of X on Y where X is the dependent variable and Y and the independent variable. In other words, we can find the value of variable X (Sales) if the value of Variable Y (Advertising) is given.
- Multiple Regression: The study of more than two variables at a time is known as multiple regression. Under this, only one variable is taken as a dependent variable and all the other variables are taken as independent variables. For example, if we consider sales as Variable X, advertising as variable Y and Income as Variable Z, then using the functional relation X = *f* (Y, Z), we can find the value of variable X (Sales) if the value of Variable Y (Advertising) and the value of variable Z (Income) is given.

2. Total and Partial Regression

a. Total Regression: Total regression analysis is one in which we study the effect of all the variables simultaneously. For example, when we want to study the effect of advertising expenditure of business represented by variable Y, income of the consumer represented by variable Z, on the amount of sales represented by variable X, we can study impact of advertising and income simultaneously on sales. This is a case of total regression analysis. In such cases, the regression equation is represented as follows:

$X = f\,(Y, Z),$

a. Partial Regression: In the case of Partial Regression one or two variables are taken into consideration and the others are excluded. For example, when we want to study the effect of advertising expenditure of business represented by variable Y, income of the consumer represented by variable Z, on the amount of sales represented by variable X, we will not study impact of both income and advertising simultaneously, rather we will first study effect of income on sales keeping advertising constant and then effect of advertising on sales keeping income constant. Partial regression can be written as

X=f (Y not Z)

3. Linear and Non-Linear Regression

a. Linear Regression: When the functional relationship between X and Y is expressed as the first-degree equations, it is known as linear regression. In other words, when the points plotted on a scatter diagram concentrate around a straight line it is the case of linear regression.

b. Non-linear Regression: On the other hand, if the line of regression (in the scatter diagram) is not a straight line, the regression is termed curved or non-linear regression. The regression equations of non-linear regression are represented by equations of a higher degree. The following diagrams show the linear and non-linear regressions:

4.5 PROPERTIES OF REGRESSION COEFFICIENTS

The regression coefficients discussed above have a number of properties which are given as under:

1. The geometric Mean of the two regression coefficients gives the coefficients of correlation i.e., r = $\sqrt{bxy\ *\ byx}$

2. Both the regression coefficients must have the same sign i.e., in other words, either both coefficients will have + signs or both coefficients will have - signs. This is due to the fact that in the first property, we have studied the geometric means of both coefficients will give us value of correlation. If one sign will be positive and other will be negative, the product of both signs will be negative. And it is not possible to find out correlation of negative value.

3. The signs of regression coefficients will give us signs of coefficient of correlation. This means if the regression coefficients are positive the correlation coefficient will be positive, and if the regression coefficients are negative then the correlation coefficient will be negative.

4. If one of the regression coefficients is greater than unity or 1, the other must be less than unity. This is because the value of the coefficient of correlation must be between ± 1. If both the regression coefficients are more than 1, then their geometric mean will be more than 1 but the value of correlation cannot exceed 1.

5. The arithmetic mean of the regression coefficients is either equal to or more than the correlation coefficient $\frac{\mathrm{bxy+byx}}{2} \geq \sqrt{bxy\ *\ byx}$

6. If the regression coefficients are given, we can calculate the value of standard deviation by using the following formula.

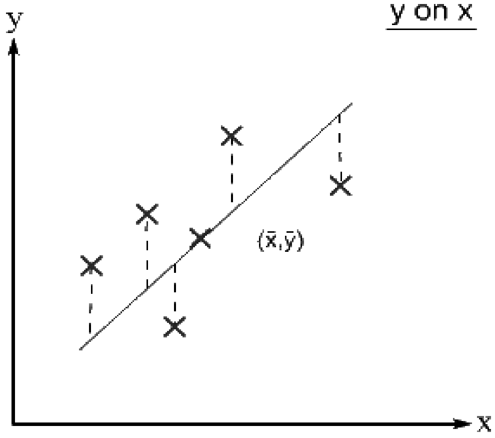
bxy = $r\,\frac{\sigma\mathrm{x}}{\sigma\mathrm{y}}$ or byx = $r\,\frac{\sigma\mathrm{y}}{\sigma\mathrm{x}}$

7. Regression coefficients are independent of change of origin but not of scale. This means that if the original values of the two variables are added or subtracted by some constant, the values of the regression coefficients will remain the same. But if the original values of the two variables are multiplied, or divided by some constant (common factors) the values of the regression equation will not remain the same.

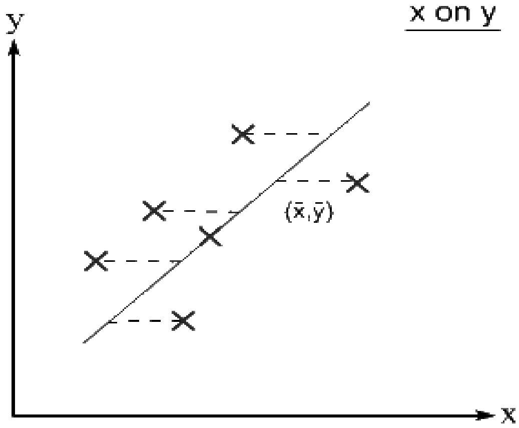
4.6 MEANING AND TYPES OF REGRESSION LINES

The lines that are used in Regression for the purpose of estimation are called regression lines. In other words, the lines that are used to study the dependence of one variable on the other variable are called regression lines. If we have two variables X and Y then there.

a. Regression Line of Y on X: Regression Line Y on X measures the dependence of Y on X and we can estimate the value of Y for the given values of X. In this line, Y is the dependent variable and X is the independent variable.

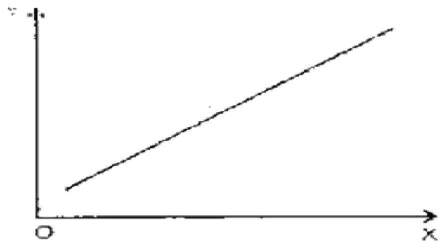


b. Regression Line of X on Y: Regression Line X on Y measures the dependence of X on Y and we can estimate the value of X for the given values of Y. In this line X is dependent variable and Y is independent variable.



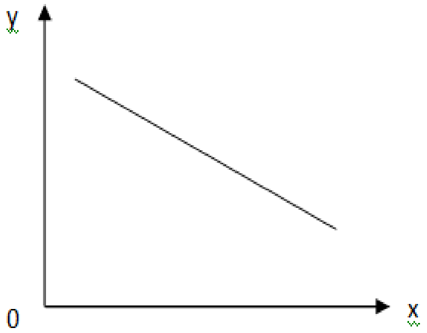
The direction of two regression equations depends upon the degree of correlation between two variables. Following can be the cases of correlation between two variables:

1. Perfect positive correlation: If there is a perfect positive correlation between two variables (i.e., r = +1), both the lines will coincide with each other and will be having a positive slope. Both the lines X on Y and Yon X will be same in this case. In other words, in that case, only one regression line can be drawn as shown in the diagram. The slope of the line will be upward.



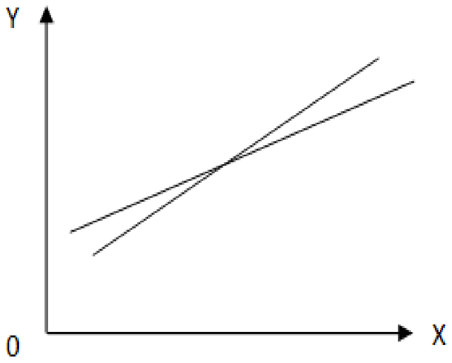
Both the line will coincide with positive slope

2. Perfect negative correlation: If there is a perfect negative correlation between two variables (i.e., r = -1), both the lines will coincide with each other and will in such case these lines will be having negative slope. Both the lines X on Y and Yon X will be same but downward sloping. In other words, in that case only one regression line can be drawn as shown in the diagram. The slope of the line will be upward.



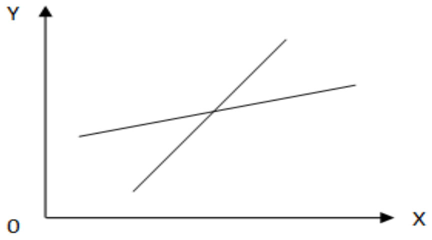
Both the line will coincide with negative slope

3. High degree of correlation: If there is a high degree of correlation between two variables, both the lines will be near to each other. In other words, these lines will be closer to each other but the lines will not coincide with each other. Both the lines will be separate. Further the direction of lines depends upon the positive or negative correlation.



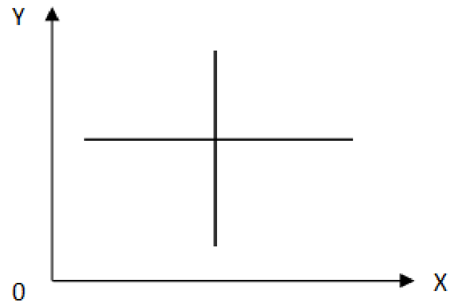
Both the line will be close to each other

4. Low degree of correlation: If there is a low degree of correlation between two variables, both the lines will be having more distance from each other. In other words, these lines will be farther to each other, that is the gap between the lines will be more. Both the lines will be separate. Further the direction of lines depends upon the positive or negative correlation.



Both the line will be farther from each other

5. No correlation: If there is a no correlation between two variables (i.e., r = 0), both the lines will be perpendicular to each other. In other words, these lines will cut each other at $90^{0.}$ This diagram depicts the perpendicular relation between the two regression lines when there is absolutely zero correlation between the two variables under the study.



Both the line will be perpendicular to each other

4.7 DIRECT METHODS TO ESTIMATE REGRESSION EQUATION

The regression equations can be obtained by 'Normal Equation Method" as follows:

1. Regression Equation of Y on X: The regression equation Y on X is in the format of Y = a + bx, where Y is a Dependent Variable and X is an Independent Variable. To estimate this regression equation, the following normal equations are used:

$$\Sigma Y = na + b_{yx} \Sigma X$$

$$\Sigma XY = a \Sigma X + b_{yx} \Sigma X^2$$

With the help of these two equations the values of ‘a' and 'b' are obtained and by putting the values of ‘a' and 'b' in the equation Y = a + b X we can predict or estimate value of Y for any value of X.

2. Regression Equation of X on Y: The regression equation X on Y is in the format of X = a + bY, where X is a Dependent Variable and Y is an Independent Variable. To estimate this regression equation, following normal equations are used:

$$\Sigma X = na + b_{xy} \Sigma Y$$

$$\Sigma XY = a \Sigma Y + b_{xy} \Sigma Y^2$$

With the help of these two equations the values of ‘a' and 'b' are obtained and by putting the values of ‘a' and 'b' in the equation X = a + bY we can predict or estimate value of Y for any value of X.

4.8 LEAST SQUARE METHOD OF FITTING REGRESSION LINES

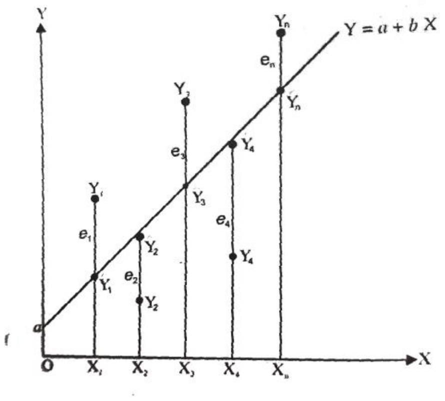
Under this method, the lines of best fit are drawn as the lines of regression. These lines of regression are known as the lines of the best fit because, with the help of these lines we can make the estimate of the values of one variable depending on the value of other variables. According to the Least Square method, regression line should be plotted in such a way that sum of the square of the difference between actual value and the estimated value of the dependent variable should be the least or minimum possible. Under this method, we draw two regression lines that are

a. Regression line Y on X – it measures the value of Y when the value of X is given. In other words, it assumes that X is an independent variable whereas the other variable Y is dependent variable. Mathematically this line is represented by

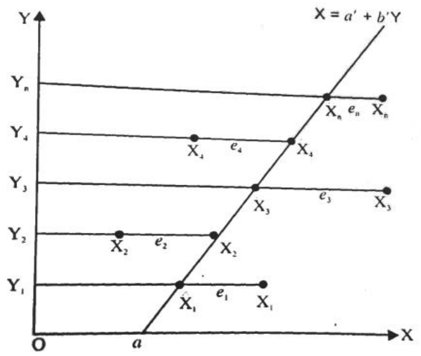
Y = a + bX
Where Y – Dependent Variable
X – Independent Variable
a & b – Constants

b. Regression line X on Y – it measures the value of X when value of Y is given. In other words, it assumes that Y is an independent variable whereas the other variable X is dependent variable. Mathematically this line is represented by

X = a + bY
Where X – Dependent Variable
Y – Independent Variable
a & b – Constants



Equation Y on X



Equation X on Y

In the above two regression lines, there are two constants represented by "a" and "b". The constant "b" is also known as regression coefficient, which are denoted as "byx" and "bxy", Where "byx" represent regression coefficient of equation Y on X and "bxy" represent regression coefficient of equation X on Y. When the value of these two variables "a" and "b" is determined we can find out the regression line.

Example 1. Find out the two regression lines for the data given below using the method of least square.

Variable X:	5	10	15	20	25
Variable Y:	20	40	30	60	50

Determination of the regression lines by the method of least square. Also, find out

a. Value of Y when value of X is 40

b. Value of X when value of Y is 80

Solution:

X	Y	X2	Y2	XY
5	20	25	400	100
10	40	100	1600	400
15	30	225	900	450
20	60	400	3600	1200
25	50	625	2500	1250
XX = 75	XY = 200	XX2=1375	XY2=9,000	XXY =3400

(i) Regression line of Y on X

This is given by Y = *a* + *bX*

where *a* and *b* are the two constants that are found by solving simultaneously the two normal equations as follows:

$\Sigma Y = na + b_{yx} \Sigma X$

$\Sigma XY = a \Sigma X + b_{yx} \Sigma X^2$

Substituting the given values in the above equations we get,

200 = 5a + 75b ……………………………………. (i)

3400 =75a + 1375b ……………………………………. (ii)

Multiplying the eqn. (i) by 15 we get

3000 = 75a + 1125b……………………………………. (iii)

Subtracting the equation (iii) from equation (ii) we get,

3400 =75a + 1375b

-3000 = -75a - 1125b

400 = 250b

or b = 1.6

Putting the above value of b in the eqn. (i) we get,

200 = 5a+ 75(1.6) or

5a =200- 120 or

a = 16

Thus, a = 16, and b = 1.6

Putting these values in the equation Y = *a* + bX we get

$Y = 16+ 1.6X$

So, when X is 40, the value of Y will be

$Y = 16+ 1.6(40) = 80$

(ii) Regression line of X on Y

This is given by X = *a* + *bY*

where *a* and *b* are the two constants that are found by solving simultaneously the two normal equations as follows:

$\Sigma X = na + b_{xy}\Sigma Y$

$\Sigma XY = a \Sigma Y + b_{xy}\Sigma Y^2$

Substituting the given values in the above equations we get,

75 = 5a+200b ……….………………………………….. (i)

3400 = 200a + 9000b …………………………………….. (ii)

Multiplying the eqn. (i) by 40 we get

3000 = 200a + 8000b…………………………………….. (iii)

Subtracting the equation (iii) from equation (ii) we get,

$3400 = 200a + 9000b$

$-3000 = -200a + -8000b$

$400 = \quad 1000b$

or $b = .4$

Putting the above value of b in the eqn. (i) we get,

$75 = 5a + 200(.4)$ or

$5a = -5$ or

$a = -1$

Thus, a = -1, and b = .4

Putting these values in the equation X = *a* + *bY* we get

$X = -1+ .4Y$

So, when Y is 80, the value of X will be

$X = -1+ .4(80) = 31$

4.9 OTHER METHODS OF ESTIMATING REGRESSION EQUATION

This method discussed above is known as direct method. This is one of the popular methods of finding the regression equation. But sometimes this method of finding regression equations becomes cumbersome and lengthy especially when the values of X and Y are very large. In this case, we can simplify the calculation by taking the deviations of X and Y than dealing with the actual values of X and Y. In such case

Regression equation Y on X

$Y = a + bX$

will be converted to $(Y - \bar{Y}) = byx\ (X - \bar{X})$

Similarly, Regression equation X on Y:

$X = a + bY$

will be converted into $(X - \bar{X}) = bxy\ (Y - \bar{Y})$

Now when we are using these regression equations, the calculations will become very simple as now we have to calculate value of only one constant which is the value of "b" which is our regression coefficient. As there are two regression equations, we need to calculate two regression coefficients that are Regression Coefficient X on Y, which is symbolically denoted as "bxy" and similarly Regression Coefficient Y on X, which is denoted as "byx". However, these coefficients can also be calculated using different methods. As we take deviations under this method, we can take deviations using actual mean, assumed mean or we can calculate it by not taking the deviations. The following formulas are used in such cases:

Method	Regression Coefficient X on Y	Regression Coefficient Y on X
When deviations are taken from actual mean	$bxy = \frac{\sum xy}{\sum y^2}$	$byx = \frac{\sum xy}{\sum x^2}$
When deviations are taken from assumed mean	$bxy = \frac{N\sum dxdy - \sum dx \sum dy}{N\sum dy^2 - (\sum dy)^2}$	$byx = \frac{N\sum dxdy - \sum dx \sum dy}{N\sum dx^2 - (\sum dx)^2}$
Direct Method: Using sum of X and Y	$bxy = \frac{N\sum XY - \sum X \sum Y}{N\sum Y^2 - (\sum Y)^2}$	$byx = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2}$
Using the correlation coefficient (r) and standard deviation (σ)	$bxy = r \cdot \frac{\sigma x}{\sigma y}$	$byx = r \cdot \frac{\sigma y}{\sigma x}$

Example 2. From the information given below obtain two regression lines X on Y and Yon X using

1. Actual Mean Method
2. Assumed Mean Method
3. Direct Method (Without taking Mean)

Number of Hrs. Machine Operated	7	8	6	9	11	9	10	12

Production (Units in 000):	4	5	2	6	9	5	7	10

Solution:

1. Actual Mean Method

Calculation of Regression Equation

X	Y	x =X - $\bar{X}$	x^2	y =Y - $\bar{Y}$	y^2	xy
7	4	-2	4	-2	4	4
8	5	-1	1	-1	1	1
6	2	-3	9	-4	16	12
9	6	0	0	0	0	0
11	9	2	4	3	9	6
9	5	0	0	-1	1	0
10	7	1	1	1	1	1
12	10	3	9	4	16	12
$\sum X = 72$	$\sum Y = 48$		$\sum x^2 = 28$		$\sum y^2 = 48$	$\sum xy = 36$

$\bar{X} = \frac{\sum X}{N} = \frac{72}{8} = 9$, $\bar{Y} = \frac{\sum Y}{N} = \frac{48}{8} = 6$

Regression equation of X on Y:

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

Where bxy = $\frac{\sum xy}{y^2}$

$= \frac{36}{48} \qquad = .75$

So $(X - 9) = .75 (Y - 6)$

$X - 9 = .75Y - 4.5$

$X = 4.5 + .75Y$

Regression equation of Y on X:

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

Where byx = $\frac{\sum xy}{x^2}$

$= \frac{36}{28} \qquad = 1.286$

So $(Y - 6) = 1.286 (X - 9)$

$Y - 6 = 1.286X - 11.57$

$Y = - 5.57 + 1.286X$

2. Assumed Mean Method

Calculation of Regression Equation

X	Y	dx =X – A (A = 8)	dx^2	dy =Y – A (A = 5)	dy^2	dx *dy
7	4	-1	1	-1	1	1
8	5	0	0	0	0	0
6	2	-2	4	-3	9	6
9	6	1	1	1	1	1
11	9	3	9	4	16	12
9	5	1	1	0	0	0
10	7	2	4	2	4	4
12	10	4	16	5	25	20
$\sum X = 72$	$\sum Y = 48$	$\sum dx = 8$	$\sum dx^2 = 36$	$\sum dy = 8$	$\sum dy^2 = 56$	$\sum xy = 44$

$\overline{X} = \frac{\sum X}{N} = \frac{72}{8} = 9$

$\overline{Y} = \frac{\sum Y}{N} = \frac{48}{8} = 6$

Regression equation of X on Y:

$(X - \overline{X}) = b_{xy} (Y - \overline{Y})$

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

$= \frac{8\,(44) - (8)\,(8)}{8\,(56) - (8)^2}$

$= \frac{352 - 64}{448 - 64} = \frac{288}{384} = .75$

So $(X - 9) = .75\,(Y - 6)$

$X - 9 = .75Y - 4.5$

$X = 4.5 + .75Y$

Regression equation of Y on X:

$(Y - \overline{Y}) = b_{xy} (X - \overline{X})$

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

$= \frac{8\,(44) - (8)\,(8)}{8\,(36) - (8)^2} = \frac{288}{224} = 1.286$

So $(Y - 6) = 1.286\,(X - 9)$

$Y - 6 = 1.286X - 11.57$

$Y = – 5.57 + 1.286X$

3. Direct Method (Without taking Mean)

Calculation of Regression Equation

X	Y	X^2	Y^2	XY
7	4	49	16	28
8	5	64	25	40
6	2	36	4	12
9	6	81	36	54
11	9	121	81	99
9	5	81	25	45
10	7	100	49	70
12	10	144	100	120
$\sum X = 72$	$\sum Y = 48$	$\sum X^{2} = 676$	$\sum Y^{2} = 336$	$\sum XY = 468$

$\overline{X} = \frac{\sum X}{N} = \frac{72}{8} = 9$

$\overline{Y} = \frac{\sum Y}{N} = \frac{48}{8} = 6$

Regression equation of X on Y:

$(X – \overline{X}) = b_{xy} (Y – \overline{Y})$

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

$= \frac{8\ (468) - (72)\ (48)}{8\ (336) - (48)^2} = \frac{3744 - 3456}{2688 - 2304} = \frac{288}{384} = .75$

So $(X – 9) = .75 (Y – 6)$

$X – 9 = .75Y – 4.5$

$X = 4.5 + .75Y$

Regression equation of Y on X:

$(Y – \overline{Y}) = b_{xy} (X – \overline{X})$

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

$= \frac{8\ (468) - (72)\ (48)}{8\ (676) - (72)^2} = \frac{3744 - 3456}{5408 - 5184} = \frac{288}{224} = 1.286$

So $(Y – 6) = 1.286 (X – 9)$

$Y – 6 = 1.286X – 11.57$

$Y = – 5.57 + 1.286X$

Example 3. Find out two Regression equations on basis of the data given below:

	X	Y
Mean	60	80
Standard Deviation (S.D.)	16	20
Coefficient of Correlation	.9	

Also find value of X when Y = 150 and value of Y when X = 100.

Solution: Regression equation of X on Y:

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

Where bxy = $r \frac{\sigma x}{\sigma y}$

$= .9 \frac{16}{20} \qquad = .72$

So $(X - 60) = .72 (Y - 80)$

$X - 60 = .72Y - 57.6$

$X = 2.4 + .72Y$

When Y = 150 than X = 2.4 + .72(150) = 110.4

Regression equation of Y on X:

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

Where byx = $r \frac{\sigma y}{\sigma x}$

$= .9 \frac{20}{16} \qquad = 1.125$

So $(Y - 80) = 1.125 (X - 60)$

$Y - 80 = 1.125X - 67.5$

$Y = 12.5 + 1.125 X$

When X = 100 than Y = 12.5 + 1.125 (100) = 125

Example 4. From the following data find out two lines of regression land also find out value of correlation.

$\sum X = 250$;	$\sum Y = 300$;	$\sum XY = 7900$;
$\sum X^2 = 6500$;	$\sum Y^2 = 10000$; n = 10	

Solution:

$\bar{X} = \frac{\sum X}{N} = \frac{250}{10} = 25$

$\overline{Y} = \frac{\sum Y}{N} = \frac{300}{10} = 30$

Regression equation of Y on X:

$(Y - \overline{Y}) = b_{xy} (X - \overline{X})$

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

$= \frac{10\ (7900) - (250)\ (300)}{10\ (6500) - (250)^2}$

$= \frac{79000 - 75000}{65000 - 62500} = \frac{4000}{2500} = 1.6$

So $(Y - 30) = 1.6\ (X - 25)$

$Y - 30 = 1.6X - 40$

$Y = -10 + 1.6\ X$

Regression equation of X on Y:

$(X - \overline{X}) = b_{xy} (Y - \overline{Y})$

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

$= \frac{10\ (7900) - (250)\ (300)}{10\ (10000) - (300)^2}$

$= \frac{79000 - 75000}{100000 - 90000}$

$= \frac{4000}{10000} = .4$

So $(X - 25) = .4\ (Y - 30)$

$X - 25 = .4Y - 12$

$X = 13 + .4Y$

Coefficients of Correlation

$r = \sqrt{bxy * byx}$

$r = \sqrt{1.6 * 0.4}$

$r = \sqrt{.64}$

$r = .8$

Example 5. From the following data find out two lines of regression land also find out value of correlation. Also find value of Y when X = 30

$\sum X = 140;$ $\sum Y = 150;$ $\sum (X - 10)(Y - 15) = 6;$

$\sum (X - 10)^2 = 180;$ $\sum (Y - 15)^2 = 215;$ $n = 10$

Solution: Let's take assumed mean of Series X = 10 and Series Y = 15.

$\sum dx = \sum (X - 10) = \sum X - 10n = 140 - 100 = 40$

$\sum dy = \sum (Y - 15) = \sum Y - 15n = 150 - 150 = 0$

$\sum dx^2 = \sum (X - 10)^2 = 180$

$\sum dy^2 = \sum (Y - 15)^2 = 215$

$\sum dxdy = \sum (X - 10)(Y - 15) = 6$

So,

$\overline{X} = A + \frac{\sum X}{N} = 10 + \frac{40}{10} = 14$

$\overline{Y} = A + \frac{\sum Y}{N} = 15 + \frac{0}{10} = 15$

Regression equation of Y on X:

$(Y - \overline{Y}) = b_{xy} (X - \overline{X})$

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

$= \frac{10(6) - (40)(0)}{10(180) - (40)^2} = \frac{60}{200} = .3$

So, $(Y - 15) = .3 (X - 14)$

$Y - 15 = .3X - 4.2$

$Y = 10.8 + .3X$

When X = 30 than $Y = 10.8 + .3(30) = 19.8$

Regression equation of Y on X:

$(Y - \overline{Y}) = b_{xy} (X - \overline{X})$

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

$= \frac{10(6) - (40)(0)}{10(25) - (0)^2} = \frac{60}{250} = .24$

So $(Y - 15) = .24 (X - 14)$

$Y - 15 = .24X - 3.36$

$Y = 11.64 + .24X$

Coefficients of Correlation

$r = \sqrt{bxy * byx} = \sqrt{.3 * .24}$

r = $\sqrt{.072}$

r = .268

Example 5. From the following data find out which equation is equation X on Y and which equation is equation Y on X. Also find $\overline{X}$, $\overline{Y}$ and r.

$3X + 2Y – 26 = 0$

$6X + Y – 31 = 0$

Solution: To find $\overline{X}$ and $\overline{Y}$, we will solve following simultaneous equations

$3X + 2Y = 26$ ……………………………. (i)

$6X + Y = 31$ ……………………………. (ii)

Multiply equation (i) with 2, we get

$6X + 4Y = 52$ ……………………………. (iii)

Deduct equation (ii) from equation (iii)

$6X + 4Y = 52$

$-6X - Y = -31$

$3Y = 21$

$Y = 7$

Or $\overline{Y} = 7$.

Put the value of Y in Equation (i), we get

$3X + 2(7) = 26$

$3X + 14 = 26$

$3X = 12$

$X = 4$

or $\overline{X} = 4$

Let $3X + 2Y = 26$ be regression equation X on Y

$3X = 26 – 2Y$

$X = \frac{26}{3} – \frac{2}{3} Y$

So $b_{xy} = – \frac{2}{3}$

Let $6X + Y = 31$ be regression equation Y on X

$Y = 31 – 6X$

So $b_{yx} = -6$

As r = $\sqrt{bxy\ *\ byx}$

r = $-\sqrt{-(\frac{2}{3})\ \times (-.6\ }\ )$

r = -2, but this is not possible as value of r always lies between -1 and +1. So, our assumption is wrong and equation are reverse.

Let 6 X + Y = 31 be regression equation X on Y

6X = 31 –Y

X = $\frac{31}{6} - \frac{1}{6}$ Y

So, $b_{xy} = -\frac{1}{6}$

Let 3X + 2Y = 26 be regression equation Y on X

2Y = 26 – 3X

Y = $\frac{26}{2} - \frac{3}{2}$ X

So $b_{yx} = -\frac{3}{2}$

As r = $\sqrt{bxy\ *\ byx}$

r = $-\sqrt{-(\frac{1}{6})\ \times -(\frac{3}{2})\ }\ )$

r = $-.5$, which is possible. So, our assumption is right.

So, $\overline{Y} = 7$; $\overline{X} = 4$;

X on Y is X = $\frac{31}{6} - \frac{1}{6}$ Y

Y on X is Y = $\frac{26}{2} - \frac{3}{2}$ X

r = $-.5$

TEST YOUR UNDERSTANDING

1. Find both regression equations:

X	6	2	10	4	8
Y	9	11	5	8	7

2. From following estimate the value of Y when X = 30 using regression equation.

X	25	22	28	26	35	20	22	40	20	18	19	25

Y	18	15	20	17	22	14	15	21	15	14	16	17

3. Fit two regression lines:

X	30	32	38	35	40
Y	10	14	16	20	15

Find X when Y = 25 and find Y when X = 36.

4. Find out two Regression equations on the basis of the data given below:

	X	Y
Mean	65	67
Standard Deviation (S.D.)	2.5	3.5
Coefficient of Correlation	.8	

5. In a data the Mean values of X and Y are 20 and 45 respectively. Regression coefficient $b_{yx} = 4$ and b_{xy} 1/9. Find

a. coefficient of correlation
b. Standard Deviation of X, if S.D. of Y = 12
c. Find two regression lines

6. You are supplied with the following information. Variance of X = 36

$12X - 51Y + 99 = 0$

$60X - 27Y = 321.$

Calculate: (a) The average values of X and Y

(b) The standard deviation of Y and

7. The lines of regression of Y on X and X on Y are Y = X + 5 and 16X = 9Y + 4 respectively

Also, $\sigma_y = 4$ Find $\bar{X}$, $\bar{Y}$, σ_x and r.

8. Given: $\sum X = 56$, $\sum Y = 40$, $\sum X^2 = 524$

$\sum Y^2 = 256$, $\sum XY = 364$, N=8, find the regression equation of X on Y

Answers

1. X = 16.4 – 1.3Y, Y = 11.9 - .65X
2. 18.875
3. Y = .46X – 1.1, X = .6Y + 26, Value of Y = 15.46, Value of X = 40.25
4. Y = 1.12X – 5.8, X = .57Y +26,81

5.	.67, 2, Y= 4X – 35 and X = 1/9 Y +15
6.	Mean of X = 13, Mean of Y= 17, S.D of Y = 8
7.	Mean of X = 7, Mean of Y= 12, S.D of X = 3, r= .75
8.	X = 1.5Y - 0.5, r = .977

4.10 RELATIONSHIP BETWEEN CORRELATION AND REGRESSION

1. Correlation is a quantitative tool that measure of the degree of relationship that is present between two variables. It shows the degree and direction of the relation between two variables. Regression helps us to find the value of a dependent variable when the value of independent variable is given.
2. Correlation between two variables is the same. For example, if we calculate the correlation between sales and advertising or advertising and sales, the value of correlation will remain the same. But this is not true for Regression. The regression equation of Advertising on sales will be different from regression equation of Sales on advertising.
3. If there is a positive correlation, the distance between the two lines will be less. That means the two regression lines will be closer to each other- Similarly, if there is a low correlation, the lines will be farther from each other. A positive correlation implies that the lines will be upward-sloping whereas a negative correlation implies that the regression lines will be downward sloping.
4. Correlation between two variables can be calculated by taking the square root of the product of the two regression coefficients.

Following is some of the differences between Correlation and Regression:

1. Correlation measures the degree and direction of relationship between two variables. Regression measures the change in value of a dependent variable given the change in value of an independent variable.
2. Correlation does not depict a cause-and-effect relationship. Regression depicts the causal relationship between two variables.
3. Correlation is a relative measure of linear relationship that exists between two variables. Regression is an absolute measure that measures the change in value of a variable.
4. Correlation between two variables is the same. In other words, Correlation between the two variables is the same. $r_{xy} = r_{yx}$. Regression is not symmetrical in formation. So, the regression coefficients of X on Y and Yon X are different.
5. Correlation is independent of Change in origin or scale. Regression is independent of Change in origin but not of scale.
6. Correlation is not capable of any further mathematical treatment. Regression can be further treated mathematically.
7. The coefficient of correlation always lies between -1 and +1. The regression coefficient can have any value.

4.11 SUM UP

- there are two regression equations X on Y and Yon X.
- Regression can be linear or nonlinear.
- It can be simple or multiple.
- Regression is based on the principle of Least Squares.
- We can also find out correlation coefficient with help of regression coefficients.
- Regression is a useful tool for forecasting.
- With the help of regression, we can predict the value of can find the value of X if the value of Y is given or the value of Y if value of X is given.
- It creates the mathematical linear relation between two variables X and Y, out of which one variable is dependent and other is independent.

4.12 QUESTIONS FOR PRACTICE

Q1. What is Regression? Explain uses of Regression.

Q2. Discuss the properties of regression analysis.

Q3. What is the relationship between Regression and correlation?

Q4. Explain different types of regressions.

Q5. How two regression lines are determined under direct method?

Q6. Explain various methods of finding regression equations.

Q7. What are limitations of regression analysis?

Q8. What are properties of regression coefficients?

4.13 SUGGESTED READINGS

- J. K. Sharma, *Business Statistics,* Pearson Education.
- S.C. Gupta, *Fundamentals of Statistics,* Himalaya Publishing House.
- S.P. Gupta and Archana Gupta, *Elementary Statistics,* Sultan Chand and Sons, New Delhi.
- Richard Levin and David S. Rubin, *Statistics for Management,* Prentice Hall of India, New Delhi. Hill Publishing Co.

M.A (ECONOMICS)

QUANTITATIVE METHODS II

SEMESTER -II

Unit 5: Meaning of Hypothesis, Characteristics of Hypothesis, Basic Concepts, Hypothesis Testing Procedures, Introduction to parametric and non-parametric tests

STRUCTURE

5.0 LEARNING OBJECTIVES

After Reading this unit, Learners can able to know about the:

- Meaning of Testing of Hypothesis
- Types of hypotheses
- Types of tails used in testing
- Types of errors in the testing of hypothesis
- Level of significance
- Power of a test

5.1 INTRODUCTION/ MEANING OF HYPOTHESIS

Testing of hypotheses is a fundamental concept in statistics and scientific research that plays a crucial role in decision-making and conclusions based on data. It involves a systematic and structured approach to evaluate and validate assumptions or claims about a population. The terms "hypo" and "thesis" combine to form the word "hypothesis." Hypo implies a subject to verification, while the thesis is a statement on how to solve a problem. Therefore, the definition of the word hypothesis is an assumption regarding how a problem might be solved. A hypothesis presents a solution to the issue that must be tested empirically and is supported by some logic.

A hypothesis is an unproven assertion of the association between two or more variables. A hypothesis is a clear, verifiable forecast of what will occur in your investigation. The population, the variables, and the relationships between the variables are necessary for the hypothesis to be complete. Hypothesis does not have to be correct. While the hypothesis forecasts what the researchers expect to see, research aims to determine whether this guess is right or wrong. When experimenting, researchers might explore several different factors to determine which ones might contribute to the outcome. In many cases, researchers may find that the results of an experiment do not support the original hypothesis. When writing up these results, the researchers might suggest other options that should be explored in future studies.

5.2 CHARACTERISTICS OF HYPOTHESIS

A good hypothesis must possess the following characteristics are as follows:

a. conceptually clear: The concepts used in the hypothesis should be clearly defined, not only formally but also, if possible, operationally. The formal definition of the concepts will clarify what a particular concept stands for, while the operational definition will leave no ambiguity about what would constitute the empirical evidence or indicator of the concept on the plane of reality.

b. should be specific: No vague or value-judgmental terms should be used in the formulation of a hypothesis. It should specifically state the posited relationship between the variables. It should include a clear statement of all the predictions and operations indicated therein and they should be precisely spelled out.

c. should be empirically testable: It should have empirical referents so that it will be possible to deduce certain logical deductions and inferences about it. Therefore, a researcher should take utmost care that his/her hypothesis embodies concepts or variables that have clear empirical correspondence and not concepts or variables that are loaded with moral judgments or values. Such statements as 'criminals are no worse than businessmen' and 'capitalists exploit their workers', in other words, a researcher should avoid using terms loaded with values or beliefs or words having moral or attitudinal connotations in his hypothesis.

d. related to available techniques: The researcher may be ignorant of the available techniques, which makes him/her weak in formulating a workable hypothesis. A hypothesis, therefore, needs to be formulated only after due thought has been given to the methods and techniques that can be used for measuring the concepts or variables incorporated in the hypothesis.

e. related to a body of theory or some theoretical orientation: A hypothesis, if tested, helps to qualify, support, correct or refute an existing theory, only if it is related to some theory or has some theoretical orientation. A hypothesis imaginatively formulated does not only elaborate and improve existing theory but may also suggest important links between it and some other theories. Thus, the exercise of deriving a hypothesis from a body of theory may also be an occasion for a scientific leap into newer areas of knowledge.

5.3 BASIC CONCEPTS OF HYPOTHESIS

The hypothesis is a fundamental concept in the scientific method and research process. It serves as a starting point for investigations and experiments, guiding researchers in their pursuit of knowledge. Here are some basic concepts related to hypotheses.

A. Null Hypothesis

The null hypothesis is a general statement that states that there is no relationship between two phenomena under consideration or that there is no association between two groups. This hypothesis is either rejected or not rejected based on the viability of the given population or sample.

In other words, the null hypothesis is a hypothesis in which the sample observations result from the chance. It is said to be a statement in which the evaluators want to examine the data. It is denoted by H_0. In statistics, the null hypothesis is usually denoted by the letter H with subscript '0' (zero), such that H_0 (pronounced as H-null or H-zero or H-nought). At the same time, the alternative hypothesis expresses the observations determined by the non-random cause. It is represented by H_1 or Ha. The main purpose of a null hypothesis is to verify/ disprove the proposed statistical assumptions.

An example of the hypothesis is as, If the hypothesis is that, "If random test scores are collected from men and women, does the score of one group differ from the other?" a possible null hypothesis will be that the mean test score of men is the same as that of the women.

H_0: $\mu_1 = \mu_2$

H_0= null hypothesis
μ_1= mean score of men
μ_2= mean score of women

Sometimes the null hypothesis is rejected too. If this hypothesis is rejected means, that research could be invalid. Many researchers will neglect this hypothesis as it is merely opposite to the alternate hypothesis. It is a better practice to create a hypothesis and test it. The goal of researchers is not to reject the hypothesis. However, a perfect statistical model is always associated with the failure to reject the null hypothesis.

B. Alternative Hypothesis

An alternative hypothesis is a statement that describes that there is a relationship between two selected variables in a study.

- An alternative hypothesis is usually used to state that a new theory is preferable to the old one (null hypothesis).
- This hypothesis can be simply termed as an alternative to the null hypothesis.
- The alternative hypothesis is the hypothesis that is to be proved that indicates that the results of a study are significant and that the sample observation does not result just from chance but from some non-random cause.
- If a study is to compare method A with method B about their relationship and we assume that method A is superior or method B is inferior, then such a statement is termed an alternative hypothesis.

The symbol of the alternative hypothesis is either H_1 or H_a while using less than, greater than, or not equal signs.

The following are some examples of alternative hypothesis:

If a researcher is assuming that the bearing capacity of a bridge is more than 10 tons, then the hypothesis under this study will be:

Null hypothesis H_0: μ= 10 tons
Alternative hypothesis H_a: μ>10 tons

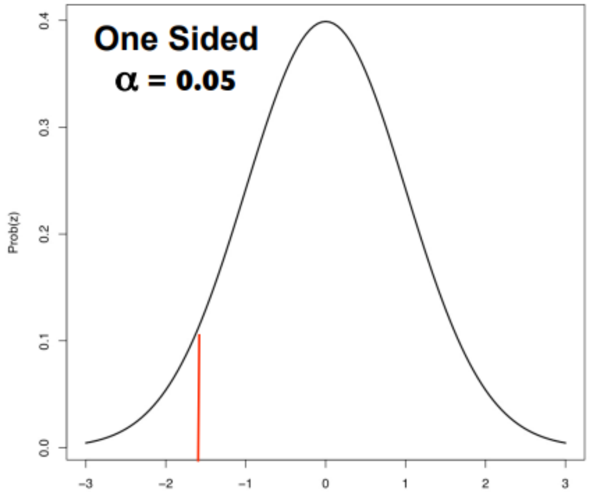
A. One-tailed & Two-tailed

A test of testing the null hypothesis is said to be a two-tailed test if the alternative hypothesis is two-tailed whereas if the alternative hypothesis is one-tailed then a test of testing the null hypothesis is said to be a one-tailed test. For example, if our null and alternative hypothesis is H_0: $\mu = \mu_0$ and H_1: $\mu \neq \mu_0$ then the test for testing the null hypothesis is two-tailed because the alternative hypothesis is two-tailed which means, the parameter μ can take value greater than μ_0 or less than μ_0. If the null and alternative hypotheses are H_0: $\mu = \mu_0$ H_1: $\mu > \mu_0$ then the test for testing the null hypothesis is right-tailed because the alternative hypothesis is right-tailed. Similarly, if the null and alternative hypotheses are H_0: $\mu = \mu_0$ H_1: $\mu < \mu_0$ then the test for testing the null hypothesis is left-tailed because the alternative hypothesis is left-tailed. The above discussion can be summarised in the Table below:

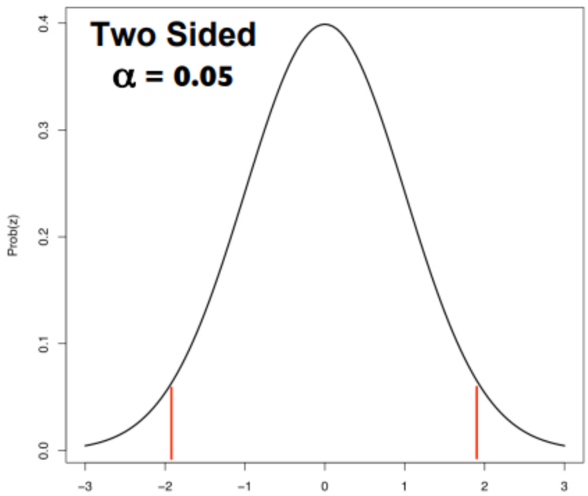
Table: Null and Alternative Hypothesis (Right and Left tailed test)

Null Hypothesis	Alternative Hypothesis	Types of Critical Region
H_0: $\mu = \mu_0$	H_1: $\mu \neq \mu_0$	Two-tailed test having critical regions under both tails
H_0: $\mu = \mu_0$	H_1: $\mu > \mu_0$	Right-tailed test having critical region under right tail only
H_0: $\mu = \mu_0$	H_1: $\mu > \mu_0$	Left-tailed test having critical region under left tail only

Let us do one example based on the type of tests.



Critical Value = -1.64



Critical Values = -1.96 and +1.96

Example 3: A company has replaced its original technology of producing electric bulbs with CFL technology. The company manager wants to compare the average life of bulbs manufactured by original technology and new technology CFL. Write appropriate null and alternate hypotheses. Also, say about one-tailed and two-tailed tests.

Solution: Suppose the average lives of original and CFL technology bulbs are denoted by: μ_1 and μ_2 respectively. If the company manager is interested just in knowing whether any significant difference exists in the average time of two types of bulbs then null and alternative hypotheses will be:

H0: $\mu_1 = \mu_2$ [average lives of two types of bulbs are same]

H1: $\mu_1 \neq \mu_2$ [average lives of two types of bulbs are different]

Since the alternative hypothesis is two-tailed therefore corresponding test will be two-tailed.

If company manager is interested just to know whether the average life of CFL is greater than original technology bulbs then our null and alternative hypotheses will be

H0: $\mu_1 \geq \mu_2$

H1: $\mu_1 < \mu_2$

Here, average life of CFL technology bulbs is greater than the average life of original technology,

Since alternative hypothesis is left-tailed therefore corresponding test will be a left-tailed test.

Difference between Null Hypothesis and Alternative Hypothesis

Now, let us discuss the difference between the null hypothesis and the alternative hypothesis.

Null Hypothesis	Alternative Hypothesis
1. Denoted by H_0	Denoted by H_1
2. The null hypothesis is a statement. There exists no relation between the two variables	An alternative hypothesis is a statement, that there exists some relationship between two measured phenomena
3. It is the hypothesis that the researcher tries to disprove.	It is a hypothesis that the researcher tries to prove.
4. The mathematical formulation of the null hypothesis is an equal sign	The mathematical formulation alternative hypothesis is an inequality sign such as greater than, less than, etc.
5. The result of the null hypothesis indicates no changes in opinions or actions.	The result of an alternative hypothesis causes changes in opinions and actions.
6. The observations of this hypothesis are the result of chance	The observations of this hypothesis are the result of real effect
7. If the null hypothesis is accepted, the results of the study become insignificant.	If an alternative hypothesis is accepted, the results of the study become significant.
8. If the p-value is greater than the level of significance, the null hypothesis is accepted.	If the p-value is greater than the level of significance, the null hypothesis is accepted.

5.3.3 ERRORS IN HYPOTHESIS

If the value of test statistic falls in rejection (critical) region then we reject the null hypothesis and if it falls in the non-rejection region then we do not reject the null hypothesis. A test statistic is calculated based on observed sample observations. But a sample is a small part of the population about which decision is to be taken. A random sample may or may not be a good representative of the population. A faulty sample misleads the inference (or conclusion) relating to the null hypothesis. For example, an engineer infers that a packet of screws is sub-standard when it is not. It is an error caused by to poor or inappropriate (faulty) sample. Similarly, a packet of screws may infer good when it is sub-standard. So, we can commit two kinds of errors while testing a hypothesis which are summarised in Table 9.1 which is given below:

Table: Types of Error

Decision	H_0 True	H_1 True
Reject H_0	Type I Error (α)	Correct decision
Do not Reject H_0	Correct Decision	Type II Error

		(Power of test) β

Let us take a situation where a patient suffering from high fever reaches a doctor. Suppose the doctor formulates the null and alternative hypotheses as

H_0: The patient has a Stomach Infection

H_1: The patient has not a Stomach Infection

The following cases arise:

Case I: Suppose that hypothesis H_0 is true, that is, the patient is a Stomach Infection and after observation, pathological and clinical examination, the doctor rejects H_0, that is, he/she declares him/her a non-Stomach Infection patient. It is not a correct decision and he/she commits an error in a decision known as a type-I error.

Case II: Suppose that hypothesis H_0 is false, that is, the patient is a non-Stomach Infection patient and after observation, the doctor rejects H_0, that is, he/she declares him/her a non-Stomach Infection patient. It is a correct decision.

Case III: Suppose that hypothesis H_0 is true, that is, the patient is a Stomach Infection patient and after observation, the doctor does not reject H_0, that is, he/she declares him/her a Stomach Infection patient. It is a correct decision.

Case IV: Suppose that hypothesis H_0 is false, that is, the patient is a non-Stomach Infection patient and after observation, the doctor does not reject H_0, that is, he/she declares him/her a Stomach Infection patient. It is not a correct decision and he/she commits an error in a decision known as a type-II error.

5.3.4 LEVEL OF SIGNIFICANCE

The level of significance is the probability of rejecting a true null hypothesis that is the probability of "Type I error" and is denoted by α. The frequently used values of α are 0.05 (i.e., 5 %); 0.01(i.e., 1 %); 0.1(i.e., 10 %), etc. When $\alpha = 0.05$ it means that the level of significance is 5%. $\alpha = 0.01$ which means 1% level of significance. $\alpha = 0.01$ which means 10% level of significance. In fact, α specifies the critical region. If the calculated value of the test statistic lies in the rejection (critical) region, then we reject the null hypothesis and if it lies in the non-rejection region, then we do not reject the null hypothesis. Also, we note that when H_0 is rejected then automatically the alternative hypothesis H_1 is accepted. Now, one point of our discussion is how to decide critical value(s) or cut-off value(s) for a known test statistic. Suppose the distribution of test statistics could be expressed into some well-known distributions like Z, χ2, t, F test etc. Then our problem will be solved and using the probability distribution of test statistics, we can find the cut-off value(s) that provides us critical area equal to 5% (or 1%). 16 Testing of Hypothesis Another viewpoint about the level of significance relates to the trueness of the conclusion. If H_0 does not reject at level, say, $\alpha = 0.05$ (5% level) then a person will be confident that "concluding statement about H_0" is true with 95% assurance. But even then, it may be false with 5%a chance. There is no percent assurance about the trueness of the statement made for H_0. As an example, if among 100 scientists, each draws a random sample and uses the same test statistic to test the same hypothesis H_0 conducting sathe me experiment, then 95 of them will reach the same conclusion about H_0. But still, 5 of them may differ (i.e., against the earlier conclusion). A similar argument can be made for, say, $\alpha = 0.01$ (=1%). It is like when H_0 is rejected at $\alpha = 0.01$ by a scientist, then out of 100 similar researchers who work together at the same time for the same problem, but with different random samples, 99 of them would reach the same conclusion however, one may differ.

5.3.5 Confidence Interval

Confidence interval is the interval marked by limits within which the population value lies by chance and the hypothesis is considered to be acceptable. If an observed value falls in the confidence interval H_0 is accepted.

5.3.6 Degree of Freedom

Degree of freedom refers to the number of values that are free to vary after we have given the number of restrictions imposed upon the data. It is commonly abbreviated by df. In statistics, it is the number of values in a study that are free to vary. The statistical formula to find out how many degrees of freedom are there is quite simple. It implies that degrees of freedom are equivalent to the number of values in a data set minus 1, and appears like this:

df=N−1

Where N represents the number of values in the data set (sample size).

That being said, let's have a look at the sample calculation.

If there is a data set of 6, (N=6).

Call the data set X and make a list with the values for each data.

For this example, data, set X of the sample size includes: 10, 30, 15, 25, 45, and 55

This data set has a mean, or average of 30. Find out the mean by adding the values and dividing by N:

(10 + 30 + 15 + 25 + 45 + 55)/6= 30

Using the formula, the degrees of freedom will be computed as df = N-1:

In this example, it appears, df = 6-1 = 5

This further implies that, in this data set (sample size), five numbers contain the freedom to vary as long as the mean remains 30.

5.3.7 Power of Test

Nowadays use of p-value is becoming more and more popular because of the following two reasons:

- most the statistical software provides a p-value rather than a critical value.

- p-value provides more information compared to critical value as far as rejection or not rejection of H_0

Moving in this direction, we note that in scientific applications one is not only interested in rejecting or not rejecting the null hypothesis but he/she is also interested in assessing how strong the data has the evidence to reject H_0.

This smallest level of significance (α) is known as the "p-value". The p-value is the smallest value of the level of significance(α) at which a null hypothesis can be rejected using the obtained value of the test statistic. The p-value is the probability of obtaining a test statistic equal to or more extreme (in the direction of sporting H_1) than the actual value obtained when null hypothesis is true.

The p-value for various tests can be obtained with the help of the tables. But unless we are dealing with the standard normal distribution, the exact p-value is not obtained with the tables as mentioned above. But if we test our hypothesis with the help of computer packages or software such as SPSS, SAS, MINITAB, STATA, EXCEL, etc. These types of computer packages or software present the p-value as part of the output for each hypothesis testing procedure. Therefore, in this block, we will also describe the procedure to decide on the null hypothesis based on critical value as well as p-value concepts.

5.4 CRITICAL REGION

Results from statistical tests will fall into one of two regions: the rejection region, which will lead you to reject the null hypothesis, or the acceptance region, where you provisionally accept the null hypothesis. The acceptance region complements the rejection region; If your result does not fall into the rejection region, it must fall into the acceptance region.

Critical values separate the values that support or reject the null hypothesis and are calculated based on alpha. We will see more examples later on and it will be clear how we choose α. Based on the alternative hypothesis, three cases of critical region arise:

Case A) Two-tailed test:

In this hypothesis testing method, the critical region lies on both sides of the sampling distribution. It is also known as a non - non-directional hypothesis testing method. The two-tailed test is used when it needs to be determined if the population parameter is assumed to be different than some value. The hypothesis can be set up as follows:

H_0: the population parameter = some value

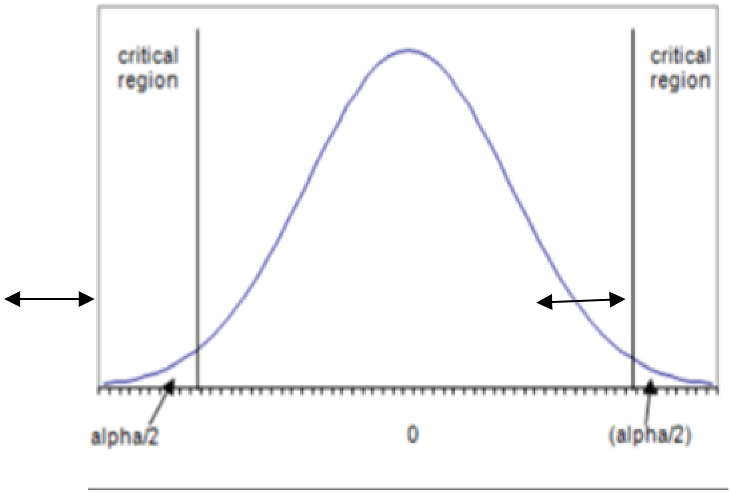
H_1: the population parameter ≠ some value

The null hypothesis is rejected if the test statistic has a value that is not equal to the critical value.

Therefore, H_0: $\mu = \mu_0$

H_1: $\mu \neq \mu_0$

Figure1: Two-tailed hypothesis testing



Case B) Left-tailed test:

The left tail test is also known as the lower tail test. It is used to check whether the population parameter is less than some value. The hypotheses for this hypothesis testing can be written as follows:

H_0: The population parameter is ≥ some value

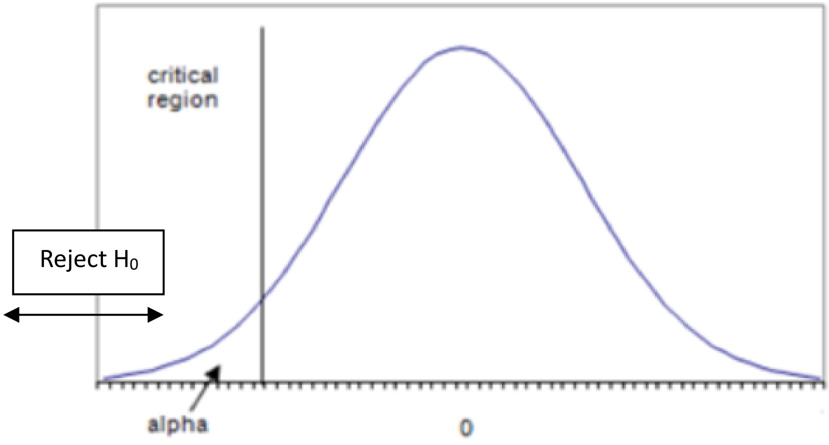
H_1: The population parameter is < some value.

The null hypothesis is rejected if the test statistic has a value lesser than the critical value.

Therefore, H_0: $\mu = \mu_0$

H_1: $\mu < \mu_0$

Figure 2: Left-tailed hypothesis testing



Case C) Right-tailed test:

The right tail test is also known as the upper tail test. This test is used to check whether the population parameter is greater than some value. The null and alternative hypotheses for this test are given as follows:

H_0: The population parameter is ≤ some value

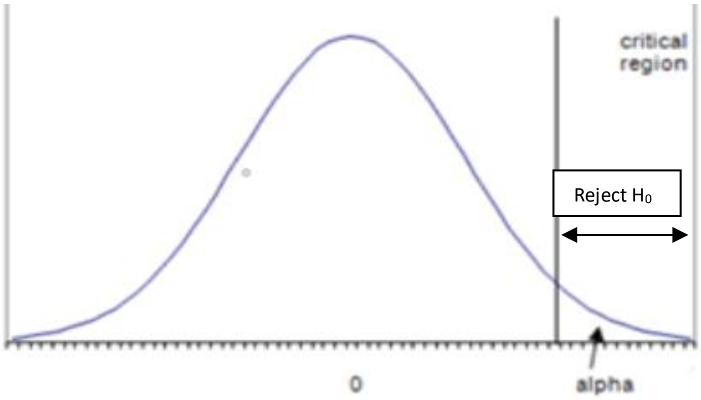
H_1: The population parameter is > some value.

If the test statistic has a greater value than the critical value then the null hypothesis is rejected.

Therefore, H_0: $\mu = \mu_0$

H_1: $\mu > \mu_0$

Figure 3: Right-tailed hypothesis testing



5.5 ACCEPTANCE REGION

The acceptance region is "the interval within the sampling distribution of the test statistic that is consistent with the null hypothesis H_0 from hypothesis testing." In more simple terms, let's say you run a hypothesis test like a z-test. The results of the test come in the form of a z-value, which has a large range of possible values. Within that range of values, some will fall into an interval that suggests the null hypothesis is correct. That interval is the acceptance region. As shown in figure 4 shows a two-tailed having 95 % acceptance region i.e., 2.5 % from the left side and 2.5 % from the right side.

Figure 4: Two-tailed hypothesis testing

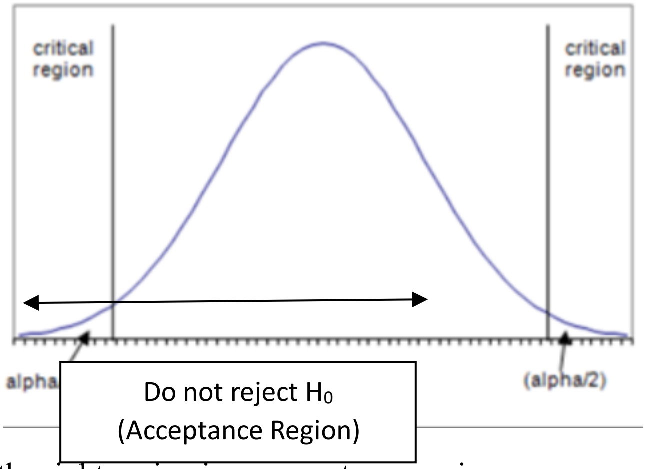
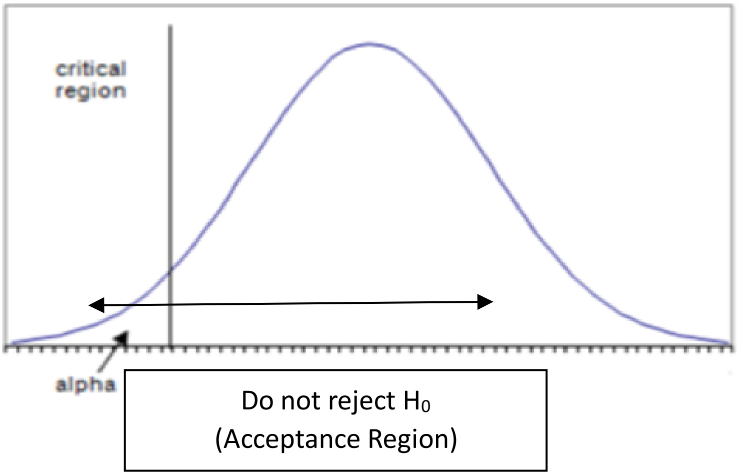


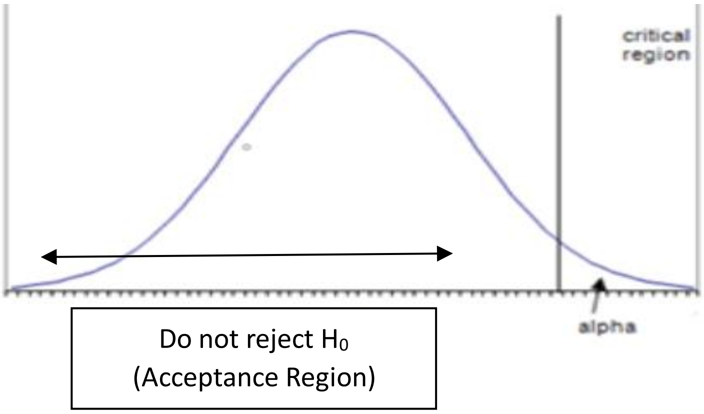
Figure 5, shows that 95 % of the right region is an acceptance region.

Figure 5: Left-tailed hypothesis testing



However, figure 6, shows 95 % of the left region as an acceptance region.

Figure 6: Right-tailed hypothesis testing



5.6 HYPOTHESIS TESTING PROCEDURE

Testing of hypothesis is a huge demanded statistical tool by many disciplines and professionals. It is a step-by-step procedure as you will see in the next three units through a large number of examples. The following steps are involved in hypothesis testing:

Step I: First of all, we have to set up null hypothesis H_0 and alternative hypothesis H_1. Suppose, we want to test the hypothetical / claimed / Testing of Hypothesis assumed value μ_0 of parameter μ.

So, we can take the null and alternative hypotheses as H_0: $\mu = \mu_0$

H_1: $\mu \neq \mu_0$ (for the two-tailed test)

While one- tail test as:

H_0: $\mu = \mu_0$ and H_1: $\mu > \mu_0$

H_0: $\mu = \mu_0$ and H_1: $\mu < \mu_0$

In case of comparing the same parameter of two populations of interest, say, μ_1 and μ_2, then our null and alternative hypotheses would be

H_0: and $\mu_1 = \mu_2$ and H1: $\mu_1 \neq \mu_2$ (for two-tailed test)

While one- tail test as:

H_0: $\mu_1 \leq \mu_2$ and H_1: $\mu_1 > \mu_2$

H_0: $\mu_1 \geq \mu_2$ and H_1: $\mu_1 < \mu_2$

Step II: After setting the null and alternative hypotheses, we establish a criteria for rejection or non-rejection of null hypothesis, that is, decide the level of significance (α), at which we want to test our hypothesis. The most common value of α is 0.05 or 5%. Other popular choices are 0.01 (1%) and 0.1 (10%).

Step III: The third step is to choose an appropriate test statistics form like Z (standard normal), $\chi 2$, t, F or any other well-known in literature.

Step IV: Obtain the critical value(s) in the sampling distribution of the test statistic and construct the rejection (critical) region of size α. Generally, critical values for various levels of significance are putted in the form of a table for various standard sampling distributions of test statistics such as Z-table, χ 2 -table, t-table, etc.

Step V: After that, compare the calculated value of test statistic obtained from Step IV, with the critical value(s) obtained in Step V and locate the position of the calculated test statistic, that is, it lies in the rejection region or non-rejection region.

Step VI: ultimately testing the hypothesis, we have to conclude.

It is done as explained below:

(i) If the calculated test statistic value lies in the rejection region at the significance level, then we reject the null hypothesis. It means that the sample data provide us sufficient evidence against the null hypothesis and there is a significant difference between hypothesized value and observed value of the parameter.

(ii) If the calculated test statistic value lies in the non-rejection region at the significance level, then we do not reject the null hypothesis. It means that the sample data fails to provide sufficient evidence against the null hypothesis and the difference between hypothesized value and observed value of the parameter due to sample fluctuation.

Nowadays the decision about the null hypothesis is taken with the help of p-value. The concept of p-value is very important because computer packages and statistical software such as SPSS, STATA, MINITAB, EXCEL, etc., all provide p-value.

Example 2: Mean average weight of men is greater than 100 kgs with a standard deviation of 15kgs. 30 men are chosen with an average weight of 112.5 Kg. Using hypothesis testing, check if there is enough evidence to support the researcher's claim. Check the significance at 5 % level.

Step 1: This is an example of a right-tailed test. Set up the null hypothesis and alternative hypothesis as

H_0: $\mu = 100$.

The alternative hypothesis is given by

H_1: $\mu > 100$.

Step 2: Level of Significance:

As this is a one-tailed test,

$\alpha = 5\%$. This can be used to determine the critical value.

$1 - \alpha = 1 - 0.05 = 0.95$

0.95 gives the required area under the curve. Now using a normal distribution table, the area 0.95 is at $z = 1.645$. A similar process can be followed for a t-test. The only additional requirement is to calculate the degrees of freedom given by $n - 1$.

Step 3: Select the statistic

Here, we must use the z statistic to test the null hypothesis since the variance is known.

Step 4: Find the critical region:

The z-value obtained from the statistical Table for z is 1.645. Hence, the critical region for a one-tailed test is: $z > 1.645$.

Step 5: Calculate the z-test statistic. This is because the sample size is 30. Furthermore, the sample and population means are known along with the standard deviation.

$Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$

μ = 100, $\bar{X}$= 112.5, n=30, σ= 15

$Z = \frac{12.5-100}{\frac{15}{\sqrt{30}}} = 4.56$

Step 6: Conclusion. As Cal Z > tab Z

i.e., 4.56 > 1.645 thus, the null hypothesis can be rejected.

Example 3 of left-tail: The average score of a class is 90. However, a teacher believes that the average score might be lower. The scores of 6 students were randomly measured. The mean was 82 with a standard deviation of 18. With a 0.05 significance level use hypothesis testing to check if this claim is true.

Solution. Step 1: Set up the null and alternative hypotheses as

H0: μ = 90,

Alternative hypothesis

H1: μ < 90 (Left-tailed)

x = 110, μ = 90, n = 6, s = 18

Step 2: level of significance:

As this is a one-tailed test,

α = 5%. This can be used to determine the critical value.

1 - α = 1 - 0.05 = 0.95, df= N-1= 6-1=5

Step 3: It is a small sample test; therefore t-test is to be determined as

$t = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$

Step 4: Determine the critical value

The critical value from the t table is -2.015

Step 5: Test Statistics

$t = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$

$t = \frac{82-90}{\frac{18}{\sqrt{6}}}$ $\quad$ t = -1.088

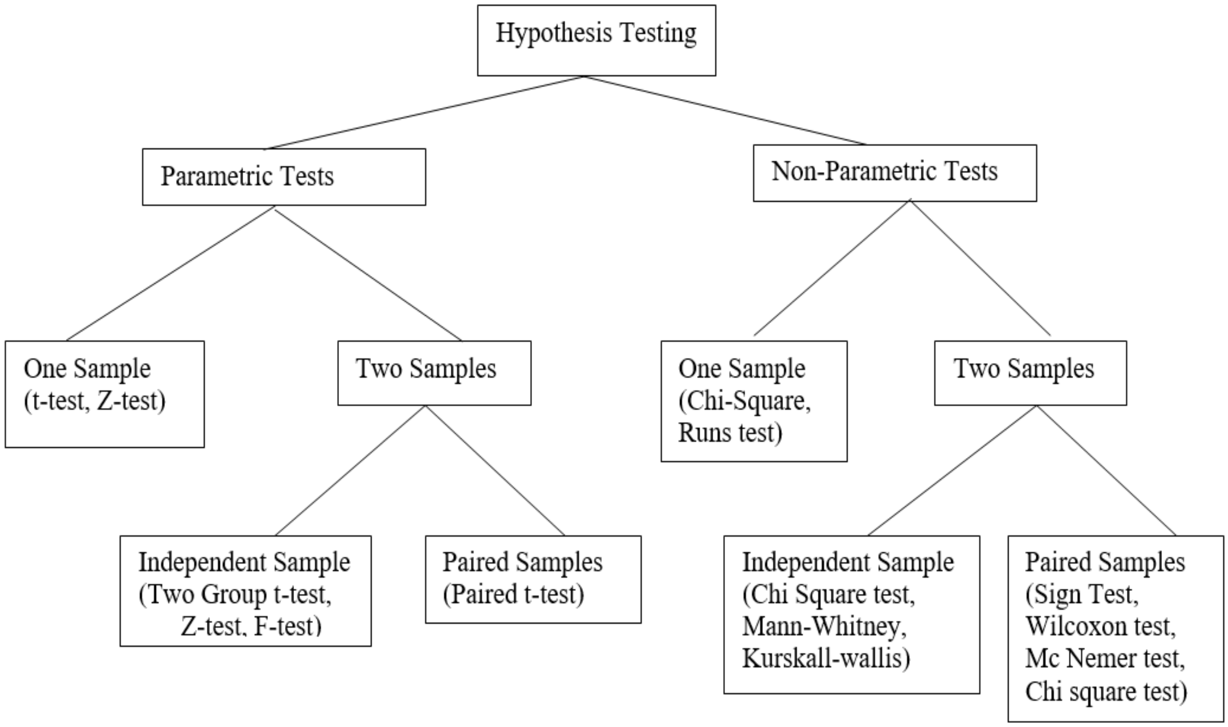
Step 6: Conclusion

As -1.088 > -2.015, we fail to reject the null hypothesis. There is not enough evidence to support the claim.

5.7 TYPES OF HYPOTHESIS TESTING

Hypothesis testing is a fundamental statistical technique used to make inferences about populations based on sample data. Several types of hypothesis tests are designed for different scenarios and research questions.

- Parametric Tests: These tests assume that the data follows a specific probability distribution, typically the normal distribution. Common parametric tests include t-tests, analysis of variance (ANOVA), and linear regression.
- Non-Parametric Tests: These tests make fewer assumptions about the data distribution. They are used when the data is not normally distributed or when you have ordinal or categorical data. Examples include the Wilcoxon signed-rank test and the Kruskal-Wallis test.


Hypothesis Testing
Parametric Tests
Non-Parametric Tests
One Sample (t-test, Z-test)
Two Samples
One Sample (Chi-Square, Runs test)
Two Samples
Independent Sample (Two Group t-test, Z-test, F-test)
Paired Samples (Paired t-test)
Independent Sample (Chi Square test, Mann-Whitney, Kurskall-wallis)
Paired Samples (Sign Test, Wilcoxon test, Mc Nemer test, Chi square test)

5.8 PARAMETRIC TEST

These tests are based on several assumptions about the parent population from which the sample was taken. The assumptions may relate to sample size, distribution type, or population characteristics like mean, standard deviation, etc. The most widely used parametric tests are the Z-test, t-test, and χ2 test (although x is considered a nonparametric test when used as a test of independence or good of fit). Since they use interval and ratio data, parametric tests are more potent than nonparametric tests. The parametric tests are based on certain assumptions.

The observations being tested should be independent so that the inclusion of one set of observations does not affect the subsequent observations,

- normality of distribution
- It requires interval or ratio measurement scales so that arithmetic operations can be applied to them

The "Z" test is used for t-distribution and binomial or Poisson distribution also when the sample size is very large on the presumption that such a distribution tends to approximate normal distribution as the sample size becomes larger. This Z value is compared with the calculated Z-statistic for judging the significance of the measure concerned. The 't' test is a univariate test that uses t-distribution for testing sample mean and proportion when the size of sample is small (i.e., less than 30). The t-distribution is a symmetrical bell-shaped curve. The variance of t-distribution approaches the variance of the standard normal distribution as the sample size increases. Hence the widely practiced rule of thumb is that $n > 30$ is considered large and for such sample size normal distribution is used and for $n < 30$ t-distribution is used. 'F'-test is based on F-distribution. It is generally used to compare the variance of two sets of observations. F-distribution is a frequency distribution that uses two sets of degrees of freedom i.e., one in numerator and one in denominator. ANOVA is a case of using F-test to compare variance. Chi-square considered as a parametric test is used to compare a sample variance to some theoretical population variance. It is based on chi-square distribution.

5.9 NON-PARAMETRIC TEST

Non-parametric tests, also known as distribution-free tests, are a category of statistical tests that do not make strong assumptions about the underlying distribution of the data. These tests are used when the data do not meet the assumptions of parametric tests, which assume that the data follow a specific distribution, such as a normal distribution. Non-parametric tests are often used when dealing with ordinal or nominal data, small sample sizes, or data that are not normally distributed.

Here are some common non-parametric tests:

Mann-Whitney U Test (Wilcoxon Rank-Sum Test): This test is used to compare two independent groups to determine if there is a significant difference between them. It is used as a non-parametric alternative to the independent samples t-test.

Wilcoxon Signed-Rank Test: This test is used to compare two related (paired) groups when the data is not normally distributed. It is an alternative to the paired samples t-test.

Kruskal-Wallis Test: This is a non-parametric alternative to the one-way ANOVA test. It is used to compare three or more independent groups to determine if there are significant differences between them.

Friedman Test: This is the non-parametric counterpart of the repeated measures ANOVA. It is used to test for differences between multiple related groups when the data are not normally distributed.

Chi-Square Test: The chi-square test is used to analyze the association between categorical variables. It can be used for tests of independence (chi-square test of independence) or tests of goodness-of-fit (chi-square goodness-of-fit test).

Mann-Whitney-Wilcoxon Test (MWW): This is an extension of the Mann-Whitney U test for comparing more than two independent groups.

Sign Test: A non-parametric test for comparing paired data. It tests whether the median of the differences between paired observations is significantly different from zero.

Runs Test: This test is used to determine whether a sequence of data points is randomly ordered or exhibits some systematic pattern.

Non-parametric tests are valuable tools in statistics when assumptions of normality or other parametric assumptions are not met. They are robust and provide a way to perform statistical analysis when the data does not conform to the assumptions of parametric tests. However, they are generally less powerful than their parametric counterparts when the assumptions of parametric tests are met, so it's important to choose the appropriate test based on the nature of your data and research questions.

5.10 DIFFERENCES BETWEEN PARAMETRIC TEST AND NON-PARAMETRIC

Properties	Parametric Test	Non-Parametric
1. Assumptions	Here assumptions are made	No assumptions are not made
2. Correlation	Pearson Correlation	Spearman Correlation
3. Probabilistic	Distribution Normal probabilistic distribution	Arbitrary probabilistic distribution
4. Use	Used for finding interval data	Used for finding nominal data
5. Application	Applicable to variables	Applicable to variables and attributes
6. Population	Knowledge Population knowledge is required	Population knowledge is not required
7. Examples	T-test, z-test	Mann-Whitney, Kruskal-Wallis

5.11 SUM UP

Testing of hypotheses means to test the assumption validity through null hypothesis. Results from statistical tests will fall into one of two regions: the rejection region and the acceptance region, rejection region leads you to reject the null hypothesis, or the acceptance region, where you provisionally accept the null hypothesis. The acceptance region is “the interval within the sampling distribution of the test statistic that is consistent with the null hypothesis H_0 from hypothesis testing. Parametric Tests assume that the data follows a specific probability distribution, typically the normal distribution. Common parametric tests include t-tests, analysis of variance (ANOVA), and linear regression. Whereas, Non-Parametric tests make fewer assumptions about the data distribution. They are used when the data is not normally distributed or when you have ordinal or categorical data-for example, the Wilcoxon signed-rank test and the Kruskal-Wallis test.

5.12 SUM UP

Hypothesis testing is a fundamental statistical method used to make inferences about population parameters based on sample data. Here is a summary of the key concepts and steps involved in hypothesis testing are as:

- Null Hypothesis (H_0): This is the default hypothesis that there is no effect or no difference in the population. It represents the status quo.
- Alternative Hypothesis (Ha or H_1): This is the hypothesis you want to test, suggesting there is an effect or difference in the population.
- Select Significance Level (α): The significance level, often denoted as α, represents the probability of making a Type I error (rejecting a true null hypothesis). Common values for α include 0.05 or 0.01.
- Determine the P-value: The p-value is the probability of obtaining a test statistic as extreme as, or more extreme than, the one calculated from the sample data, assuming the null hypothesis is true. A smaller p-value indicates stronger evidence against the null hypothesis.
- Make a Decision: If the p-value is less than α, you reject the null hypothesis and If the p-value is greater than α, you fail to reject the null hypothesis.
- Draw a Conclusion: If you reject the null hypothesis, you conclude that there is evidence to support the alternative hypothesis.
- If you fail to reject the null hypothesis, you do not have enough evidence to support the alternative hypothesis.

5.13 SUGGESTED READINGS

- Anderson, D.R.; Sweeney, D.J. and Williams, T.A., “Statistics for Business and Economics”, 2nd edition (2011), Thompson, New Delhi.
- Gupta SC: Fundamental of statistics, S. Chand & Company. New Delhi
- Gupta, SP: Statistical Methods, S. Chand & Company. New Delhi
- Kothari, C. R., “Research Methodology”, 2nd Edition (2008), New Age International.
- Meyer, P.L. (1990): Introductory Probability and Statistical Applications, Oxford & IBH Pub.

- Monga, GS: Mathematics and Statistics for Economics, Vikas Publishing House, New Delhi.
- Rohatgi, V. K. and Saleh, A.K.M.E. (2010): An Introduction to Probability Theory and Mathematical Statistics, Wiley Eastern.

M.A (ECONOMICS)

QUANTITATIVE METHODS II

SEMESTER -II

Unit 6: Sampling distributions of a Statistics- Small Sample test or student-t test and its applications: t-test for single mean, difference of means, Paired t-test

STRUCTURE

6.0 OBJECTIVES

After studying this unit, the learner should be able to:

- know the procedure of t-test for testing a hypothesis
- describe testing of the hypothesis for the population mean for using a t-test
- explain the testing of the hypothesis for the difference between two population means
- when samples are independent using a t-test
- describe the procedure for paired t-test for testing of hypothesis
- difference of two populations means when samples are dependent or paired
- testing of the hypothesis for the population correlation coefficient using a t-test.

6.1 INTRODUCTION

t-test was developed in 1908 by “Willian Sealy Gosset” he was working with “Guinners Son & Company- A Dublin Brewery, in Ireland” and company did not permit employees to publish their research findings under their names, so he published his findings under the pen name “Student”. So, it is also called as “Student” t-test. It is a statistical hypothesis testing tool that is used to determine whether there is a significant difference between the means of two groups or samples. t-test is commonly used when the sample size is small or n<30 (n, means number of observations) and population standard deviation is unknown. It is based on t-distribution which is similar to the normal distribution but less peaked than normal distribution and has a higher tail than normal distribution. The shape of the t-distribution varies with the change in the degree of freedom, it is less peaked than the normal distribution at centre and more peaked in the tails. The value of t-distribution ranges from $-\infty$ to $+\infty$ ($-\infty < t < +\infty$).

The following are the standard t-tests:

- One-sample: Compares a sample mean to a reference value.
- Two-sample: Compares two sample means.
- Paired: Compares the means of matched pairs, such as before and after scores.

To choose the correct t-test, you must know whether you are assessing one or two group means. If you’re working with two groups means, do the groups have the same or different items/people? Use the table below to choose the proper analysis.

Number of Group Means	Group Type	t-test
One	--------	One sample t-test
Two	Different items in each group	Two sample t-test
Two	The same items in both groups	Paired t-test

6.3 PROCEDURE OF T-TEST FOR TESTING A HYPOTHESIS

Let us give you similar details here. For this purpose, let X_1, X_2, …, Xn be a random sample of small size n (< 30) selected from a normal population having parameter of interest, say, θ which is unknown but its hypothetical value, say, θ_0 estimated from some previous study or some other way is to be tested. t-test involves the following steps for testing this hypothetical value:

Step I: First of all, we have to set up null hypothesis H_0 and alternative hypothesis H_1.

Suppose, we want to test the hypothetical / Testing of the Hypothesis assumed value μ_0 of parameter μ. So, we can take the null and alternative hypotheses as H_0: $\mu = \mu_0$

H_1: $\mu \neq \mu_0$ (for the two-tailed test)

While one- tail test as:

H_0: $\mu = \mu_0$ and H_1: $\mu > \mu_0$ (Right-tailed)

H_0: $\mu = \mu_0$ and H_1: $\mu < \mu_0$ (Left-tailed)

In case of comparing the same parameter of two populations of interest, say, μ_1 and μ_2, then our null and alternative hypotheses would be

H_0: and $\mu_1 = \mu_2$ and H1: $\mu_1 \neq \mu_2$ (for two-tailed test)

While one- tail test as:

H_0: $\mu_1 \leq \mu_2$ and H_1: $\mu_1 > \mu_2$ (Right-tailed)

H_0: $\mu_1 \geq \mu_2$ and H_1: $\mu_1 < \mu_2$ (Left-tailed)

Step II: After setting the null and alternative hypotheses, we establish a criterion for rejection or non-rejection of null hypothesis, that is, decide the level of significance (α), at which we want to test our hypothesis. The most common value of α is 0.05 or 5%. Other popular choices are 0.01 (1%) and 0.1 (10%).

Step III: The third step is to choose an appropriate test statistics form like t-test of any application.

Step IV: Obtain the critical value(s) in the sampling distribution of the test statistic and construct the rejection (critical) region of size α. Generally, critical values for various levels of significance are put in the form of a table for various standard sampling distributions of test statistics such as t-table of respective d.f, etc.

Step V: After that, compare the calculated value of test statistic obtained from Step IV, with the critical value(s) obtained in Step V and locate the position of the calculated test statistic, that is, it lies in the rejection region or non-rejection region.

Step VI: ultimately testing the hypothesis, we have to conclude.

It is done as explained below:

(i) If the calculated test statistic value lies in the rejection region at the significance level, then we reject the null hypothesis. It means that the sample data provide us sufficient evidence against the null hypothesis and there is a significant difference between hypothesized value and observed value of the parameter.

(ii) If the calculated test statistic value lies in the non-rejection region at the significance level, then we do not reject the null hypothesis. It means that the sample data fails to provide sufficient evidence against the null hypothesis and the difference between hypothesized value and observed value of the parameter due to sample fluctuation.

Note: The decision about the null hypothesis is taken with the help of the p-value. The concept of p-value is very important because computer packages and statistical software such as SPSS, STATA, MINITAB, EXCEL, etc., all provide p-value.

6.4 TESTING OF HYPOTHESIS FOR POPULATION MEAN USING t-TEST

ASSUMPTIONS

When the standard deviation of a population is not known and the sample size is small so in this situation, we use a t-test provided the population under study is normal. Virtually every test has some assumptions which must be met before the application of the test. This t-test needs the following assumptions to work:

(i) The characteristic under study follows a normal distribution. In other words, populations from which a random sample is drawn should be normal for the characteristic of interest

(ii) Sample observations are random and independent.

(iii) Population variance σ 2 is unknown.

For describing this test, let X_1, X_2, ……, Xn be a random sample of small size n (< 30) selected from a normal population with mean μ and unknown variance σ_2. Now, follow the same procedure as we have discussed, that is, first of all, we set up the null and alternative hypotheses. Here, we want to test the claim about the specified value μ_0 of population means μ so we can take the null and alternative hypotheses as

take the null and alternative hypotheses as H_0: $\mu = \mu_0$

H_1: $\mu \neq \mu_0$ (for the two-tailed test)

While one- tail test as:

H_0: $\mu = \mu_0$ and H_1: $\mu > \mu_0$(Right-tailed)

H_0: $\mu = \mu_0$ and H_1: $\mu < \mu_0$(Left-tailed)

For testing the null hypothesis, the test statistic t is given by

$$t = \frac{\bar{X}-\mu_0}{s\ /\ \sqrt{n}}$$

where $\bar{X} = \frac{\sum X}{n}$ is the sample mean

$\bar{X}$ can be solved by A + $\frac{1}{n}\sum d$

$\sum d = \sum$ (X-A)

A= assumed mean

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}\ )^2$ is the sample variance

s^2 can be solved by $\frac{1}{n-1}\ \left(\sum d^2 - \frac{(\sum d)^2}{n}\right)$

Here, t-distribution with (n − 1) degrees of freedom.

After substituting values of X, S and n, we get the calculated value of test statistic t. Then we look for the critical value of test statistic t from the t-table. On comparing the calculated value and critical value(s), we decide on the null hypothesis.

Applications:

- Testing whether the average score of students in a small class differs from the national average.
- Determining if the average weight of a sample of products matches the specified weight standard.

Example 1: An electric tube producer claims that the average life of a particular category of electric tubes is 18000 km when used under normal driving conditions. A random sample of 16 electric tubes was tested. The mean and SD of life of the electric tubes in the sample were 20000 km and 6000 km respectively. Assuming that the life of the electric tubes is normally distributed, test the claim of the producer at a 1% level of significance.

Solution: Here, we are given that

$$n= 16,\ \mu_0 = 18000,\ \bar{X} = 20000,\ s = 6000$$

Here, we want to test that the producer's claim is true that the average life (μ) of electric tubes is 18000 km.

Step 1: Set up the null and alternative hypotheses as

H_0: $\mu_0 = 18000$ (average life of electric tubes is 18000 km)

H_1: $\mu_0 \neq 18000$ two-tailed (average life of electric tubes is not 18000 km)

Step 2: level of significance: As this is a one-tailed test,

$\alpha = 5\%$. This can be used to determine the critical value.

$1 - \alpha = 1 - 0.05 = 0.95$, df= N-1= 6-1=5

Step 3: It is a small sample test; therefore t-test is to be determined as

$$t = \frac{\bar{X}-\mu_0}{s\ /\ \sqrt{n}}$$

Step 4: Determine the critical value

The critical value of test statistic t for two-tailed test corresponding (n-1) = 15 df at 1% level of significance are $\pm\, t_{(15),\,1\,\%} = \pm\, 2.947$. Here, t-distribution with (n −1) degrees of freedom.

Step 5: Test Statistics

$$t = \frac{20000-18000}{6000/\sqrt{16}}$$

$$t = \frac{2000}{1500} = 1.33$$

Step 6: Conclusion

Since the calculated value of test statistic t (=1.33) is less than the critical (tabulated) value (= 2.947) and greater than the critical value (= − 2.947), that means a calculated value of test statistic lies in the non-rejection region, thus we do not reject the null hypothesis i.e. we support the producer's claim at 1% level of significance. Therefore, we conclude that the sample fails to provide sufficient evidence against the claim so we may assume that the producer's claim is true.

Example 2: (left-tail) The average score of a class is 90. However, a teacher believes that the average score might be lower. The scores of 6 students were randomly measured. The mean was 82 with a standard deviation of 18. With a 0.05 significance level use hypothesis testing to check if this claim is true.

Solution: Step 1: Set up the null and alternative hypotheses as

H_0: $\mu = 90$,

Alternative hypothesis

H_1: $\mu < 90$ (Left-tailed)

$n = 6$, $s = 18$

Step 2: level of significance:

As this is a one-tailed test,

$\alpha = 5\%$. This can be used to determine the critical value.

$1 - \alpha = 1 - 0.05 = 0.95$, df= N-1= 6-1=5

Step 3: It is a small sample test; therefore t-test is to be determined as

$$t = \frac{\bar{X}-\mu_0}{\frac{s}{\sqrt{n}}}$$

Step 4: Determine the critical value

The critical value from the t table is -2.015

Step 5: Test Statistics

$$t = \frac{\bar{X}-\mu_0}{\frac{s}{\sqrt{n}}} \qquad t = \frac{82-90}{\frac{18}{\sqrt{6}}} \qquad t = -1.088$$

Step 6: Conclusion

As Cal t > Tab t (-1.088 > -2.015), therefore, fail to reject the null hypothesis. There is not enough evidence to support the claim.

6.5 TESTING OF HYPOTHESIS FOR DIFFERENCE OF TWO POPULATION MEANS USING T-TEST

When standard deviations of both populations are not known, in real-life problem t-test is more suitable compared to the Z-test.

Assumptions

This test works under the following assumptions:

a) It follows a normal distribution in both populations. Both populations from which random samples are drawn should be normal for the characteristics of interest.

b) Samples and their observations both are independent of each other.

c) Population variances σ_1^2 and σ_2^2 are both unknown but equal.

Let's we have to draw two independent random samples, $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Yn$ of sizes n_1 and n_2 from these normal populations. Let $\bar{X}$ and $\bar{Y}$ be the means of first and second sample respectively. Further, suppose the variances of both the populations are unknown but are equal, i.e., $\sigma_1^2 = \sigma_2^2 = \sigma$. In this case, σ^2 is estimated by value of pooled sample variance S^2

$$s_p^2 = \frac{1}{n_1+n_2-2}[(n_1-1)s_1^2 + (n_2-1)s_2^2]$$

$$s_1^2 = \frac{1}{n_1-1}\sum(X-\overline{X})^2 \text{ and } s_2^2 = \frac{1}{n_2-1}\sum(Y-\overline{Y})^2$$

$$s_p^2 = \frac{1}{n_1+n_2-2}[\sum(X-\overline{X})^2 + \sum(Y-\overline{Y})^2]$$

$$\bar{X} = A + \frac{\sum d_1}{n_1}$$

$$\bar{Y} = A + \frac{\sum d_2}{n_2}$$

d_1= (X-A_1), d_2= (Y-A_2)

A_1 is assumed mean from series X, A_2 is assumed mean from series Y

$$s_p^2 = \frac{1}{n_1+n_2-2}[\sum d_1^2 - \frac{(\sum d_1)2}{n_1} + \sum d_2^2 - \frac{(\sum d_2)2}{n_2}]$$

Here, the hypotheses for a difference in two population means are similar to those for a difference in two population proportions. The null hypothesis, H_0, is again a statement of "no effect" or "no difference."

- H_0: $\mu_1 - \mu_2 = 0$, which is the same as H_0: $\mu_1 = \mu_2$

The alternative hypothesis, H_a, can be any one of the following.

- H_a: $\mu_1 - \mu_2 < 0$, which is the same as H_a: $\mu_1 < \mu_2$
- H_a: $\mu_1 - \mu_2 > 0$, which is the same as H_a: $\mu_1 > \mu_2$
- H_a: $\mu_1 - \mu_2 \neq 0$, which is the same as H_a: $\mu_1 \neq \mu_2$

For testing the null hypothesis, the test statistic t is given by $t = \frac{\bar{X}-\bar{Y}}{S\sqrt{\frac{1}{n_1}-\frac{1}{n_2}}}$

After substituting values of X, Y, S, n_1 and n_2 we get the calculated value of test statistic t. Then we look for critical (or tabulated) value(s) of test statistic t from the t-table. On comparing calculated value and critical value(s), we decide the null hypothesis either to accept or reject.

Applications:

- Comparing the performance of two groups of students taught using different teaching methods.
- Evaluating the effectiveness of two medications in separate patient groups.
- Analyzing differences in average salaries between two departments in a company.

Example 3: Two different types of drugs A and B were tried on some patients for increasing their weights. Six persons were given drug A and other 7 persons were given drug B. The gain in weights (in ponds) is given below:

Drug A:	5	8	7	10	9	6	–
Drug B:	9	10	15	12	14	8	12

Assuming that increase in the weights due to both drugs follow normal distributions with equal variances, do the both drugs differ significantly with regard to their mean weights increment at 5% level of significance?

Solution: If μ_1 and μ_2 denote the mean weight increment due to drug A and drug B respectively then our claim is $\mu_1 = \mu_2$ and its complement is $\mu_1 \neq \mu_2$.

Since the claim contains the equality sign so we can take the claim as the null hypothesis and complement as the alternative hypothesis. Thus,

H_0: $\mu_1 = \mu_2$ [effect of both drugs is same]

H_1: $\mu_1 \neq \mu_2$ [effect of both drugs is not same]

Since the alternative hypothesis is two-tailed so the test is two-tailed test. Since it is given that increments in the weight due to both drugs follow normal distributions with equal and unknown variances and other assumptions of t-test for testing a hypothesis about difference of two population means also meet. So, we can go for this test. For testing the null hypothesis, the test statistic t is given by

$$t = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{n_1} - \frac{1}{n_2}}}$$

Assume, a = 8, b = 12 and use short-cut method to find X, Y and S

Drug A (X)			Drug B (Y)		
X	$d_1 = (X - A_X)$ A_X=8		Y	$d_2 = (Y - A_y)$ A_y= 12	
5	-3	9	9	-3	9

8	0	0	10	-2	4
7	-1	1	15	3	9
10	2	4	12	0	0
9	1	1	14	2	4
6	-2	4	8	-4	16
			12	0	0
$\sum X$ = 45	$\sum d_1$ = -3	$\sum d_1^2$ =19	$\sum Y$ =80	$\sum d_2$ = -4	$\sum d_2^2$ =42

$$\bar{X} = \frac{\sum X}{n} = \frac{45}{6} = 7.5, \bar{Y} = \frac{\sum Y}{n} = \frac{80}{7} = 11.43$$

$$S^2 = \frac{1}{n_1+n_2-2}\left(\sum {d_1}^2 - \frac{(\sum d1)^2}{n_1}\right) + \left(\sum {d_2}^2 - \frac{(\sum d2)^2}{n_1}\right)$$

$$= \frac{1}{6+7-2}\left(19 - \frac{(-3)^2}{6}\right) + \left(42 - \frac{(-4)^2}{7}\right), = \frac{1}{11}(17.5 - 39.71), = \sqrt{5.20}$$

S = 2.28

$$t = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{n_1} - \frac{1}{n_2}}} \qquad t = \frac{7.5 - 11.43}{2.28\sqrt{\frac{1}{6} - \frac{1}{7}}} = \frac{-3.93}{2.28 \times 0.56} = \frac{-3.93}{1.28} = -3.07$$

The critical values of test statistic t for two-tailed test corresponding (n_1 + n_2 -2) = 11 df at 5% level of significance are ± t (11) ,0.025 = ± 2.201. Since calculated value of test statistic t (= −3.07) is less than the critical values (= ± 2.201) that means calculated value of test statistic t lies in rejection region, so we reject the null hypothesis i.e. we reject the claim at 5% level of significance. Thus, we conclude that samples provide us sufficient evidence against the claim so drugs A and B differ significantly. Any one of them is better than other.

6.6 PAIRED t-TEST

Paired t-test gives a hypothesis examination of the difference between population means for a set of random samples whose variations are almost normally distributed. Subjects are often tested in a before-after situation or with subjects as alike as possible. The paired t-test is a test that the differences between the two observations are zero. For instance, a pharmaceutical company might create a new blood pressure-lowering medication. Twenty individuals had their blood pressure taken both before and after the medicine is administered for a month. To determine whether there is a statistically significant difference between pressure readings taken before and after taking the medication, analysts utilize a paired t-test.

Let us assume two paired sets, such as Xi and Yi for i = 1, 2, …, n such that their paired difference is independent which is identically and normally distributed. Then the paired t-test concludes whether they notably vary from each other. Here,

- Null hypothesis: The mean difference between pairs equals zero in the population ($\mu_D = 0$).
- Alternative hypothesis: The mean difference between pairs does not equal zero in the population ($\mu_D \neq 0$).

Assumptions

This test works under following assumptions:

- The population of differences follows normal distribution.
- Samples are not independent.
- Size of both the samples is equal.
- Population variances are unknown but not necessarily equal.

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a paired random sample of size n and the difference between paired observations Xi & Yi be denoted by D_i, that is, $D_i = X_i - Y_i$ for I = 1, 2, …. n. Hence, we can assume that $D_1, D_2, \ldots, D_n$ be a random sample from normal population of differences with mean μ_D and unknown variance σ^2_D. This is same as the case of testing of hypothesis for population mean when population variance is unknown.

Here, we want to test that there is an effect of a diet, training, treatment, medicine, etc.

A paired samples t-test always uses the following null hypothesis:

- H_0: $\mu_1 = \mu_2$ or H_0**:** $\mu_{D} = \mu_1 - \mu_2$ (the two-population means are equal)

The alternative hypothesis can be either two-tailed, left-tailed, or right-tailed:

- H_1 (two-tailed): $\mu_1 \neq \mu_2$ or $\mu_D \neq 0$ (the two-population means are not equal)
- H_1 (left-tailed): $\mu_1 < \mu_2$ (population 1 mean is less than population 2 mean)
- H_1 (right-tailed): $\mu_1 > \mu_2$ (population 1 mean is greater than population 2 mean)

For testing null hypothesis, paired t statistic is given as:

$$t = \frac{\bar{D}}{s_D \,/\, \sqrt{n}}$$

$$\bar{D} = \frac{1}{n} \sum_{i=1}^{n} D_i$$

$$S_D^2 = \frac{1}{n-1} \sum_{i=1}^{n} (D_i - \bar{D})^2 = \frac{1}{n-1} \left[\sum D^2 - \frac{(\sum D)2}{n} \right]$$

After substituting values of D, S and n D we get calculated value of test statistic t. Then we look for critical (or cut-off or tabulated) value(s) of test statistic t from the t-table. On comparing calculated value and critical value(s), we take the decision about the null hypothesis.

Applications:

- Assessing the impact of a training program by comparing test scores before and after the training.
- Measuring weight loss before and after a diet program for the same individuals.

- Evaluating the effect of a new teaching method by comparing students' scores on pre-tests and post-tests.

Example 4: To verify whether the training programme improved performance of the laborers, a similar test was given to 10 laborers both before and after the programme. The original marks out of 100 (before training) recorded in an alphabetical order of the participants are

before training: 42, 46, 50, 36, 44, 60, 62, 43, 70 53

After training: 45, 46, 60, 42, 60, 72, 63, 43, 80 65 Assuming that performance of the before training and after follows normal distribution. Test whether the training programme has improved the performance of the laborers at 5% level of significance?

Solution**:** Here, we want to test whether the training programme has improved the performance of the laborer. Thus,

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 < \mu_2$ (left-tailed)

Since the alternative hypothesis is left-tailed so the test is left-tailed test.

It is a situation of before and after. Also, the marks of the students before and after the training programme follows normal distributions. Therefore, population of differences will also be normal. Also, all the assumptions of paired t-test meet. So, we can go for paired t-test. For testing the null hypothesis, the test statistic t is given by

t = $\frac{\bar{D}}{S_D \,/\, \sqrt{n}}$

Labours	X	Y	D = (X-Y)	D^2
1	42	45	-3	9
2	46	46	0	0
3	50	60	-10	100
4	36	42	-6	36
5	44	60	-16	256
6	60	72	-12	144
7	62	63	-1	1
8	43	43	0	0
9	70	80	-10	100
10	53	65	-12	144
			$\sum D$ = -70	$\sum D^2 = 790$

$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$, $= \frac{1}{10}\,(-70)$ = -7

$S_D^2 = \frac{1}{10-1}\left[\sum D^2 - \frac{(\sum D)2}{n}\right]$

$= \frac{1}{9}\left[790 - \frac{(-70)2}{10}\right]$, $= \frac{1}{9}[300]$, $=33.33$

$S_D = \sqrt{33.33}$

$S_D = 5.77$

Putting the values in t-test, we have

$t = \frac{\bar{D}}{S_D / \sqrt{n}}$

$t = \frac{-7}{5.77 / \sqrt{10}}$, $t = -3.38$

The critical value of test statistic t for left-tailed test corresponding (n-1) = 9 df at 5% level of significance is – t (9), 0.05 = −1.833. Since calculated value of test statistic t (= −3.83) is less than the critical (tabulated) value (= −1.833), that means calculated value of test statistic t lies in rejection region, so we reject the null hypothesis and support the alternative hypothesis i.e. we support our claim at 5% level of significance. Thus, we conclude that samples fail to provide us sufficient evidence against the claim so we may assume that the participants have significant improvement after training programme.

6.7 TESTING OF HYPOTHESIS FOR POPULATION CORRELATION COEFFICIENT USING T-TEST

if two variables are related in such a way that change in the value of one variable affects the value of another variable then the variables are said to be correlated or there is a correlation between these two variables. Correlation can be positive, which means the variables move together in the same direction, or negative, which means they move in opposite directions. And correlation coefficient is used to measure the intensity or degree of linear relationship between two variables. The value of correlation coefficient varies between −1 and +1, where −1 representing a perfect negative correlation, 0 Small Sample Tests representing no correlation, and +1 representing a perfect positive correlation. Sometime, the sample data indicate for non-zero correlation but in population they are uncorrelated (ρ = 0). For example, price of tomato in Delhi (X) and in London (Y) are not correlated in population (ρ = 0). But paired sample data of 20 days of prices of tomato at both places may show correlation coefficient (r) ≠ 0. In general, in sample data r ≠ 0 does not ensure in population ρ ≠ 0 holds. here, we will know how we test the hypothesis that population correlation coefficient is zero.

Assumptions

This test works under following assumptions:

(i) The characteristic under study follows normal distribution in both the populations. In other words, both populations from which random samples are drawn should be normal with respect to the characteristic of interest.

(ii) Samples observations are random.

Let us consider a random sample (X_1, Y_1), (X_2, Y_2), …, (Xn, Yn) of size n taken from a bivariate normal population. Let ρ and r be the correlation coefficients of population and sample data respectively. Here, we wish to test the hypothesis about population correlation coefficient (ρ), that is, linear correlation between two variables X and Y in the population, so we can take the null hypothesis as

H_0: ρ = 0 and H_1: ρ ≠ 0 (two-tailed)

H_0: ρ ≤ 0 and H_1: ρ > 0 (right -tailed)

H_0: ρ ≥ 0 and H_1: ρ < 0 (left -tailed)

Here t statistic is as: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

here n-2 degrees of freedom

After substituting values of r and n, we find out calculated value of t-test statistic. Then we look for critical (or cut-off or tabulated) value(s) of test statistic t from the t-table. On comparing calculated value and critical value(s), we take the decision about the null hypothesis. Let us do some examples of testing of hypothesis that population correlation coefficient is zero.

Example 5: 20 families were selected randomly from Area A group to determine that correlation exists between family income and the amount of money spent per family member on food each month. The sample correlation coefficient (r) was computed as r = 0.40. By follow the normal distributions, test that there is a positive linear relationship between the family income and the amounts of money spent per family member on food each month in Area A group at 1% level of significance.

Solution: here, n = 20, r = 0.40, and to test that there is a positive linear relationship between the family income and the amounts of money spent per family member on food each month in area A group. If ρ denote the correlation coefficient between the family income and the amounts of money spent per family member then the claim is ρ > 0 and its complement is ρ ≤ 0. Since complement contains the equality sign so we can take the complement as the null hypothesis and the claim as the alternative hypothesis.

Thus, H_0: ρ ≤ 0

H_1: ρ > 0 (right -tailed)

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad t = \frac{0.40\sqrt{20-2}}{\sqrt{1-(0.40)^2}} \qquad t = \frac{0.40 \times 4.24}{0.92} = 1.84$$

The critical value of test statistic t for right-tailed test corresponding (n-2) = 18 df at 1% level of significance is $t_{(n-2),\alpha} = t_{(18),0.01} = 2.552$. Since calculated value of test statistic t (=1.84) is less than the critical value (= 2.552), it calculated value of test statistic t lies in non-rejection region, so we do not reject the null hypothesis and reject the alternative hypothesis *i.e.* we reject our claim at 1% level of significance. Thus, we conclude that sample provide us sufficient evidence against the claim so there is no positive linear correlation between the family income and the amounts of money spent per family member on food each month in area A group.

6.8 SUM UP

Since in many of the problems it becomes necessary to take a small size sample, considerable attention has been paid in developing suitable tests for dealing with problems of small samples. The greatest contribution to the theory of small samples is that of Sir William Gosset and Prof. R.A. Fisher. Sir William Gosset published his discovery in 1905 under the pen name 'Student' and later on developed and extended by Prof. R.A. Fisher. He gave a test popularly known as 't-test'. The t-distribution has a number of applications in statistics, t-test for significance of single mean, t-test for significance of the difference between two sample means, independent samples, paired t-test.

6.9 QUESTIONS FOR PRACTICE

Q1. Explain the need of small sample tests.

Q2. List out the assumptions of t-test.

Q3. List out the Procedure of testing a hypothesis for t-test.

Q4. Explain the testing of hypothesis for population mean using t-test.

Q5. Describe testing of hypothesis for difference of two population means when samples are independent using t-test.

Q6. Explain the procedure of paired t-test for testing of hypothesis for difference of two population means when samples are dependent or paired.

6.10 SUGGESTED READINGS

- C.R. Kothari (1990) Research Methodology. Vishwa Prakasan. India.
- Gupta SC: Fundamental of statistics, S. Chand & Company. New Delhi.
- Gupta, SP: Statistical Methods, S. Chand & Company. New Delhi.
- A.M Goon, M.K Gupta and B. Dasgupta, fundamental of statistics Vol-I, World press Calcutta.
- Anderson, D.R.; Sweeney, D.J. and Williams, T.A., "Statistics for Business and Economics", 2nd edition (2011), Thompson, New Delhi.

M.A (ECONOMICS)

SEMESTER -II

QUANTITATIVE METHODS II

Unit 7: Large Sample test: Introduction, Sampling of Attributes- test for Single Proportion, test for difference in proportion and F-test

STRUCTURE

7.0 OBJECTIVES

After studying this unit, learners should be able to know:

- Meaning of large sample test

- Applying the Z-test to test the hypothesis about the population mean and the difference between the two population means
- proportion and difference of two population proportions
- Applying the Z-test to test the hypothesis about the population variance and two population variances.

7.1 INTRODUCTION

Sometimes in our studies in economics, psychology, medicine, etc., we take a sample of objects/units/participants/patients, etc. such as 70, 500, 1000, 10,000, etc. This situation comes under the category of large samples. As a thumb rule, a sample of size n is treated as a large sample only if it contains more than 30 units (or observations, n > 30). We know that for large samples (n > 30), one statistical fact is that almost all sampling distributions of the statistic(s) are closely approximated by the normal distribution. Therefore, the test statistic, a function of sample observations based on n > 30, could be assumed to follow the normal distribution approximately (or exactly). In other words, we have seen that for large values of n, the number of trials, almost all the distributions e.g., Binomial, Poisson, etc. are very closely approximated by Normal distribution and in this case, we apply Normal Deviate test (Z-test). In cases where the population variance (s) is/are known, we use Z-test. The distribution of Z is always normal with mean zero and variance one. In statistics, a sample is said to be large if its size exceeds 30.

7.2 ASSUMPTIONS FOR A SINGLE SAMPLE Z-TEST

Every statistical method has assumptions. Assumptions mean that your data must satisfy certain properties for statistical method results to be accurate.

The assumptions for the Single Sample Z-Test include:

1. Population Standard Deviation is Known: The z-test assumes that the standard deviation (σ) of the population is known. If σ is unknown, a t-test may be more appropriate.

2. Data is Normally Distributed:

- If the sample size is small (n<30), the data should follow a normal distribution.
- For larger samples (n≥30), the Central Limit Theorem applies, and the sampling distribution of the mean will approximate normality regardless of the population distribution.

3. Random Sampling: The sample should be drawn randomly from the population to ensure that it is representative.

4. Data is Continuous: The test is designed for data measured on a continuous scale (e.g., height, weight, or temperature).

5. Observations are Independent: Each observation in the sample must be independent of others, meaning the value of one observation does not affect another.

6. No Significant Outliers: Extreme outliers can distort the results of the test. It's important to check for and address outliers before conducting the test.

7.3 STEPS OF t-TEST TESTING OF HYPOTHESIS

Suppose X_1, X_2, …, X_n is a random sample of size n (> 30) selected from a population having unknown parameter θ and we want to test the hypothesis about the hypothetical / claimed/assumed value θ_0 of parameter θ. For this, a test procedure is required. We discuss it step by step as follows:

t-test involves the following steps for testing this hypothetical value:

Step I: First of all, we have to set up null hypothesis H_0 and alternative hypothesis H_1.

Suppose, we want to test the hypothetical / Testing of the Hypothesis assumed value μ_0 of parameter θ.

So, we can take the null and alternative hypotheses as H_0: $\theta = \theta_0$

H_1: $\theta \neq \theta_0$ (for the two-tailed test)

While one- tail test as:

H_0: $\theta = \theta_0$ and H_1: $\theta > \theta_0$ (Right-tailed)

H_0: $\theta = \theta_0$ and H_1: $\theta < \theta_0$ (Left-tailed)

In case of comparing the same parameter of two populations of interest, say, θ_1 and θ_2, then our null and alternative hypotheses would be

H_0: and $\theta_1 = \theta_2$ and H1: $\theta_1 \neq \theta_2$ (for two-tailed test)

While one- tail test as:

H_0: $\theta_1 \leq \theta_2$ and H_1: $\theta_1 > \theta_2$ (Right-tailed)

H_0: $\theta_1 \geq \theta_2$ and H_1: $\theta_1 < \theta_2$ (Left-tailed)

Step II: After setting the null and alternative hypotheses, we establish a criterion for rejection or non-rejection of null hypothesis, that is, decide the level of significance (α), at which we want to test our hypothesis. The most common value of α is 0.05 or 5%. Other popular choices are 0.01 (1%) and 0.1 (10%).

Step III: Third step is to determine an appropriate test statistic, say, Z in case of large samples. Suppose Tn is the sample statistic such as sample mean, sample proportion, sample variance, etc. for the parameter θ then for testing the null hypothesis, test statistic is given by

$$Z = \frac{t - E\ (t)}{S.E.(t)}$$

Step IV: Obtain the critical value(s) in the sampling distribution of the test statistic and construct the rejection (critical) region of size α. Generally, critical values for various levels of significance are put in the form of a table for various standard sampling distributions of test statistics such as Z-table.

Step V: After that, we obtain the critical (cut-off or tabulated) value(s) in the sampling distribution of the test statistic Z corresponding to α assumed in Step II. These critical values are given in Table A (Z-table) this course corresponding to different levels of significance (α). For convenience, some useful critical values at $\alpha = 0.1$, 0.01 and 0.05 for Z-test are given in Table A (mentioned below). After that, we construct a rejection (critical) region of size α in the probability curve of the sampling distribution of test statistic Z.

Step VI: ultimately testing the hypothesis, we have to conclude.

(i) Case I: When H_0: $\theta \leq \theta_0$ and H_1: $\theta > \theta_0$ (right-tailed test)

Now, if z (calculated value) $\geq Z\alpha$ (tabulated value), that means the calculated value of test statistic Z lies in the rejection region, then we reject the null hypothesis H_0 at α the level of significance. Therefore, we conclude that sample data provides us sufficient evidence against the null hypothesis and there is a significant difference between hypothesized or specified value and the observed value of the parameter. If $Z < Z\alpha$, that means the calculated value of test statistic Z lies in non-rejection region, then we do not reject the null hypothesis H_0 at α level of significance. Therefore, we conclude that the sample data fails to provide us sufficient evidence against the null hypothesis and the difference between hypothesized value and observed value of the parameter due to the fluctuation of a sample.

so, the population parameter θ may be θ_0.

Case II: When H_0: $\theta \geq \theta_0$ and H_1: $\theta < \theta_0$ (left-tailed test)

In this case, the rejection (critical) region falls under the left tail of the probability curve of the sampling distribution of test statistic Z. If $Z \leq -Z\alpha$, that means the calculated value of test statistic Z lies in the rejection region, then we reject the null hypothesis H_0 at α level of significance. If $Z > -Z\alpha$, that means the calculated value of test statistic Z lies in the non-rejection region, then we do not reject the null hypothesis H_0 at α level of significance.

In case of two-tailed test: H_0: $\theta = \theta_0$ and H_1: $\theta \neq \theta_0$

In this case, the rejection region falls under both tails of the probability curve of the sampling distribution of the test statistic Z. Half the area (α) i.e. $\alpha/2$ will lie under the left tail and the other half under the right tail. If $Z \geq Z_{\alpha/2}$ or $Z \leq - Z_{\alpha/2}$, that means the calculated value of test statistic Z lies in the rejection region, then we reject the null hypothesis H_0 at α level of significance. If $-Z < Z < - Z_{\alpha/2}$, that means the calculated value of test statistic Z lies in the non-rejection region, then we do not reject the null hypothesis H0 at α level of significance.

Table A: Critical Values of Z-test

Level of Significance (%)	Two-tailed	Right-tailed	Left-tailed
1	2.58	2.33	-2.33

5	1.96	1.645	-1.645
10	1.645	1.28	-1.28

7.4 APPLICATIONS OF LARGE SAMPLE TEST

- Test for a single proportion.
- Test for significance of difference of proportions.
- Test of significance for a single mean.
- Test of significance for difference of means.

7.5 TEST FOR SINGLE PROPORTION

For this purpose, let X_1, X_2, ..., X_n be a random sample of size n taken from a population with population proportion P. Also, let X denote the number of observations or elements that possess a certain attribute (number of successes) out of n observations of the sample then sample proportion p can be defined as

$$p = \frac{X}{n}$$

here mean and variance of the sampling distribution of sample proportion are E(p) = P and Var(p) = $\frac{\mathrm{PQ}}{\mathrm{n}}$ where, Q = 1– P.

Now, two cases arise: Large Sample Tests

Case I: When the sample size is not sufficiently large i.e. either of the condition's np > 5 or nq > 5 does not meet, then we use the exact binomial test. However, an exact binomial test is beyond the scope of this course.

Case II: When the sample size is sufficiently large, such that np > 5 and nq > 5 then by central limit theorem, the sampling distribution of sample proportion p is approximately normally distributed with mean and variance as

$$E(p) = P \text{ and } Var(p) = \frac{\mathrm{PQ}}{\mathrm{n}}$$

But we know that standard error = Variance

$$SE(p) = \sqrt{\frac{\mathrm{PQ}}{\mathrm{n}}}$$

Now, first of all, we set up null and alternative hypotheses. Here we want to test the hypothesis about the specified value P_0 of the population proportion. So, we can take the null and alternative hypotheses as

For two-tailed

H_0: $P = P_0$

H_1: $P \neq P_0$

For one-tailed

H_0: $P = P_0$

H_1: $P > P_0$

Or

H_1: $P < P_0$

$Z = \frac{p-P_0}{\sqrt{\frac{PQ}{n}}}$

After that, we calculate the value of the test statistic and compare it with the critical value(s) given in below Table A at a prefixed level of significance α.

Table A: Critical Values of Z-test

Level of Significance (%)	Two-tailed	Right-tailed	Left-tailed
1	2.58	2.33	-2.33
5	1.96	1.645	-1.645
10	1.645	1.28	-1.28

Example 1: A die is thrown 9000 times and a draw of 2 or 5 is observed 3100 times. Can we regard that die as unbiased at a 5% level of significance?

Solution: Let getting a 2 or 5 be our success, and getting a number other than 2 or 5 be a failure then in usual notions, we have n = 9000, X = number of successes = 3100, p = 3100/9000 = 0.3444

Here, we want to test that the die is unbiased and we know that if the die is Large Sample Tests

unbiased then the proportion or probability of getting 2 or 5 is

P = Probability of getting a 2 or 5

= Probability of getting 2 + Probability of getting 5

$\frac{1}{6} + \frac{1}{3} + \frac{1}{3} = 0.3333$

So, our claim is P = 0.3333 and its complement is P ≠ 0.3333. Since the claim contains the quality sign. Thus, we can take the claim as the null hypothesis and complement as the alternative hypothesis. Thus, H_0: $P = P_0 = 0.3333$ and

H_1 :$P \neq 0.3333$

Since the alternative hypothesis is two-tailed so the test is two-tailed. Before proceeding further, first, we have to check whether the condition of normality meets or not.

$np = 9000 \times 0.3444 = 3099.6 > 5$

$nq = 9000 \times (1 - 0.3444) = 9000 \times 0.6556 = 5900.4 > 5$

We see that the condition of normality meets, so we can go for Z-test. So, for testing the null hypothesis, the test statistic Z is given by

$$Z = \frac{p-P_0}{\sqrt{\frac{PQ}{n}}} \qquad Z = \frac{0.3444-0.3333}{\sqrt{\frac{0.3333 \times 0.6667}{9000}}} \qquad = \frac{0.0111}{0.005} = 2.22$$

Since the test is two-tailed so the critical values at 5% level of significance are $\pm Z_{\alpha/2} = \pm Z_{0.025} = \pm 1.96$. Since calculated value of Z (= 2.22) is greater than the critical value (= 1.96), that means it lies in the rejection region, so we reject the null hypothesis i.e. we reject our claim.

Example 2: A shop claims that only, 1% of its products are imperfect. Out of sample of 500 units 10 are found to be imperfect. Check the claim of the shop.

Solution: Let us set up the null hypothesis that the proportion of defective products is j against the alternative hypothesis that it is not equal to 1%. i.e.

H_0: P=0.01

H_1: $P \neq 0.01$

Q =1-P=1-0.01= 0.99

Size of sample = n = 500

Number of imperfect items found =X = 10

Therefore, sample proportion of imperfect items = $p = \frac{X}{n} = \frac{10}{500} = = 0.02$

The test statistic is,

$$Z = \frac{p-P_0}{\sqrt{\frac{PQ}{n}}} \qquad = \frac{0.02-0.01}{\sqrt{\frac{0.01 \times 0.99}{500}}} \qquad = \frac{0.01}{0.00445} \qquad = 2.25$$

The calculated value of Z is more than 1.96 (Critical value at 5% level), and H_0 is rejected. Thus, the claim of the shop is rejected.

7.6 TEST FOR SIGNIFICANCE OF DIFFERENCE OF PROPORTIONS

If we have two populations and each item of a population belongs to either of the two classes C_1 and C_2. A person is often interested to know whether the proportion of items in class C_1 in both the populations is the same or not that is we want to test the hypothesis.

H_0: $P_1=P_2$

H_1: $P_1 \neq P_2$ or $P_1>P_2$ or $P_1<P_2$

where P_1 and P_2 are the proportions of items in the two populations belonging to class C_1.

Let X_1, X_2 be the number of items belonging to class C_1 in random samples of sizes n_1 and n_2 from the two populations respectively. Then the sample proportion

$$p_1 = \frac{X_1}{n_2}, p_2 = \frac{X_2}{n_2}$$

If P_1 and P_2 are the proportions then

$$E\ (P_1) = P_1,\ E\ (P_2) = P_2$$

$$\text{Var}\ (p_1) = \frac{P_1Q_1}{n_1},\ \text{Var}\ (p_2) = \frac{P_2Q_2}{n_2}$$

Since $P_1 = P_2 = P$ and $Q_1 = Q_2 = Q$, therefore

$$Z = \frac{p_1 - p_2}{\sqrt{P \times Q\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

If the population proportion P_1 and P_2 are given to be distinctly different that is $P_1 \neq P_2$, then

$$Z= \frac{(p_1-p_2)-(P_1-P_2)}{\sqrt{\frac{P_1Q_2}{n_1}+\frac{P_1Q_2}{n_2}}}$$

In general P, the common population proportion (under H_o) is not known, then an unbiased estimate of population proportion P based on both the samples is used and is given by

$$\hat{p} = \frac{X_1+X_2}{n_1+n_2} = \frac{n_1p_2+n_1p_2}{n_1+n_2}$$

Example 3. A machine turns out 16 imperfect items in a sample of 500. After overhauling it turns 3 imperfect articles in a batch of 100. Has the machine improved after overhauling?

Solution: We are given n_1=500and n_2=100

p_1= Proportions of imperfect items before overhauling of machine =16/500=0.032

p_2= Proportions of imperfect items after overhauling of machine =3/100=0.03

H_0: P_1=P_2 i.e. the machine has not improved after overhauling.

H_1: P_1>P_2 or P_2<P_1

Here, $\hat{p} = \frac{X_1+X_2}{n_1+n_2} = \frac{n_1p_2+n_1p_2}{n_1+n_2}$

$$= \frac{16+3}{500\ +100} = 0.032$$

$$Z = \frac{p_1 - p_2}{\sqrt{P \times Q\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \qquad Z = \frac{0.032-0.03}{\sqrt{0.032\ \times 0.968}\left(\frac{1}{500}+\frac{1}{100}\right)}$$

$$Z = \frac{0.002}{\sqrt{0.031}\left(\frac{1+5}{500}\right)} \quad = \frac{0.002}{0.01878} = 0.106$$

Since Z<1.645 (Right-tailed test), it is not significant at 5% level of significance. Hence, we accept the null hypothesis and conclude that the machine has not improved after overhauling.

7.7 TEST OF SIGNIFICANCE FOR A SINGLE MEAN

We have seen that if X_i (i=1, 2, ..., n) is a random sample of size n from a normal population with mean and variance σ^2, then the sample mean $\bar{X}$ is distributed normally with mean μ and variance σ^2/n i.e., $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.Thus for large samples normal variate corresponding to $\bar{X}$ is

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

In test of significance for a single mean we deal the following situations

1) To test if the mean of the population has a specified value (μ_0) and null hypothesis in this case will be H_0: $\mu=\mu_0$ i.e., the population has a specified mean value.
2) To test whether the sample mean differs significantly from the hypothetical value of population mean with null hypothesis as there is no difference between sample mean ($\bar{X}$) and population mean (μ).
3) To test if the given random sample has been drawn from a population with specified mean μ_0 and variance σ^2 with null hypothesis the sample has been drawn from a normal population with specified mean μ_0 and variance σ^2

In all the above three situations the test statistic is given by

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

If |Z| <1.96, H_o is not rejected at 5% level of significance which implies that there is no significant difference between sample mean and population mean and whatever difference is there, it exists due to fluctuation of sampling.

|Z| >1.96, H_o is rejected at 5% level of significance which implies that there is a significant difference between sample mean and population mean.

Example 4. A random sample of 100 workers gave a mean weight of 64 kg with a standard deviation of 16 kg. Test the hypothesis that the mean weight in the population is 60 kg.

Solution: H_0: μ =60 kg., i.e. the mean weight in the population is 60 kg.

H_1: $\mu \neq 60$ kg., i.e. the mean weight in the population is not 60 kg.

here, n=100, μ=60 kg., $\bar{X}$ =64 kg.,

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \quad = \frac{64 - 60}{16 / \sqrt{100}} \quad = 2.5$$

Since calculated value of Z statistic is more than 1.96, it is significant at 5% level of significance. Therefore, H_0 is rejected at all levels of significance which implies that mean weight of population is not 60 kg.

Example 5: A sample of 900 rods has a mean length 3.4 cm. Is the sample regarded to be taken from a large population of rods with mean length 3.25 cm and S.D 2.61 cm at 5% level of significance?

Solution: Here, n = 900, $\bar{X}$ =3.4 cm, μ =3.25 cm and σ = 2.61 cm

Thus, H_0: $\mu = \mu_0 = 3.25$

H_1: $\mu \neq 3.25$ (two-tailed)

Here, we want to test the hypothesis regarding population mean when Large Sample Tests population SD is unknown, so we should use t-test if the population of rods known to be normal. But it is not the case. Since the sample size is large (n > 30) so we can go for Z-test instead of t-test as an approximate. So, test statistic is given by

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

$$= \frac{3.40 - 3.25}{2.61 / \sqrt{900}} \quad = \frac{0.15}{0.087} \quad = 1.72$$

The critical (tabulated) values for two-tailed test at 5% level of significance are $\pm Z_{\alpha/2} = \pm Z_{0.02}5 = \pm 1.96$. Since calculated value of test statistic Z (=1.72) is less than the critical value (=1.96) and greater than critical value (= −1.96), that means it lies in non-rejection region, so we do not reject the null hypothesis i.e. we support the claim at 5% level of significance.

Thus, we conclude that sample does not provide us sufficient evidence against the claim so we may assume that the sample comes from the population of rods with mean 3.25cm.

7.8 TEST OF SIGNIFICANCE FOR DIFFERENCE OF MEANS

Let $\overline{X_1}$ be the mean of a sample of size n_1 drawn from a population with mean μ_1 and variance σ_1^2 and let $\overline{X_2}$ be the mean of an independent sample of size n_2 drawn from another population with mean μ_2 and variance σ_2^2. Since sample sizes are large.

The co-variance terms vanish, since the sample means $\overline{X_1}$,$\overline{X_2}$ are independent.

Thus, under H_o: $\mu_1 = \mu_2$, the Z statistic is given by

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Here σ_1^2 and σ_2^2 are assumed to be known. If they are unknown then their estimates provided by corresponding sample variances s_1^2 and s_2^2 respectively are used, i.e., $\widehat{\sigma_1^2} = s_1^2$ and $\widehat{\sigma_2^2} = s_2^2$, thus, in this case the test statistic becomes

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Remarks**:** If we want to test whether the two independent samples have come from the same population i.e., if $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (with common S.D. σ), then under H_o : $\mu_1 = \mu_2$

$$Z=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

If the common variance σ^2 is not known, then we use its estimate based on both the samples which is given by

$$\widehat{\sigma}^2=\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2}$$

Example 6: A university conducts both face to face and regular classes for a particular course indented both to be identical. A sample of 50 students of face-to-face mode yields examination results mean and SD respectively as: $\bar{X}_1$ = 80.4, S_1 = 12.8 and other sample of 100 regular students yields mean and SD of their examination results in the same course respectively as: $\bar{X}_2$= 74.3, S_2 = 20.5, Are both educational methods statistically equal at 5% level?

Solution: Here, we are given that

n_1 = 50, $\bar{X}_1$=80.4, S_1 =12.8

n_2 = 100, $\bar{X}_2$=74.3, S_2 =20.5

We wish to test that both educational methods are statistically equal. If μ_1 and μ_2 denote the average marks of face to face and distance mode students respectively then our claim is $\mu_1 = \mu_2$ and its complement is $\mu_1 \neq \mu_2$. Since the claim contains the equality sign so we can take the claim as the null hypothesis and complement as the alternative hypothesis. Thus,

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$ (two-tailed)

We want to test the null hypothesis regarding two population means when σ standard deviations of both populations are unknown. So, we should go for t-test if population of difference is known to be normal. But it is not the case.

Since sample sizes are large (n_1, and n_2 > 30) so we go for Z-test. For testing the null hypothesis, the test statistic Z is given by

$$Z=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \quad =\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{12.8^2}{50}+\frac{20.5^2}{100}}} \quad =\frac{6.1}{\sqrt{3.28+4.20}} \quad = 2.23$$

The critical (tabulated) values for two-tailed test at 5% level of significance are $\pm Z_{\alpha/2} = \pm Z_{0.025}$ = ± 1.96. Since calculated value of Z (=2.23) is greater than the critical values (= ±1.96), that means it lies in rejection region, so we reject the null hypothesis i.e. we reject the claim at 5% level of significance.

Example 7: Two research laboratories have identically produced medicines that provide relief to thyroid patients. The first medicine was tested on a group of 50 thyroid patients and produced an average 8.3 hours of relief with a standard deviation of 1.2 hours. The second medicine was tested on 100 patients, producing an average of 8.0 hours of relief with a standard deviation of 1.5 hours. Do the first medicines provide a significant longer period of relief at a significant level of 5%?

Solution: $n_1 = 50$, $\bar{X}_1 = 8.3$, $S_1 = 1.2$

$n_2 = 100$, $\bar{X}_2 = 8.0$, $S_2 = 1.5$

Here, we want to test that the first medicines provide a significant longer period of relief than the other. If μ_1 and μ_2 denote the mean relief time due to first and second medicines respectively then our claim is $\mu_1 > \mu_2$ and its complement is $\mu_1 \le \mu_2$. Since complement contains the equality sign so we can take the complement as the null hypothesis and the claim as the alternative hypothesis.

Thus, $H_0: \mu_1 = \mu_2$

$H_1: \mu_1 > \mu_2$

Since the alternative hypothesis is right-tailed so the test is right-tailed test.

We want to test the null hypothesis regarding equality of two population means. The standard deviations of both populations are unknown. So, we should go for t-test if population of difference is known to be normal. But it is not the case. Since sample sizes are large (n_1, and $n_2 > 30$) so we go for Z-test. So, for testing the null hypothesis, the test statistic Z is given by

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{8.3 - 8.0}{\sqrt{\frac{1.2^2}{50} + \frac{1.5^2}{100}}} \qquad = \frac{0.3}{\sqrt{0.0288 + 0.0255}} \qquad = \frac{0.3}{0.2265}$$

$Z = 1.32$

The critical (tabulated) value for right-tailed test at 5% level of significance is $Z_\alpha = Z_{0.05} = 1.645$.

Since calculated value of test statistic Z (=1.32) is less than the critical value (=1.645), that means it lies in non-rejection region, so we do not reject the null hypothesis and reject the alternative hypothesis i.e. we reject the claim at 5% level of significance.

Thus, the samples provide sufficient evidence against the claim so the first medicines do not have longer period of relief than the other.

7.9 COMPARISON OF Z-TEST (Large Sample Test) AND t-TEST (Small Sample Test)

Basis for comparison	**t-test**	**Z-test**

Definition	When the population's standard deviation is unknown, the t-test is a statistical test that is used to evaluate hypotheses about the mean of a small sample taken from the population.	A statistical technique called the z-test is used to compare or assess the significance of various statistical measures, most notably the mean in a sample taken from a population that is normally distributed or between two independent samples.
Sample size	$n \leq 30$	$n > 30$
Assumptions	A t-test is not based on the assumption that all key points on the sample are independent.	z-test is based on the assumption that all key points on the sample are independent.
Population variance	Unknown	known
Variance or standard deviation	Variance or standard deviation is not known in the t-test.	Variance or standard deviation is known in z-test.
Distribution	The sample values are to be recorded or calculated by the researcher.	In a normal distribution, the average is considered 0 and the variance as 1.
Population parameters	In addition, to the mean it compares partial or simple correlations among two samples.	In addition, to mean, it compares the population proportion.

7.10 ASSUMPTIONS OF F-TEST

1. **Normality**: The data within each group or sample should be approximately normally distributed. This assumption is more critical when sample sizes are small.
2. **Homogeneity of Variances**: The variances of the populations from which the samples are drawn should be equal. This is crucial for the validity of the F-test results. In one-way ANOVA, this assumption is specifically about the equality of variances across groups.
3. **Independence**: Observations within each group should be independent of each other. The values in one group should not be dependent on or related to the values in another group.
4. **Random Sampling**: The data should be collected through a random sampling process to ensure that the sample is representative of the population.

5. **Independence of Errors**: this implies that the variation of each item around the group should be independent for each value.

7.11 MAIN PROPERTIES OF F DISTRIBUTION

The F-distribution depends on the degrees of freedom and is usually defined as the ratio of variances of two populations normally distributed and therefore it is also called Variance Ratio Distribution**.**

1. The F-distribution is positively skewed and with the increase in the degrees of freedom m and n, its skewness decreases.
2. The value of the F-distribution is always positive, or zero since the variances are the square of the deviations and hence cannot assume negative values. Its value lies between 0 and ∞.
3. The shape of the F-distribution depends on its parameters ν_1 and ν_2 degrees of freedom.
4. Mean $=\frac{n}{n-2}$

   Mean is defined for n>2 and is independent of m. Mean of F is always positive

5. Variance $=\frac{2n^2(m+n+2)}{m\,(n-2)^2\,(n-4)}$, n > 4

   Variance is always positive and is defined for n > 4.

6. Mode $= \frac{(\mathrm{m}-2)}{m\,(n+2)}$

   Mode is defined for n > 2 and is always less than 1.

7. Karl Pearson's coefficient of skewness

   $S_k = \frac{\text{Mean}-\text{Mode}}{S.D} > 0$

   Since mean > 1 and Mode < 1.

8. F-distribution is positively skewed.
9. The probability curve of F firstly increases rapidly and reaches its maximum at mode (which is less than 1). Then it falls slowly and becomes an asymptote to the X-axis.
10. If a statistic F follows F distribution with degrees of freedom (m, n), then its reciprocal, 1/F follows F distribution with (n, m) degrees of freedom.

11. distribution tends to be the normal distribution for large (m, n).
12. Critical region of F distribution: The F-test is always a right-tailed test and as such the critical region always lies on the right tail of the distribution.

7.12 THE PROCEDURE FOR F-TEST FOR EQUALITY OF POPULATION VARIANCE

Let X_1, X_2, …….. X_n is a random sample of size n1 from a normal population with mean μ_1 and variance σ_1^2. Similarly, Y_1, Y_2…….. Y_n is a random sample of size n_2 from another normal population with mean μ_2 and variance σ_2^2. Here, we want to test the hypothesis about the two population variances so we can take our alternative null and hypotheses as

1. Set up the null hypothesis

H_0: $\sigma_1^2 = \sigma_2^2$

2. Set up alternative hypothesis

In case of one tailed:

H_1: $\sigma_1^2 > \sigma_2^2$ (Right-tailed)

In this case, the rejection (critical) region falls at the right side of the probability curve of the sampling distribution of test statistic F.

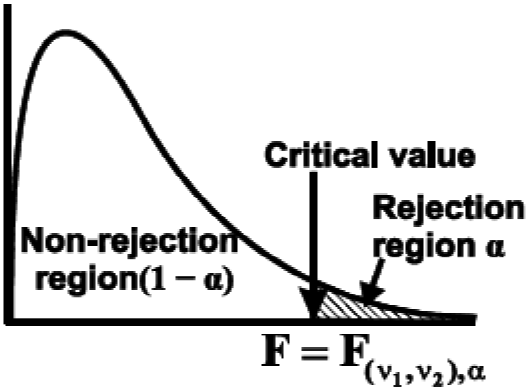

Critical value
Rejection region α
Non-rejection region(1 − α)
$\mathbf{F = F_{(v_1,v_2),\alpha}}$

Fig. 1: Right-tailed

H_1: $\sigma_1^2 < \sigma_2^2$ (Left-tailed)

In this case, the rejection (critical) region falls at the left side of the probability curve of the sampling distribution of test statistic F.

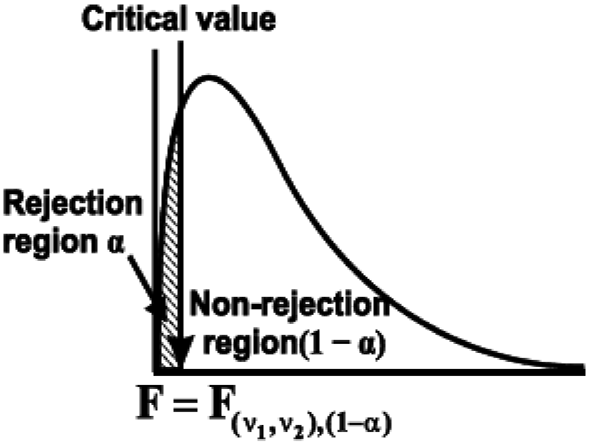


Fig. 2. Left-tailed

In case of two tailed

H1: $\sigma_1^2 \neq \sigma_2^2$

In this case, the rejection (critical) region falls at both sides of the probability curve of the sampling distribution of test statistic F and half the area(α) i.e. α/2 of rejection (critical) region lies at left tail and other half on the right tail.

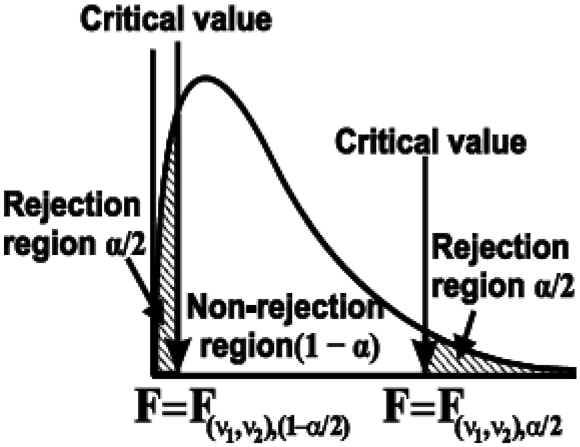

Critical value
Critical value
Rejection region α/2
Rejection region α/2
Non-rejection region(1 − α)
$F=F_{(v_1,v_2),(1-\alpha/2)}$
$F=F_{(v_1,v_2),\alpha/2}$

Fig 3. Two-tailed

3. Compute the test statistic

$$F=\frac{s_1^2}{s_2^2}$$

Where, $s_1^2 = \frac{1}{n_1-1} \sum(X - \overline{X})^2$

and $s_2^2 = \frac{1}{n_2-1}\sum(Y-\overline{Y})^2$

Note: Always take larger variance in the numerator of F. If sample standard deviations are given, then

$$s_1^2 = \frac{n_1\, s_1^2}{n_1 - 1}$$

$$s_2^2 = \frac{n_1\, s_2^2}{n_2 - 1}$$

4. Choose the appropriate level of significance (α) 90 %, 95% and 99%

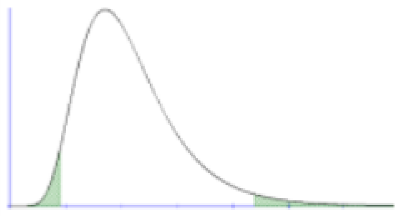
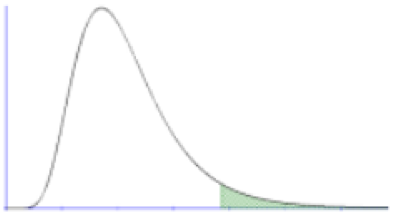
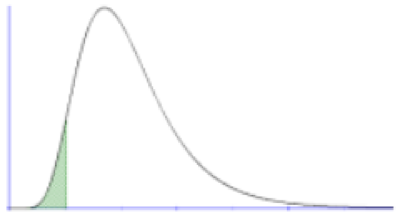
5. Procedure of deciding the null hypothesis based on p-value.

To decide on the null hypothesis based on p-value, the p-value is compared with the level of significance (α). Compare the calculated value of F with the critical value. If $F > F_\alpha$, then reject H_0 where F_α is the critical value of F at a level of significance.

Note: With the help of computer packages and software such as SPSS, SAS, MINITAB, EXCEL, etc. we can find the exact p-value for F-test.

7.13 HYPOTHESIS TEST FOR TWO VARIANCES

Sometimes we will need to compare the variation or standard deviation between two groups. For example, let’s say that the average delivery time for two locations of the same company is the same but we hear complaint of inconsistent delivery times for one location. We can use an F-test to see if the standard deviations for the two locations was different.

Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 < \sigma_2^2$
		
$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ $H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$	$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ $H_1: \frac{\sigma_1^2}{\sigma_2^2} > 1$	$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ $H_1: \frac{\sigma_1^2}{\sigma_2^2} < 1$

7.14 F-TEST FOR EQUALITY OF POPULATION VARIANCES FORMULA

Suppose we have two random samples of size n_1 and n_2 from two independent populations. We want to test whether the variances of two populations are equal or not. variances of two s are same i.e.,

H_0: $\sigma_1^2 = \sigma_2^2$

H_1: $\sigma_1^2 \neq \sigma_2^2$

Under H_0 the F-statistic is

$$F = \frac{s_1^2}{s_2^2}$$

When $s_1^2 \; and \; s_2^2$ are unbiased estimates of population variances

$$s_1^2 = \frac{1}{n_1 - 1} \sum (X - \overline{X})^2$$

$$s_2^2 = \frac{1}{n_2 - 1} \sum (Y - \overline{Y})^2$$

F follows F-distribution with n_1-1 and n_2-1, d.f It should be noted that the alternative hypothesis in this case is $\sigma_1^2 > \sigma_2^2$ (Right tail). numerical problems we take greater of the variances s_1^2 or s_2^2 in the numerator and adjust the d.f accordingly.

7.15 APPLICATION OF F-DISTRIBUTION

F-distribution has the following applications in statistical theory:

1. F-test for equality of population variances
2. F-test for testing the significance of an observed multiple correlation coefficient
3. F-test for equality of several means.

1. F-test for equality of population variances:

Suppose we are interested to find if two normal populations have same variance. Let $X_1, X_2, ..., X_{n1}$ be a random sample of size n_1, from the first normal population with variance $\sigma_1^2$and $Y_1, Y_2, ..., Y_{n2}$ be a random sample of size n_2 from the second normal population with variance σ_2^2 Obviously the two samples are independent. Set up the Null Hypothesis H_0: $\sigma_1^2 = \sigma_2^2 = \sigma^2$, population variances are same. In other words, H_0 is that the two independent estimates of the common population variance do not differ significantly.

Therefore, Under H_0 the test statistic is

$$F=\frac{s_1^2}{s_2^2}$$

F (n_1-1, n_2-1)

where s_1^2, and s_2^2 are unbiased estimates of the common population variance σ^2 and are given by:

$$s_1^2 = \frac{1}{n_1-1} \sum(X - \overline{X})^2$$

$$s_2^2 = \frac{1}{n_2-1} \sum(Y - \overline{Y})^2$$

F distribution with v_1= n_1-1, v_2= n_2-1 d.f F (v_1, v_2),

2. **F-test for testing the significance of an observed multiple correlation coefficient**

In multiple regression, the coefficient of determination, R^2, is the squared correlation between the observed values of the outcome variable y, and its predicted values. To test whether the population coefficient of determination, denoted ρ2, is 0, an F-test is used. Suppose k is the number of predictors in the regression model, and N is the sample size, the F-test is computed as

$$F = \frac{R^2/k}{(1-R^2)/(N-k-1)} \quad \ldots\ldots\ldots\ldots (1)$$

which, under the assumption of normality of the errors, has an F-distribution with k numerator degrees of freedom, and N−k−1 denominator degrees of freedom.

When investigators want to test a large model with k_2 predictors against a smaller model with k_1 (k1 < k_2) predictors, an F-test may be used for testing the change in R^2 for significance, denoted ΔR^2. Suppose that R_1^2 is the R^2 of the smaller model and R_2^2 is the R^2 of the larger model. The F-test for testing ΔR^2 for significance is given by

$$F = \frac{(R_2^2 - R_1^2\big)/\big(k_2 - k_1\big)}{(1-R_2^2\big)/N - k_2 - 1}. \quad \ldots\ldots\ldots\ldots (2)$$

For an overview of regression and its statistical tests, see Chatterjee and Hadi (Citation1999).

The computation of both (Δ)R^2 and the F-tests may be complicated by missing data. A highly recommended technique to handle missing data is multiple imputation.

The complete multiple imputation process consists of three steps:

- the missing data are estimated several times (M) using a stochastic model that accurately describes the data, creating M plausible complete versions of the incomplete data set,
- each completed data set is analyzed using the same statistical analysis, resulting in M different outcomes of this analysis, and
- the M analyses are combined into one analysis, using specific formulas that take into account the additional uncertainty due to the missing data in the standard errors and statistical tests. Such formulas for obtaining overall statistics from multiply imputed data sets are henceforth denoted combination rules.

3. F-test for equality of several means

Analysis of variance (ANOVA) can determine whether the means of three or more groups are different. ANOVA uses F-tests to statistically test the equality of means.

Example 1: In a sample of 8 observations, the sum of the squared deviations of items from their mean was 94.5. ln another sample of 10 observations, the value was found to be 101.7 Test whether the difference is significant at 5% level. (Given that at 5% level, critical value of F for v_1=7 and v_2=9 degrees of freedom are 3.29 and for v_1=8 and v_2=10 degrees of freedom, its value is 3.07)

Solution: Since we are given the critical values of F-statistic we shall apply F-test for equality of population variances.

Null Hypothesis H_0 = $\sigma_1^2 = \sigma_2^2$, i.e., the sample variances do not differ significantly

Alternative Hypothesis H_1 = $\sigma_1^2 \neq \sigma_2^2$, i.e, the sample variances differ significantly

n_1= 8, n_2= 10

$\sum(x - \overline{x})^2$= 94.5

$\sum(y - \overline{y})^2$ = 101.7

$s_x^2 = \frac{1}{n_1 - 1} \sum(X - \overline{X})^2$

$= \frac{94.5}{7} = 13.5$

$s_y^2 = \frac{1}{n_2-1}\sum(Y-\overline{Y})^2$

$= \frac{101.7}{9} = 11.3$

Now, $s_x^2 > s_y^2$

$F = \frac{s_x^2}{s_y^2} = 1.195$

f-distribution with d.f is (8-1, 10-1) i.e., (7,9) d.f tab $F_{0.05}$ = (7,9) = 3.29.

Since the calculated value is less than the table value (Cal F < Tab F), it is not significant. Hence H_0 is accepted and concluded that the difference in sample variability is not significant and may be due to sample fluctuations.

Example 2: In a study of wheat productivity in a sample of common ten subdivisions of equal area of agricultural plots it was seen that $\sum(x-\overline{x})^2$= 0.92 and $\sum(y-\overline{y})^2$= 0.26. Test at 5% significance level whether samples taken from two random populations have the same variance.

Solution: H_0: $\sigma_1^2 = \sigma_2^2$, null hypothesis states that there is no difference between the variance of two populations.

H_1: $\sigma_1^2 \neq \sigma_2^2$, alternative hypothesis states that there is a difference between the variance of two populations.

F-test is calculated as

$F = \frac{s_x^2}{s_y^2}$

$s_x^2 = \frac{1}{n_1-1}\sum(X-\overline{X})^2$

$= \frac{0.92}{10-1} = .102$

$s_y^2 = \frac{1}{n_2-1}\sum(Y-\overline{Y})^2$

$= \frac{0.26}{10-1} = .028$

$$F=\frac{.102}{.028} = 3.64$$

Degree of freedom for sample 1 = (n-1) = 9

Degree of freedom for sample 2 = (n-1) = 9

The table value of F for v_1 = 9, v_2 = 9 at 5% significance level is 3.18.

Since the calculated value is more than the table value (Cal F> Tab F), hence the null hypothesis is rejected. i.e. the samples have been drawn from populations having different variance.

Example 3: Two random sample has been drawn from two normal populations:

Sample A:	**75**	**68**	**65**	**70**	**84**	**66**	**55**
Sample B:	**42**	**44**	**56**	**52**	**46**		

Test using variance ratio of 5 % level of significance that whether two populations have same variance

Solution: H_0: $\sigma_1^2 = \sigma_2^2$, null hypothesis states that there is no difference between the variance of sample A and B.

H_1: $\sigma_1^2 \neq \sigma_2^2$, alternative hypothesis states that there is a difference between the variance of sample A and B.

x	(x− $\bar{x}$)	$(x-\overline{x})^2$	(y- $\bar{y}$)	$(y-\overline{y})^2$	
75	6				
68	-1	1	42	-6	36
65	-4	16	44	-4	16
70	1	1	56	8	64
84	15	225	52	-4	16
66	-3	9	46	-2	4
55	14	196			
$\sum x = 483$		$\Sigma(x-\overline{x})^2$= 484	$\sum y = 240$		$\Sigma(y-\overline{y})^2$ = 136

$$\bar{x} = \frac{\Sigma x}{n1} = \frac{483}{7} = 69$$

$\bar{y} = \frac{\Sigma y}{n2} = \frac{240}{5} = 48$

$s_x^2 = \frac{1}{n_1 - 1} \sum(X - \overline{X})^2$

$= \frac{484}{7-1} = 80.67$

$s_y^2 = \frac{1}{n_2 - 1} \sum(Y - \overline{Y})^2$

$= \frac{136}{5-1} = 34$

Now, $s_x^2 > s_y^2$

$F = \frac{s_x^2}{s_y^2} = \frac{80.67}{34} = 2.37$

Degree of freedom for sample 1 = $(n_1\text{-}1) = 6$

Degree of freedom for sample 2 = $(n_2\text{-}1) = 4$

The table value of F for $v_1 = 6$, $v_2 = 4$ at 5% significance level is 6.16

Since the calculated value is less than the table value (Cal F < Tab F), hence the null hypothesis is accepted. i.e. null hypothesis accepted and samples have been drawn from populations having same variance.

7.16 SUM UP

Z-test is a statistical test that is used to determine whether the mean of a sample is significantly different from a known population mean when the population standard deviation is known. It is particularly useful when the sample size is large (>30). Z-test can also be defined as a statistical method that is used to determine whether the distribution of the test statistics can be approximated using the normal distribution or not. It is the method to determine whether two sample means are approximately the same or different when their variance is known and the sample size is large (should be >= 30). The Z-test compares the difference between the sample mean and the population means by considering the standard deviation of the sampling distribution.As, before applying t-test for difference of two population means, one of the requirements is to check the equality of variances of two populations. This assumption can be checked with the help of F-test for two population variances. For example, an economist may want to test whether the variability in incomes differ in two populations, a quality controller may want to test whether the quality of the product is changing over time, etc.

7.19 QUESTIONS FOR PRACTICE

Q1. Two sources of raw materials are under consideration by a tubes manufacturing company. Both sources seem to have similar characteristics but the company is not sure about their respective uniformity. A sample of 12 lots from source A yields a variance of 125 and a sample of 10 lots from source B yields a variance of 112. Is it likely that the variance of source A significantly differs to the variance of source B at significance level $\alpha = 0.01$?

Ans: F-test= 0.28. (do not reject the null hypothesis and reject the alternative hypothesis i.e. we reject the claim at 5% level of significance.)

Q2. Two random samples drawn from two normal populations gave the following results:

	Size	Mean	Sum of Squares of Deviation from the Mean
Sample A	9	59	26
Sample B	11	60	32

Test whether both samples are from the same normal populations?

Ans: F-test = 0.88 (do not reject the null hypothesis)

Q3. In a sample of 10 observations, the sum of square of observations is 120 and in another sample of 12 observations it is 314. Test the significance defence at 5 % level.

Ans: F= 2.14, not significant at 5%

7.18 SUGGESTED READINGS

- C.R. Kothari (1990) Research Methodology. Vishwa Prakasan. India.
- Gupta SC: Fundamental of statistics, S. Chand & Company. New Delhi
- Gupta, SP: Statistical Methods, S. Chand & Company. New Delhi
- A.M Goon, M.K Gupta and B. Dasgupta, fundamental of statistics Vol-I, World press Calcutta
- Anderson, D.R.; Sweeney, D.J. and Williams, T.A., “Statistics for Business and Economics”, 2nd edition (2011), Thompson, New Delhi.

M.A (ECONOMICS)

SEMESTER -II

QUANTITATIVE METHODS II

Unit 8: Interpolation and Extrapolation

STRUCTURE

8.0 LEARNING OBJECTIVES

After Reading this Unit, Learner will be able to know:

- Meaning of interpolation and extrapolation
- Assumptions of validity
- About the methods of interpolation and extrapolation
- Advantages and Disadvantages

8.1 INTRODUCTION

In real-life economic situations, decision-making must be judicious, operative, and capable. It is well well-known fact that decision-making activity is extremely intricate given the kind and nature of economic variables. Decision-making surroundings, necessities information to scrutinize, comprehend, and develop an appropriate model for effective and efficient management. Information is critical to the decision-making process. Sometimes information relating to the decision ecosystem is available and sometimes essential information is either absent or not available for several reasons. Interpolation and Extrapolation are both statistical methods that enable the estimation of unfamiliar values in a given series. Both extrapolation and interpolation are valuable approaches to calculating or estimating the hypothetical values for an unidentified variable based on the reflection of other data points. Nevertheless, it can be difficult to distinguish among these techniques and comprehend how they differ from each other.

In the regular management of any industry or for doing forecasting for decision-making, information is gathered regularly. But it may be difficult or not possible to collect information for every time point. In the real-life world and business, several times we come across situations where we have to make an estimation of a value that is either missing or not available in the given series of information or forecast a prospective value. To handle these types of situations instead of being subject to some guesswork, the techniques of interpolation and extrapolation are quite helpful. For instance, the census of the populace in India takes place every 10 years, i.e., we have the survey statistics for 1951, 1961, 1971, 1981, 1991, and 2001. Taking the support of this accessible information, if anyone desires to know the survey data for the years 1996 or 2007 then with the help of the technique of interpolation and extrapolation the same can be estimated and arrived at. The requirement for interpolating missing information or making predictions or estimates arises in some fields like economics, commerce, social sciences, actuarial work, population studies, etc. Therefore, the practice of interpolation and extrapolation are very supportive in assessing the missing values or forecasting future values. The present chapter is focused on Interpolation and Extrapolation.

8.2 MEANING AND DEFINITION OF INTERPOLATION AND EXTRAPOLATION

In simple words, interpolation is done to estimate a value inside the specified range of the series while extrapolation on the other hand, it deals with obtaining the prediction or estimates of the required information in the earlier or future outside the specified range of the series. Interpolation, therefore, talks about the insertion of an in-between value in a sequence of items while extrapolation denotes anticipating a value for the future. Every time the technique of interpolation or extrapolation is used, it is based on the supposition that the variable whose value is to be projected is the function of the other variable. A variable is said to be the function of the other, if for any values of the independent variable (say x) we can always find a certain value of the dependent variable (say y).

Explanation of the concept with an example: Let us assume that there are two variables x and y, x being the independent variable and y the dependent variable. Further, let the given values of x be X_0, X_1, X_2….. Xn, and let the resultant values of y be Y_0. Y_1, Y_2, …. Yn, respectively. If there is a requirement to guess the value of y, for any value of x between the limits, X_0 and this can be done by using the method of Interpolation. For instance, imagine we have been given the sale values figures for the years (x) 2010, 2012, 2013, 2015, and 2018 and we want to know the sale values for any year between 2010 and 2018, say, 2017, 2014, etc. This can be done by the technique of interpolation. On the other hand, if we have to estimate the sale values for the period outside the range 2010-2018, say, for 2008 or 2020, the method is known as extrapolation.

One of the easiest methods to distinguish these dissimilarities is to know the prefix of each term. Extra- denotes "in addition to," while inter- means "in between." Consequently, extrapolation points out a user is attempting to obtain a value in addition to available values, at the same time interpolation indicates that they want to ascertain a new value in between existing values.

“Interpolation is the estimation of a most likely estimate in given conditions. The Technique of estimating a past figure is termed as interpolation, while that of estimating a probable figure for the future is called extrapolation.” by M. Harper

8.3 ASSUMPTIONS OF INTERPOLATION AND EXTRAPOLATION

As has been stated above, the interpolation or extrapolation technique is applied based on particular suppositions. The following are some of the assumptions that are taken into consideration while using the techniques of interpolation and extrapolation.

i. **No sudden or violent fluctuations in the intervening period**: At the time of interpolating or extrapolating a value, it is at all times assumed that there are no sudden deviations in the given data. In other words, the values should communicate the periods of normal and steady economic state of affairs. To put it differently, the given data on which the interpolation or extrapolation technique is to be applied should be free from all types of anomalies and all categories of unsystematic and uneven variations. If, for instance, we are interpolating the data of sales figures of a company for the year 2020 and we are given the figures of sales data for the year 2017. 2018, 2019, and 2021 we would assume that the sales of the company under consideration have grown up evenly and there are no aggressive ups and downs in these sales figures. There are a number of cases like earthquakes, wars, floods, labor strikes, lockouts, economic boom depression and political disturbances, etc., which may lead to violent ups and downs in the values, which should not be considered while applying the techniques of interpolation or extrapolation.

ii. **The percentage of change of figures from one period to another is uniform**: The second supposition is that the degree of variation of the data is uniform. Therefore, in the example of sales data given above, if we want to interpolate or extrapolate the sales figure, it is assumed that the data from a period from 2017 to 2021 has witnessed evenly growth, i.e. free from all the types of abnormalities. Taking into account these assumptions, missing data can be interpolated with a reasonable degree of precision.

8.4 ACCURACY OF INTERPOLATION AND EXTRAPOLATION

As the interpolation and extrapolation techniques are based on particular postulations which may sometimes pose some difficulties in practice, the values so estimated, may not at all times be precise or dependable, and it is difficult to ascertain the degree of error of the estimate. So, the accuracy of the interpolated or extrapolated values is affected due to:

a) likely variations in the values of the trend under investigation, which is given by the existing information at our disposal.

b) A known fact around the sequence of happenings that may disturb the value of the observable fact under consideration. If it is known that the expected value of the specified event at a specific period is influenced by random circumstances, like political disturbances, floods, etc., then the interpolation or extrapolation is disturbed and these known facts should be taken into account while reaching a certain conclusion for making any estimation for missing values.

8.5 NEED AND IMPORTANCE OF INTERPOLATION AND EXTRAPOLATION

The methods of interpolation and extrapolation are of immense real-world use, because of:

(i) Non-availability of data: Interpolation may also be necessary in case the data are inadequate because of gaps in the data or are ineptly gathered while collecting the information.

(ii) Loss of data: Data from some of the periods may be deleted, damaged, or missing due to several causes like wrong management or random and natural causes like fire, floods, etc. Such types of data may be acquired with the help of the interpolation method. The interpolation technique is therefore supportive in filling up the data gaps in accessible data.

(iii) To estimate the intermediate values: Owing to several financial and organisational teething troubles, information may not be accumulated on a survey basis and random sampling practices may be used to find the appropriate data. The in-between differences are then satisfied by interpolation methods.

(iv) To bring uniformity in the data: From time to time, it so happens that the data relating to a particular event are assembled by diverse working groups working in different categories of groups and to draw any inferences from this data is difficult for evaluation. To achieve equality in the groups, the interpolation method is resorted to. If for instance, the information is gathered for two diverse dates, for doing a comparison in them, they have to be brought at one point in time. For instance, in a nation the population survey was done in 2020, and in India the survey was done in 2021. For doing a comparison between the populations of the two nations either India’s population is to be interpolated for 2020 or the other nation’s populace is to be projected by extrapolation for 2021.

(v) For doing forecasts: Projection of future data is a fundamental necessity in any policy formation or economic planning. The extrapolation technique is used in making predictions. For instance, a company wants to project for the next financial year based on records. This can easily be done with the help of extrapolation technique.

(vi) To ascertain the positional averages in **continuous frequency spreading**: The interpolation technique has been used to develop the formulae for the working out of the median, quartiles, quintiles, cortiles, deciles, percentiles, and mode in case of continuous frequency distribution.

8.6 INTERPOLATION

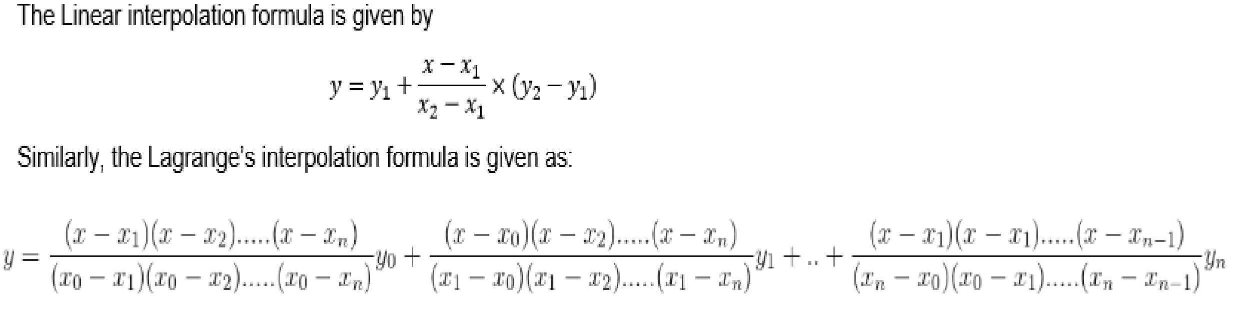
It is a technique of fitting the data points to denote the value of a function. It has some applications in engineering, commerce, industry, and science that are used to build new data points within the range of a discrete data set of known data points or can be used for finalizing a formula of the function that will pass from the given set of points (x, y). In this study material, we will be discussing the concept of interpolation in Statistics, its formulas, and its uses in detail.

Interpolation is a technique of deriving a simple function from a particular discrete data set such that the function passes through the provided data points. This supports to conclusion of the data points in between the given data. This process is at all times required to figure out the value of a function for an in-between value of the independent function. To summarize, interpolation is a method to determine the unidentified values that lie in between the known data points. It is frequently used to forecast the unknown values for any ecologically connected data points such as noise level, rainfall, elevation, and so on.

Hirach "Interpolation is the art of understanding between the lines of the table."

Interpolation Formula

The unknown value on the data points can be found using the linear interpolation and Lagrange's interpolation formula.



The Linear interpolation formula is given by

$$y = y_1 + \frac{x - x_1}{x_2 - x_1} \times (y_2 - y_1)$$

Similarly, the Lagrange's interpolation formula is given as:

$$y = \frac{(x - x_1)(x - x_2).....(x - x_n)}{(x_0 - x_1)(x_0 - x_2).....(x_0 - x_n)} y_0 + \frac{(x - x_0)(x - x_2).....(x - x_n)}{(x_1 - x_0)(x_1 - x_2).....(x_1 - x_n)} y_1 + .. + \frac{(x - x_1)(x - x_1).....(x - x_{n-1})}{(x_n - x_0)(x_0 - x_1).....(x_n - x_{n-1})} y_n$$

Interpolation explained with an example

Interpolation means ascertaining a value from the existing values in a given data set. In other words, it is a process of putting in or interjecting a middle value between two other values.

In data science or mathematics, interpolation is something like calculating a function's value based on the value of other data points in a given sequence. This function may be represented as f(x), and the known x values may range from X_0 to X_n.

For instance, imagine we have a regression line y = 3x + 4. We know that, to produce this "best-fit" line, the value of x must be between 0 and 10. Supposing we choose x = 5 Based on this best-fit line and equation, we can estimate the value of y as the following:

$$y = 3(5) + 4 = 19$$

Our x value (5) is within the range of adequate x values used to make the line of finest fit, so this is a valid y value, which we have computed by interpolation.

8.7 INTERPOLATION METHODS

Three of the most common interpolation methods are the following:

- Linear Interpolation
- Polynomial Interpolation
- Spline Interpolation

1. Linear Interpolation

Linear interpolation is amongst the simplest interpolation techniques. At this point, a straight line is drawn amid two points on a graph to control the other unidentified values. This simple technique frequently produces wrong estimates.

2. Polynomial Interpolation

While using the polynomial interpolation method, polynomial roles are used on a graph to estimate the values in the set of data that has been misdirected. It is a somewhat more comprehensive, perfect method. The polynomial graph fills in the curve amongst identified points to find the missing data between those points.

There are multiple methods of polynomial interpolation:

- Lagrange interpolation
- Newton polynomial interpolation
- Spline interpolation

The Newton method is also identified as Newton's divided differences interpolation polynomial. The Lagrange and Newton interpolation techniques outcome in the smallest polynomial function, i.e., the polynomial of the lowermost potential point that goes across the data points in the data set. Both methods produce the same outcome but to arrive at the results both use different types of calculations.

3. Spline Interpolation: In spline interpolation, piecewise functions are employed to make an estimate of the missing values and fill the gaps in a data set. In its place of assessing one polynomial for the whole of the data set as takes place in the Lagrange and Newton methods, spline interpolation describes multiple simpler polynomials for subgroups of the data. For this purpose, it commonly delivers more precise results and is believed to be a more trustworthy method.

- **Nearest Neighbour Method**– This technique introduces the value of an interpolated point to the value of the most nearby data point. Consequently, this technique does not create any new data points.
- **Cubic Spline Interpolation Method**– This process fits a diverse cubic polynomial between each pair of data points for curves or between sets of three points for surfaces.
- **Shape-Preservation Method**– This method is also known as Piecewise Cubic Hermite Interpolation (PCHIP). It maintains the monotonicity and the shape of the data. It is for curves only.
- **Thin-plate Spline Method**– This technique contains smooth surfaces that also extrapolate well. It is only for surfaces only
- **Biharmonic Interpolation Method**– This method is applied to the surfaces only.

8.8 EXTRAPOLATION

In Statistics, **Extrapolation** is a method of assessing the value outside the different range of the specified variable based on its connection with another variable. It is a very essential notion not only in Mathematics but also in other fields like Psychology, Sociology, Statistics, etc., with some definite data. Now, we will examine in detail regarding definition, formula, and examples of extrapolation. Another more significant concept is an **interpolation,** which has been discussed above as it is an estimation between the given data.

Extrapolation is described as an estimation of a value based on expanding the identified series or factors outside the range that is known. In other words, extrapolation is a method in which the data values are studied as points such as x_1, x_2, …., x_n. It normally occurs in statistical data very regularly, if that data is sampled intermittently and it approximates the next data point. One such instance is when a driver is driving a car, he ordinarily **extrapolates** about road conditions beyond his vision.

Extrapolation is a statistical method that is used in comprehending the unidentified data from the known data. It tries to forecast future data based on past data. For instance, estimating the size of the population of a country for policy making by the government after a few years based on the existing population size and its rate of growth. Another example is forecasting the sale of a particular product in the future based on the past sales record of a company.

8.9 EXTRAPOLATION METHODS

Extrapolation is categorized into three types, namely

- Linear extrapolation
- Conic extrapolation
- Polynomial Extrapolation

Let us briefly talk about these three kinds of extrapolation methods.

1. Linear Extrapolation

For any linear function, the linear extrapolation method delivers a good result when the point to be projected is not excessively far off from the given data. It is typically done by sketching the tangent line at the endpoint of the given graph and that will be extended beyond the limit.

2. Conic Extrapolation

A conic section can be formed with the assistance of five points closer to the end of the given i.e. known data. The conic segment will curve back on itself if it is a circle or ellipse. But for parabola or hyperbola, the curve will not back on itself as it is relative to the X-axis.

3. Polynomial Extrapolation

A polynomial curve can be shaped with the assistance of the whole of the identified data or near the endpoints. This technique is normally performed using Lagrange interpolation or Newton's system of finite series that arranges for the data. The final polynomial is used to extrapolate the data using the connected endpoints.

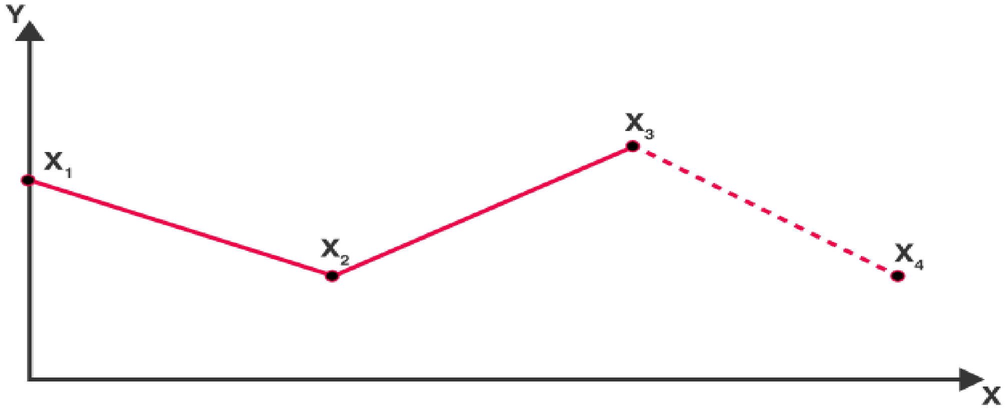
Extrapolation Formula

Let us consider the two endpoints in a linear graph (x_1, y_1) and (x_2, y_2) where the value of the point "x" is to be extrapolated, and then the extrapolation formula is given

$$y(x) = y_1 + \frac{x-x_1}{x_2-x_1}(y_2 - y_1)$$

Extrapolation Graph

As it is known, extrapolation is a method of forecasting the data point about the exterior of a curve when normally a few points are given. In the model given below, the identified data are x_1, x_2, x_3. Locating the point x_4 is acknowledged as an extrapolation point.



8.10 ADVANTAGES AND DISADVANTAGES OF INTERPOLATION

Advantages of Interpolation

- Can assess extreme changes in terrain such as Cliffs and Fault Lines.
- Thick evenly spaced points are well interpolated (flat areas with cliffs).

- Can augment or decline the amount of sample points to influence cell values.

Disadvantages of Interpolation

- Cannot make an estimation above maximum or below minimum values.
- Not very good for peaks or mountainous areas.

8.11 ADVANTAGES AND DISADVANTAGES OF EXTRAPOLATION

Advantages of Extrapolation

- Extrapolation is the analysis of data based on past trends. As past data is easily available based on which forecasting can be done to make informed decisions.
- It is an uncomplicated technique of forecasting as the rough judgment of the data can be done based on the earlier data.
- Not much data information is required as past data of a very near period is relevant for extrapolating the data.
- It is prompt and low-priced as the cost involved in collecting the past data is not high.
- It can encourage staff if aims are high. With the help of extrapolation, the staff can be encouraged by educating them that the targets have been fixed on a scientific basis.

Disadvantages of Extrapolation

- It can be uneven if there have been no changes in rise and fall with past data
- It undertakes that historical trends will always carry on into the future, which is doubtful in numerous business environments. The results will not be accurate, if there is a change in the policy of the government about a particular industry. There might be chances that the circumstances may change on a global basis like war or abnormal currency fluctuations etc.
- The technique of extrapolation overlooks the qualitative factors which cannot easily be quantifiable. If for instance, there is a change of fashion or liking of the consumers in the clothing industry, it will be difficult to extrapolate the data based on records. Extrapolation done based on the past data may not produce accurate results.
- It is a possibility that high-pitch targets can strain staff members. If targets are too high or too low based on the forecast of the previous data, it can promote dissatisfaction among the labour. As high targets will unnecessarily pressure them and low targets will result in suboptimal use of the resources.
- Extrapolating beyond a reasonable range is quite a difficult task. Sometimes a company wants to modernize the machinery to increase its productivity. It may be difficult to do a cost-benefit analysis based on future production.

8.12 APPLICATION OF EXTRAPOLATION AND INTERPOLATION

Interpolation time and again delivers a legitimate estimation of an unknown value, because of this, it's deliberated as a more dependable assessment method than extrapolation. Both approaches are advantageous for different purposes. Interpolation is particularly valuable to guess missing values or lost records to complete the records for deciding on doing any project or activity. Extrapolation is done to make forecasts about an event or occurrence based on a set of known or past values. In the real world, interpolation and extrapolation are applied in numerous fields, including the following:

- **Mathematics** to ascertain function values to reveal unidentified values to solve real-world problems;
- **Science** to make weather prediction models, forecast rainfall, or predict unknown chemical concentration values; and
- **Statistics** to forecast prospective data, such as population growth or the spread of a disease.

8.13 COMPARISON BETWEEN INTERPOLATION AND EXTRAPOLATION

Interpolation	Extrapolation
The interpretation of the values between two points in a data set. It is the prediction of a most likely assessment in the given circumstances.	Assessing a value that's outside the data set. Assessing a likely figure for the future is called extrapolation.
Predominantly used to ascertain missing past values. When data from some past periods are missing, information connecting to such projects may be assessed to finish the records by interpolation	Performs a most important role in predicting. It plays an important part in economic forecasting. For financial forecasting, prediction of future data is indispensable. This is done by extrapolation.
The expected information is more likely to be accurate. Interpolation has a preference because it has a better probability of finding an acceptable assessment.	The projected values are only possibilities, so they may not be completely accurate. In extrapolation, we normally assume that our perceived trend lasts for values of x outside the range. We worked to form our model. This may not be the case. So proper care should be taken while doing extrapolation.
It can be computed graphically. It is one of the easiest methods of interpolation.	The graphic method is not useful for extrapolation.
The technique of estimating a past figure is termed interpolation.	The technique of estimating a Future figure is termed as interpolation.

8.14 SUM UP

It is well known fact that decision-making activity is extremely intricate given the kind and nature of economic variables. Information is critical to the decision-making process. The number of times information relating to the decision ecosystem is available and sometimes essential information is either absent or not available for several reasons. Interpolation and Extrapolation are both statistical methods that enable the estimation of unfamiliar values in a given series. In simple words, interpolation is done to estimate a value inside the specified range of the series while extrapolation on the other hand deals with obtaining the prediction or estimates of the required information in the earlier or future outside the specified range of the series. At the time of interpolating or extrapolating a value, it is at all times assumed that there are no sudden deviations in the given data. The second supposition is that the degree of variation of the data is uniform. Interpolation and extrapolation are useful in the non-availability of data and loss of data, for estimating the intermediate values, bringing uniformity in the data, doing forecasts and ascertaining the positional averages in continuous frequency spreading.

8.15 QUESTIONS FOR PRACTICE

A. Short Answer Type Questions

Q1. Write a short note on interpolation with an example.
Q2. Write a short note on extrapolation with an example.
Q3. What are the different methods of interpolation?
Q4. What are the main differences between interpolation and extrapolation?

B. Long Answer Type Questions

Q1. What are the applications of interpolation and extrapolation in forecasting information for businesses?
Q2. What is the Need and Importance of Interpolation and Extrapolation?
Q3. What is extrapolation? Discuss its various methods.
Q4. What are the advantages and disadvantages of interpolation and extrapolation?

Q5. Why Interpolation and Extrapolation are required to be done? Discuss its importance.

Q6. Discuss the important assumption of interpolation and extrapolation

PRACTICAL QUESTIONS WITH SOLUTION

Q1. If (20, 60) and (40, 100) are the two points on a straight line, find the value of y, when x = 60 using linear extrapolation.

Solution: Given x1=20 and y1=60 similarly x2= 40 and y2=100, x=60

We know that $y(x) = y_1 + ((x - x_1) / (x_2 - x_1) (y_2 - y_1))$.

Now putting the value from the above we get

$$y(60) = 60 + \frac{(60 - 20)}{(40 - 20)} \times (100 - 60)$$

On solving the equation, we get $y(60) = 60 + 2 \times 40$

Thus, by solving we get $y(60) = 140$

Q2. If (60, 30) and (80, 90) are two points on a straight line, find the value of ‘y’ when x = 120 using linear extrapolation.

Solution: Given x1=60 and y1=30 similarly x2= 80 and y2=90, x=120

We know that $y(x) = y_1 + ((x - x_1) / (x_2 - x_1) (y_2 - y_1))$.

Now putting the value from the above we get

$$y(120) = 30 + \frac{(120-60)}{(80-60)} \times (90-30)$$

On solving the equation, we get $y(120) = 30 + 3 \times 60$

Thus, by solving we get $y(120) = 210$.

Q3. Find the value of y when x=10 by Lagrange's interpolation method.

x	**5**	**6**	**9**	**11**
y= f(x)	**12**	**13**	**14**	**16**

Solution:

Firstly, we will write Lagrange's interpolation method formula as given below.

$$\frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} \times f(x_0) + \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \times f(x_1) + \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} \times f(x_2) +$$

$$\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \times f(x_3)$$

The given value of x and y are depicted in the table

x	x_0= 5	x_1= 6	x_2= 9	x_3= 11
y= f(x)	12	13	14	16

Putting the values in the formula we get f(x)=

$$\frac{(x-6)(x-9)(x-11)}{(5-6)(5-9)(5-11)} \times 12 + \frac{(x-5)(x-9)(x-11)}{(6-5) 6-9) (6-11)} \times 13 + \frac{(x-5)(x-6)(x-11)}{(9-5)(9-6)(9-11)} \times 14 +$$

$$\frac{(x-5)(x-6)(x-9)}{(11-5)(11-6)(11-9)} \times 16$$

$$\text{Now } f(10) = \frac{(10-6)(10-9)(10-11)}{(5-6)(5-9)(5-11)} \times 12 + \frac{(10-5)(10-9)(10-11)}{(6-5)(6-9)(6-11)} \times 13 + \frac{(10-5)(10-6)(10-11)}{(9-5)(9-6)(9-11)} \times 14$$

$$+ \frac{(10-5)(10-6)(10-9)}{(11-5)(11-6)(11-9)} \times 16$$

Solving the equation, we get

$$\frac{(4)\ (1)\ (-1))}{(-1)\ (-4)\ (--6)} \times 12 + \frac{(5)\ (1)\ (-1)}{(1)(-3)(-5)} \times 13 + \frac{(5)\ (4)\ (-1)}{(4)(3)(-2)} \times 14 + \frac{(5)\ (4)\ (1)}{(6)(5)(2)} \times 16$$

=14.666.

Hence the value of y=f(x) will be 14.666 when x=10.

- Interpolate the value f(x) when x=4 in the following table.

x	3	5	7	9
y =f(x)	180	150	120	90

Solution:

x	x_0=3	x_1=5	x_2=7	x_3=9
y =f(x)	180	150	120	90

As the value of x lies between x_0 and x_1

First, we calculate the value of 'h', which is h=2.

Then u= $\frac{x-x_0}{h}$, where x=4 and x_0 = 3 hence u= $\frac{4-3}{2}$ = $\frac{1}{2}$ =.05.

Difference Table

x	y	Δy	Δ^2 y	Δ^3 y
3	180			
		-30		
5	150		0	
		-30		0
7	120		0	
		-30		
9	90			

The formula for interpolating the above problem is

$$y=y_0 + \frac{u}{1!} \Delta\ y_0 + \frac{u(u-1)}{2!} \Delta^2\ y_0 + \frac{u(u-1)\ (u-2)}{3!} \Delta^3\ y_0$$

$$y = 180+ \frac{(.05)}{1!} (-30) + \frac{(.05)\ (.05-1)}{2!} (0) + 0$$

y =180 -15 = 165

Hence y=165 is the answer.

Q4. Estimate the turnover of business enterprises for the year 2021 from the under mentioned data.

Year	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
Turnover in Rs. lacs	**50**	**?**	**100**	**150**	**260**	**?**

Solution:

Year	Turnover In Rs. lacs	
2016	50	y_0
2017	?	y_1
2018	100	y_2
2019	150	y_3
2020	260	y_4
2021	?	y_5

As the known values are 4

$\therefore (y - 1)^4 = 0$

$y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0$ …… (1)

The second equation can be calculated by increasing the suffixes of each term of 'y' by one and let the coefficients same.

then we get

$y_5 - 4y_4 + 6y_3 - 4y_2 + y_1 = 0$ ….(2)

Substitute y values in equation (1) we get

$y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0$

$260 - 4(150) + 6(100) - 4y_1 + 50 = 0$

$260 - 600 + 600 - 4y_1 + 50 = 0$

$-4y_1 + 310 = 0$

$y_1 = \underline{-310/ -4} = 77.50$

Hence sales in 2017 are Rs.77.50 lacs.

Now substitute y values in equation (2) we get

$y_5 - 4y_4 + 6y_3 - 4y_2 + y_1 = 0$

$y_5 - 4\,(260) + 6\,(150) - 4\,(100) + 77.50 = 0$

$y_5 - 1040 + 900 - 400 + 77.50 = 0$

$y_5 - 462.50 = 0$

$y_5 = 462.50$

Hence sales in 2021 are Rs.462.50 lacs