DIRECTORATE OF DISTANCE & CONTINUING EDUCATION

MANONMANIAM SUNDARANAR UNIVERSITY

TIRUNELVELI- 627 012

BBA Course Material

# BUSINESS STATISTICS

Prepared by

**Dr. A. Jafar Sathic**
**Assistant Professor**
**Department of Business Administration**
**Manonmaniam Sundaranar University College**
**Naduvakurichi, Sankarankovil - 627862**

## SYLLABUS

| UNIT | Details |
|---|---|
| I | **Introduction – Meaning and Definition of Statistics** – Collection and Tabulation of Statistical Data – Presentation of Statistical Data – Graphs and Diagrams- |
| II | **Measures of Central Tendency** – Arithmetic Mean,Median and Mode – Harmonic Mean and Geometric Mean. |
| III | **Measures of Variation** — Quartile deviation Mean deviation – Standard Deviation |
| IV | **Simple Correlation** – Scatter Diagram – Karl Pearson‘s Correlation – Rank Correlation – Regression. |
| V | **Testing of hypothesis** – Chi-Square test, T Test, F Test, ANOVA |

# UNIT I

## INTRODUCTION - MEANING AND DEFINITION OF STATISTICS

### 1.1 Introduction

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data. In other words, it is a mathematical discipline to collect, summarize data.

The word ‘Statistics ‘ is derived from a Latin term “Status’ or Italian term ‘Statistics’ or the German term ‘Statistick’ is the French term ‘Statistique’ each of which means a political state. The term statistics was applied to mean facts and figures and figures which were needed the state in respect of the division of the state, their respective population birth rate, income and the like.

**Meaning of statistics**

Statistics refers to numerical facts and figures collected in a systematic manner with a specific purpose in any field of study. In this sense, statistics is also aggregates of facts expressed in numerical form.

In singular sense, statistics refers to a science which comprises methods that are used in the collection, analysis, interpretation and presentation of numerical data. These methods are used to draw conclusion about the population parameters.

**Definition of statistics**

- **AL Bowley** defines statistics as "Statistics is numerical statement of facts in any development of enquiry placed in relation to each other"
- According to **Croxton and Cowden**, Statistics may be defined as the science of collection, presentation, analysis and interpretation of numerical data.
- **Gottfried Achenwall** defined statistics as "Statistics are collection of noteworthy facts concerning state both historical and descriptive".
- According to **Yule and Kendall**, “Statistics means quantitative data affected to a marked extent by multiplicity of causes.”
- “Statistics” - as defined by the **American Statistical Association (ASA**) - “is the science of learning from data, and of measuring, controlling and communicating uncertainty.”

**Nature of statistics**

- Statistics is both a science and an art
- As a science statistical methods are generally systematic and based on fundamental ideas and processes

- It also works as a base for all other sciences.
- As an art it explores the merits and demerits, guides about the means to achieve the objective

**Scope of statistics**

- Statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation and presentation of data.
- It provides tools for predicting and forecasting the economic activities.
- It is useful for an academician, government, business etc.

**Characteristics of statistics**

- Statistics are Aggregate of facts
- It is numerically expressed
- The statistical Data affected by multiplicity of causes
- It is enumerated according to reasonable standard of accuracy
- It is collected in systematic accuracy
- It is collected for pre-determined purpose
- It is placed in relation to other

**Uses of statistics**

Statistics helps in

- Providing a better understanding
- Exact description
- efficient planning of a statistical inquiry in any field of study
- Collecting appropriate quantitative data
- Business forecasting
- Decision making
- Quality control
- Search of new ventures
- Study of market
- Study of business cycles
- Useful for planning
- Useful for finding averages
- Useful for bankers, brokers, insurance, etc.

**Limitations of statistics**

- It is not useful for individual cases
- It ignores qualitative aspects
- It deals with average only
- Improper use of statistics can be dangerous
- It is only a mean, not an end
- It do not distinguish between cause and effect
- Its results are not always dependable

## 1.2 Collection and tabulation of data

## Collection of data

In Statistics, data collection is a process of gathering information from all the relevant sources to find a solution to the research problem. It helps to evaluate the outcome of the problem. The data collection methods allow a person to conclude an answer to the relevant question. Most of the organizations use data collection methods to make assumptions about future probabilities and trends. Once the data is collected, it is necessary to undergo the data organization process.

The main sources of the data collections methods are "Data". Data can be classified into two types, namely primary data and secondary data. The primary importance of data collection in any research or business process is that it helps to determine many important things about the company, particularly the performance. So, the data collection process plays an important role in all the streams. Depending on the type of data, the data collection method is divided into two categories namely,

a) Primary Data Collection methods

b) Secondary Data Collection methods

## Primary Data Collection Methods

Primary data or raw data is a type of information that is obtained directly from the first-hand source through experiments, surveys or observations. There are several methods to collect this type of data. They are

**Observation Method**

Observation method is used when the study relates to behavioural science. This method is planned systematically. It is subject to many controls and checks. The different types of observations are:

- Structured and unstructured observation
- Controlled and uncontrolled observation
- Participant, non-participant and disguised observation

### Interview Method

The method of collecting data in terms of verbal responses. It is achieved in two ways, such as

- **Personal Interview** – In this method, a person known as an interviewer is required to ask questions face to face to the other person. The personal interview can be structured or unstructured, direct investigation, focused conversation, etc.
- **Telephonic Interview** – In this method, an interviewer obtains information by contacting people on the telephone to ask the questions or views, verbally.

### Questionnaire Method

In this method, the set of questions are mailed to the respondent. They should read, reply and subsequently return the questionnaire. The questions are printed in the definite order on the form. A good survey should have the following features:

- Short and simple
- Should follow a logical sequence
- Provide adequate space for answers
- Avoid technical terms
- Should have good physical appearance such as colour, quality of the paper to attract the attention of the respondent

### Schedule method

This method is similar to the questionnaire method with a slight difference. The enumerations are specially appointed for the purpose of filling the schedules. It explains the aims and objects of the investigation and may remove misunderstandings, if any have come up. Enumerators should be trained to perform their job with hard work and patience.

## Secondary Data Collection Methods

Secondary data is data collected by someone other than the actual user. It means that the information is already available, and someone analyses it. The secondary data includes magazines, newspapers, books, journals, etc. It may be either published data or unpublished data.

Published data are available in various resources including

- Government publications
- Public records
- Historical and statistical documents
- Business documents
- Technical and trade journals

Unpublished data includes

- Diaries
- Letters
- Unpublished biographies, etc.

## 1.3 Presentation of statistical data

## Presentation of data

Presentation of data refers to an exhibition or putting up data in an attractive and useful manner such that it can be easily interpreted. The three main forms of presentation of data are:

a) Textual presentation

b) Tabular presentation

c) graphical and Diagrammatic presentation

**a) Textual presentation**

When presenting data in this way, researcher use words to describe the relationship between information. Textual presentation enables researchers to share information that cannot display on a graph. An example of data, the researcher present textually is findings in a study. When a researcher wants to provide additional context or explanation in their presentation, they may choose this format because, in text, information may appear more clear.

Textual presentation is common for sharing research and presenting new ideas. It only includes paragraphs and words, rather than tables or graphs to show data.

**b) Tabular presentation**

Tabular presentation is using a table to share large amounts of information. When using this method, researcher organise and classify the data in rows and columns according to the characteristics of the data. Tabular presentation is useful in comparing data, and it helps visualise information. Researches use this type of presentation in analysis, such as classify and tabulate them.

## Classification of data

Classification is a process of arranging things or data in groups or classes according to the common characteristics. It is based on

- Geographical (i.e. on the basis of area or region wise)
- Chronological (On the basis of Historical, i.e. with respect to time)
- Qualitative (on the basis of character / attributes)
- Numerical, quantitative (on the basis of magnitude)

1) **Geographical Classification**

In geographical classification, the classification is based on the geographical regions.

**Ex**: Sales of the company (In Million Rupees) (region – wise)

| **Region** | **Sales** |
|---|---|
| North | 285 |
| South | 300 |
| East | 185 |
| West | 235 |

2) **Chronological Classification**

If the statistical data are classified according to the time of its occurrence, the type of classification is called chronological classification.

S**ales reported by a departmental store**

| **Month** | **Sales (Rs.) in lakhs** |
|---|---|
| January | 22 |
| February | 26 |
| March | 32 |
| April | 25 |
| May | 27 |
| June | 30 |

3) **Qualitative Classification**

In qualitative classifications, the data are classified according to the presence or absence of attributes in given units. Thus, the classification is based on some quality characteristics / attributes.

**Ex**: Sex, Literacy, Education, Class grade etc.

Further, it may be classified as

a) Simple classification b) Manifold classification

i) Simple classification:

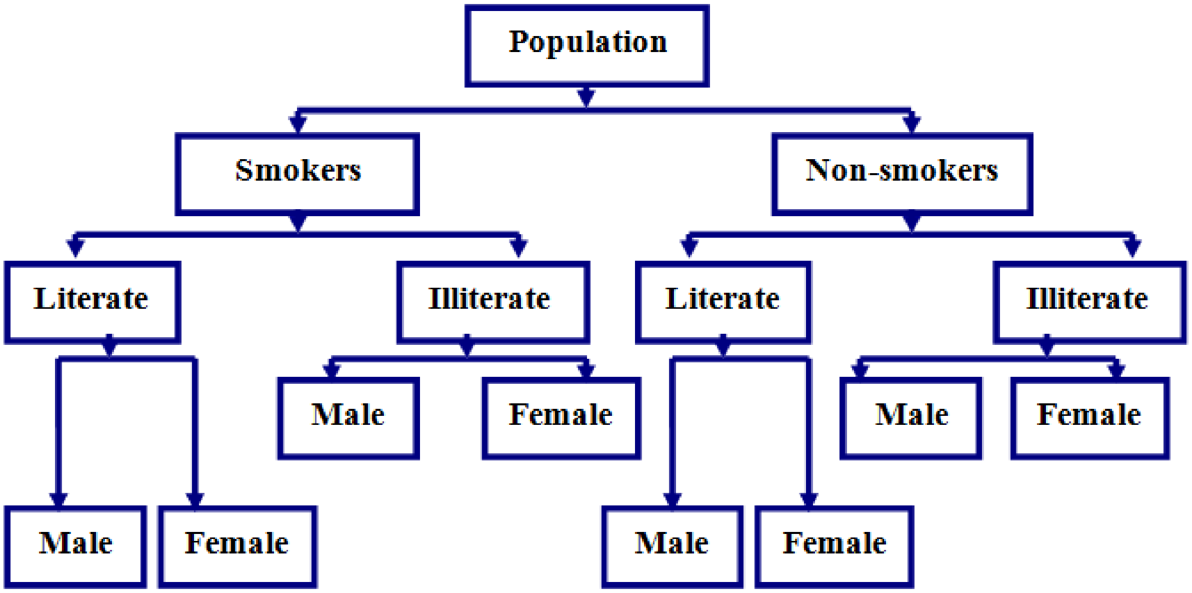
If the classification is done into only two classes then classification is known as simple classification.

**Ex:** a) Population in to Male / Female

b) Population into Educated / Uneducated

ii) Manifold classification:

In this classification, the classification is based on more than one attribute at a time.



4) **Quantitative Classification**

In Quantitative classification, the classification is based on quantitative measurements of some characteristics, such as age, marks, income, production, sales etc. The quantitative phenomenon under study is known as variable and hence this classification is also called as classification by variable.

**Ex**:

For a 50 marks test, Marks obtained by students as classified as follows

| **Marks** | **No.of students** |
|---|---|
| 0 – 10 | 5 |
| 10 – 20 | 7 |
| 20 – 30 | 10 |
| 30 – 40 | 25 |
| 40 – 50 | 3 |
| **Total Students = 50** | |

In this classification marks obtained by students is variable and number of students in each class represents the frequency.

## Tabulation of data

Tabulation may be defined, as systematic arrangement of data is column and rows. It is designed to simplify presentation of data for the purpose of analysis and statistical inferences.

**Objectives of Tabulation**

- To simplify the complex data
- To facilitate comparison
- To economise the space
- To draw valid inference / conclusions
- To help for further analysis

**Components of Data Tables**

- Table Number: Each table should have a specific table number for ease of access and locating.
- Title: A table must contain a title that clearly tells the readers about the data.
- Head notes: A head note further aids in the purpose of a title and displays more information about the table.
- Stubs: These are titles of the rows in a table.
- Caption: A caption is the title of a column in the data table. Body or field: The body of a table is the content of a table in its entirety. Each item in a body is known as a ‘cell’.
- Footnotes: Footnotes are rarely used. In effect, they supplement the title of a table if required.

**Types of tabulation**

In general, the tabulation is classified in two parts, that is simple tabulation, and a complex tabulation. Simple tabulation, gives information regarding one or more independent questions complex tabulation gives information regarding two manually dependent questions.

**Simple tabulation**

Data are classified based on only one characteristic.

**Distribution of marks**

| Class Marks | No. of students |
|---|---|
| 30 – 40 | 20 |
| 40 – 50 | 20 |
| 50 – 60 | 10 |
| **Total** | **50** |

**Complex tabulation**

Data are classified based on two or more characteristics. Two-way table: Classification is based on two characteristics.

| Class Marks | Number of students | | |
|---|---|---|---|
| | Boys | Girls | Total |
| 30 – 40 | 10 | 10 | 20 |
| 40 – 50 | 15 | 5 | 20 |
| 50 – 60 | 3 | 7 | 10 |
| **Total** | **28** | **22** | **50** |

## 1.4 Graphs and diagrams

Graphical and diagrammatic representations of data are visual aids that can help people understand data more easily. This method of displaying data uses diagrams, graphs and images. Graphs and diagrams are both visual representations of data, but they have different purposes and uses.

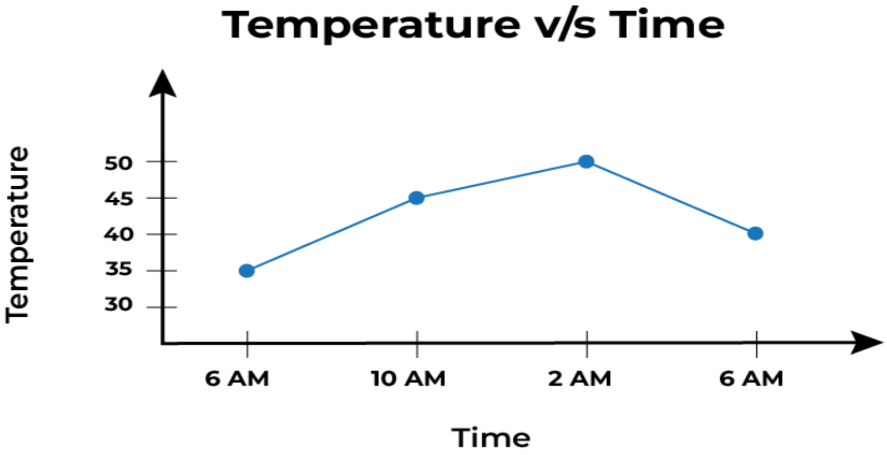
### Graphs

Data can also be effectively presented by means of graphs. A graph consists of curves or straight lines. Graphs provide a very good method of showing fluctuations and trends in statistical data. Graphs can also be used to make predictions and forecasts.

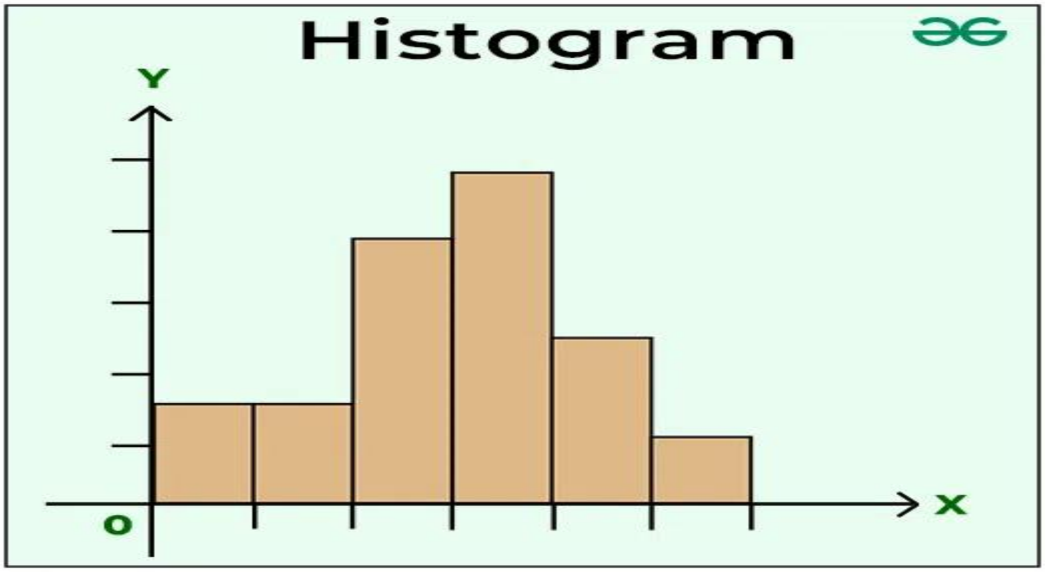
- Line graphs
- Histogram
- Frequency Polygon
- Frequency Curve

- Cumulative Frequency Polygon (Ogive)

**Line graphs**: Show changes and trends in data over time by connecting data points with lines.

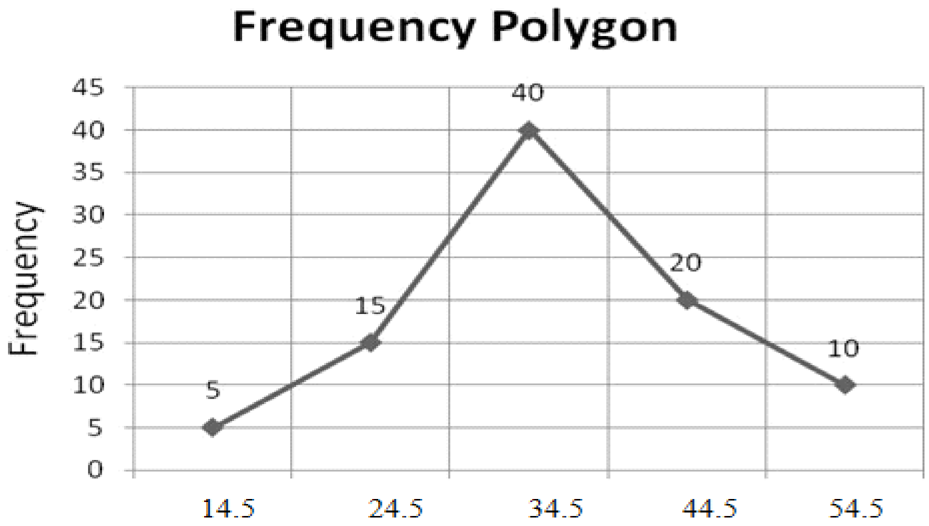


**Histograms:** Similar to column charts, histograms are made up of vertical columns whose length is proportional to the frequency of a variable.



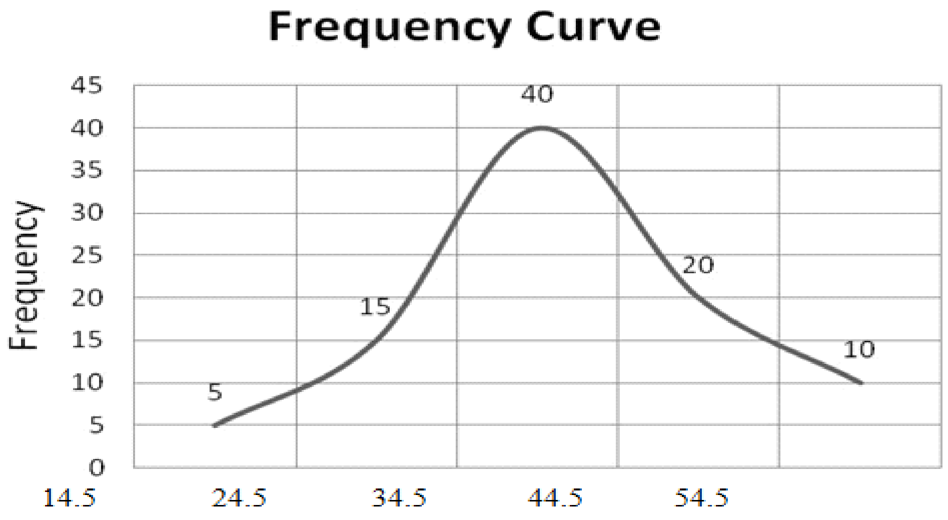
**Frequency polygon**

A frequency polygon is a many sided closed figure. It can also be obtained by joining the mid-points of the tops of rectangles in the histograms.



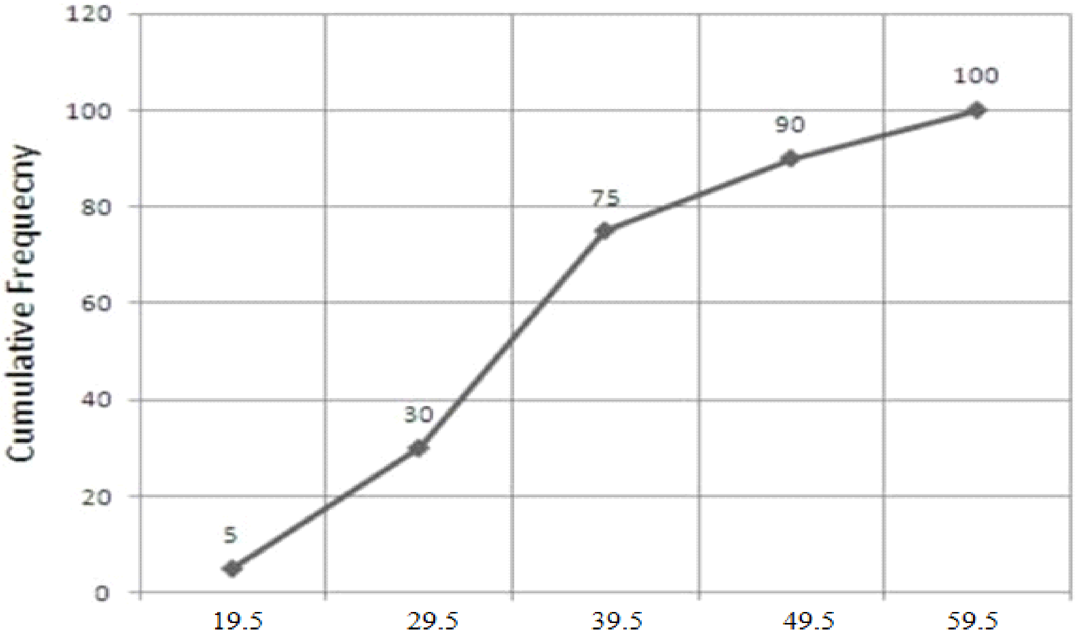
**Frequency Curve**

When the frequency polygon is smoothed out as a curve then it becomes frequency curve. OR when the mid-points are potted against the frequencies then a smooth curve passes through these points is called a frequency curve.



**Cumulative Frequency Polygon (Ogive)**

When a curve is based on cumulative frequencies then it is called a cumulative frequency polygon or ogive.



## Diagrams

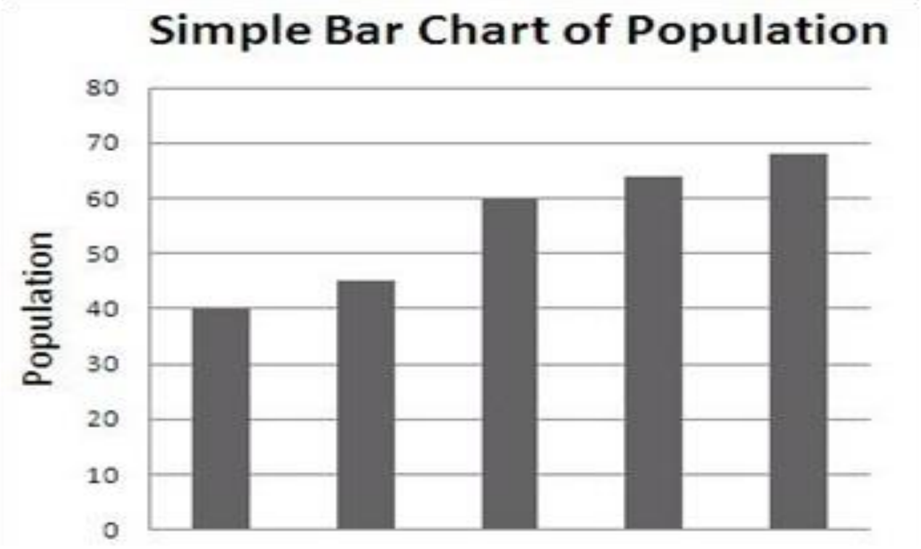
It is a technique of presenting numeric data through pictograms, cartograms, bar diagrams, and pie diagrams. It is the most attractive and appealing way to represent statistical data. Diagrams help in visual comparison and they have a bird's eye view. Diagrams classified as;

- Bar diagram/Charts
- Rectangle and Sub-divided Rectangle
- Pie diagram/Chart
- scatter plots
- Cartograms
- Pictograms

### Bar diagram/Charts

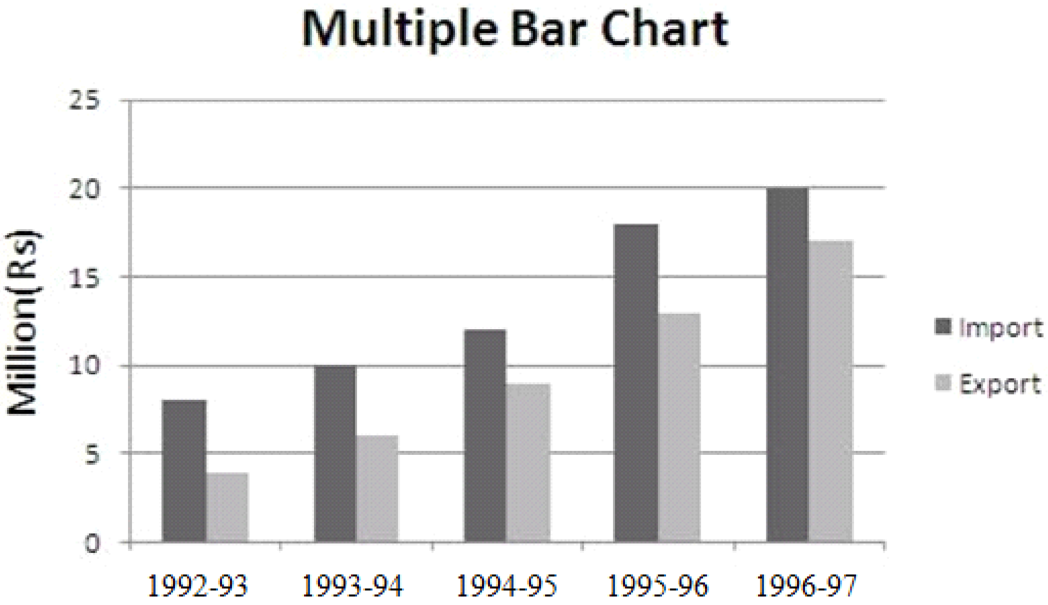
- **Simple bar chart**

This chart consists of vertical or horizontal bars of equal width.



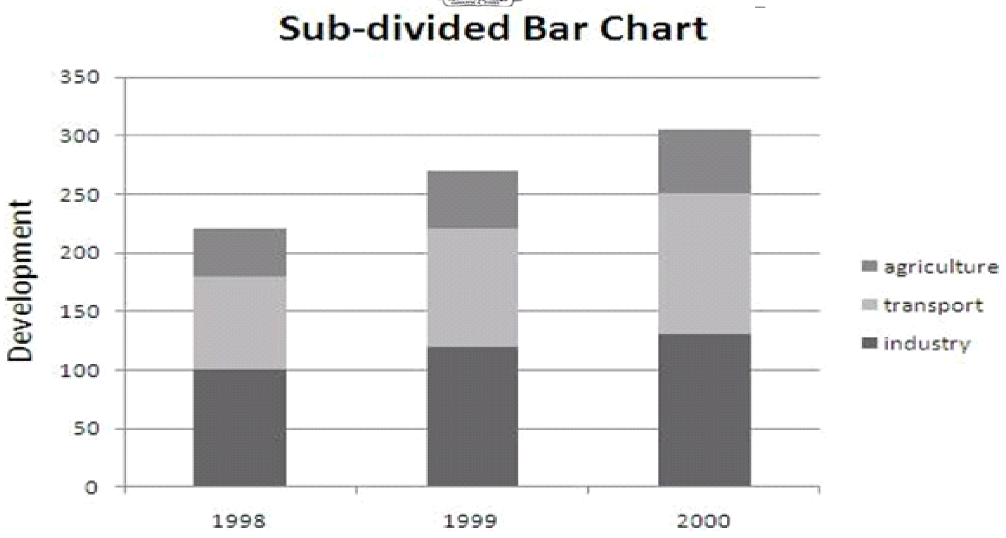
- **Multiple bar charts or cluster charts**

By multiple bar charts two or more sets of inter-related data are represented. Multiple bar charts facilities comparison between more than one phenomenon.



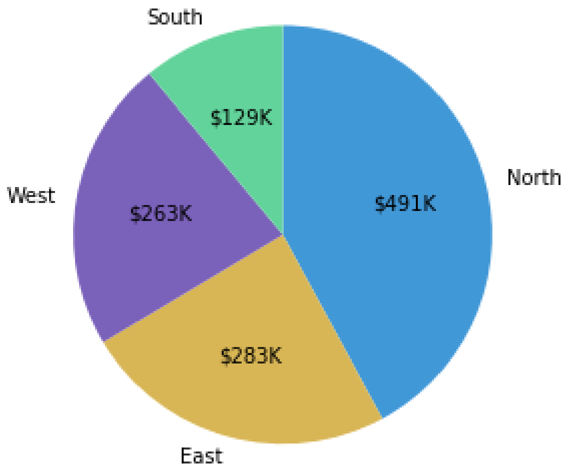
**Component bar chart or sub divided bar chart**

A component bar chart is an effective technique in which each bar is sub-divided into two or more parts. The component parts are shaded or coloured differently to increase the overall effectiveness of the diagram.



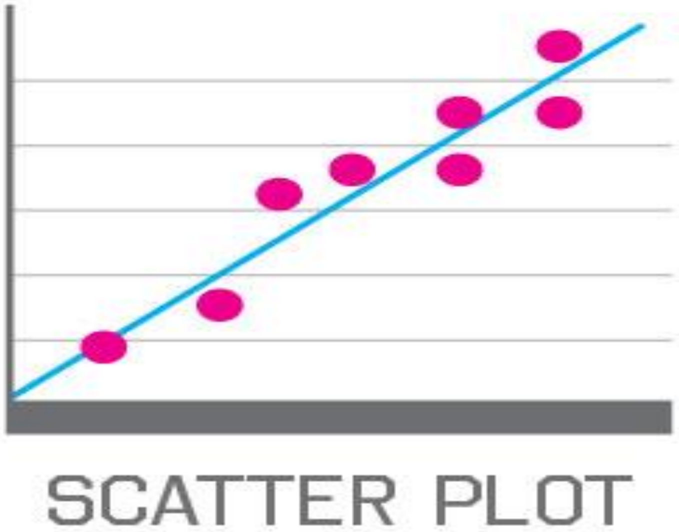
**Pie chart**

A pie chart is a type of a chart that visually displays data in a circular graph. It is one of the most commonly used graphs to represent data using the attributes of circles, spheres, and angular data to represent real-world information.



**Scatter plots**

A scatter plot, also called a scatter plot, scatter graph, scatter chart, scattergram, or scatter diagram. It uses dots to represent values for two different numeric variables. Scatter plots are used to observe relationships between variables.

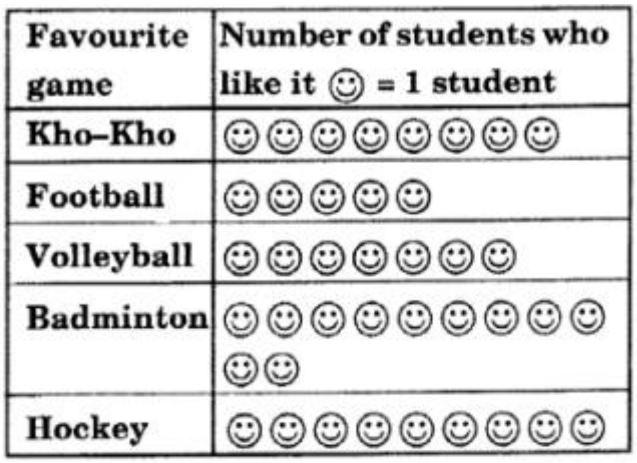


**Cartograms**

This includes any type of map that shares the location of a person, place or object. For example, cartograms help navigate theme parks so you can find attractions, food and gift shops.

INDIA
POPULATION DENSITY (Per Sq. Km)
2001
LEGEND

**Pictograms:** This diagram uses images to represent data. For example, to show the number students and their favourite games shown below using pictogram.



| Favourite game | Number of students who like it ☺ = 1 student |
|---|---|
| Kho–Kho | ☺☺☺☺☺☺☺☺ |
| Football | ☺☺☺☺☺ |
| Volleyball | ☺☺☺☺☺☺☺ |
| Badminton | ☺☺☺☺☺☺☺☺☺<br>☺☺ |
| Hockey | ☺☺☺☺☺☺☺☺☺ |

**Applications of statistics in Business Decisions:**

The field of statistics has numerous applications in business. Because of technological advancements, large amounts of data are generated by business these days. These data are now being used to make decisions. These better decisions we make help us improve the running of a department, a company, or the entire economy.

***"Statistics is extensively used to enhance Business performance through Analytics"***

- **Marketing:** As per Philip Kotler and Gary Armstrong marketing ― identifies customer needs and wants , determine which target markets the organisations can serve best, and designs appropriate products, services and Programs to serve these markets.

  Marketing is all about creating and growing customers profitably. Statistics is used in almost every aspect of creating and growing customers profitably. Statistics is extensively used in making decisions regarding how to sell products to customers. Also, intelligent use of statistics helps managers to design marketing campaigns targeted at the potential customers. Marketing research is the systematic and objective gathering, recording and analysis of data about aspects related to marketing. IMRB international, TNS India, RNB Research, The Nielson, Hansa Research and Ipsos Indica Research are some of the popular market research companies in India. Web analytics is about the tracking of online behaviour of potential customers and studying the behaviour of browsers to various websites. Use of Statistics is indispensable in forecasting sales, market share and demand for various types of Industrial products. Factor analysis, conjoint analysis and multidimensional scaling are invaluable tools which are based on statistical concepts, for designing of products and services based on customer response.

- **Finance:** Uncertainty is the hallmark of the financial world. All financial decisions are based on ‒Expectation‖ that is best analysed with the help of the theory of probability and statistical techniques. Probability and statistics are used extensively in designing of new insurance policies and in fixing of premiums for insurance policies. Statistical tools and technique are used for analysing risk and quantifying risk, also used in valuation of derivative instruments, comparing return on investment in two or more instruments or companies. Beta of a stock or equity is a statistical tool for comparing volatility, and is highly useful for selection of portfolio of stocks. The most sophisticated traders in today‘s stock markets are those who trade in ‒derivatives‥ i.e financial instruments whose underlying price depends on the price of some other asset.

- **Economics:** Statistical data and methods render valuable assistance in the proper understanding of the economic problem and the formulation of economic policies. Most economic phenomena and indicators can be quantified and dealt with statistically sound logic. In fact, Statistics got so much integrated with Economics that it led to development of a new subject called Econometrics which basically deals with economics issues involving use of Statistics.

- **Operations:** The field of operations is about transforming various resources into product and services in the place, quantity, cost, quality and time as required by the customers. Statistics plays a very useful role at the input stage through sampling inspection and inventory management, in the process stage through statistical quality control and six sigma method, and in the output stage through sampling inspection. The term Six Sigma quality refers to situation where there is only 3.4 defects per million opportunities.

- **Human Resource Management or Development:** Human Resource departments are inter alia entrusted with the responsibility of evaluating the performance, developing rating systems, evolving compensatory reward and training system, etc. All these functions involvedesigning forms, collecting, storing, retrieval and analysis of a mass of data. All these functions can be performed efficiently and effectively with the help of statistics.

- **Information Systems:** Information Technology (IT) and statistics both have similar systematic approach in problem solving. IT uses statistics in various areas like, optimisation of server time, assessing performance of a program by finding time taken as well as resources used by the program. It is also used in testing of the software.

- **Data Mining:** Data Mining is used in almost all fields of business. In Marketing, Data mining can be used for market analysis and management, target marketing, CRM, market basket analysis, cross selling, market segmentation, customer profiling and managing web based marketing, etc. In Risk analysis and management, it is used for forecasting, customer retention, quality control, competitive analysis and detection of unusual patterns.

In Finance, it is used in corporate planning and risk evaluation, financial planning and asset evaluation, cash flow analysis and prediction, contingent claim analysis to evaluate assets, cross sectional and time series analysis, customer credit rating, detecting of money laundering and other financial crimes. In Operations, it is used for resource planning, for summarising and comparing the resources and spending. In Retail industry, it is used to identify customer behaviours, patterns and trends as also for designing more effective goods transportation and distribution policies, etc.

## Important questions

**Choose the correct answer:**

1. Review of performance appraisal, labour turnover rates, planning of incentives, and training programs are the examples of which of the following?

   a) Statistics in production

   b) Statistics in marketing

   c) Statistics in finance

   **d) Statistics in personnel management**

2. When an investigator uses the data which has already been collected by others, such data iscalled _____________.

   a) Primary data

   b) Collected data

   c) Processed data

   **d) Secondary data**

3. In the case of the questionnaire method of gathering data, it should be made certain that allthe questions have been_____.

   a) Read

   b) Interpreted

   c) **Answered**

   d) All of the above

4. ______means separating items according to similar characteristics and groupingthem into various classes.

   a) Tabulation

   b) Editing

   c) Separation

   d) **Classification**

5. ________is one which is collected by the investigator himself for thepurpose of a specific inquiry or study.

   a) Secondary data

   b) **Primary data**

   c) Statistical data

   d) Published data

6 In chronological classification, the data is classified on the basis of:

a) **Time**
b) Money
c) Location
d) Quality

7. The classification of data according to location is what classification:

a) Chronological
b) Quantitative
c) Qualitative
d) **Geographical**

8. A systematic arrangement of data in rows and columns is:

a) **Table**
b) Tabulation
c) Body
d) All the above

9. In the statistical table column headings are called:

a) Stubs
b) **Captions**
c) Source note
d) None of these

10. One dimensional diagram is:

a) **Line diagram**
b) Rectangles
c) Cubes
d) Squares

11. A pie diagram is also called:

a) Pictogram
b) **Angular diagram**
c) Line diagram
d) Bar diagram

12. The median of a frequency distribution is found graphically with the help of:

    a) Histogram

    b) Frequency curve

    c) Frequency polygon

    **d) Ogive**

13. The mode of a frequency distribution can be determined graphically by:

    **a) Histogram**

    b) Frequency curve

    c) Frequency polygon

    d) Ogive

**Theory questions:**

1. Explain the functions of statistics.
2. State the limitations of statistics
3. State the uses of statistics.
4. Describe the applications of statistics in Business & Industry?
5. Explain Classification.
6. Explain different types of primary data collection methods.
7. What are the major sources for secondary data?
8. Distinguish between classification and tabulation.
9. State the scope and limitations of statistics.
10. Explain briefly the important types of diagrams.
11. Explain briefly the different types of graphic presentation of frequency distribution.
12. Differentiate between Bar charts and line Graphs.

# UNIT II

## MEASURES OF CENTRAL TENDENCY

Measures of central tendency are a typical value of the entire group or data. It describes the characteristics of the entire mass of data. It reduces the complexity of data and makes them to compare. Human mind is incapable of remembering the entire mass of unwieldy data. So a simple figure is used to describe the series which must be a representative number. It is generally called, "a measure of central tendency or the average".

A central tendency is a central or typical value for a probability distribution. It may also be called a center or location of the distribution. Colloquially, measures of central tendency are often called averages. The term central tendency dates from the late 1920s. If a large volume of data is summarized and given is one simple term. Then it is called as the _Central Value' or an _average'. In other words an average is a single value that represents group of values.

**Characteristics of Ideal Measures:**

A measure of central tendency is a typical value around which other figures congregate. Average condenses a frequency distribution in one figure. According to the statisticians, an average will be termed good or efficient if possesses the following characteristics:

- It should be rigidly defined. It means that the definition should be so clear that the interpretation of the definition does not differ from person to person.
- It should be easy to understand and simple to calculate.
- It should be such that it can be easily determined.
- The average of a variable should be based on all the values of the variable. This means that in the formula for average all the values of the variable should be incorporated.
- The value of average should not change significantly along with the change in sample. This means that the values of the averages of different samples of the same size drawn from the same population should have small variations.

- It should be amenable to algebraic treatment.
- It should be unduly affected by extreme values. i.e, the formula for average should be such that it does not show large due to the presence of one or two very large or very small values of the variable.
- It should be properly defined, preferably by a mathematical formula, so that different individuals working with the same data should get the same answer unless there are mistakes in calculations.
- It should be based on all the observations so that if we change the value of any observation, the value of the average should also be changed.
- It should not be unduly affected by extremely large or extremely small values.
- It should be capable of algebraic manipulation. By this we mean that if we are given the average heights for different groups, then the average should be such that we can find the combined average of all groups taken together.
- It should have quality of sampling stability. That is, it should not be affected by the fluctuations of sampling. For example, if we take ten or twelve samples of twenty students' each and find the average height for each sample, we should get approximately the same average height for each sample.

## 2.1 MEAN

Mean is one of the types of averages. Mean is further divided into three kinds, which are the arithmetic mean, the geometric mean and the harmonic mean. These kinds are explained as follows;

### i) Arithmetic Mean: Simple Arithmetic Average:

### A. Individual Observation: Direct Method:

The arithmetic mean is most commonly used average. It is generally referred as the average or simply mean. The arithmetic mean or simply mean is defined as the value obtained by dividing the sum of values by their number or quantity. It is denoted as $\bar{\mathbf{X}}$(read as X-bar). Therefore, the mean for the values $\mathbf{X_1}$**,** $\mathbf{X_2}$**,** $\mathbf{X_3}$**,………..,** $\mathbf{X_n}$ shall be denoted by $\bar{\mathbf{X}}$Following is the mathematical representation for the formula for the arithmetic mean or simply, the mean.

.

$$\bar{\boldsymbol{X}} = \frac{\mathbf{X1+X2+X3+\cdots+Xn}}{\mathbf{N}} = \frac{\mathbf{\Sigma X}}{\mathbf{N}}$$

*Where,* $\bar{X}$*= Arithmetic Mean; Σx = Sum of all the values of the variables i.e.,* $X_1 + X_2 + X_3 + \ldots + Xn$

*N = Number of observations.*

**Illustration 1:** Calculate mean from the following data:

| **Roll Numbers** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Marks** | 40 | 50 | 55 | 78 | 58 | 60 | 73 | 35 | 43 | 48 |

**Solution:** Calculation of mean

| **Roll Numbers** | **Marks (x)** |
|---|---|
| 1 | 40 |
| 2 | 50 |
| 3 | 55 |
| 4 | 78 |
| 5 | 58 |
| 6 | 60 |
| 7 | 73 |
| 8 | 35 |
| 9 | 43 |
| 10 | 48 |
| N = 10 | $\boldsymbol{\Sigma X}$ = 540 |

$$\bar{\boldsymbol{X}} = \frac{\boldsymbol{\Sigma X}}{\boldsymbol{N}}$$

$$= \frac{540}{10}$$

= 54 marks.

**Short cut method:**

The arithmetic mean can also be calculated by short cut method. This method reduces the amount of calculation. Formula for calculation

$$\bar{X} = \mathbf{A} \pm \frac{\Sigma \mathbf{d}}{\mathbf{N}}$$

*Where,* $\bar{X}$*= Arithmetic Mean; A = Assumed mean;* $\Sigma d$ *= Sum of the deviations; N = Number of items.*

**Illustration 2: (Solving the previous problem)**

| **Roll Numbers** | **Marks (X)** | **d = X - A** |
|---|---|---|
| 1 | 40 | -10 |
| 2 | 50 | 0 |
| 3 | 55 | 5 |
| 4 | 78 | 28 |
| 5 | 58 | 8 |
| 6 | 60 | 10 |
| 7 | 73 | 23 |
| 8 | 35 | -15 |
| 9 | 43 | -7 |
| 10 | 48 | -2 |
| **N = 10** | | $\boldsymbol{\Sigma d}$ **= 40** |

Let the assumed mean, A = 50

$$\bar{X} = \text{A} \pm \frac{\Sigma \text{d}}{\text{N}}$$

$$= 50 + \frac{40}{10}$$

= 54 marks.

### B. Discrete Series: Direct Method:

To find out the total of items in discrete series, frequency of each value is multiplied with the respective size. The values so obtained are totaled up. This total is then divided by the total number of frequencies to obtain the arithmetic mean. The formula is

$$\bar{\mathbf{X}}= \frac{\boldsymbol{\Sigma f x}}{\boldsymbol{N}}$$

*Where,* $\bar{X}$*= Arithmetic Mean;* $\Sigma f x$ *= the sum of products; N = Total frequency.*

**Illustration 3:** Calculate mean from the following data:

| **Value** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Frequency** | 21 | 30 | 28 | 40 | 26 | 34 | 40 | 9 | 15 | 57 |

**Solution:** Calculation of Mean

| ***x*** | **f** | **Fx** |
|---|---|---|
| 1 | 21 | 21 |
| 2 | 30 | 60 |
| 3 | 28 | 84 |
| 4 | 40 | 160 |
| 5 | 26 | 130 |
| 6 | 34 | 204 |
| 7 | 40 | 280 |
| 8 | 9 | 72 |
| 9 | 15 | 135 |
| 10 | 57 | 570 |
| | $\boldsymbol{\Sigma f} = \mathbf{N} = \mathbf{300}$ | **Σfx = 1716** |

$$\bar{X}= \frac{\Sigma \text{fx}}{N} = \frac{1716}{300} = 5.72$$

**Short cut Method: Formula:**

$$\bar{X} = A \pm \frac{\Sigma fd}{N}$$

*Where,* $\bar{X}$*= Arithmetic Mean; A = Assumed mean;* $\Sigma fd$ *= Sum of total deviations; N = Total frequency.*

**Illustration: 4** (Solving the previous problem)

| **X** | **F** | **d = X - A** | **fd** |
|---|---|---|---|
| 1 | 21 | -4 | -84 |
| 2 | 30 | -3 | -90 |
| 3 | 28 | -2 | -56 |
| 4 | 40 | -1 | -40 |
| 5 | 26 | 0 | 0 |
| 6 | 34 | 1 | 34 |
| 7 | 40 | 2 | 80 |
| 8 | 9 | 3 | 27 |
| 9 | 15 | 4 | 60 |
| 10 | 57 | 5 | 285 |
|  | **Σf** = N = 300 |  | **Σfd** = + 216 |

Let the assumed mean, A = 5

$$\bar{X} = A \pm \frac{\Sigma fd}{N}$$

$$\bar{X} = 5 + \frac{216}{300} = 5.72$$

**C. Continuous Series**

In continuous frequency distribution, the value of each individual frequency distribution is unknown. Therefore an assumption is made to make them precise or on the assumption that the frequency of the class intervals is concentrated at the centre that the midpoint of each class interval has to be found out. In continuous frequency distribution, the mean can be calculated by any of the following methods:

1. Direct Method
2. Short cut method
3. Step Deviation Method

**1. Direct Method:** The formula is $\bar{X} = \frac{\Sigma fm}{N}$

*Where,* $\bar{X}$*= Arithmetic Mean;* $\Sigma fm$ *= Sum of the product of f & m; N = Total frequency.*

**Illustration 5:** From the following find out the mean:

| **Class Interval** | 0 – 10 | 10 – 20 | 20 – 30 | 30 – 40 | 40 - 50 |
|---|---|---|---|---|---|
| **Frequency** | 6 | 5 | 8 | 15 | 7 |

**Solution: Calculation of Mean**

| **Class Interval** | **Mid Point (m)** | **Frequency (f)** | **fm** |
|---|---|---|---|
| 0 – 10 | $\frac{0+10}{2} = 5$ | 6 | 30 |
| 10 – 20 | $\frac{10+20}{2} = 15$ | 5 | 75 |
| 20 – 30 | $\frac{20+30}{2} = 25$ | 8 | 200 |
| 30 – 40 | $\frac{30+40}{2} = 35$ | 15 | 525 |
| 40 - 50 | $\frac{40+50}{2} = 45$ | 7 | 315 |
| | | **Σf = N = 41** | **Σfm = 1145** |

$$\bar{X} = \frac{\Sigma fm}{N}$$

$$= \frac{1145}{41} = 27.93$$

**2. Short cut method: Formula:**

$$\bar{\boldsymbol{X}} = \mathrm{A} \pm \frac{\boldsymbol{\Sigma}\mathbf{fd}}{\mathbf{N}}$$

*Where,* $\bar{X}$*= Arithmetic Mean; A = Assumed mean*; $\Sigma fd$ *= Sum of total deviations; N = Total frequency.*

**Illustration: 6** (Solving the previous problem)

| Class Interval | M | d = m - A | F | fd |
|---|---|---|---|---|
| 0 – 10 | $\frac{0+10}{2} = 5$ | 5 – 25 = -20 | 6 | -120 |
| 10 – 20 | $\frac{10+20}{2} = 15$ | 15 – 25 = - 10 | 5 | -50 |
| 20 – 30 | $\frac{20+30}{2} = 25$ | 25 – 25 = 0 | 8 | 0 |
| 30 – 40 | $\frac{30+40}{2} = 35$ | 35 – 25 = 10 | 15 | 150 |
| 40 - 50 | $\frac{40+50}{2} = 45$ | 45 – 25 = 20 | 7 | 140 |
|  |  |  | $\boldsymbol{\Sigma}\mathbf{f}$ = N = 41 | $\boldsymbol{\Sigma}\mathbf{fd}$ = +120 |

d = m – A; here A = 25

$$\bar{\boldsymbol{X}} = \mathbf{A} \pm \frac{\boldsymbol{\Sigma}\mathbf{fd}}{\mathbf{N}}$$

$$= 25 + \frac{120}{41}$$

$$= 25 + 2.93$$

$$= 27.93$$

**3. Step Deviation Method**

**Formula:**

$$\bar{\boldsymbol{X}} = \mathrm{A} \pm \frac{\boldsymbol{\Sigma}\mathbf{fd}'}{\mathbf{N}} \text{ x C}$$

*Where,* $\bar{X}$*= Arithmetic Mean; A = Assumed mean;* $\Sigma fd'$ *= Sum of total deviations;*

*N = Total frequency; C = Common Factor*

**Illustration: 7** (Solving the previous problem)

| Class Interval | Mid Point (m) | Frequency (f) | d = m - A | d' = $\frac{m-A}{C}$ | fd’ |
|---|---|---|---|---|---|
| 0 – 10 | $\frac{0+10}{2} = 5$ | 6 | 5 – 25 = -20 | -2 | -12 |
| 10 – 20 | $\frac{10+20}{2} = 15$ | 5 | 15 – 25 = -10 | -1 | -5 |
| 20 – 30 | $\frac{20+30}{2} = 25$ | 8 | 25 – 25 = 0 | 0 | 0 |
| 30 – 40 | $\frac{30+40}{2} = 35$ | 15 | 35 – 25 = 10 | 1 | 15 |
| 40 - 50 | $\frac{40+50}{2} = 45$ | 7 | 45 – 25 = 20 | 2 | 14 |
| | | **Σf = N = 41** | | | **$\Sigma fd'$ = +12** |

Here A = 25; C = 10

$$\bar{X} = A \pm \frac{\Sigma fd'}{N} \times C$$

$$= 25 + \frac{12}{41} \times 10$$

$$= 25 + \frac{120}{41}$$

$$= 25 + 2.93$$

$$= 27.93$$

## 2.2 MEDIAN

Median is the value of item that goes to divide the series into equal parts. It may be defined as the value of that item which divides the series into equal parts, one half containing values greater that it and the other half containing values less than it. Therefore, the series has to be arranged in ascending or descending order, before finding the median. If the items of a series are arranged in ascending or descending order of magnitude, the item which falls in the middle of it is called median. Hence it is the –middle most‖ or –most central‖ value of a set of number.

**Calculation of Median – Individual Series:**

**Illustration 1:** Find out the median of the following items. X: 10, 15, 9, 25, 19.

**Solution: Computation of Median**

| S. No. | Size of ascending order | Size of descending order |
|---|---|---|
| 1 | 9 | 25 |
| 2 | 10 | 19 |
| 3 | 15 | 15 |
| 4 | 19 | 10 |
| 5 | 25 | 9 |

$$\text{Median} = \text{Size of } \frac{(N+1)}{2}^{\text{th}} \text{ item}$$

$$= \text{Size of } \frac{(5+1)}{2}^{\text{th}} \text{ item}$$

$$= 3^{\text{rd}} \text{ item} = 15.$$

**Illustration 2:** Find out the median of the following items. X: 8, 10, 5, 9, 12, 11.

**Solution: Computation of Median**

| S. No. | X |
|---|---|
| 1 | 5 |
| 2 | 8 |
| 3 | 9 |
| 4 | 10 |
| 5 | 11 |
| 6 | 12 |

$$\text{Median} = \text{Size of } \frac{(N+1)}{2}^{\text{th}} \text{ item}$$

$$= \text{Size of } \frac{(6+1)}{2}^{\text{th}} \text{ item}$$

= Size of 3.5th item

= Size of $\frac{(3\text{rd item} + 4\text{th item})}{2}$

= $\frac{9+10}{2}$ = 9.5

**Calculation of Median – Discrete Series**

**Illustration 3:** Locate median from the following:

| Size of shoes | 5 | 5.5 | 6 | 6.5 | 7 | 7.5 | 8 |
|---|---|---|---|---|---|---|---|
| Frequency | 10 | 16 | 28 | 15 | 30 | 40 | 34 |

**Solution: Computation of Median**

| Size of shoes | F | c.f |
|---|---|---|
| 5 | 10 | 10 |
| 5.5 | 16 | 26 |
| 6 | 28 | 54 |
| 6.5 | 15 | 69 |
| 7 | 30 | 99 |
| 7.5 | 40 | 139 |
| 8 | 34 | 173 |

Median = Size of $\frac{(N+1)^{\text{th}}}{2}$ item

= Size of $\frac{(173+1)^{\text{th}}}{2}$ item

= Size of 87th item

= 7

**Median – Continuous Series**

**Illustration 4:** Calculate the median of the following table:

| Marks | 10 – 25 | 25 - 40 | 40 – 55 | 55 - 70 | 70 – 85 | 85 - 100 |
|---|---|---|---|---|---|---|
| Frequency | 6 | 20 | 44 | 26 | 3 | 1 |

**Solution: Computation of Median**

| *x* | F | c.f |
|---|---|---|
| 10 – 25 | 6 | 6 |
| 25 – 40 | 20 | 26 |
| 40 – 55 | 44 | 70 |
| 55 – 70 | 26 | 96 |
| 70 – 85 | 3 | 99 |
| 85 - 100 | 1 | 100 |

$$\text{Median} = L + \frac{\frac{N}{2} - cf}{f} x\ i$$

$$\frac{N}{2} = \frac{100}{2} = 50;$$

$$L = 40;\ f = 44;\ cf = 26;\ i = 15$$

$$\text{Median} = 40 + \frac{50-26}{44}\ x\ 15$$

$$= 40 + 8.18$$

$$= 48.\ 18 \text{ marks}$$

**Merits:**

1. It is easy to compute and understand.
2. It eliminates the effect of extreme items.
3. The value of median can be located graphically.
4. It is amenable to further algebraic process as it is used in the measurement of dispersion.
5. It can be computed even if the items at the extremes are unknown.

**Demerits:**

1. For calculating median, it is necessary to arrange the data; other averages do not need any arrangement.
2. Typical representative of the observations cannot be computed if the distribution of item is irregular.
3. It is affected more by fluctuation of sampling than the arithmetic mean.

## 2.3 MODE

Mode is the value which occur the greatest number of frequency in a series. It is derived from the French word “La mode” meaning the fashion. It is the most fashionable or typical value of a distribution, because it is repeated the highest number of times in the series.

Mode or the modal value is defined as the value of the variable which occur more number of times or most frequently in a distribution.

**Types of Mode:**

**i) Unimodal:**

If there is only one mode in series, it is called unimodal.

Eg., 10, 15, 20, 25, 18, 12, 15 (Mode is 15)

**ii) Bi – modal:**

If there are two modes in the series, it is called bi - modal.

Eg., 20, 25, 30, 30, 15, 10, 25 (Modes are 25, 30)

**iii) Tri – modal**:

If there are three modes in the series, it is called Tri - modal.

Eg., 60, 40, 85, 30, 85, 45, 80, 80, 55, 50, 60 (Modes are 60, 80, 85)

**iv) Multi – modal:**

If there are more than three modes in the series it is called multi-modal.

**Merits:**

1. It can be easily ascertained without much mathematical calculation.
2. It is not essential to know all the items in a series to compute mode.
3. Open – end classes do not disturb the position of the mode.
4. Its values can be ascertained graphically as well as empirically.

5. It may be very well applied to qualitative as well quantitative data.
6. It is not affected by extreme values as in the average.

**Demerits:**

1. The mode becomes less useful as an average which the distribution is bi-modal.
2. It is not suitable for further mathematical treatment.
3. It is stable only when the sample is large.
4. Mode is influenced by magnitude of the class-intervals.

**Mode - Individual Series**

**Illustration : 1.** Calculate the mode from the following data of the marks obtain by 10 students.

| Serial No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|
| Marks obtained | 60 | 77 | 74 | 62 | 77 | 77 | 70 | 68 | 65 | 80 |

**Solution:**

Marks obtained by 10 students 60, 77, 74, 62, 77, 77, 70, 68, 65, and 80.

Here 77 is repeated three times.

∴The Mode mark is 77.

**DISCRETE SERIES:**

A grouping Table has six columns

**Column 1:** In column 1 rite the actual frequencies and mark the highest frequency.

**Column 2:** Frequencies are grouped in twos, adding frequencies of items 1 and 2; 3 and 4; 5 and 6; and so on.

**Column 3:** Leave the first frequency and then add the remaining in twos.

**Column 4:** Group of frequencies in threes.

**Column 5:** Leave the first frequency and group the remaining in threes.

**Column 6:** Leave the first two frequencies and then group the remaining the threes.

The maximum frequencies in all six columns are marked with a circle and an analysis table is prepared as follows:

1. Put column number on the left – hand side
2. Put the various probable values of mode on the right – hand side.
3. Enter the highest marked frequencies by means of a bar in the relevant box corresponding to the values they represent.

**Illustration: 2.** Calculate the mode from the following:

| **Size** | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|
| **Frequency** | 10 | 12 | 15 | 19 | 20 | 8 | 4 | 3 | 2 |

**Solution: Grouping Table**

| **Size** | **Frequency** | | | | | |
|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** |
| **10** | 10 | | | | | |
| | | 22 | | | | |
| **11** | 12 | | | 37 | | |
| | | | 27 | | | |
| **12** | 15 | | | | 46 | |
| | | 34 | | | | |
| **13** | 19 | | | | | 54 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | 39 | | | |
| **14** | 20 | | | 47 | | |
| | | 28 | | | | |
| **15** | 8 | | | | 32 | |
| | | | 12 | | | |
| **16** | 4 | | | | | 15 |
| | | 7 | | | | |
| **17** | 3 | | | 9 | | |
| | | | 5 | | | |
| **18** | 2 | | | | | |

**Analysis Table**

| **Column No.** | **Size of item containing maximum frequency** | | | | |
|---|---|---|---|---|---|
| | **11** | **12** | **13** | **14** | **15** |
| **1** | | | | 1 | |
| **2** | | 1 | 1 | | |
| **3** | | | 1 | 1 | |
| **4** | | | 1 | 1 | 1 |
| **5** | 1 | 1 | 1 | | |
| **6** | | 1 | 1 | 1 | |
| | 1 | 3 | 5 | 4 | 1 |

The mode is 13, as the size of item repeats 5 times. But through inspection, we say the mode is 14, because the size 14 occurs 20 times. But this wrong decision is revealed by analysis table.

## Calculation of Mode – Continuous Series

$$\mathbf{Z = L_1 + \frac{f1-f0}{2f1-f0-f2} \; x \; i}$$

*Where,*

*Z = Mode; $L_1$ = Lower limit of the modal class; $f_1$ = Frequency of the modal; $f_0$ = Frequency of the class preceding the modal class; $f_2$ = Frequency of the class succeeding the modal class; i = Class interval;*

**Illustration: 3.**

Calculate the mode from the following:

| Size of item | Frequency |
|---|---|
| 0 – 5 | 20 |
| 5 – 10 | 24 |
| 10 – 15 | 32 |
| 15 – 20 | 28 |
| 20 – 25 | 20 |
| 25 – 30 | 16 |
| 30 – 35 | 34 |
| 35 – 40 | 10 |
| 40 – 45 | 8 |

**Solution:**

**Grouping Table**

| Size of item | Frequency | | | | | |
|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** |
| **0 – 5** | 20 | | | | | |
| | | 44 | | 76 | | |
| **5 – 10** | 24 | | | | | |
| | | | 56 | | | |
| **10 – 15** | 32 | | | | 84 | |
| | | 60 | | | | 80 |
| **15 – 20** | 28 | | | | | |
| | | | 48 | | | |
| **20 – 25** | 20 | | | 64 | | |
| | | 36 | | | | |
| **25 – 30** | 16 | | | | 70 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | 50 | | | |
| **30 – 35** | 34 | | | | | 60 |
| | | 44 | | | | |
| **35 – 40** | 10 | | | 52 | | |
| | | | 18 | | | |
| **40 - 45** | 8 | | | | | |

**Analysis Table**

| **Column No.** | **Size of item containing maximum frequency** | | | | | |
|---|---|---|---|---|---|---|
| | **0 – 5** | **5 – 10** | **10 – 15** | **15 – 20** | **20 - 25** | **30 - 35** |
| **1** | | | | | | 1 |
| **2** | | | 1 | 1 | | |
| **3** | | 1 | 1 | | | |
| **4** | 1 | 1 | 1 | | | |
| **5** | | 1 | 1 | 1 | | |
| **6** | | | 1 | 1 | 1 | |
| | 1 | 3 | **5** | 3 | 1 | 1 |

$$Z = L_1 + \frac{\mathbf{f1-f0}}{\mathbf{2f1-f0-f2}} \text{ x i}$$

$L_1 = 10$; $f_1 = 32$; $f_0 = 24$; $f_2 = 28$; $i = 5$

$$Z = 10 + \frac{32-24}{2\,x\,32-24-28} \text{ x } 5$$

$$= 10 + \frac{40}{12} = 10 + 3.\ 33$$

∴ The Mode is 13. 33

**The relationship among mean, median and mode:**

The three averages (mean, median and mode) are identical, when the distribution is symmetrical. In an asymmetrical distribution, the values of mean, median and mode are not equal. In a moderately asymmetrical distribution the distance between the mean and median is about one-third of the distance between mean and mode.

Mean - Median = 1/3 (Mean - Mode)

Mode = 3 Median - 2 Mean

Median = Mode + 2/3 (Mean - Mode)

## 2.4 GEOMATRIC MEAN

The geometric mean, G, of a set of *n* positive values $X_1$, $X_2$, ……,$X_n$ is the N$^{th}$ root ofthe product of N items. Mathematically the formula for geometric mean will be as follows;

$$\text{G. M} = \sqrt[n]{\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n}} = (\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots \ldots, \boldsymbol{X_n}^{1/n}$$

G.M = Geometric Mean; n = number of items; $X_1$, $X_2$, $X_3$,….. = are various values

**Illustration1:** The geometric mean of the values 2, 4 and 8 is the cubic root of 2 x 4 x 8 or

$$\sqrt[3]{2x4x8} = \sqrt[3]{64} = 4$$

In practice, it is difficult to extract higher roots. The geometric mean is, therefore, computed using logarithms. Mathematically, it will be represented as follows;

Geometric Mean = Antilog of $\frac{\log X1+\log X2+\log X3 \ldots\log Xn}{N}$ (or) G. M. = Antilog of $\frac{\log X}{N}$

Here we assume that all the values are positive, otherwise the logarithms will be not defined.

## Geometric Mean – Individual Series

**Illustration 2:** Calculate the geometric mean of the following:

50 72 54 82 93

**Solution: Calculation of Geometric Mean**

| X | log of X |
|---|---|
| 50 | 1.6990 |
| 72 | 1.8573 |
| 54 | 1.7324 |
| 82 | 1.9138 |
| 93 | 1.9685 |
| **N = 5** | $\boldsymbol{\sum \log X = 9.1710}$ |

G.M. = $\surd 50 \ \times 72 \times 54 \times 82 \times 93$ or

G. M. = Antilog of $\frac{\log \text{X}}{\text{N}}$

G.M = Antilog of $\frac{9.1710}{5}$

= Antilog of 1.8342

= 68.26

## Geometric Mean – Discrete Series

**G.M. = antilog of $\boldsymbol{\frac{\sum f log X}{N}}$**

Where, f = frequency value; log x = logarithm of each value; N = Total frequencies

**Illustration 3:** The following table gives the weight of 31 persons in a sample survey. Calculate geometric mean.

| Weight (lbs) | 130 | 135 | 140 | 145 | 146 | 148 | 149 | 150 | 157 |
|---|---|---|---|---|---|---|---|---|---|
| No. of persons | 3 | 4 | 6 | 6 | 3 | 5 | 2 | 1 | 1 |

Solution: Calculation of Geometric Mean

| Size of item (X) | Frequency (f) | log X | f log X |
|---|---|---|---|
| 130 | 3 | 2.1139 | 6.3417 |
| 135 | 4 | 2.1303 | 8.5212 |
| 140 | 6 | 2.1461 | 12.8766 |
| 145 | 6 | 2.1614 | 12.9684 |
| 146 | 3 | 2.1644 | 6.4932 |
| 148 | 5 | 2.1703 | 10.8515 |
| 149 | 2 | 2.1732 | 4.3464 |
| 150 | 1 | 2.1761 | 2.1761 |
| 157 | 1 | 2.1959 | 2.1959 |
|  | N = $\sum \mathbf{f} = \mathbf{31}$ |  | $\sum \mathbf{flogX} = \mathbf{66.7710}$ |

$$\text{G.M.} = \text{antilog of } \frac{\sum \boldsymbol{f} \boldsymbol{log} \boldsymbol{X}}{\boldsymbol{N}}$$

$$\text{G.M.} = \text{antilog of } \frac{66.7710}{31} = \text{antilog of } 2.1539$$

G.M. Weight = 142.5 lbs

**Geometric Mean – Continuous Series**

$$\textbf{G.M. = antilog of } \frac{\sum \mathbf{f} \textbf{ log m}}{\mathbf{N}}$$

Where, f = frequency; m = mid value; log m = logarithm of each mid value; N = Total frequencies

**Illustration 4:** Find out the geometric mean:

| Yield of wheat (mounds) | No. of farms |
|---|---|
| 7.5 – 10.5 | 5 |
| 10.5 – 13.5 | 9 |
| 13.5 – 16.5 | 19 |
| 16.5 – 19.5 | 23 |
| 19.5 – 22.5 | 7 |
| 22.5 – 25.5 | 4 |
| 25.5 – 28.5 | 1 |

Solution: Calculation of Geometric Mean

| Yield of wheat (mounds) | No. of farms (f) | m | log m | f log m |
|---|---|---|---|---|
| 7.5 – 10.5 | 5 | 9 | 0.9542 | 4.7710 |
| 10.5 – 13.5 | 9 | 12 | 1.0792 | 9.7128 |
| 13.5 – 16.5 | 19 | 15 | 1.1761 | 22.3459 |
| 16.5 – 19.5 | 23 | 18 | 1.2553 | 28.8719 |
| 19.5 – 22.5 | 7 | 21 | 1.3222 | 9.2554 |
| 22.5 – 25.5 | 4 | 24 | 1.3802 | 5.5208 |
| 25.5 – 28.5 | 1 | 27 | 1.4314 | 1.4314 |
|  | N = $\sum \mathbf{f = 68}$ |  |  | $\sum \mathbf{flogm\ =\ 81.9092}$ |

G.M. = antilog ~~of~~ $\frac{\sum \boldsymbol{flogm}}{\boldsymbol{N}}$

G.M. = antilog $\frac{81.9092}{68}$ = antilog of 1.2045

= 16.02 maunds

**Uses:**

- Geometric mean is highly useful in averaging ratios, percentages and rate of increase between two periods.

- Geometric mean is important in the construction of index numbers.

- In economic and social sciences, where we want to give more weight to smaller items and smaller weight to large items, geometric mean is appropriate.

- It is the only useful average that can be employed to indicate rate of change.

### Merits:

- Every item in the distribution is included in the calculation.
- It can be calculated with mathematical exactness, provided that all the quantities aregreater than zero and positive.
- Large items have less effect on it than the arithmetic average.
- It is amenable to further algebraic manipulation.

### Demerits:

- It is very difficult to calculate.
- It is impossible to use it when any item is zero or negative.
- The value of the geometric mean may not correspond with any actual value in thedistribution.
- If cannot be used in the series in which the end values of the classes are left open.

## 2.5 HARMONIC MEAN

Harmonic Mean, like geometric mean is a measure of central tendency in solving special types of problems. Harmonic Mean is the reciprocal of the arithmetic average of the reciprocal ofvalues of various items in the variable. The reciprocal of a number is that value, which is obtained dividing one by the value.

**For example**, the reciprocal of 5 is 1/5. The reciprocal can be obtained from logarithm tables.

Harmonic Mean – Individual Series

$$\textbf{H.M.} = \frac{N}{\frac{1}{X1}+\frac{1}{X2}+\frac{1}{X3}+\cdots\frac{1}{Xn}} \text{ or } \textbf{H.M.} = \frac{N}{\Sigma\frac{1}{X}}$$

$X_1$, $X_2$, $X_3$ .........$X_n$, refer to the various in the observations

**Illustration 5:** The monthly incomes of 10 families in rupees are given below:

| **Family:** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Income:** | 85 | 70 | 10 | 75 | 500 | $^{-}8$ | 42 | 250 | 40 | 36 |

Solution: Calculation of Harmonic Mean

| **Family** | **Income (x)** | **Reciprocals**$(\frac{1}{X})$ |
|---|---|---|
| 1 | 85 | 1/85 = 0. 0118 |
| 2 | 70 | 1/70 = 0. 0143 |
| 3 | 10 | 1/10 = 0. 1000 |
| 4 | 75 | 1/75 = 0. 0133 |
| 5 | 500 | 1/500 = 0. 0020 |
| 6 | 8 | 1/8 = 0. 1250 |
| 7 | 42 | 1/42 = 0. 0232 |
| 8 | 250 | 1/250 = 0. 0040 |
| 9 | 40 | 1/40 = 0. 0250 |
| 10 | 36 | 1/36 = 0. 0278 |
| **N =10** | | $\boldsymbol{\sum\frac{1}{X} = 0.3464}$ |

$$\text{H.M.} = \frac{N}{\sum\frac{1}{X}} = \frac{10}{0.3464}$$

= Rs. 28.87/-

**Harmonic Mean – Discrete Series**

$$\textbf{H.M.} = \frac{\mathbf{N}}{\sum \mathbf{f}(\frac{\mathbf{1}}{\mathbf{X}})}$$

**Illustration 6:** Calculate harmonic mean from the following data.

| **Size of items:** | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|
| **Frequency:** | 4 | 6 | 9 | 5 | 2 | 8 |

Solution: Calculation of harmonic mean

| **Size of items (x)** | **Frequency (f)** | **Reciprocal** $(\frac{1}{X})$ | $\boldsymbol{f\,(\frac{1}{X})}$ |
|---|---|---|---|
| 6 | 4 | 0.1667 | 0.6668 |
| 7 | 6 | 0.1429 | 0.8574 |
| 8 | 9 | 0.1250 | 1.1250 |
| 9 | 5 | 0.1111 | 0.5555 |
| 10 | 2 | 0.1000 | 0.2000 |
| 11 | 8 | 0.0909 | 0.7272 |
|  | **N =** $\boldsymbol{\sum f = 34}$ |  | $\boldsymbol{\sum f(\frac{1}{X}) = 4.1319}$ |

$$\text{H.M.} = \frac{N}{\sum f(\frac{1}{X})} = \frac{34}{4.1319} = 8.23$$

## Harmonic Mean - Continuous Series

$$\text{H.M.} = \frac{\mathbf{N}}{\sum \mathbf{f}\left(\frac{\mathbf{1}}{\mathbf{m}}\right)}$$

**Illustration 7:** Calculate H.M. of the following data:

| **Size:** | 0 – 10 | 10 – 20 | 20 – 30 | 30 – 40 | 40 – 50 |
|---|---|---|---|---|---|
| **Frequency:** | 5 | 8 | 12 | 6 | 4 |

| **Size (x)** | **Frequency (f)** | **Mid Value (m)** | $\left(\frac{\mathbf{1}}{\mathbf{m}}\right)$ | $\mathbf{f}\left(\frac{\mathbf{1}}{\mathbf{m}}\right)$ |
|---|---|---|---|---|
| 0 – 10 | 5 | 5 | 1/5 = 0.2000 | 1.0000 |
| 10 – 20 | 8 | 15 | 1/15 = 0.0667 | 0.5336 |
| 20 – 30 | 12 | 25 | 1/25 = 0.0400 | 0.4800 |
| 30 – 40 | 6 | 35 | 1/35 = 0.0286 | 0.1716 |
| 40 - 50 | 4 | 45 | 1/45 = 0.0222 | 0.0888 |
| | **N = ∑f = 35** | | | $\sum \mathbf{f}\left(\frac{\mathbf{1}}{\mathbf{m}}\right) = \mathbf{2.274}$ |

$$\text{H.M.} = \frac{\mathbf{N}}{\sum \mathbf{f}\left(\frac{\mathbf{1}}{\mathbf{m}}\right)} = \frac{35}{2.274} = 15.39$$

## Important Questions

**Choose the correct answer:**

1. Which average is affected most by extreme observations?

(a) Mode (b) Median **(C) Geometric Mean** (d) Arithmetic mean

2. Which of the following is the most unstable average?

(a) **Mode** (b) Median (c) Geometric mean (d) Harmonic mean

3. For dealing with qualitative data the best average is:

(a) Arithmetic mean (b) Geometric mean (c) Harmonic mean (d) **Median**

4. The sum of deviations taken from arithmetic mean is:

(a) Minimum (b) **Zero** (c) Maximum (d) Equal

5. The sum of squares of deviations from arithmetic mean is:

(a) Zero (b) Maximum (c) **Minimum** (d) Equal

6. When calculating the average growth of economy, the correct mean to use is?

(a) Weighted mean (b) **Geometric Mean** (c) Arithmetic mean (d) Harmonic Mean

7. When an observation in the data is zero, then its geometric mean is?

(a) Negative **(b) Zero** (c) Positive (d) Cannot be calculated.

8. The best measure of central tendency is:

**(a) Arithmetic Mean** (b) Geometric mean (c) Harmonic mean (d) Weighted mean

9. The point of intersection of the 'less than 'and 'more than ' ogives corresponds to:

(a) Mean **(b) Median** (c) Geometric mean (d) Mode

**Exercises:**

1. The monthly income of 10 families of a certain locality is given in rupees as below. Calculate the arithmetic average.

| Families | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Income in rupees | 85 | 70 | 10 | 75 | 500 | 8 | 42 | 250 | 40 | 36 |

(Mean = Rs. 111. 60)

2. The coins are tossed 1024 times. The theoretical frequencies of 10 heads to 0 head are given below. Calculate the mean number of heads per tossing.

| No. of heads | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency | 1 | 10 | 45 | 120 | 210 | 252 | 210 | 120 | 45 | 10 | 1 |

(Mean = Rs. 5)

3. Find mean from the following frequency distribution:

| Class Interval | 15 – 25 | 25 – 35 | 35 – 45 | 45 – 55 | 55 – 65 | 65 - 75 |
|---|---|---|---|---|---|---|
| Frequency | 4 | 11 | 19 | 14 | 0 | 2 |

(Mean = Rs. 40. 2)

4. The following are the marks scored by 7 students; find out the median marks:

| Roll Numbers | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Marks | 45 | 32 | 18 | 57 | 65 | 28 | 46 |

(Median marks = 45)

5. Find out the median from the following:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 57 | 58 | 61 | 42 | 38 | 65 | 72 | 66 |

(Median = 59.5)

6. Find the median

| *X* | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F | 37 | 162 | 343 | 390 | 256 | 433 | 161 | 355 | 65 | 85 | 49 | 46 | 40 |

(Median = 18)

7. Find the median:

| Wages Rs. | 60 – 70 | 50 – 60 | 40 – 50 | 30 – 40 | 20 - 30 |
|---|---|---|---|---|---|
| No. of labourers | 5 | 10 | 20 | 5 | 3 |

(Median = 46.75)

8. 10 persons have the following income:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Rs. | 850 | 750 | 600 | 825 | 850 | 725 | 600 | 850 | 640 | 530 |

(Mode = 850)

9. Calculate the mode from the following series:

| Size | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency | 40 | 48 | 52 | 57 | 60 | 63 | 57 | 55 | 50 | 52 | 41 | 57 | 63 | 52 | 48 | 40 |

(Mode = 9)

10. Find the mode:

| Size | 0 – 10 | 10 – 20 | 20 – 30 | 30 – 40 | 40 – 50 | 50 – 60 | 60 – 70 |
|---|---|---|---|---|---|---|---|
| Frequency | 5 | 7 | 12 | 18 | 16 | 10 | 5 |

(Mode = 37.5)

11. Calculate mean, median and mode form the following frequency distribution of marks at a test in statistics:

| Marks | 5 | 10 | 15 | 20 | 25 | 30 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|
| No. of students | 20 | 43 | 75 | 76 | 72 | 45 | 9 | 8 | 50 |

(Mean = 22. 16; Median = 20; mode = 20)

12. Calculate the mean, median and mode for the following data.

| Profits per shop | 0 – 10 | 10 – 20 | 20 – 30 | 30 – 40 | 40 – 50 | 50 - 60 |
|---|---|---|---|---|---|---|
| No. of shops | 12 | 18 | 27 | 20 | 17 | 6 |

(Mean = 28; Median = 27.4; mode = 25.62)

# UNIT III

## DISPERSION (Measures of Variation)

Dispersion is studied to have an idea of the homogeneity or heterogeneity of the distribution. Measures of dispersion are the measures of scatter or spread about an average. Measures of dispersion are called the averages of the second order.

**Methods of Measuring Dispersion:**

There are various methods of studying variation or dispersion important methods studying dispersion are as follows:

1. Range
2. Inter - quartile range
3. Mean Deviation
4. Standard Deviation
5. Lorenz curve

### 1. Range

Range is the simplest and crudest measure of dispersion. It is a rough measure of dispersion. It is the difference between the highest and the lowest value in the distribution.

**Range = L – S**

Where, L = Largest Value; S = Smallest Value.

The Relative measure of range is called as the Co – efficient of Range.

**Co – efficient of Range =** $\frac{\mathbf{L-S}}{\mathbf{L+S}}$

**Illustration 1:**

Find the range of weights of 7 students from the following.

27, 30, 35, 36, 38, 40, 43

**Solution:**

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

$$\text{Here } L = 43;\ S = 27$$

$$\therefore \text{Range} = 43 - 27 = 16$$

$$\text{Co – efficient of Range} = \frac{\mathbf{L-S}}{\mathbf{L+S}}$$

$$= \frac{43-27}{43+27} = \frac{16}{70} = 0.23$$

**Practical utility of Range**

1. It is used in industries for the statistical quality control of the manufactured product.
2. It is used to study the variations such as stock, shares and other commodities.
3. It facilitates the use of other statistical measures.

**Advantages**

1. It is the simplest method
2. It is easy to understand and the easiest to compute.
3. It takes minimum time to calculate and accurate.

**Disadvantages**

1. Range is completely dependent on the two extreme values.
2. It is subject to fluctuations of considerable magnitude from sample to sample.
3. Range cannot tell us anything about the character of the distribution.

## 3.1 Quartile Deviation (Q.D)

Quartile deviation is an absolute measure of dispersion. Co-efficient of quartile deviation is known as relative measure of dispersion.

In the series, four quartiles are there. By eliminating the lowest items (25%) and the highest items (25%) of a series we can obtain a measure of dispersion and can find out the half of the distance between the first and the third quartiles. That is, [Q3 (third quartiles) – Q1 (first quartiles). The inter-quartile range is reduced to the form of the semi – inter quartile range (or) quartile deviation by dividing it by 2.

$$\text{Inter quartile range} = Q_3 - Q_1$$

$$\text{Inter quartile range or Quartile deviation} = \frac{\mathbf{Q3-Q1}}{\mathbf{2}}$$

$$\text{Coefficient of Quartile deviation} = \frac{\mathbf{Q3-Q1}}{\mathbf{Q3+Q1}}$$

**Quartile Deviation – Individual Series**

**Illustration 2:** Find out the value of Quartile Deviation and its coefficient from the following data:

| Roll No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Marks | 20 | 28 | 40 | 30 | 50 | 60 | 52 |

**Solution: Calculation of Q.D.**

| Marks arranged in ascending order: | 20 | 28 | 30 | 40 | 50 | 52 | 60 |
|---|---|---|---|---|---|---|---|

$$Q_1 = \text{Size of } \frac{\boldsymbol{N+1}}{\mathbf{4}}^{\text{th}} \text{ item}$$

$$Q_1 = \text{size of } \frac{7+1}{4}^{\text{th}} \text{ item}$$

$$= \text{size of } \frac{8}{4}^{\text{ th}} \text{ item}$$

$$= \text{size of } 2^{\text{nd}} \text{ item}$$

$$= 28$$

$$Q_3 = \text{Size of } 3\left(\frac{\boldsymbol{N+1}}{\mathbf{4}}\right)^{\text{th}} \text{ item}$$

$Q_3$ = Size of $3(\frac{7+1}{4})^{th}$ item

= Size of $3(\frac{8}{4})^{th}$ item

= size of $\frac{24}{4}$ th item

= size of 6th item

= 52

Q. D. = $\frac{\mathbf{Q3-Q1}}{\mathbf{2}}$

= $\frac{52-28}{2}$

= $\frac{24}{2}$

= 12

Coefficient of Q.D = $\frac{\mathbf{Q3-Q1}}{\mathbf{Q3+Q1}}$

= $\frac{52-28}{52+28}$

= $\frac{24}{80}$

= 0.3

**Quartile Deviation – Discrete Series**

**Illustration 3:** Find out the value of Quartile Deviation and its coefficient from the following data:

| **Age in years** | 20 | 30 | 40 | 50 | 60 | 70 | 80 |
|---|---|---|---|---|---|---|---|
| **No. of members** | 3 | 61 | 132 | 153 | 140 | 51 | 3 |

**Solution:**

**Calculation of Q.D.**

| x | F | c.f. |
|---|---|---|
| 20 | 3 | 3 |
| 30 | 61 | 64 |
| 40 | 132 | 196 |
| 50 | 153 | 349 |
| 60 | 140 | 489 |
| 70 | 51 | 540 |
| 80 | 3 | 543 |

$$Q_1 = \text{Value of } \frac{\boldsymbol{N+1}}{\mathbf{4}}^{\text{th}} \text{ item}$$

$$Q_1 = \text{value of } \frac{543+1}{4}^{\text{th}} \text{ item} = \text{value of } \frac{544}{4}^{\text{th}} \text{ item}$$

$$= \text{value of } 136^{\text{th}} \text{ item} = 40 \text{ years}$$

$$Q_3 = \text{Value of } 3\left(\frac{\boldsymbol{N+1}}{\mathbf{4}}\right)^{\text{th}} \text{ item}$$

$$Q_3 = \text{value of } 3\left(\frac{543+1}{4}\right)^{\text{th}} \text{ item} = \text{value of } 3\left(\frac{544}{4}\right)^{\text{th}} \text{ item}$$

$$= \text{value of } 3(136)^{\text{th}} \text{ item} = \text{value of } 408^{\text{th}} \text{ item} = 60 \text{ years}$$

$$\text{Q. D.} = \frac{\mathbf{Q3-Q1}}{\mathbf{2}} = \frac{60-40}{2} = \frac{20}{2} = 10 \text{ years}$$

$$\text{Coefficient of Q.D} = \frac{\mathbf{Q3-Q1}}{\mathbf{Q3+Q1}}$$

$$= \frac{60-40}{60+40} = \frac{20}{100} = 0.2$$

**Quartile Deviation – Continuous Series**

**Illustration 4:** Find out the value of Quartile Deviation and its coefficient from the following data:

| **Wages (Rs.)** | 30 – 32 | 32 – 34 | 34 - 36 | 36 - 38 | 38 - 40 | 40 - 42 | 42 - 44 |
|---|---|---|---|---|---|---|---|
| **Labourers** | 12 | 18 | 16 | 14 | 12 | 8 | 6 |

**Solution: Calculation of Q.D.**

| Wages (x) | Labourers (f) | c.f. |
|---|---|---|
| 30 – 32 | 12 | 12 |
| 32 – 34 | 18 | 30 |
| 34 – 36 | 16 | 46 |
| 36 – 38 | 14 | 60 |
| 38 – 40 | 12 | 72 |
| 40 – 42 | 8 | 80 |
| 42 – 44 | 6 | 86 |

$Q_1$ = size of $\frac{N}{4}^{th}$ item

= size of $\frac{86}{4}^{th}$ item

= 21.5th item

∴ $Q_1$ lies in the group 32 - 34

$$Q_1 = L + \frac{\frac{N}{4} - cf}{f} \text{ x i}$$

$$= 32 + \frac{21.5-12}{18} \text{ x } 2 = 32 + \frac{9.5}{18} \text{ x } 2$$

$$Q_1 = 32 + \frac{19}{18} = 32 + 1.06 = 33.06$$

$Q_3$ = size of $\frac{3N}{4}^{th}$ item

= size of $\frac{3 X 86}{4}^{th}$ item

= 64.5 th item

∴ $Q_3$ lies in the group 38 – 40.

$$Q_3 = L + \frac{\frac{3N}{4} - cf}{f} \text{x i}$$

$$= 38 + \frac{64.5-60}{12} \text{ x } 2 = 38 + \frac{4.5}{12} \text{ x } 2$$

$$Q_3 = 38 + \frac{9}{12} = 32 + 0.75 = 38.75$$

$$\text{Q. D.} = \frac{\mathbf{Q3-Q1}}{\mathbf{2}} = \frac{38.75-33.06}{2} = \frac{5.69}{2} = 2.85$$

$$\text{Coefficient of Q.D} = \frac{\mathbf{Q3-Q1}}{\mathbf{Q3+Q1}} = \frac{38.75-33.06}{38.75+33.06} = \frac{5.69}{71.81} = 0.08$$

**Merits:**

**1.** It is simple to calculate.

**2.** It is easy to understand.

**3.** Risk of excrement item variation is eliminated, as it depends upon the central 50 percent items.

**Demerits**

**1.** Items below Q1 and above Q3 are ignored.

**2.** It is not capable of further mathematical treatment.

**3.** It is affected much by the fluctuations of sampling.

**4.** It is not calculated from a computed average, but from a positional average.

## 3.2 Mean Deviation

The mean deviation is also known as the average deviation. It is the average difference between the items in a distribution computed from the mean, median or mode of that series counting all such deviation as positive. Median is preferred to the average because the sum of deviation of items from median is minimum when signs are ignored. But, the arithmetic mean is more frequently used in calculating the value of average deviation. Hence, it is commonly called Mean deviation.

**Mean Deviation – Individual Series**

$$\text{M. D. (mean or median or mode)} = \frac{\Sigma|\boldsymbol{D}|}{\boldsymbol{N}}$$

Coefficient of Mean Deviation: $\frac{\textbf{Mean Deviation}}{\textbf{Mean or median or mode}}$

**Illustration 5:** Calculate mean deviation from mean and median for the following data:

| 100 | 150 | 200 | 250 | 360 | 490 | 500 | 600 | 671 |
|---|---|---|---|---|---|---|---|---|

**Solution: Calculation of Mean Deviation**

| **X** | $\lvert \mathbf{D} \rvert$ **=** $\mathbf{X - \bar{X}}$<br>**X – 369** | $\lvert \mathbf{D} \rvert$ **= X – median;**<br>**X - 360** |
|---|---|---|
| 100 | 269 | 260 |
| 150 | 219 | 210 |
| 200 | 169 | 160 |
| 250 | 119 | 110 |
| 360 | 9 | 0 |
| 490 | 121 | 130 |
| 500 | 131 | 140 |
| 600 | 231 | 240 |
| 671 | 302 | 311 |
| $\mathbf{\sum X = 3321}$ | $\mathbf{\sum \lvert D \rvert = 1570}$ | $\mathbf{\sum \lvert D \rvert = 1561}$ |

Mean $\bar{X} = \frac{\Sigma \mathbf{X}}{\mathbf{N}}$

$= \frac{3321}{9} = 369$

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

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

= Value of 5th item = 360

M.D. from mean = $\frac{\Sigma \lvert \boldsymbol{D} \rvert}{\boldsymbol{N}}$

$= \frac{1570}{9} = 174.44$

M.D. from median = $\frac{\Sigma \lvert \boldsymbol{D} \rvert}{\boldsymbol{N}}$

$= \frac{1561}{9} = 173.44$

Coefficient of M.D. = $\frac{\mathbf{M.D.}}{\bar{\mathbf{X}}}$

$= \frac{174.44}{369} = 0.47$

Coefficient of M.D. = $\frac{\mathbf{M.D.}}{\mathbf{Median}}$

$= \frac{173.44}{360} = 0.48$

**Mean Deviation – Discrete Series**

$$\textbf{M. D.} = \frac{\sum f\,|D|}{N}$$

**Illustration 6:** Calculate mean deviation from mean from the following data:

| *X* | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|
| F | 1 | 4 | 6 | 4 | 1 |

**Solution: Calculation of Mean Deviation**

| ***x*** | **F** | **f*x*** | $\|D\| = x - \bar{X}$ | $f\|D\|$ |
|---|---|---|---|---|
| 2 | 1 | 2 | 4 | 4 |
| 4 | 4 | 16 | 2 | 8 |
| 6 | 6 | 36 | 0 | 0 |
| 8 | 4 | 32 | 2 | 8 |
| 10 | 1 | 10 | 4 | 4 |
| | **N = ∑f = 16** | **∑fx = 96** | | $\sum f\|D\| = 24$ |

$$\text{Mean } \bar{\mathbf{X}} = \frac{\sum \mathbf{fX}}{\mathbf{N}} = \frac{96}{16} = 6$$

$$\text{M.D. from mean} = \frac{\sum f\,|D|}{N} = \frac{24}{16} = 1.5$$

$$\text{Coefficient of M.D.} = \frac{\mathbf{M.D.}}{\bar{\mathbf{X}}} = \frac{1.5}{6} = 0.25$$

**Mean Deviation – Continuous Series**

$$\textbf{M. D.} = \frac{\sum f\,|D|}{N}$$

**Illustration 7:**

Calculate mean deviation from mean from the following data:

| **Class interval** | 2 - 4 | 4 - 6 | 6 - 8 | 8 - 10 |
|---|---|---|---|---|
| **Frequency** | 3 | 4 | 2 | 1 |

**Solution:**

**Calculation of Mean Deviation**

| $x$ | **M** | **f** | **Fm** | $\left\|D\right\| = m - \bar{X}$ | $f\left\|D\right\|$ |
|---|---|---|---|---|---|
| 2 – 4 | 3 | 3 | 9 | 2.2 | 6.6 |
| 4 – 6 | 5 | 4 | 20 | 0.2 | 0.8 |
| 6 – 8 | 7 | 2 | 14 | 1.8 | 3.6 |
| 8 - 10 | 9 | 1 | 9 | 3.8 | 3.8 |
| | | **N = ∑f = 10** | **∑fm = 52** | | $\sum f\left\|D\right\| = 14.8$ |

$$\text{Mean } \bar{X} = \frac{\sum \text{fm}}{\text{N}} = \frac{52}{10} = 5.2$$

$$\text{M.D. from mean} = \frac{\sum f\left|D\right|}{N} = \frac{14.8}{10} = 1.48$$

$$\text{Coefficient of M.D.} = \frac{\text{M.D.}}{\bar{X}} = \frac{1.48}{5.2} = 0.29$$

**Merits**

1. It is clear and easy to understand.
2. It is based on each and every item of the data.
3. It can be calculated from any measure of central tendency and as such is flexible too.
4. It is not disturbed by the values of extreme items as in the case of range.

**Demerits:**

1. It is not suitable for further mathematical processing.
2. It is rarely used in sociological studies.

## 3.3 Standard Deviation

Karl Pearson introduced the concept of Standard deviation in 1893. Standard deviation is the square root of the means of the squared deviation from the arithmetic mean. So, it is called as Root - Mean Square Deviation or Mean Error or Mean Square Error. The Standard deviation is denoted by the small Greek letter ‗$\sigma$‘ (read as sigma)

**Standard Deviation – Individual Observation**

**Deviation taken from Actual Mean**

$$\sigma = \sqrt{\frac{\sum x^2}{N}} \text{ or } \sigma = \sqrt{\frac{\sum (X-\bar{X})^2}{N}} \text{ or } \sigma = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2}$$

**Illustration 8:** Calculate the standard deviation from the following data;

14, 22, 9, 15, 20, 17, 12, 11

**Solution: Calculation of standard deviation from actual mean**

| **Values (X)** | **$X^2$** | **$X - \bar{X}$; (X - 15)** | **$(X - \bar{X})^2$** |
|---|---|---|---|
| 14 | 196 | -1 | 1 |
| 22 | 484 | 7 | 49 |
| 9 | 81 | -6 | 36 |
| 15 | 225 | 0 | 0 |
| 20 | 400 | 5 | 25 |
| 17 | 289 | 2 | 4 |
| 12 | 144 | -3 | 9 |
| 11 | 121 | -4 | 16 |
| **$\sum X = 120$** | **$\sum X^2 = 1940$** | | **$\sum (X - \bar{X})^2 = 140$** |

$$\mathbf{N = 8;\ \bar{X} = \frac{\sum X}{N} = \frac{120}{8} = 15}$$

$$\sigma = \sqrt{\frac{\sum x^2}{N}} \text{ or } \sigma = \sqrt{\frac{\sum(X-\bar{X})^2}{N}}$$

$$= \sqrt{\frac{140}{8}}$$

$$= \sqrt{17.5}$$

$$= 4.18$$

**Alternatively:**

We can find out standard deviation by using variables directly, i.e., no deviation is found out.

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

$$= \sqrt{\frac{1940}{8} - \left(\frac{120}{8}\right)^2}$$

$$= \sqrt{242.5 - 225}$$

$$= \sqrt{17.5}$$

$$= \mathbf{4.18}$$

**Deviation taken from Assumed Mean**

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

Where d = X – A

**Illustration 9:**

Calculate the standard deviation from the following data;

30, 43, 45, 55, 68, 69, 75.

**Solution:**

**Calculation of standard deviation from assumed mean**

| X | d = X – A= X - 55 | $d^2$ |
|---|---|---|
| 30 | -25 | 625 |
| 43 | -12 | 144 |
| 45 | -10 | 100 |
| 55 | 0 | 0 |
| 68 | 13 | 169 |
| 69 | 14 | 196 |
| 75 | 20 | 400 |
| **N = 7** | **$\sum d = 0$** | **$\sum d^2 = 1634$** |

$$\sigma = \sqrt{\frac{\sum \mathbf{d^2}}{\mathbf{N}} - (\frac{\sum \mathbf{d}}{\mathbf{N}})^2}$$

$$= \sqrt{\frac{1634}{7} - (\frac{0}{7})^2}$$

$$= \sqrt{233.429}$$

$$= 15.28$$

**Standard Deviation – Discrete Series: Actual Mean Method:**

$$\sigma = \sqrt{\frac{\sum \mathbf{fd^2}}{\mathbf{N}}}$$

**Illustration 10:**

Calculate the standard deviation from the following data;

| **Marks** | 10 | 20 | 30 | 40 | 50 | 60 |
|---|---|---|---|---|---|---|
| **No. of students** | 8 | 12 | 20 | 10 | 7 | 3 |

**Solution:**

**Calculation of standard deviation (from actual mean)**

| x | F | Fx | d = x -$\bar{X}$<br>x – 30.8 | $d^2$ | $fd^2$ |
|---|---|---|---|---|---|
| 10 | 8 | 80 | -20.8 | 432.64 | 3461.12 |
| 20 | 12 | 240 | -10.8 | 116.64 | 1399.68 |
| 30 | 20 | 600 | -0.8 | 0.64 | 12.80 |
| 40 | 10 | 400 | 9.2 | 84.64 | 846.40 |
| 50 | 7 | 350 | 19.2 | 368.64 | 2580.48 |
| 60 | 3 | 180 | 29.2 | 852.64 | 2557.92 |
|  | $\mathbf{N = \sum f = 60}$ | $\mathbf{\sum fx = 1850}$ |  |  | $\mathbf{\sum fd2 = 10858.40}$ |

$$\text{Mean: } \bar{\mathbf{X}} = \frac{\sum \mathbf{fx}}{\mathbf{N}}$$

$$= \frac{1850}{60}$$

$$= 30.8$$

$$\text{Standard Deviation: } \sigma = \sqrt{\frac{\sum \mathbf{fd^2}}{\mathbf{N}}}$$

$$= \sqrt{\frac{10858.40}{60}}$$

$$= 13.45$$

**Assumed Mean Method:**

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

Where d = X – A

**Illustration 11: (Solving the previous problem)**

**Solution:**

**Calculation of standard deviation (from assumed mean)**

| *x* | f | d = x -30 | $d^2$ | fd | $fd^2$ |
|---|---|---|---|---|---|
| 10 | 8 | -20 | 400 | -160 | 3200 |
| 20 | 12 | -10 | 100 | -120 | 1200 |
| 30 | 20 | 0 | 0 | 0 | 0 |
| 40 | 10 | 10 | 100 | 100 | 1000 |
| 50 | 7 | 20 | 400 | 140 | 2800 |
| 60 | 3 | 30 | 900 | 90 | 2700 |
| | $\mathbf{N = \sum f = 60}$ | | | $\sum fd = 50$ | $\sum fd^2 = 10900$ |

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

$$= \sqrt{\frac{\mathbf{10900}}{\mathbf{60}} - \left(\frac{\mathbf{50}}{\mathbf{60}}\right)^{\mathbf{2}}}$$

$$= \sqrt{181.67 - 0.69}$$

$$= \sqrt{180.98}$$

$$\mathbf{= 13.45}$$

**Step Deviation Method**

$$\sigma = \sqrt{\frac{\sum \mathbf{fd^2}}{\mathbf{N}} - \left(\frac{\sum \mathbf{fd}}{\mathbf{N}}\right)^{\mathbf{2}}} \text{ x C}$$

Where d‘ = $\frac{X - A}{C}$; C = Common Factor

**Illustration 12:**

**(Solving the previous problem)**

**Solution:**

**Calculation of standard deviation (from step deviation)**

| $x$ | f | d' = $\frac{X-30}{10}$ | d'$^2$ | fd' | fd'$^2$ |
|---|---|---|---|---|---|
| 10 | 8 | -2 | 4 | -16 | 32 |
| 20 | 12 | -1 | 1 | -12 | 12 |
| 30 | 20 | 0 | 0 | 0 | 0 |
| 40 | 10 | 1 | 1 | 10 | 10 |
| 50 | 7 | 2 | 4 | 14 | 28 |
| 60 | 3 | 3 | 9 | 9 | 27 |
|  | $\mathbf{N = \sum f = 60}$ |  |  | **∑fd’ = 5** | **∑fd’$^2$ = 109** |

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \text{ x C}$$

$$= \sqrt{\frac{109}{60} - \left(\frac{5}{60}\right)^2} \text{ x } 10$$

$$= \sqrt{1.817 - 0.0069} \text{ x } 10$$

$$= \sqrt{1.81} \text{ x } 10 = 1.345 \text{ x } 10$$

$\sigma$ = **13.45**

**Standard Deviation – Continuous Series**

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \text{ x C}$$

Where d = $\frac{m - A}{C}$; C = Common Factor

**Illustration13:**

Compute the standard deviation from the following data:

| Class | 0 - 10 | 10 - 20 | 20 - 30 | 30 - 40 | 40-50 |
|---|---|---|---|---|---|
| Frequency | 5 | 8 | 15 | 16 | 6 |

**Solution:**

**Computation of standard deviation**

| x | M | F | d = $\frac{m-25}{10}$ | d$^2$ | fd | fd$^2$ |
|---|---|---|---|---|---|---|
| 0 - 10 | 5 | 5 | -2 | 4 | -10 | 20 |
| 10 - 20 | 15 | 8 | -1 | 1 | -8 | 8 |
| 20 - 30 | 25 | 15 | 0 | 0 | 0 | 0 |
| 30 - 40 | 35 | 16 | 1 | 1 | 16 | 16 |
| 40 - 50 | 45 | 6 | 2 | 4 | 12 | 24 |
| | | $\mathbf{N = \sum f = 50}$ | | | $\mathbf{\sum fd}$ **= 10** | $\mathbf{\sum fd^2}$ **= 68** |

$$\sigma = \sqrt{\frac{\sum \mathbf{fd^2}}{\mathbf{N}} - \left(\frac{\sum \mathbf{fd}}{\mathbf{N}}\right)^{2}} \text{ x C}$$

$$\sigma = \sqrt{\frac{68}{50} - \left(\frac{10}{50}\right)^2} \text{ x } 10 = \sqrt{1.36 - (0.2)^2} \text{ x } 10$$

$$= \sqrt{1.36 - 0.04} \text{ x } 10 = \sqrt{1.32} \text{ x } 10$$

= 1.1489 x 10 = **11.49**

**Merits:**

1. It is rigidly defined determinate.
2. It is based on all the observations of a series.
3. It is less affected by fluctuations of sampling and hence stable.
4. It is amenable to algebraic treatment and is less affected by fluctuations of sampling most other measures of dispersion.
5. The standard deviation is more appropriate mathematically than the mean deviation, since the negative signs are removed by squaring the deviations rather than by ignoring

**Demerits:**

1. It lacks wide popularity as it is often difficult to compute, when big numbers are involved, the process of squaring and extracting root becomes tedious.
2. It attaches more weight to extreme items by squaring them.
3. It is difficult to calculate accurately when a grouped frequency distribution has extreme groups with no definite range.

**Uses:**

1. Standard deviation is the best measure of dispersion.
2. It is widely used in statistics because it possesses most of the characteristics of an ideal measure of dispersion.
3. It is widely used in sampling theory and by biologists.
4. It is used in coefficient of correlation and in the study of symmetrical frequency distribution.

**Co - efficient of variation (Relative Standard Deviation)**

The Standard deviation is an absolute measure of dispersion. The corresponding relative measure is known as the co - efficient of variation. It is used to compare the variability of two or more than two series. The series for which co-efficient or variation is more is said to be more variable or conversely less consistent, less uniform less table or less homogeneous.

**Variance:**

Square of standard deviation is called variance.

$$\text{Variance} = \sigma^2; \sigma = \sqrt{\textbf{Variance}}$$

$$\text{Co – efficient of standard deviation} = \frac{\sigma}{\bar{x}}$$

$$\text{Co – efficient of variation (C.V.)} = \frac{\sigma}{\bar{x}} \text{ x } 100$$

**Illustration 14:** The following are the runs scored by two batsmen A and B in ten innings:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| A | 101 | 27 | 0 | 36 | 82 | 45 | 7 | 13 | 65 | 14 |
| B | 97 | 12 | 40 | 96 | 13 | 8 | 85 | 8 | 56 | 15 |

Who is the more consistent batsman?

**Solution: Calculation of Co-efficient of Variation**

| Batsman A | | | Batsman B | | |
|---|---|---|---|---|---|
| **Runs Scored X** | **dx = X - $\bar{X}$** | **$dx^2$** | **Runs Scored Y** | **dx = Y - $\bar{Y}$** | **$dy^2$** |
| 101 | 62 | 3844 | 97 | 54 | 2916 |
| 27 | -12 | 144 | 12 | -31 | 961 |
| 0 | -39 | 1521 | 40 | -3 | 9 |
| 36 | -3 | 9 | 96 | 53 | 2809 |
| 82 | 43 | 1849 | 13 | -30 | 900 |
| 45 | 6 | 36 | 8 | -35 | 1225 |
| 7 | -32 | 1024 | 85 | 42 | 1764 |
| 13 | -26 | 676 | 8 | -35 | 1225 |
| 65 | 26 | 676 | 56 | 13 | 169 |
| 14 | -25 | 625 | 15 | -28 | 784 |
| **$\sum X = 390$** | | **$\sum dx^2 = 10404$** | **$\sum Y = 430$** | | **$\sum dy^2 = 12762$** |

**Batsman A**

$\bar{X} = \frac{\sum X}{N} = \frac{390}{10} = 39$

$\sigma_X = \sqrt{\frac{\sum dx^2}{N}} = \sqrt{\frac{10404}{10}} = 32.26$

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

$= \frac{32.26}{39}$ x 100

= 82.72%

**Batsman B**

$\bar{Y} = \frac{\sum Y}{N} = \frac{430}{10} = 43$

$\sigma_Y = \sqrt{\frac{\sum dy^2}{N}} = \sqrt{\frac{12762}{10}} = 35.72$

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

$= \frac{35.72}{43}$ x 100

= 83.07%

Batsman A is more consistent in his batting, because the co-efficient of variation of runs is less for him.

## Important questions

**Choose the correct answer:**

1. Sum of absolute deviations about median is :

**(a) The Least** (b) The greatest (c) Zero (d) Equal

2. The sum of squares of deviations is least when measured from:

(a) Median (b) Zero **(c) Mean** (d) Mode

3. The appropriate measure whenever the extreme items are to be disregarded and when the distribution contains indefinite classes at the end is :

(a) Median (b) Mode **(c) Quartile Deviation** (d) Mean Deviation

4. The quartile deviation includes the:

(a) First 50% (b) Last 50% **(c) Central 50%** (d) Atleast 50%

5. Which of the following is a relative measure of dispersion:

(a) Variance **(b) Coefficient of Variance** (c) Standard Deviation (d) Mean Deviation

6. The square of the variance of a distribution is the :

(a) Median (b) Mean (c) Mode **(d) None of these**.

**Questions:**

1. Why is that standard deviation is considered to be the most popular measure of dispersion?

2. What is coefficient of variation? What purpose does it serve? Also distinguish between ' variance' and 'coefficient of variation'.

3. Define coefficient of variation? In what situation would you prefer this as a measure of dispersion?

4. Define mean deviation. How does it differ from standard deviation?

5. Dispersion is known as the second average of the second order. Discuss.

6. What are the properties of a good measure of variation?

10. What are the requisites of a good measure of dispersion?

10. What are quartiles? How are they used for measuring dispersion?

11. How do you define coefficient of variation and what are its uses?

12. What are the objectives of measuring dispersion of a frequency distribution? Explain.

13. "Average and measures of dispersion are useful in understanding a frequency distribution". Elucidate the statement giving illustrations.

**Exercises:**

1. Calculate Range, Q.D, M.D (from mean), S.D and C.V of the marks obtained by 10 students given below:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 50 | 55 | 57 | 49 | 54 | 61 | 64 | 59 | 59 | 56 |

(Range = 15, Q.D = 2.75, M. D = 3.6, S.D = 4.43 and C.V = 7.85%)

2. Compute Q.D from the following data:

| Height in inches | 58 | 59 | 60 | 61 | 62 | 63 | 64 | 65 | 66 |
|---|---|---|---|---|---|---|---|---|---|
| No. of students | 15 | 20 | 32 | 35 | 33 | 22 | 20 | 10 | 8 |

(Q.D = 1.5)

3. Compute quartile deviation from the following data.

| *x:* | 4 – 8 | 8 – 12 | 12 - 16 | 16 - 20 | 20 - 24 | 24 - 28 | 28 - 32 | 32 – 36 | 36 - 40 |
|---|---|---|---|---|---|---|---|---|---|
| *f:* | 6 | 10 | 18 | 30 | 15 | 12 | 10 | 6 | 2 |

(Q.D = 5.21)

4. Calculate mean deviation (from mean) from the following data:

| *x:* | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|
| *f:* | 1 | 4 | 6 | 4 | 1 |

(M.D = 1.5)

5. Calculate the mean deviation from the following data.

| *x:* | 0 - 5 | 5 - 10 | 10 - 15 | 15 - 20 | 20 - 25 | 25 - 30 | 30 - 35 | 35 - 40 |
|---|---|---|---|---|---|---|---|---|
| *f:* | 449 | 705 | 507 | 281 | 109 | 52 | 16 | 4 |

(M.D = 5.25)

6. Calculate the S.D of the following:

| Size of the item | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|
| Frequency | 3 | 6 | 9 | 13 | 8 | 5 | 4 |

(S.D = 1.6)

7. Compute standard deviation of the following data:

| Wages(Rs. Per day) | 1 – 3 | 3 – 5 | 5 – 7 | 7 – 9 | 9 – 11 |
|---|---|---|---|---|---|
| No. of workmen | 15 | 18 | 27 | 10 | 6 |

(S.D = 2.33)

8. Two cricketers scored the following runs in the several innings. Find who is a better run getter and who more consistent player is?

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| A: | 42 | 17 | 83 | 59 | 72 | 76 | 64 | 45 | 40 | 32 |
| B: | 28 | 70 | 31 | 0 | 59 | 108 | 82 | 14 | 3 | 95 |

(C.V (A) = 38% and C.V (B) = 75.6%)

9. Two brands of types are tested with the following results:

| Life („000 miles) | 20 – 25 | 25 – 30 | 30 – 35 | 35 – 40 | 40 – 45 | 45 – 50 |
|---|---|---|---|---|---|---|
| Brand X | 8 | 15 | 12 | 18 | 13 | 9 |
| Brand Y | 6 | 20 | 32 | 30 | 12 | 0 |

(C.V (X) = 21.82% and C.V (Y) = 16.11%)

# UNIT IV

# SIMPLE CORRELATION

**Meaning:**

Correlation refers to the relationship of two or more variables. For example, there exists some relationship between the height of a mother and the height of a daughter, sales and cost and so on. Hence, it should be noted that the detection and analysis of correlation between two statistical variables requires relationship of some sort which associates the observation in pairs, one of each pair being a value of the two variables. The word relationship is of important and indicates that there is some connection between the variables under observation. Thus, the association of any two variates is known as correlation.

**Significance:**

Correlation is useful in physical and social sciences. We can study the uses of correlation in business and economics. The following are the significance of study of correlation:

- Correlation is very useful to economics to study the relationship between variables, like price and quantity demanded. To the businessmen, it helps to estimate costs, sales, price and other related variables.
- Some variables show some kind of relationship; correlation analysis helps in measuring the degree of relationship between the variables like supply and demand, price and supply, income and expenditure, etc.
- The relation between variables can be verified and tested for significance, with the help of the correlation analysis. The effect of correlation is to reduce the range of uncertainty of our prediction.
- The coefficient of correlation is a relative measure and we can compare the relationship between variables which are expressed in different units.
- Sampling error can also be calculated.
- Correlation is the basis for the concept of regression and ratio of variation.

## Types of Correlation:

Correlation is classified into many types but the important are:

1. Positive and Negative Correlation
2. Simple and Multiple Correlations
3. Partial and Total Correlation
4. Linear and Non-linear Correlation.

### 1. Positive and Negative Correlation :

The correlation is said to be positive when the values of two variables move in the same direction, so that an increase in the value of one variable is accompanied by an increase in the value of the other variable or a decrease in the value of one variable is followed by a decrease in the value of the other variable. Example: Height and weight, rainfall and yield of crops, etc.,

The correlation is said to be negative when the values of two variables move in opposite direction, so that an increase or decrease in the values of one variable is followed by a decrease or increase in the value of the other. Example: Price and demand, yield of crops and price, etc.,

### 2. Simple and multiple Correlation :

When we study only two variables, the relationship is described as simple correlation; Example: The study of price and demand of an article.

When more than two variables are studied simultaneously, the correlation is said to be multiple correlation. Example: the relationship of price, demand and supply of a commodity.

### 3. Partial and total Correlation:

Partial correlation coefficient provides a measure of relationship between a dependent variable and a particular independent variable when all other variables involved are kept constant. i.e., when the effect of all other variables are removed.

**Example:** When we study the relationship between the yield of rice per acre and both the amount of rainfall and the amount of fertilizers used. In these relationship if we limit our correlation analysis to yield and rainfall. It becomes a problem relating to simple correlation.

**4. Linear and Non-linear Correlation :**

The correlation is said to be linear, if the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable.

The correlation is non-linear, if the amount of change in one variable does not bear a constant ratio to the amount of change in the other related variable.

## 4.1 Scatter Diagram Method

### Methods of Studying Correlation:

The following correlation methods are used to find out the relationship between two variables.

**A.** Graphic Method :

- **i.** Scatter diagram (or) Scattergram method.
- **ii.** Simple Graph or Correlogram method.

**B.** Mathematical Method :

- **i.** Karl Pearson‘s Coefficient of Correlation.
- **ii.** Spearman‘s Rank Correlation of Coefficient
- **iii.** Coefficient of Concurrent Deviation
- **iv.** Method of Least Squares.

**C.** Graphic Method

### i. Scatter diagram (or) Scattergram method

This is the simplest method of finding out whether there is any relationship present between two variables by plotting the values on a chart, known as scatter diagram. In this method, the given data are plotted on a graph paper in the form of dots. X variables are plotted on the horizontal axis and Y variables on the vertical axis. Thus we have the dots and we can know the scatter or concentration of various points.

If the plotted dots fall in a narrow band and the dots are rising from the lower left hand corner to the upper right-hand corner it is called high degree of positive correlation.

If the plotted dots fall in a narrow band from the upper left hand corner to the lower right hand corner it is called a high degree of negative correlation.

If the plotted dots line scattered all over the diagram, there is no correlation between the two variables.

**Merits:**

1. It is easy to plot even by beginner.
2. It is simple to understand.
3. Abnormal values in a sample can be easily detected.
4. Values of some dependent variables can be found out.

**Demerits:**

1. Degree of correlation cannot be predicted.
2. It gives only a rough idea.
3. The method is useful only when number of terms is small.

### ii. Simple Graph Method of Correlation:

In this method separate curves are drawn for separate series on a graph paper. By examining the direction and closeness of the two curves we can infer whether or not variables are related. If both the curves are moving in the same direction correlation is said to be positive. On the other hand, if the curves are moving in the opposite directions is said to be negative.

**Merits:**

1. It is easy to plot
2. Simple to understand
3. Abnormal values can easily be deducted.

**Demerits:**

1. This method is useless when number of terms is very big.
2. Degree of correlation cannot be predicted.

**B. Mathematical Method:**

## 4.2 Karl Pearson's Coefficient of Correlation

Karl Pearson, a great biometrician and statistician, introduced a mathematical method for measuring the magnitude of linear relationship between two variables. This method is most widely used in practice. This method is known as Pearsonian Coefficient of Correlation. It is denoted by the symbol $'r'$; the formula for calculating Pearsonian r is:

(i) r = $\frac{\boldsymbol{Covariance\ of\ xy}}{\boldsymbol{\sigma x \times \sigma y}}$, (ii) r = $\frac{\boldsymbol{\Sigma xy}}{\boldsymbol{N\sigma x \times \sigma y}}$, (iii) r = $\frac{\boldsymbol{\Sigma xy}}{\sqrt{\boldsymbol{\Sigma x^2 \Sigma y^2}}}$

$$\boldsymbol{x = (X - \bar{X}),\ y = (Y - \bar{Y})}$$

$$\boldsymbol{\sigma x = Standard\ deviation\ of\ series\ x}$$

$$\boldsymbol{\sigma y = Standard\ deviation\ of\ series\ y}$$

The value of the coefficient of correlation shall always lie between **+1 and -1**.

When *r* = +1, then there is perfect positive correlation between the variables.

When *r* = -1, then there is perfect negative correlation between the variables.

When *r* = 0, then there is no relationship between the variables.

Illustration 1: calculate Karl Pearson coefficient of correlation from the following.

| **X** | 100 | 101 | 102 | 102 | 100 | 99 | 97 | 98 | 96 | 95 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Y** | 98 | 99 | 99 | 97 | 95 | 92 | 95 | 94 | 90 | 91 |

**Solution: Calculation of coefficient of correlation**

| **X** | **$x = X - \bar{X}$<br>$=X – 99$** | **$x^2$** | **Y** | **$y = Y - \bar{Y}$<br>$=Y - 95$** | **$y^2$** | **$Xy$** |
|---|---|---|---|---|---|---|
| 100 | 1 | 1 | 98 | 3 | 9 | 3 |
| 101 | 2 | 4 | 99 | 4 | 16 | 8 |
| 102 | 3 | 9 | 99 | 4 | 16 | 12 |
| 102 | 3 | 9 | 97 | 2 | 4 | 6 |
| 100 | 1 | 1 | 95 | 0 | 0 | 0 |
| 99 | 0 | 0 | 92 | -3 | 9 | 0 |
| 97 | -2 | 4 | 95 | 0 | 0 | 0 |
| 98 | -1 | 1 | 94 | -1 | 1 | 1 |
| 96 | -3 | 9 | 90 | -5 | 25 | 15 |
| 95 | -4 | 16 | 91 | -4 | 16 | 16 |
| **$\sum X =$ 990** | | **$\sum x^2 = 54$** | **$\sum Y =$ 950** | | **$\sum y^2 = 96$** | **$\sum xy =$ 61** |

$\bar{X} = \frac{\Sigma X}{N} = \frac{990}{10} = 99$; $\bar{Y} = \frac{\Sigma Y}{N} = \frac{950}{10} = 95$;

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

$$= \frac{61}{\sqrt{54 \times 96}} = \frac{61}{\sqrt{5184}}$$

$$= \frac{61}{72}$$

$$= +\ 0.85$$

Illustration 2: calculate Karl Pearson coefficient of correlation from the following.

| **X:** | 6 | 2 | 10 | 4 | 8 |
|---|---|---|---|---|---|
| **Y:** | 9 | 11 | 5 | 8 | 7 |

**Solution: Calculation of coefficient of correlation**

| X | $X^2$ | Y | $Y^2$ | XY |
|---|---|---|---|---|
| 6 | 36 | 9 | 81 | 54 |
| 2 | 4 | 11 | 121 | 22 |
| 10 | 100 | 5 | 25 | 50 |
| 4 | 16 | 8 | 64 | 32 |
| 8 | 64 | 7 | 49 | 56 |
| **$\sum X = 30$** | **$\sum X^2 = 220$** | **$\sum Y = 40$** | **$\sum Y^2 = 340$** | **$\sum XY = 214$** |

$$r = \frac{\mathbf{N\Sigma XY - (\Sigma X.\Sigma Y)}}{\sqrt{\mathbf{N\Sigma}\boldsymbol{X}^{\mathbf{2}}\mathbf{-(\Sigma}\boldsymbol{X}\mathbf{)^2}}\,\mathbf{x}\sqrt{\mathbf{N\Sigma}\boldsymbol{Y}^{\mathbf{2}}\mathbf{-(\Sigma}\boldsymbol{Y}\mathbf{)^2}}}$$

$$= \frac{(5 \text{ x } 214)-(30 \text{ x } 40)}{\sqrt{5 \text{ x } 220-(30)^2}\ \text{x}\sqrt{5 \text{ x } 340-(40)^2}}$$

$$= \frac{1070 - 1200}{\sqrt{1100 - 900}\ \text{x}\sqrt{1700 - 1600}}$$

$$\boldsymbol{r} = \frac{-130}{\sqrt{200}\ \text{x}\sqrt{100}} \mathbf{= -\ 0.92}$$

## 4.3 RANK CORRELATION CO-EFFICIENT

**Spearman’s Rank Correlation Co-efficient:**

In 1904, a famous British psychologist Charles Edward Spearman found out the method of ascertaining the coefficient of correlation by ranks. This method is based on rank. Rank correlation is applicable only to individual observations. This measure is useful in dealing with qualitative characteristics such as intelligence, beauty, morality, character, etc.,

The formula for Spearman‘s rank correlation which is denoted by P is;

$$P = 1- \frac{6\Sigma D^{2}}{N(N^2-1)}$$

or

$$P = 1- \frac{6\Sigma D^{2}}{(N^3-N)}$$

Where, P = Rank co-efficient of correlation

D = Difference of the two ranks

$\sum D^2$ = Sum of squares of the difference of two ranks

N = Number of paired observations

Like the Karl Pearson‘s coefficient of correlation, the value of **P** lies between **+ 1** and **– 1**.

**Where ranks are given**

**Illustration 3:** Following are the rank obtained by 10 students in two subjects, Statistics and Mathematics. To what extent the knowledge of the students in the two subjects is related?

| **Statistics** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Mathematics** | 2 | 4 | 1 | 5 | 3 | 9 | 7 | 10 | 6 | 8 |

**Solution:**

**Calculation of Pearman’s rank correlation coefficient**

| **Rank of Statistics (*x*)** | **Rank of Mathematics (*y*)** | **D = *x* – *y*** | **$D^2$** |
|---|---|---|---|
| 1 | 2 | -1 | 1 |
| 2 | 4 | -2 | 4 |
| 3 | 1 | +2 | 4 |
| 4 | 5 | -1 | 1 |

| | | | |
|---|---|---|---|
| 5 | 3 | +2 | 4 |
| 6 | 9 | -3 | 9 |
| 7 | 7 | 0 | 0 |
| 8 | 10 | -2 | 4 |
| 9 | 6 | +3 | 9 |
| 10 | 8 | +2 | 4 |
| **N = 10** | | | $\boldsymbol{\Sigma D^2 = 40}$ |

$$P = 1- \frac{6\Sigma D^{2}}{N(N^2-1)}$$

$$P = 1- \frac{6 \text{ x } 40}{10(10^2 - 1)}$$

$$= 1- \frac{240}{10(100 \ -1)}$$

$$= 1- \frac{240}{990}$$

$$= 1 – 0.24$$

$$\mathbf{= + \ 0.76}$$

**Where Ranks are not given:**

**Illustration 4:**

A random sample of 5 college students is selected and their grades in Mathematics and Statistics are found to be:

| **Mathematics** | 85 | 60 | 73 | 40 | 90 |
|---|---|---|---|---|---|
| **Statistics** | 93 | 75 | 65 | 50 | 80 |

**Solution:**

**Calculation of spearman's rank correlation coefficient**

| Mathematics ($x$) | Rank x | Statistics ($y$) | Rank y | D = $x - y$ | $D^2$ |
|---|---|---|---|---|---|
| 85 | 2 | 93 | 1 | +1 | 1 |
| 60 | 4 | 75 | 3 | +1 | 1 |
| 73 | 3 | 65 | 4 | -1 | 1 |
| 40 | 5 | 50 | 5 | 0 | 0 |
| 90 | 1 | 80 | 2 | -1 | 1 |
| | | | | | $\Sigma D^2 = 4$ |

$$P = 1- \frac{6\Sigma D^2}{N(N^2-1)}$$

$$= 1- \frac{6 \text{ x } 4}{5(5^2-1)}$$

$$= 1- \frac{24}{5(25-1)}$$

$$= 1- \frac{24}{5(24)}$$

$$= 1- \frac{1}{5} = 1 - 0.2$$

$$= + \mathbf{0.8}$$

**Equal or Repeated Ranks:**

When two or more items have equal values, it is difficult to give ranks to them. In that case the items are given the average of the ranks they would have received, if they are not tied. A slightly different formula is used when there is more than one item having the same value.

$$\mathbf{P = 1 - 6\{\frac{\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \cdots}{N^3 - N}\}}$$

*m = the number of items whose ranks are common*

**Illustration 5:** From the following data calculate the rank correlation coefficient after making adjustment for tied ranks.

| **X** | 48 | 33 | 40 | 9 | 16 | 16 | 65 | 24 | 16 | 57 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Y** | 13 | 13 | 24 | 6 | 15 | 4 | 20 | 9 | 6 | 19 |

**Solution: Calculation of spearman's rank correlation coefficient**

| **X** | **Rank *x*** | **Y** | **Rank *y*** | **D =R(*x*) – R(*y*)** | **D$^2$** |
|---|---|---|---|---|---|
| 48 | 8 | 13 | 5.5 | 2.5 | 6.25 |
| 33 | 6 | 13 | 5.5 | 0.5 | 0.25 |
| 40 | 7 | 24 | 10 | -3.0 | 9.00 |
| 9 | 1 | 6 | 2.5 | -1.5 | 2.25 |
| 16 | 3 | 15 | 7 | 4.0 | 16.00 |
| 16 | 3 | 4 | 1 | 2.0 | 4.00 |
| 65 | 10 | 20 | 9 | 1.0 | 1.00 |
| 24 | 5 | 9 | 4 | 1.0 | 1.00 |
| 16 | 3 | 6 | 2.5 | 0.5 | 025 |
| 57 | 9 | 19 | 8 | 1.0 | 1.00 |
| | | | | | $\boldsymbol{\Sigma D^2 = 41}$ |

$$\mathbf{P = 1 - 6\{}\frac{\boldsymbol{\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m)\ldots}}{\boldsymbol{N^3 - N}}\mathbf{\}}$$

$$= 1 - 6\{\frac{41 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)\ldots}{10^3 - 10}\}$$

$$= 1 - \{\frac{6\ (41+2+0.5+05)}{990}\}$$

$$= 1 - \{\frac{264}{990}\} = 1 - 0.267$$

$$= \mathbf{+\ 0.\ 733}$$

**Merits:**

1. It is simple to understand and easier to apply.
2. It can be used to any type of data, qualitative or quantitative.
3. It is the only method that can be used where we are given the ranks and not the actual data.
4. Even where actual data are given, rank method can be applied for ascertaining correlation by assigning the ranks to each data.

**Demerits:**

1. This method is not useful to find out correlation in a grouped frequency distribution.
2. For large samples it is not convenient method. If the items exceed 30 the calculations become quite tedious and require a lot of time.
3. It is only an approximately calculated measure as actual values are not used for calculations.

## 4.4 REGRESSION ANALYSIS

The statistical method which helps us to estimate the unknown value of one variable from the known value of the related variable is called Regression. The dictionary meaning of the word regression is “return or going back”. In 1877, Sir Francis Galton, first introduced the word ‘Regression’. The tendency to regression or going back was called by Galton as the ‘Line of Regression’. The line describing the average relationship between two variables is known as the line of regression. The regression analysis confined to the study of only two variables at a time is termed as simple regression. The regression analysis for studying more than two variables at a time is known as multiple regressions.

**Regression Vs Correlation:**

| S. No. | Regression | Correlation |
|---|---|---|
| 1 | It is a mathematical measure showing the average relationship between two variables. | It is the relationship between two or more variable, which vary in sympathy with the other in the same or the opposite direction. |
| 2 | Here x is a random variable and y is a fixed variable. | Both x and y are random variables. |
| 3 | In indicates the cause and effect relationship between the variables. | It finds out the degree of relationship between two variables. |

|  |  |  |
|---|---|---|
| 4 | It is the prediction of one value, in relationship to the other given value. | It is used for testing and verifying the relation between two variables. |
| 5 | It is an absolute figure. | It is a relative measure. The range of relationship lies between ± 1. |
| 6 | Here there is no such nonsense regression | There may be nonsense correlation between two variables. |
| 7 | It has wider application, as it studies linear and non-linear relationship between the variables. | It has limited application, because it is confined only to linear relationship between the variables. |
| 8 | It is widely used for further mathematical treatment. | It is not very useful for mathematical treatment. |
| 9 | It explains that the decrease in one variable is associated with the increase in the other variable. | If the coefficient of correlation is positive, then the two variables are positively correlated and vice - versa. |
| 10 | There is a functional relationship between the two variables so that we may identify between the independent and dependent variables. | It is immaterial whether X depends upon Y or Y depends upon X. |

**Linear Regression:**

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable *causes* the other (for example, higher SAT scores do not *cause* higher college grades), but that there is some significant association between the two variables. A scatter plot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatter plot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

A linear regression line has an equation of the form $\boldsymbol{Y = a + bX}$, where $\boldsymbol{X}$ is the explanatory variable and $\boldsymbol{Y}$ is the dependent variable. The slope of the line is $\boldsymbol{b}$, and $\boldsymbol{a}$ is the intercept (the value of $\boldsymbol{y}$ when $\boldsymbol{x} = 0$).

**Regression lines:**

If we take two variables X and Y we have two regression lines:

i) Regression of X on Y and

ii) Regression of Y on X

The regression line of X on Y gives the most probable value of X for any given value of Y. The regression of Y on X gives the most probable value of Y for any given value of X. There are two regression lines in the case of two variables.

**Regression Equations:**

The algebraic expressions of the two regression lines are called regression equations.

**Regression Equation of X on Y:**

$$X_c = a + by$$

To determine the values of _a‘ and _b‘, the following two normal equations are to be solved simultaneously.

$$\sum X = Na + b\sum Y$$

$$\sum XY = a\sum Y + b\sum Y^2$$

**Regression Equation of Y on X:**

$$Y_c = a + bx$$

To determine the value of _a‘ and _b‘, the following two normal equations are to be solved simultaneously.

$$\sum Y = Na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

We can call these equations as normal equations.

**Illustration 1:** Determine the two regression equations of a straight line which best fits the data.

| **X** | 10 | 12 | 13 | 16 | 17 | 20 | 25 |
|---|---|---|---|---|---|---|---|
| **Y** | 10 | 22 | 24 | 27 | 29 | 33 | 37 |

**Solution: Calculation of Regression**

| **X** | **$X^2$** | **Y** | **$Y^2$** | **XY** |
|---|---|---|---|---|
| 0 | 100 | 10 | 100 | 100 |
| 12 | 144 | 22 | 484 | 264 |
| 13 | 169 | 24 | 576 | 312 |
| 16 | 256 | 27 | 729 | 432 |
| 17 | 289 | 29 | 841 | 493 |
| 20 | 400 | 33 | 1089 | 660 |
| 25 | 625 | 37 | 1369 | 925 |
| **$\sum X = 113$** | **$\sum X^2 = 1983$** | **$\sum Y = 182$** | **$\sum Y^2 = 5188$** | **$\sum XY = 3186$** |

**Regression Equation of Y on X:**

The two normal equations are:

$$\sum Y = Na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

Substituting the values,

$N = 7;$ $\sum X = 113;$ $\sum X^2 = 1983;$ $\sum Y = 182;$ $\sum XY = 3186;$

$$7a + 113b = 182 \qquad \ldots(1)$$

$$113a + 1983b = 3186 \qquad \ldots(2)$$

Multiplying (1), by 113,

$$791a + 12769b = 20566 \qquad \ldots(3)$$

Multiplying (2), by 7,

$791a + 13881b = 22302 \qquad \ldots(4)$

Subtracting (4) from (3)

$-1112\ b = -1736$

$b = \frac{-1736}{-1112}$ => $\mathbf{b = 1.56}$

Put b = 1.56 in (1) we get

$7a + 113(1.56)b = 182$

$7a + 176.28 = 182$ => $7a = 5.72$

$a = \frac{5.72}{7}$ => $\mathbf{a = 0.82}$

The equation of straight line is $Y_c = a + bX$

Put a = 0.82, b = 1.56

∴The equation of the required straight line is $Y_c = 0.82 + 1.56\ X$

This is called regression of y on x

**Regression Equation of X on Y:**

The two normal equations are:

$$\sum X = Na + b\sum Y$$

$$\sum XY = a\sum Y + b\sum Y^2$$

Substituting the values,

$N = 7; \quad \sum X = 113; \quad \sum Y^2 = 5188; \quad \sum Y = 182; \quad \sum XY = 3186;$

$7a + 182b = 113$ …(1)

$182a + 5188b = 3186$ …(2)

Multiplying (1), by 182,

$1274a + 33124b = 20566$ …(3)

Multiplying (2), by 7,

$1274a + 36316b = 22302$ …(4)

Subtracting (4) from (3)

$3192\ b = 1736$

$b = \frac{1736}{3192}$ => **$b = 0.54$**

Put $b = 0.54$ in (1) we get

$7a + 182(0.54) = 113$

$7a + 98.28 = 113$ => $7a = 14.72$

$a = \frac{14.72}{7}$ => **$a = 2.1$**

The equation of straight line is $X_c = a + bY$

Put $a = 2.1$, $b = 0.54$

∴The equation of the required straight line is $X_c = 2.1 + 0.54\ Y$

This is called regression of x on y

**Deviation taken from Actual Means:**

Regression equation of X on Y:

$$X - \bar{X} = r\,\frac{\sigma x}{\sigma y}\,(\mathbf{Y} - \bar{Y})$$

Where, X = the value of $x$ to be estimated for the given y value. $\bar{X}$ = Mean value of X variable. Y = the value of $y$ given in the problem; $\bar{Y}$ = Mean value of y variables.

$$r\,\frac{\sigma x}{\sigma y} = \frac{\sum xy}{\sum y2}$$ = Regression co - efficient of X on Y. x = X - $\bar{X}$, y = Y- $\bar{Y}$

**Regression equation of Y on X:**

$$Y - \bar{Y} = r\,\frac{\sigma x}{\sigma y}\,(X - \bar{X})$$

$$r\,\frac{\sigma x}{\sigma y} = \frac{\sum xy}{\sum y2}$$ = Regression co-efficient of Y on X.

**Illustration 2:** Find regression lines from the following data:

| X | 3 | 5 | 6 | 8 | 9 | 11 |
|---|---|---|---|---|---|---|
| Y | 2 | 3 | 4 | 6 | 5 | 10 |

And also estimate Y when X is 15.

**Solution: Calculation of Regression Equations (by actual mean)**

| **X** | **x =**<br>**X -** $\bar{X}$ | $\mathbf{x^2}$ | **Y** | **y =**<br>**Y –** $\bar{Y}$ | $\mathbf{y^2}$ | **xy** |
|---|---|---|---|---|---|---|
| 3 | -4 | 16 | 2 | -3 | 9 | 12 |
| 5 | -2 | 4 | 3 | -2 | 4 | 4 |
| 6 | -1 | 1 | 4 | -1 | 1 | 1 |
| 8 | 1 | 1 | 6 | 1 | 1 | 1 |
| 9 | 2 | 4 | 5 | 0 | 0 | 0 |
| 11 | 4 | 16 | 10 | 5 | 25 | 20 |
| **∑X = 42** | **∑x = 0** | **∑ $\mathbf{x^2}$ = 42** | **∑Y = 30** | **∑y = 0** | **∑ $\mathbf{y^2}$ = 40** | **∑xy = 38** |

$\bar{X} = \frac{\Sigma X}{N} = \frac{42}{6} = 7$ $\qquad\qquad$ $\bar{Y} = \frac{\Sigma Y}{N} = \frac{30}{6} = 5$

**Regression equation of X on Y:** **Regression equation of Y on X:**

$X - \bar{X} = r\frac{\sigma x}{\sigma y}(Y - \bar{Y})$ $\qquad\qquad$ $Y - \bar{Y} = r\frac{\sigma x}{\sigma y}(X - \bar{X})$

$r\frac{\sigma x}{\sigma y} = \frac{\sum xy}{\sum y2} = \frac{38}{40} = 0.95$ $\qquad\qquad$ $r\frac{\sigma x}{\sigma y} = \frac{\sum xy}{\sum y2} = \frac{38}{42} = 0.90$

X – 7 = 0.95(Y – 5) Y – 5 = 0.90(X – 7)

X – 7 = 0.95Y – 4.75 Y – 5 = 0.90X – 6.30

X = 0.95Y + 2.25 Y = 0.90X – 1.30

When X is 15, Y will be, Y = 0.90 x 15 – 1.30

= 13.5 – 1.30

Y = 14.8

**Deviation taken from the Assumed Mean:**

**Regression equation of X on Y:**

$$X - \bar{X} = r\frac{\sigma x}{\sigma y}(Y - \bar{Y})$$

Where, $r\frac{\sigma x}{\sigma y} = \frac{N\Sigma dxdy - (\Sigma dx)(\Sigma dy)}{N\Sigma dy^2 - (\Sigma dy)^2}$

dx = X - A; dy = Y - A; ( A = assumed mean)

**Regression equation of Y on X :**

$$Y - \bar{Y} = r\frac{\sigma y}{\sigma x}(X - \bar{X})$$

$$r\frac{\sigma x}{\sigma y} = \frac{N\Sigma dxdy - (\Sigma dx)(\Sigma dy)}{N\Sigma d\,x^2 - (\Sigma dx)^2}$$

**Illustration: 3.** Find regression lines from the following data:

| **X** | 40 | 38 | 35 | 42 | 30 |
|---|---|---|---|---|---|
| **Y** | 30 | 35 | 40 | 36 | 29 |

Also calculate Karl Pearson‘s coefficient of correlation.

**Solution: Calculation of Regression Equations (by assumed mean)**

| **X** | **d*x* = X – A** | **dx$^2$** | **Y** | **d*y* = Y – A** | **dy$^2$** | **d*x*.d*y*** |
|---|---|---|---|---|---|---|
| 40 | 5 | 25 | 30 | 0 | 0 | 0 |
| 38 | 3 | 9 | 35 | 5 | 25 | 15 |
| 35 | 0 | 0 | 40 | 10 | 100 | 0 |
| 42 | 7 | 49 | 36 | 6 | 36 | 42 |
| 30 | -5 | 25 | 29 | -1 | 1 | 5 |
| | **∑ dx = 10** | **∑ dx$^2$ =108** | | **∑ dy =1020** | **∑ dy$^2$ =162** | **∑dx.dy = 62** |

$\bar{X} = A \pm \frac{\Sigma dx}{N} = 35 \pm \frac{10}{5} = 37$

$\bar{Y} = A \pm \frac{\Sigma dy}{N} = 30 \pm \frac{20}{5} = 34$

Regression equation of X on Y

$X - \bar{X} = bxy\ (Y - \bar{Y})$

$bxy = \frac{N\Sigma dxdy - (\Sigma dx)(\Sigma dy)}{N\Sigma dy^2 - (\Sigma dy)^2}$

$= \frac{5 \times 62 - (10)(20)}{5 \times 162 - (20)^2}$

$= \frac{310 - 200}{810 - 400} = \frac{110}{410}$

*bxy* = 0.27

Regression equation of Y on X

$Y - \bar{Y} = byx\ (X - \bar{X})$

$byx = \frac{N\Sigma dxdy - (\Sigma dx)(\Sigma dy)}{N\Sigma d\,x^2 - (\Sigma dx)^2}$

$= \frac{5 \times 62 - (10)(20)}{5 \times 108 - (10)^2}$

$= \frac{310 - 200}{540 - 100} = \frac{110}{440}$

*byx* = 0.25

$X – 37 = 0.27(Y – 34)$

$X – 37 = 0.27Y – 9.18$

$X = 0.27Y + 27.82$

$Y – 34 = 0.25(X – 37)$

$Y – 34 = 0.25X – 9.25$

$Y = 0.25X + 24.75$

$$r = \sqrt{bxy\ x\ byx}$$

$$= \sqrt{2.2 \text{ x } 0.37}$$

$$= \sqrt{0.814}$$

$$r = \mathbf{0.9}$$

**Illustration 4:** Given the following data, calculate the expected value of Y when X = 12.

| | **X** | **Y** |
|---|---|---|
| Arithmetic Mean $(\bar{X})$ | 7.6 | 14.8 |
| Standard Deviation (σ) | 3.6 | 2.5 |
| Coefficient of correlation*(r) = 0.99* | | |

**Solution:**

**Regression of Y on X**

$$Y - \bar{\boldsymbol{Y}} = r\ \frac{\sigma y}{\sigma x}\ (X - \bar{\boldsymbol{X}})$$

$$Y – 14.8 = 0.99 \text{ x } \frac{2.5}{3.6} (X – 7.6)$$

$$Y – 14.8 = 0.688\ (X – 7.6)$$

$$Y – 14.8 = 0.688\ X – 5.23$$

$$Y = 0.688\ X – 5.23 + 14.8$$

$$Y = 0.688\ X + 9.57$$

When X = 12 => Y = 0.688 (12) + 9.57 = 17.826

Hence the expected value of Y is 17.83.

**Standard Error of Estimate:**

We found it necessary to supplement an average for a series with a measure of dispersion or variation to show how representative the average is. The regression equations help us to predict the values of Y for values X or the value of X far values of Y . These are only estimations or predictions; but cannot be treated as a precise value. If we have a wide scatter or variation of the dots about the regression line, then it would be considered a poor representative of the relationship. The more closely the dots cluster around the line, the more representative it is and better the estimate based on the equation for this line. This variation about the line of average relationship can be measured in a manner analogous to the measuring of the variation of the items about an average. Thus, we use here a measure of variation similar to the standard deviation - the standard error of estimates. It is computed as is a standard, being also a square root of the mean of squared deviations. But the deviations here are not the deviations of the items from the arithmetic mean, they are rather the vertical distances of every dot from the line of average relationship.

It measures the scattering of the observations the regression line. It is calculated as follows :

Standard Error of X values from $X_c$ $[S_{xy}]$ = $\sqrt{\frac{\Sigma(X-X_c)^2}{N}}$

Standard Error of Y values from $Y_c$ $[S_{yx}]$ = $\frac{\sqrt{\Sigma(Y\text{-}Y_C)^2}}{N}$

**Interpretation of Standard Error of Estimate :**

1. Smaller the value, precision of the estimate is better.
2. Larger the value, lesser is correctness of the estimate.
3. If it is zero, there is no variation about the line and both the lines will coincide and correlation will be perfect.

**Illustration:**

Given the regression equation of Y on X as Y = 3 + 9X for the following data series, calculate (i) Standard error of estimate (ii) Explained variation in Y (iii) unexplained variation in Y.

| X | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Y | 10 | 20 | 30 | 50 | 40 |

**Solution :**

| X | Y | $y = Y - \overline{Y}$ | $y^2$ | $y_c$<br>$[3 + 9X]$ | $[Y - Y_c]$ | $[Y - Y_c]^2$ |
|---|---|---|---|---|---|---|
| 1 | 10 | -20 | 400 | 12 | -2 | 4 |
| 2 | 20 | -10 | 100 | 21 | -1 | 1 |
| 3 | 30 | 0 | 0 | 30 | 0 | 0 |
| 4 | 50 | 20 | 400 | 39 | 11 | 121 |
| 5 | 40 | 10 | 100 | 48 | -8 | 64 |
|  | $\sum Y = 150$ | 0 | $\sum Y^2 =$ 1000 |  |  | $\sum[Y - Y_c]^2$ =190 |

$$\overline{Y} = \frac{\sum Y}{N} = \frac{150}{5} = 30$$

(i) Standard error of estimate $S_{yx} = \sqrt{\frac{\sum[Y - Y_c]^2}{N}} = \sqrt{\frac{190}{5}} = \sqrt{38} = 6.164$

(ii) Unexplained variation in Y = $\sum[Y - Y_c]^2$ = 190

(iii) Total variation $y^2 = 1000$

Explained variation = Total variation– Unexplained variation

$= y^2 - \sum[Y - Y_c]^2$

$= 1000 - 190$

$= 810.$

## Important Questions:

**Choose the correct answer :**

1. The coefficient of correlation :

a) has no limits

b) can be less than 1

c) can be more than 1

d) **Varies Between +- 1**

2. The value of r2 for a particular situation is 0.81. What is coefficient of correlation :

(a) 0.81 (b) **0.9** (c) 0.09 (d) 9

3. Which of the following is the highest range of r?

(a) 0 and 1 (b) -1 and 0 (c) **-1 and 1** (d) 0 and -1

4. The coefficient of correlation is independent of :

(a) change of scale only

(b) change of origin only

(c) **Both Change of Scale and Origin**

(d) No change

5. The coefficient of correlation :

(a) cannot be positive

(b) cannot be negative

(c) **Can be either Positive or Negative**

(d) zero

6. The greater the value of r:

**(a) The better are estimates, obtain through Regression Analysis**

(b) The worst are the estimates

(c) Really makesvno difference.

7. Where r is zero the regression lines cut each other making an angle of:

(a) 30° (b) 60° (c) 90° (d) **None of these**

8. The further the two regression lines cut each other :

(a) Greater will be degree of correlation

**(b) The Lessor will be Degree of Correlation**

(c) Does not matter.

9. The regression lines cut each other at the point of:

**(a) Average of X and Y**

(b) Average of X only

(c) Average of Y only.

10. When the two regression coincide, then

r is :(a) 0 (b) -1 **(c) +1** (d) 0.5

**Questions:**

1. What is meant by correlation? What are the properties of the coefficient of correlation?

2. (a) Distinguish coefficient of correlation from coefficient of variation.

(b) What is scatter diagram? How does it helps us in studying the correlation between two variables, in respect of both their nature and extent?

3. Define Karl Pearson's coefficient of correlation. What is it intended to measure?

4. Distinguish between:

(a) Positive and negative correlation.

(b) Linear and non-linear correlation,

(c) Simple, partial and multiple correlation.

5. What are the methods of calculating coefficient of correlation?

6. Explain the assumption on which Karl Pearson coefficient of correlation, is based.

7. Explain the concepts of correlation and regression, bringing out the inter-relationship between them. Also state their numerical measures.

8. Explain clearly why there are usually two lines of regression. Point out the case when there is one line of regression. Illustrate your answer by diagram.

9. Explain the concepts of regression and ratio of variation and State their utility in the field of economic enquiries.

10. What is regression? How is this concept useful to business forecasting?

11. Explain the meaning of regression coefficient and the regression lines.

12. What are the properties of the regression coefficients?

**Exercises:**

1. Calculate Pearson‟s coefficient of correlation between advertisement cost and sales from the following data:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Advertisement Cost(„000) Rs. | 39 | 65 | 62 | 90 | 82 | 75 | 25 | 98 | 36 | 78 |
| Sales (‟00,000) | 47 | 53 | 58 | 86 | 62 | 68 | 60 | 91 | 51 | 84 |

(*r = + 0.78*)

2. Compute the coefficient of correlation of the following score of A and B.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| A: | 5 | 10 | 5 | 11 | 12 | 4 | 3 | 2 | 7 | 1 |
| B: | 1 | 6 | 2 | 8 | 5 | 1 | 4 | 6 | 5 | 2 |

(*r = + 0.58*)

3. Ten competitors in a voice contest are ranked by 3 judges in the following orders:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| I Judge | 1 | 6 | 5 | 10 | 3 | 2 | 4 | 9 | 7 | 8 |
| II Judge | 3 | 5 | 8 | 4 | 7 | 10 | 2 | 1 | 6 | 9 |
| III Judge | 6 | 4 | 9 | 8 | 1 | 2 | 3 | 10 | 5 | 7 |

Use the rank correlation to gauge which pair of judges have the nearest approach to common likings in voice. (I & II = - *0.212; II & III = - 0.297; I & III = + 0.636*)

4. Calculate the rank correlation coefficient for the following table of marks of students in two subjects:

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Major I | 80 | 64 | 54 | 49 | 48 | 35 | 32 | 29 | 20 | 18 | 15 | 10 |
| Major II | 36 | 38 | 39 | 41 | 27 | 43 | 45 | 52 | 51 | 42 | 40 | 52 |

(*r* = *- 0.685*)

5. The following table gives the score obtained by 11 students in Mathematics and Statistics. Find the rank correlation coefficient.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Mathematics | 40 | 46 | 54 | 60 | 70 | 80 | 82 | 85 | 85 | 90 | 95 |
| Statistics | 45 | 45 | 50 | 43 | 40 | 75 | 55 | 72 | 65 | 42 | 70 |

(*r* = + *0.36*)

6. Calculate the coefficient of concurrent deviation from the data given below:

| Year | 1976 | 1977 | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 | 1984 |
|---|---|---|---|---|---|---|---|---|---|
| Supply | 160 | 164 | 172 | 182 | 166 | 170 | 178 | 192 | 186 |
| Price | 292 | 290 | 260 | 234 | 266 | 254 | 230 | 130 | 200 |

(*r* = *-1*)

7. Obtain the equations of lines of regression between the indices.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X: | 78 | 77 | 85 | 88 | 87 | 82 | 81 | 77 | 76 | 83 | 97 | 93 |
| Y: | 84 | 82 | 82 | 85 | 89 | 90 | 88 | 92 | 83 | 89 | 98 | 99 |

(X = 0.79 Y + 13.82; Y = 0.59 X + 39.05)

8. Calculate the two regression equations of X on Y and Y on X from the data given below, takingdeviations from actual means of X and Y.

| | | | | | | |
|---|---|---|---|---|---|---|
| Price (Rs.) | 10 | 12 | 13 | 12 | 16 | 15 |
| Amount demanded | 40 | 38 | 43 | 45 | 37 | 43 |

Estimate the likely demand when the price is Rs. 20.

(X = -0.12 Y + 17.92; Y = -0.25 X + 44.25; Y = 39.25)

9. The correlation coefficient between two variable X and Y is *r* = 0.6. If $\sigma_x$ = 1.5 $\sigma_y$= 2.0, = 10and = 20, find the regression lines of Y on X and X on Y.

(X = 0.45 Y + 1; Y = 0.8 X + 12)

10. By using the following data. Find out the two lines regression and from them compute the KarlPearson‟s Coefficient of Correlation:

$\sum X = 250; \sum Y = 300; \sum XY = 7900; \sum X^2 = 6500; \sum Y^2 = 10,000; N -= 10.$

($r$ = - *0.8*)

11. Given the following data, estimate the marks in mathematics obtained by a student who hasscored 60 marks in statistics.

| | |
|---|---|
| Mean marks of mathematics | 80 |
| Mean marks of statistics | 50 |
| S.D marks in mathematics | 15 |
| S.D marks in statistics | 10 |
| Coefficient of correlation | 0.4 |

(X = 0.6 Y + 50; X = 86.)

# UNIT V

# TESTING OF HYPOTHESIS

## Hypothesis

Hypothesis is a precise, testable statement of what the researchers predict will be outcome of the study. Hypothesis usually involves proposing a relationship between two variables: the independent variable (what the researchers change) and the dependent variable (what the research measures).

Hypothesis is usually considered as the principal instrument in research. The main goal in many research studies is to check whether the data collected support certain statements or predictions. A statistical hypothesis is an assertion or conjecture concerning one or more populations. Test of hypothesis is a process of testing of the significance regarding the parameters of the population on the basis of sample drawn from it. Thus, it is also termed as "Test of Significance'.

In short, hypothesis testing enables us to make probability statements about population parameter. The hypothesis may not be proved absolutely, but in practice it is accepted if it haswithstood a critical testing.

**Points to be considered while formulating Hypothesis**

- Hypothesis should be clear and precise.
- Hypothesis should be capable of being tested.
- Hypothesis should state relationship between variables.
- Hypothesis should be limited in scope and must be specific.
- Hypothesis should be stated as far as possible in most simple terms so that the same is easily understandable by all concerned.
- Hypothesis should be amenable to testing within a reasonable time.
- Hypothesis must explain empirical reference.

## Procedure of Testing a Hypothesis

After having completed collection, processing and analysis of data a test procedure has to be followed for determining if the null hypothesis is to be accepted orrejected. The test procedure or the rule is based upon a test statistic and a rejection region. The procedure of testing hypothesis is briefly described below:

- **Setting up a hypothesis:**

At the very outset, we take certain hypothesis with regard to related variables under the assumptions defined for the study. Generally, there are two forms of hypotheses which must be constructed; and if one hypothesis is accepted, the other one is rejected.

i. **Null Hypothesis**: It is very useful tool to test the significance of difference. Any hypothesis concerned to a population is called statistical hypothesis. In the process of statistical test, the rejection or acceptance of hypothesis depends on sample drawnfrom population. The statistician tests the hypothesis through observation and gives a probability statement. The simple hypothesis states that the statistical measures of sample and those of the population under study do not differ significantly. Similarly it may assume no relationship or association between two variables or attributes. In case of assessing the effectiveness of a literacy campaign on the awareness of rural people we assume "There is no effect of the campaign on public awareness". It is denoted by $H_O$

For example, if we want to find out whether extra coaching has benefitted the students or not: the null hypothesis would be;

Ho (1): The extra coaching has not benefitted the students.

Similarly, if we want to find out whether a particular drug is effective in curingMalaria we will take the null hypothesis:

Ho (2): The drug used under experimentation is not effective in curing Malaria. Similarly for testing the significance of difference between two sample, nullhypothesis would be:

Ho (3): "There is no significant difference between the variation in data of twosamples taken from the same parent population." i.e. σ1 = σ2

The rejection of the null hypothesis indicates that the differences have statistical significance and the acceptance of null hypothesis indicates that the differences are due to chance and arised because of sampling fluctuation. Since many practical problems aim at establishment of statistical significance of differences, rejection of the null hypothesis may thus indicate success in statistical project.

ii. **Alternative Hypothesis**: As against the null hypothesis, the alternative hypothesis specifies those values that the researcher believes to hold true, and, of course, he hopes that the sample data lead to acceptance of this hypothesis as true.

Rejection of Null hypothesis $H_O$ leads to the acceptance of alternative hypothesis,which is denoted by $H_1$.

With respect to the three null hypotheses as stated above, researcher might establish the following alternative hypotheses:

Thus H1 (1): The extra coaching has benefitted the students.

H1 (2): The drug used under experimentation is effective in curing Malaria

$H_1$ (3): “There is significant difference between the variation in data of two samples taken from the same parent population.” i.e. $\sigma_1 \neq \sigma_2$

*The null and alternative hypothses can also be written as :*

$$H_0: \ (\sigma_1 - \sigma_2 = 0)$$

$$H_1: \ (\sigma_1 - \sigma_2 \neq 0)$$

$$H_O : \mu_1 - \mu_2 = 0 H_1: \mu_1 - \mu_2 \neq 0$$

## Type I and Type II Errors

When two hypotheses are set up, the acceptance or rejection of a null hypothesis is based on a sample study. While we make a decision on the basis of the data analysis and testing of the significance difference, it may lead to wrong conclusions in two ways

(i.) Rejecting a true null hypothesis

(ii) Accepting $a$ false hypothesis. This can be presented in the following table:

| | **Decision** | |
|---|---|---|
| | **Accepted $H_O$** | **Rejected $H_O$** |
| **$H_O$ true** | Correct Decision | Type I error<br>(α error) |
| **$H_O$ false** | Type II error<br>(β error) | Correct Decision |

By rewriting;

Reject $H_O$ when it is true (Type I error) = α

Accept $H_O$ when it is false (Type II error) = β

Accept $H_O$= when it is true (Correct decision)

Reject $H_O$= when it is false (Correct decision)

- **Setting up a Suitable Significance Level:**

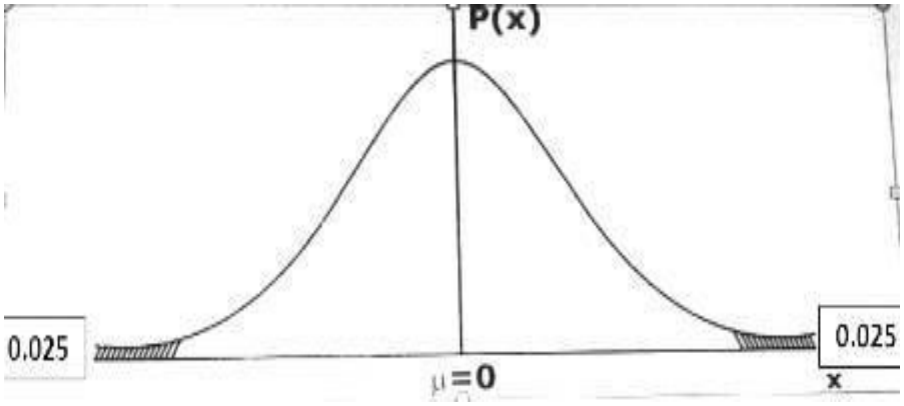
The maximum possibility of committing type I error, which we use to specify in a test, is known as the level of significance. Generally, 5% level of significance is fixed in statistical tests. This implies that we can have 95% confidence in acceptinga hypothesis or we could be wrong 5% in taking the decision.

The range of variation has two regions-acceptance region and critical region or rejection region. If the sample statistic falls in critical region, we reject the hypothesis, as it leads to false decision. We go with

H1, if the computed value of sample statistic falls in the rejection region.

The critical region under a normal curve, as stated earlier can be divided into two ways; (a) two sides under a curve (Two Tailed Test) (b) one side under a curve; either on the right tail or left tail (One Tailed Test).

Acceptance and rejection regions in case of a two-tailed Test (with 5% significance level)



- **Setting a Test Criterion:**

The third step in hypothesis testing procedure is to construct a test criterion. This involves selecting an appropriate probability distribution for the particular test, that is, a probability distribution which can properly be applied. Some probability distributions that are commonly used in testing procedures are Z, t, F and $\chi^2$.

- **Computation:**

After completing first three steps we have completely designed a statistical test. We now proceed to the fourth step- computing various measures from a random sample of size n, which are necessary for applying the test. These calculations include the test statistic and the standard error of the test statistic.

- **Making a decision or Conclusion:**

Finally we come to a conclusion stage where either we accept or reject the null hypothesis. The decision is based on computed value of test statistic, whether it lies in the acceptance region or rejection region.

If the computed value of the test statistic falls in the acceptance region (it means computed value is less than critical value), the null hypothesis is accepted. On the contrary, if the computed value of the test statistic is greater than the critical value, thecomputed value of the statistic falls in the rejection region and the null hypothesis is rejected.

**Tests of Hypotheses**

Hypothesis testing determines the validity of the assumption (technically described as null hypothesis) with a view to choose between two conflicting hypotheses about the value of a population parameter. Hypothesis testing helps to decide on the basis of a sample data, whethera hypothesis about the population is likely to be true or false. Statisticians have developed several tests of hypotheses (also known as the tests of significance) for the purpose of testing of hypotheses which can be classified as:

a) Parametric tests or standard tests of hypotheses; and

b) Non-parametric tests or distribution-free test of hypotheses.

Parametric tests usually assume certain properties of the parent population from which we drawsamples. Assumptions like observations come from a normal population, sample size is large, assumptions about the population parameters like mean, variance, etc., must hold good before parametric tests can be used. But there are situations when the researcher cannot or does not want to make such assumptions. In such situations we use statistical methods for testing hypotheses which are called non-parametric tests because such tests do not depend on any assumption about the parameters of the parent population. Besides, most non-parametric testsassume only nominal or ordinal data, whereas parametric tests require measurement equivalent to at least an interval scale. As a result, non-parametric tests need more observations than parametric tests to achieve the same size of Type I and Type II errors.

**Non parametric Tests**

Non parametric tests are used when the data isn't normal. Therefore, the key is to figure out ifyou have normally distributed data. The only non-parametric test you are likely to come acrossin elementary stats is the chi-square test. However, there are several others.

## 5.1 Chi-Square Test ($\chi^2$)

The chi-square test is an important test amongst the several tests of significance developedby statisticians. Chi-square, symbolically written as $\chi^2$ (Pronounced as Ki-square), is a statistical measure used in the context of sampling analysis for comparing a variance to atheoretical variance. As a non-parametric test, it "can be used to determine if categoricaldata shows dependency or the two classifications are independent. It can also be used to make comparisons between theoretical populations and actual data when categories are used." Thus, the chi-square test is applicable in large number of problems. The test is, in fact, a technique through the use of which it is possible for all researchers to

- test the goodness of fit;
- test the significance of association between two attributes, and

- test the homogeneity or the significance of population variance.

**Chi-Square Test for Goodness of Fit**

The Chi-Square Test for Goodness of Fit tests claims about population proportions. It is a non-parametric test that is performed on categorical (nominal or ordinal) data.

**Illustration: 1**

In the 2000 US Census, the ages of individuals in a smalltown were found to be the following:

| Less than 18 | 18–35 | Greater than 35 |
|---|---|---|
| 20% | 30% | 50% |

In 2010, ages of n = 500 individuals from the same small town were sampled. Below are the results:

| Less than 18 | 18–35 | Greater than 35 |
|---|---|---|
| 121 | 288 | 91 |

Using 5% level of significance (alpha = 0.05), would you conclude that the population distribution of ages has changed in the last 10 years?

**Solution:**

Using our sample size and expected percentages, we can calculate how many people we expected to fall within each range. We can then make a table separating observed values versus expected values:

| | Less than 18 | 18–35 | Greater than 35 |
|---|---|---|---|
| Expected | 20% | 30% | 50% |

| | Less than 18 | 18–35 | Greater than 35 |
|---|---|---|---|
| Observed | 121 | 288 | 91 |
| Expected | 500*0.20 = 100 | 500*0.30 = 150 | 500*50 = 250 |

| | Less than 18 | 18-35 | Greater than 35 |
|---|---|---|---|
| Observed | 121 | 288 | 91 |
| Expected | 100 | 150 | 250 |

Let's perform a hypothesis test on this new table to answer the original question.

**Steps for Chi-Square Test for Goodness of Fit**

1. Define Null and Alternative Hypotheses
2. State Alpha

3. Calculate Degrees of Freedom
4. State Decision Rule
5. Calculate Test Statistic
6. State Results
7. State Conclusion

**1. Define Null and Alternative Hypotheses**

$H_0$; the data meet the expected distribution

$H_1$; the data do not meet the expected distribution

**1. State Alpha**

Alpha = 0.05

**2. Calculate Degrees of Freedom**

df = k – 1, where k = your number of groups.df = 3 – 1 = 2

**3. State Decision Rule**

Using our alpha and our degrees of freedom, who look up a critical value in the Chi-Square Table. We find our critical value to be 5.99.

If $\chi^2$ is greater than 5.99, reject $H_0$.

**4. Calculate Test Statistic**

The Chi-Square statistic is found using the following equation, where observed values are compared to expected values:

| | **Less than 18** | **18–35** | **Greater than 35** |
|---|---|---|---|
| Observed | 121 | 288 | 91 |
| Expected | 100 | 150 | 250 |

$\chi^2 = \sum(O-E)^2/O$

$\chi^2 = (121 - 100)^2 / 100 + (288 - 150)^2 /150 + (91 - 250)^2 /250$

$\chi^2 = 232.494$

**6. State Results**

If $\chi^2$ is greater than 5.99, reject $H_0$.

$\chi^2$ = 232.494

Reject the null hypothesis.

**7. State Conclusion**

The ages of the 2010 population are different than those expected based on the2000 population.

**Chi-Square Test for Independence**

The Chi-Square Test for Independence evaluates the relationship between two variables. It is a non-parametric test that is performed on categorical (nominal or ordinal) data.

**Illustration: 2**

500 elementary school boys and girls are asked which their favourite colour is: blue, green, or pink. Results are shown below:

| | **Blue** | **Green** | **Pink** | |
|---|---|---|---|---|
| Boys | 100 | 150 | 20 | 300 |
| Girls | 20 | 30 | 180 | 200 |
| | 120 | 180 | 200 | N = 500 |

Using 5% level of significance (alpha = 0.05), would you conclude that there is a relationship between genderand favourite colour?

Let’s perform a hypothesis test to answer this question.

**Steps for Chi-Square Test for Independence**

1. Define Null and Alternative Hypotheses
2. State Alpha
3. Calculate Degrees of Freedom
4. State Decision Rule
5. Calculate Test Statistic
6. State Results
7. State Conclusion

**1. Define Null and Alternative Hypotheses**

$H_0$; For the population of elementary school students, gender and favouritecolour are not related.

$H_1$; For the population of elementary school students, gender and favouritecolour are related.

**2. State Alpha**

Alpha = 0.05

**3. Calculate Degrees of Freedom**df = (rows – 1)(columns – 1)df = (2 – 1)(3 – 1)
df = (1)(2) = 2

**4. State Decision Rule**

Using our alpha and our degrees of freedom, who look up a critical value in the Chi-Square Table. We find our critical value to be 5.99.

If $\chi^2$ is greater than 5.99, reject $H_0$.

**5. Calculate Test Statistic**

First, we need to calculate our expected values using the equation below. We find the expected values by multiplying each row total by each column total, and then diving by the total number of subjects. The calculations are shown.

EXPECTED = ROW TOTAL * COLUMN TOTAL / GRAND TOTAL

E(Boys, Blue) = 300*120/500 = 72

E(Boys, Green) = 300*180/500 = 108

E(Boys, Pink) = 300*200/500 = 120

E(Girls, Blue) = 200*120/500 = 48

E(Girls, Green) = 200*180/500 = 72

E(Girls, Pink) = 200*200/500 = 80

| **Expected** | **Blue** | **Green** | **Pink** | |
|---|---|---|---|---|
| Boys | 72 | 108 | 120 | 300 |
| Girls | 48 | 72 | 80 | 200 |
| | 120 | 180 | 200 | N = 500 |

**6. State Results**

If $\chi^2$ is greater than 5.99, reject $H_0$.

$\chi^2$ = 266.389

Reject the null hypothesis

**7.** State Conclusion

In the population, there is a relationship between gender and favourite colour.

**Important parametric tests**

The important parametric tests are: (1) *t*-test; and (2) *F*-test. All these tests are based on the assumption of normality i.e., the source of data is considered to be normally distributed.

**t- test:** It is based on t-distribution and is considered an appropriate test for judging the significance of a sample mean or for judging the significance of difference between the means of two samples in case of small sample(s) when population variance is not known (in which case we use variance of the sample as an estimate of the population variance). In case two samples are related, we use paired t-test (or what is known as difference test) for judging the significance of the mean of difference between the two related samples. It can also be used for judging the significance of the coefficients of simple and partial correlations.

**F-test**: It is based on *F*-distribution and is used to compare the variance of the two- independent samples. This test is also used in the context of analysis of variance (ANOVA) for judging the significance of more than two sample means at one and the same time. It is also used for judging the significance of multiple correlation coefficients.

## 5.2 ‘t’- test

‘t’ Test was developed by Gossett around 1900. He published his theoretical ideas about this test in the pen name of ‘Student’ and so this test in also called Student Test. Statistical method for the comparison of the mean of the two groups of the normally distributed sample(s).

It is used when:

- Population parameter (mean and standard deviation) is not known
- Sample size (number of observations) < 30

**Type of t-test**

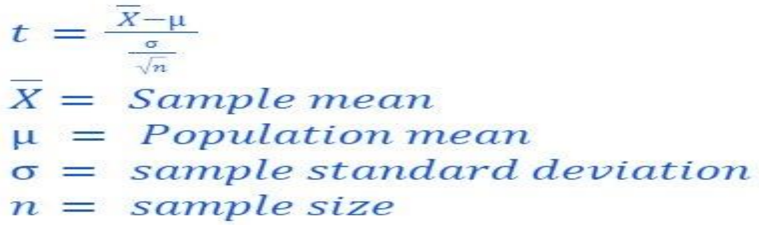
The T-test is mainly classified into 3 parts:

- One sample
- Independent sample
- Paired sample

**One Sample**

In one sample t-test, we compare the sample mean with the population mean.

**Formula:**



$$t = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

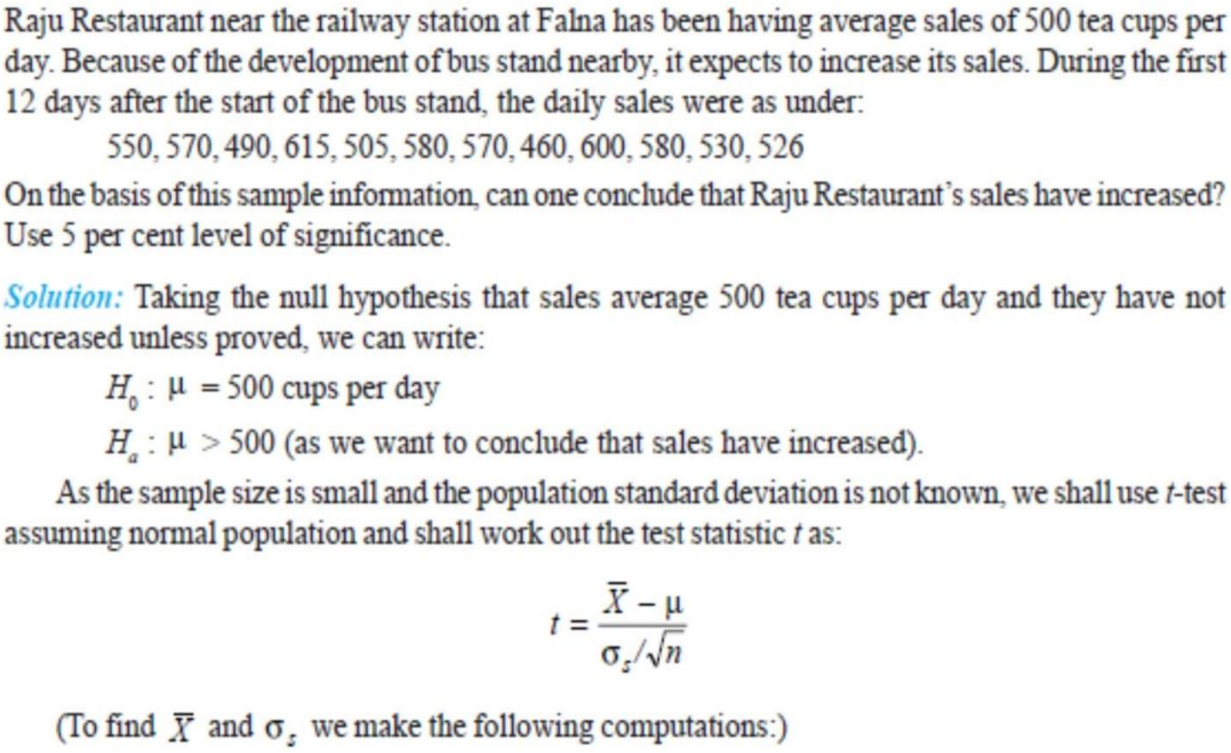
$\overline{X}$ = *Sample mean*

$\mu$ = *Population mean*

$\sigma$ = *sample standard deviation*

$n$ = *sample size*

**Illustration: 3**



Raju Restaurant near the railway station at Falna has been having average sales of 500 tea cups per day. Because of the development of bus stand nearby, it expects to increase its sales. During the first 12 days after the start of the bus stand, the daily sales were as under:

550, 570, 490, 615, 505, 580, 570, 460, 600, 580, 530, 526

On the basis of this sample information, can one conclude that Raju Restaurant's sales have increased? Use 5 per cent level of significance.

*Solution:* Taking the null hypothesis that sales average 500 tea cups per day and they have not increased unless proved, we can write:

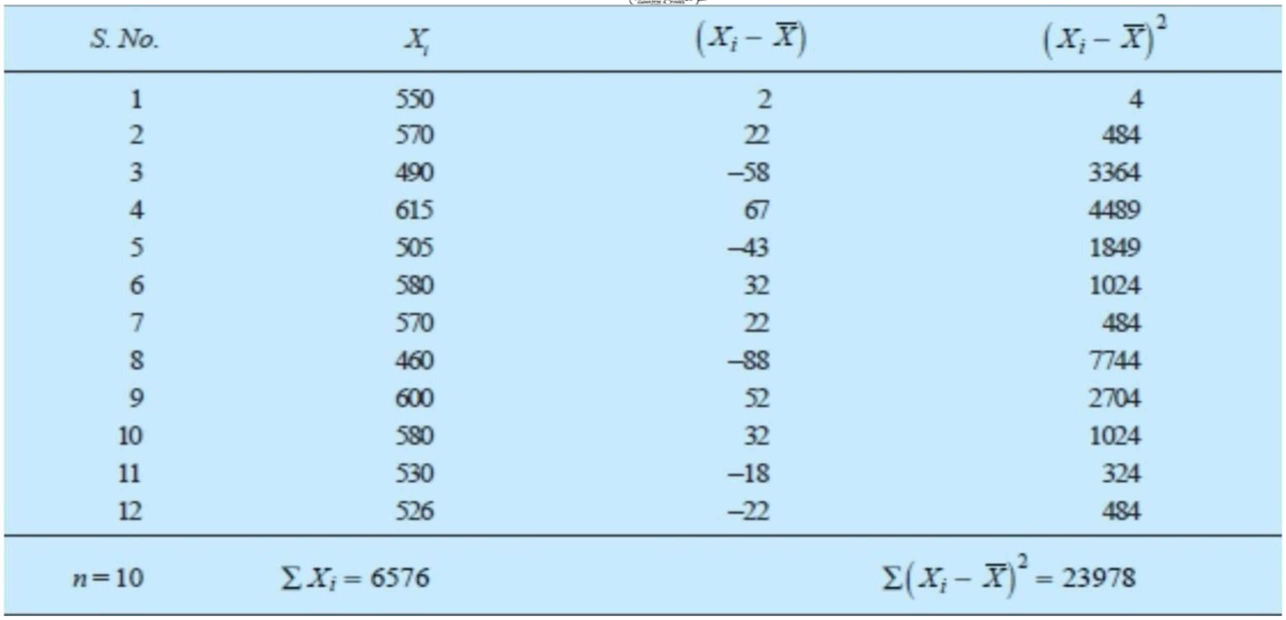
$H_0 : \mu = 500$ cups per day

$H_a : \mu > 500$ (as we want to conclude that sales have increased).

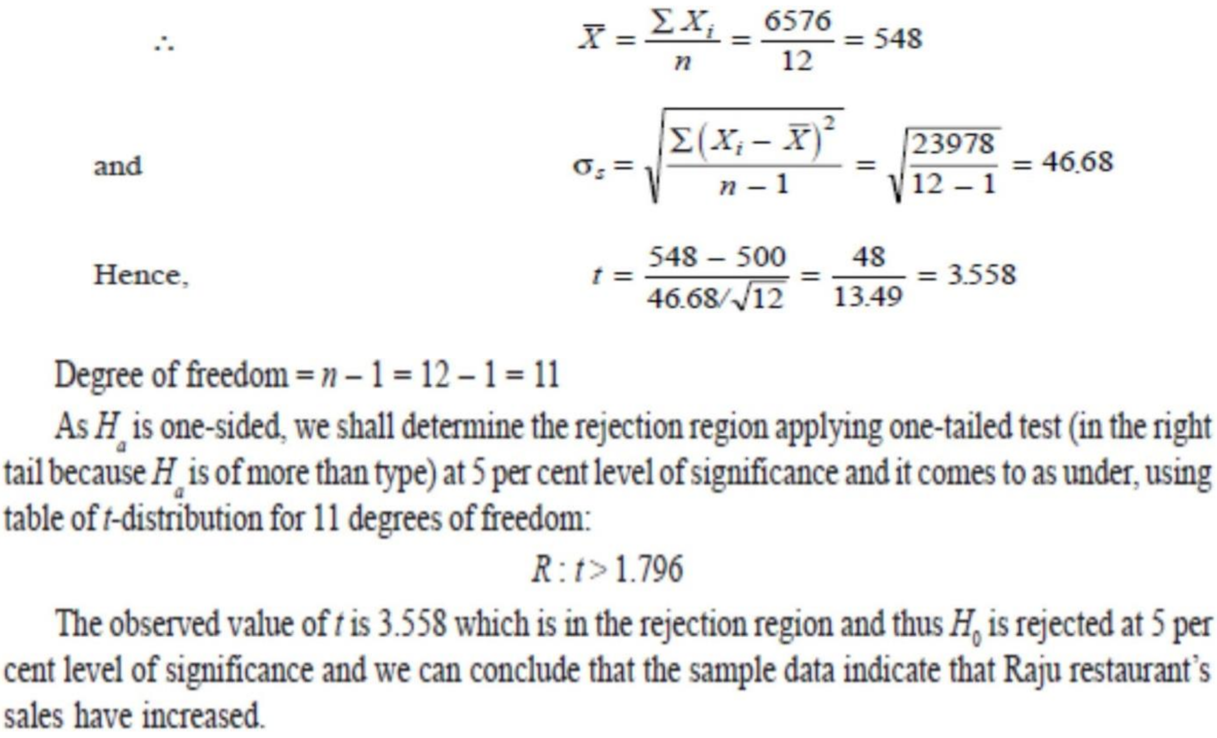
As the sample size is small and the population standard deviation is not known, we shall use *t*-test assuming normal population and shall work out the test statistic *t* as:

$$t = \frac{\overline{X} - \mu}{\sigma_s / \sqrt{n}}$$

(To find $\overline{X}$ and $\sigma_s$ we make the following computations:)



| S. No. | $X_i$ | $(X_i - \overline{X})$ | $(X_i - \overline{X})^2$ |
|---|---|---|---|
| 1 | 550 | 2 | 4 |
| 2 | 570 | 22 | 484 |
| 3 | 490 | –58 | 3364 |
| 4 | 615 | 67 | 4489 |
| 5 | 505 | –43 | 1849 |
| 6 | 580 | 32 | 1024 |
| 7 | 570 | 22 | 484 |
| 8 | 460 | –88 | 7744 |
| 9 | 600 | 52 | 2704 |
| 10 | 580 | 32 | 1024 |
| 11 | 530 | –18 | 324 |
| 12 | 526 | –22 | 484 |
| $n = 10$ | $\Sigma X_i = 6576$ | | $\Sigma(X_i - \overline{X})^2 = 23978$ |



$\therefore$
$$\overline{X} = \frac{\Sigma X_i}{n} = \frac{6576}{12} = 548$$

and
$$\sigma_s = \sqrt{\frac{\Sigma(X_i - \overline{X})^2}{n-1}} = \sqrt{\frac{23978}{12-1}} = 46.68$$

Hence,
$$t = \frac{548 - 500}{46.68/\sqrt{12}} = \frac{48}{13.49} = 3.558$$

Degree of freedom = $n - 1 = 12 - 1 = 11$

As $H_a$ is one-sided, we shall determine the rejection region applying one-tailed test (in the right tail because $H_a$ is of more than type) at 5 per cent level of significance and it comes to as under, using table of $t$-distribution for 11 degrees of freedom:

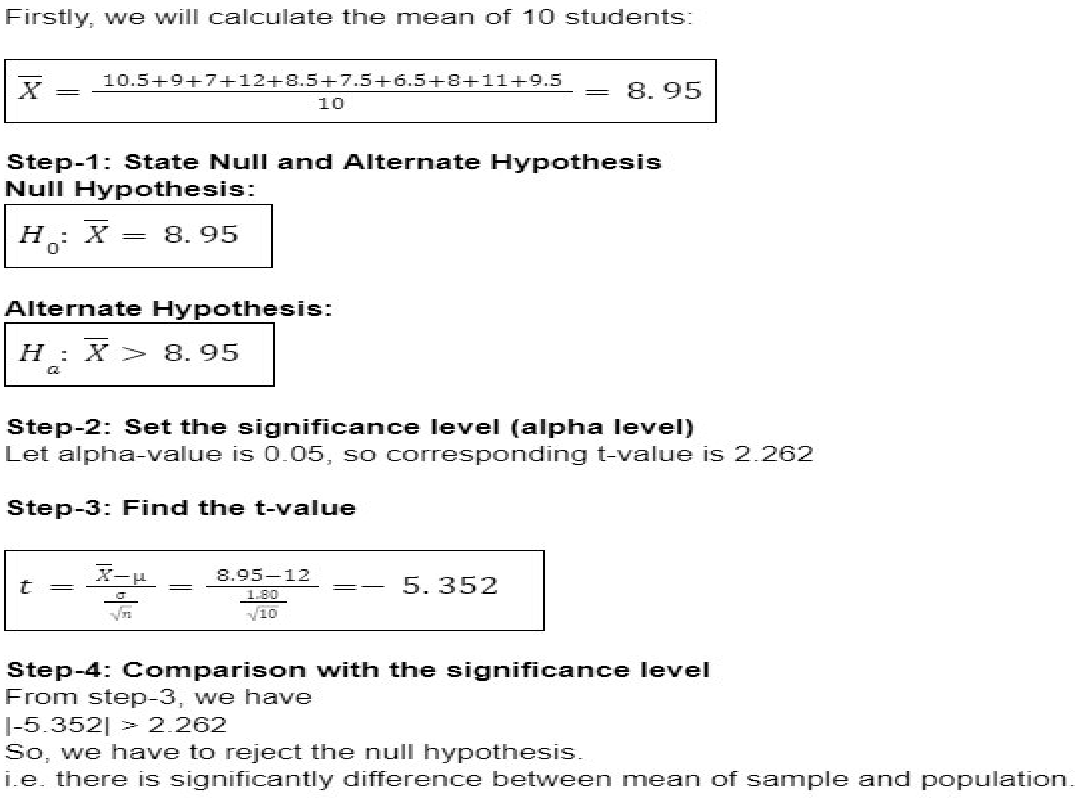
$$R : t > 1.796$$

The observed value of $t$ is 3.558 which is in the rejection region and thus $H_0$ is rejected at 5 per cent level of significance and we can conclude that the sample data indicate that Raju restaurant's sales have increased.

**Illustration: 4**

Marks of student are 10.5, 9, 7, 12, 8.5, 7.5, 6.5, 8, 11 and 9.5. Mean population score is 12 and standard deviation is 1.80. Is the mean value for student significantly differing from the mean population value?

**Solution:**



Firstly, we will calculate the mean of 10 students:

$$\boxed{\overline{X} = \frac{10.5+9+7+12+8.5+7.5+6.5+8+11+9.5}{10} = 8.95}$$

**Step-1: State Null and Alternate Hypothesis**
**Null Hypothesis:**

$$\boxed{H_0: \overline{X} = 8.95}$$

**Alternate Hypothesis:**

$$\boxed{H_a: \overline{X} > 8.95}$$

**Step-2: Set the significance level (alpha level)**
Let alpha-value is 0.05, so corresponding t-value is 2.262

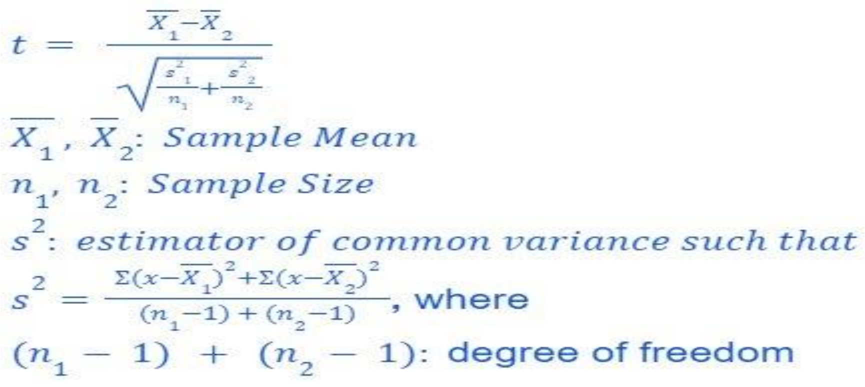
**Step-3: Find the t-value**

$$\boxed{t = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{8.95-12}{\frac{1.80}{\sqrt{10}}} =- 5.352}$$

**Step-4: Comparison with the significance level**
From step-3, we have
|-5.352| > 2.262
So, we have to reject the null hypothesis.
i.e. there is significantly difference between mean of sample and population.

**Independent (two-sample t-test):**

In this test, we compare the means of two different samples. The formula is



$$t = \frac{\overline{X}_1-\overline{X}_2}{\sqrt{\frac{s^2_1}{n_1}+\frac{s^2_2}{n_2}}}$$

$\overline{X}_1, \overline{X}_2$: *Sample Mean*

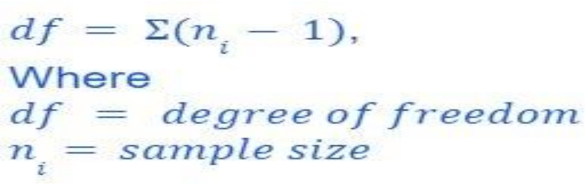
$n_1, n_2$: *Sample Size*

$s^2$: *estimator of common variance such that*

$s^2 = \frac{\Sigma(x-\overline{X}_1)^2+\Sigma(x-\overline{X}_2)^2}{(n_1-1)+(n_2-1)}$, where

$(n_1 - 1) + (n_2 - 1)$: degree of freedom

**Degree of Freedom:** Degree of freedom is defined as the number of independent variables. It is given by:



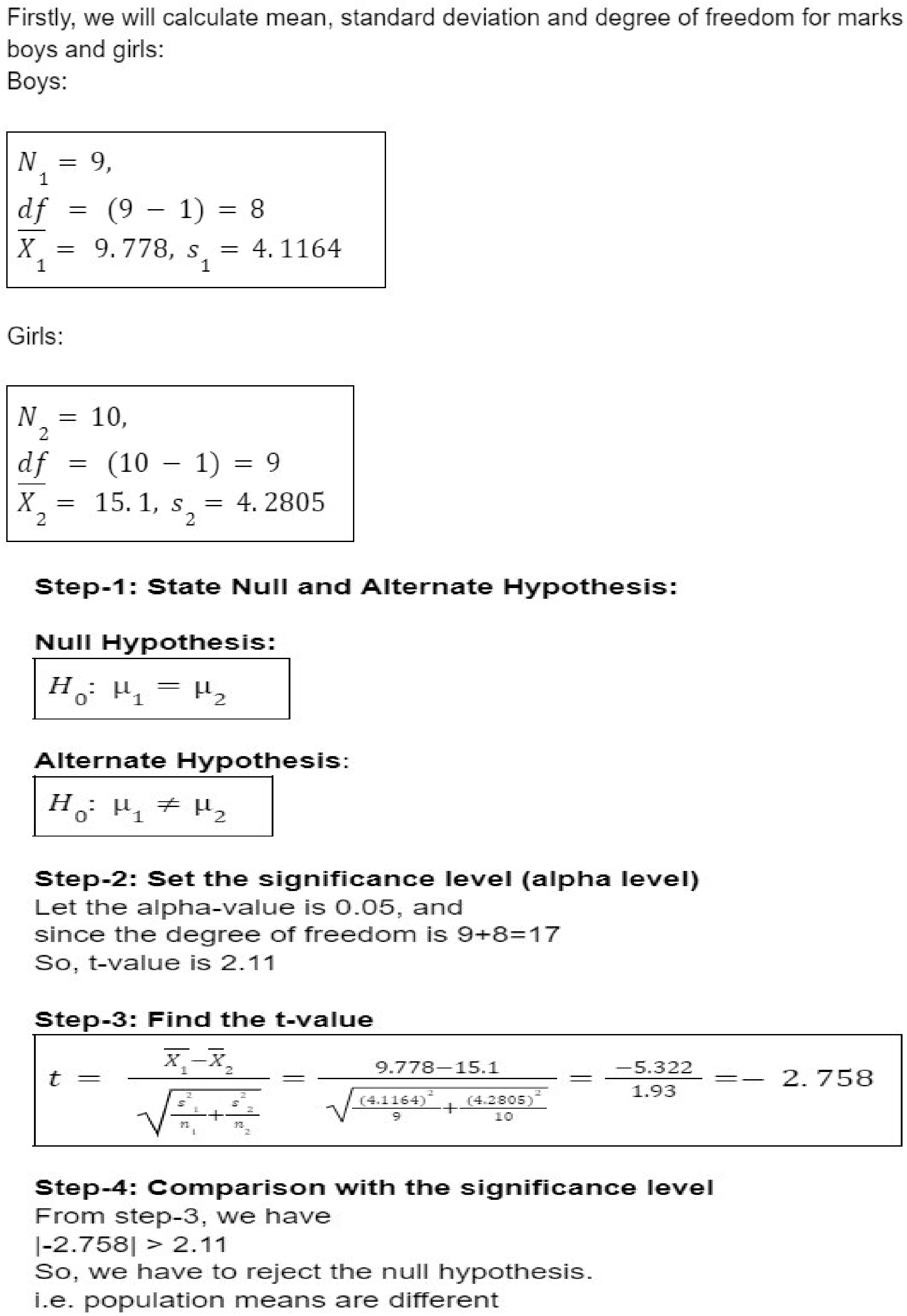
$df = \Sigma(n_i - 1)$,
Where
$df = $ *degree of freedom*
$n_i = $ *sample size*

Let’s understand two-sample t-test by an example:

**Illustration: 5**

The marks of boys and girls are given: Boys: 12, 14, 10, 8, 16, 5, 3, 9, and 11 Girls: 21, 18, 14, 20, 11, 19, 8, 12, 13, and 15. Is there any significant differnece between marks of males and females i.e. population means are different.

**Solution:**



Firstly, we will calculate mean, standard deviation and degree of freedom for marks of boys and girls:

Boys:

$$N_1 = 9,$$
$$df = (9 - 1) = 8$$
$$\overline{X}_1 = 9.778,\ s_1 = 4.1164$$

Girls:

$$N_2 = 10,$$
$$df = (10 - 1) = 9$$
$$\overline{X}_2 = 15.1,\ s_2 = 4.2805$$

**Step-1: State Null and Alternate Hypothesis:**

**Null Hypothesis:**

$$H_0\colon\ \mu_1 = \mu_2$$

**Alternate Hypothesis**:

$$H_0\colon\ \mu_1 \neq \mu_2$$

**Step-2: Set the significance level (alpha level)**

Let the alpha-value is 0.05, and
since the degree of freedom is 9+8=17
So, t-value is 2.11

**Step-3: Find the t-value**

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{9.778 - 15.1}{\sqrt{\frac{(4.1164)^2}{9} + \frac{(4.2805)^2}{10}}} = \frac{-5.322}{1.93} =- 2.758$$

**Step-4: Comparison with the significance level**

From step-3, we have
|-2.758| > 2.11
So, we have to reject the null hypothesis.
i.e. population means are different

**Paired t-test:**

In this test, we compare the means of two related or same group at two different time. Formula:

$$t = \frac{m}{\frac{s}{\sqrt{n}}}$$

$m:$ *mean of difference between each pair of values*

$s:$ *standard deviation of difference between each pair of values*

$n:$ *sample size*

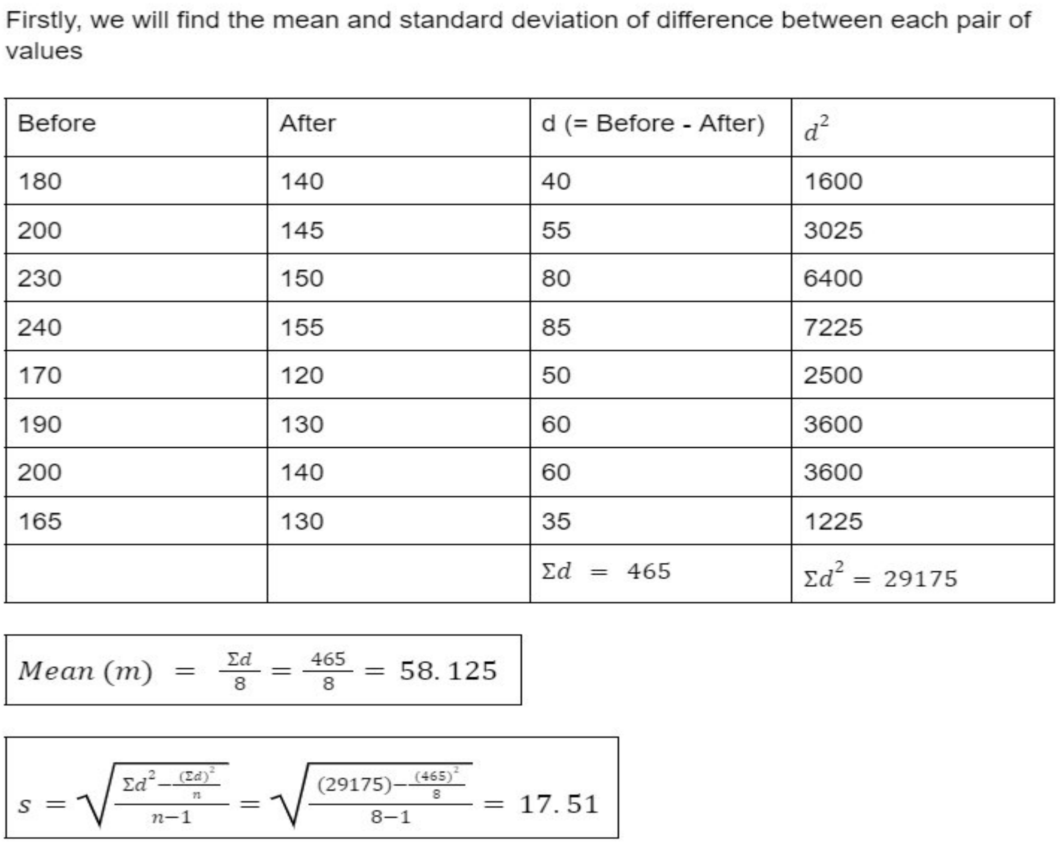
***Degree of freedom is n-1.***

Let's understand two-sample t-test by an example:

**Illustration: 6**

Blood pressures of 8 patients are before and after are recorded: Before: 180, 200, 230, 240, 170, 190, 200, and 165 after: 140, 145, 150, 155, 120, 130, 140, and 130. Is there any significant difference between BP reading before and after?

**Solution:**

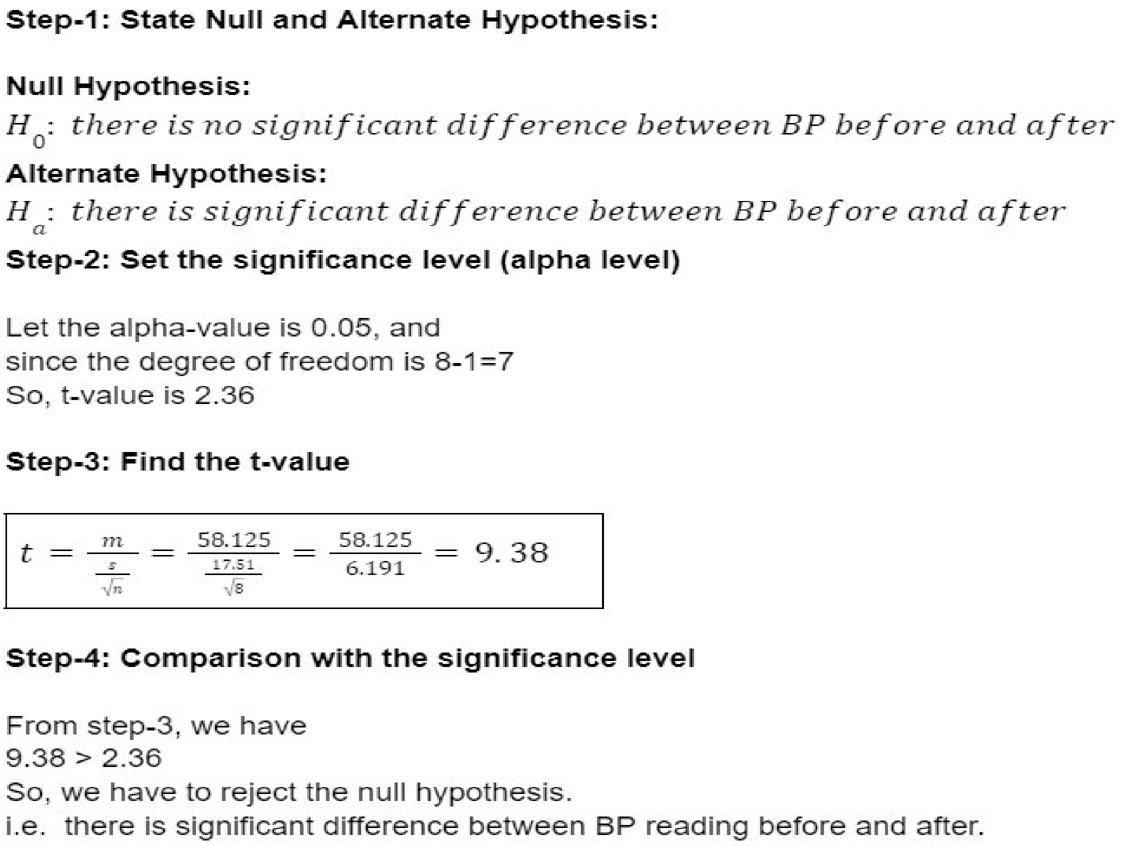


Firstly, we will find the mean and standard deviation of difference between each pair of values

| Before | After | d (= Before - After) | $d^2$ |
|---|---|---|---|
| 180 | 140 | 40 | 1600 |
| 200 | 145 | 55 | 3025 |
| 230 | 150 | 80 | 6400 |
| 240 | 155 | 85 | 7225 |
| 170 | 120 | 50 | 2500 |
| 190 | 130 | 60 | 3600 |
| 200 | 140 | 60 | 3600 |
| 165 | 130 | 35 | 1225 |
| | | $\Sigma d = 465$ | $\Sigma d^2 = 29175$ |

$$\boxed{Mean\ (m) = \frac{\Sigma d}{8} = \frac{465}{8} = 58.125}$$

$$\boxed{s = \sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n-1}} = \sqrt{\frac{(29175) - \frac{(465)^2}{8}}{8-1}} = 17.51}$$



**Step-1: State Null and Alternate Hypothesis:**

**Null Hypothesis:**

$H_0$: *there is no significant difference between BP before and after*

**Alternate Hypothesis:**

$H_a$: *there is significant difference between BP before and after*

**Step-2: Set the significance level (alpha level)**

Let the alpha-value is 0.05, and
since the degree of freedom is 8-1=7
So, t-value is 2.36

**Step-3: Find the t-value**

$$\boxed{t = \frac{m}{\frac{s}{\sqrt{n}}} = \frac{58.125}{\frac{17.51}{\sqrt{8}}} = \frac{58.125}{6.191} = 9.38}$$

**Step-4: Comparison with the significance level**

From step-3, we have
9.38 > 2.36
So, we have to reject the null hypothesis.
i.e. there is significant difference between BP reading before and after.

## 5.3 F Tests

F-tests are named after the name of Sir Ronald Fisher. The F-statistic is simply a ratio of two variances. Variance is the square of the standard deviation. For a common person, standard deviations are easier to understand than variances because they're in the same units as the data rather than squared units. F-statistics are based on the ratio of mean squares. The term "mean squares" may sound confusing but it is simply an estimate of population variance that accounts for the degrees of freedom used to calculate that estimate. For carrying out the test of significance, we calculate the ratio F, which is defined as:

$F = \frac{S_1^2}{S_2^2}$ , where $S_1^2 = \frac{(X_1 - \bar{X}_1)^2}{n_1 - 1}$

And $S_2^2 = \frac{(X_2 - \bar{X}_2)^2}{n_2 - 1}$

It should be noted that $S_1^2$ is always the larger estimate of variance, i.e., $S_1^2 > S_2^2$

$F = \frac{\textit{Larger estimate of variance}}{\textit{Smaller estimate of variance}}$

$\nu_1 = n_1$-1 and $\nu_2 = n_2$-1

$\nu_1$= degrees of freedom for sample having larger variance.

$\nu_2$= degrees of freedom for sample having smaller variance.

The calculated value of F is compared with the table value for $v_1$ □and $v_2$ at 5% or 1% level of significance. If calculated value of F is greater than the table value then the F ratio is considered significant and the null hypothesis is rejected. On the other hand, if the calculated value of F is less than the table value the null hypothesis is accepted and it is inferred that both the samples have come from the population having same variance.

**Illustration: 7**

Two random samples were drawn from two normal populations and their values are:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| A | 65 | 66 | 73 | 80 | 82 | 84 | 88 | 90 | 92 | | |
| B | 64 | 66 | 74 | 78 | 82 | 85 | 87 | 92 | 93 | 95 | 97 |

Test whether the two populations have the same variance at the 5% level of significance.(Given: F=3.36 at 5% level for $v_1$=10 and $v_2$=8.)

**Solution:** Let us take the null hypothesis that the two populations have not the same variance. Applying F-test:

$$F=\frac{S_1^2}{S_2^2}$$

| A<br>$X_1$ | $(X_1\text{-}\bar{X}_1)$<br>$x_1$ | $x_1^2$ | B<br>$X_2$ | $(X_2\text{-}\bar{X}_2)$<br>$x_2$ | $x_2^2$ |
|---|---|---|---|---|---|
| 65 | -15 | 225 | 64 | -19 | 361 |
| 66 | -14 | 196 | 66 | -17 | 289 |
| 73 | -7 | 49 | 74 | -9 | 81 |
| 80 | 0 | 0 | 78 | -5 | 25 |
| 82 | 2 | 4 | 82 | -1 | 1 |
| 84 | 4 | 16 | 85 | 2 | 4 |
| 88 | 8 | 64 | 87 | 4 | 16 |
| 90 | 10 | 100 | 92 | 9 | 81 |
| 92 | 12 | 144 | 93 | 10 | 100 |
| | | | 95 | 12 | 144 |
| | | | 97 | 14 | 196 |
| $\sum X_1$= 720 | $\sum x_1$=0 | $\sum x_1^2$-=798 | $\sum X_2$=913 | $\sum x_2$=0 | $\sum x_2^2$=1298 |

$$\bar{X}_1=\frac{\sum X_1}{n_1}=\frac{720}{9}=80;$$

$$\bar{X}_2=\frac{\sum X_2}{n_2}=\frac{913}{11}=83$$

$$S_1^2=\sum x_1^2/n_1\text{-}1=\frac{798}{9-1}=99.75$$

$$S_2^2=\sum x_2^2/n_2\text{-}1=\frac{734}{11-1}=129.8$$

$$F=\frac{S_1^2}{S_2^2}=\frac{99.75}{129.8}=0.768$$

At 5 percent level of significance, for $v_1$=10 and $v_2$=8, the table value of $F_{0.05 = 3.36.}$ The calculated value of F is less than the table value. The hypothesis is accepted. Hence the twopopulations have not the same variance.

## 5.4 Analysis of Variance (ANOVA)

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. The method is based upon an unusual result that the equality of several population means can be tested by comparing the sample variances using F distribution. In t statistic we test whether two population means are equal. The analysis of variance is an extension of the t test for the case of more than two means.

### One-Way ANOVA Versus Two-Way ANOVA

There are two types of ANOVA: one-way (or unidirectional) and two-way. One-way or two- way refers to the number of independent variables in the analysis of variance test. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. It determines whether all the samples are the same. The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.

A two-way ANOVA is an extension of the one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents. For example, a two-way ANOVA allows a company to compare worker productivity based on two independent variables, such as salary and skill set. It is utilized to observe the interaction between the two factors and tests the effect of two factors at the same time.

### Example of One Way ANOVA

**Illustration: 8**

The following table shows the retail prices (Rs. per kg.) of a commodity in some shops selected at random in four cities:

| A | B | C | D |
|---|---|---|---|
| 34 | 29 | 27 | 34 |
| 37 | 33 | 29 | 36 |
| 32 | 30 | 31 | 38 |
| 33 | 34 | 28 | 35 |

Carry out the analysis of variance to test the significance of the differences between prices of the commodity in the four cities. [Given, $F_{0.05}$ = 3.49 for (3, 12) degrees of freedom].

**Solution:**

Each observation is reduced by 39, and shown below:Calculation for Analysis of Variance

| A | B | C | D |
|---|---|---|---|
| -5 | -10 | -12 | -5 |
| -2 | -6 | -10 | -3 |
| -7 | -9 | -8 | -1 |
| -6 | -5 | -11 | -4 |

| Total | $T_1$= -20 | $T_2$= -30 | $T_3$= -41 | $T_4$= -13 | T = -104 |
|---|---|---|---|---|---|
| Total of Squares | 114 | 242 | 429 | 51 | $\sum\sum xij^2$= 836 |
| sample size | $n_1$= 4 | $n_2$= 4 | $n_3$= 4 | $n_4$= 4 | N= 16 |

Correction Factor (C.F.)= $T^2/N = (-104)^2/16 = 10816/16 = 676$

Total Sum of Squares (SS) = $\sum\sum xij^2$ - C.F. = 836- 676 = 160

Sum of Squares Between Groups (SSB) = $\sum (T_i^2 /n_i)$ - C.F.

$= (-20^2/4 + -30^2/4 +-41^2/4 + -13^2/4) – 676$

= 787.50- 676

= 111.50

Sum of Squares due to Errors (SSE) = Total SS –SSB = 160-111.50 = 48.50

**Analysis of Variance Table**

| **Source of Variation** | **S.S** | **d.f** | **Mean Squares (M.S)** | **Observed F Value** | **Tabulated F Value** |
|---|---|---|---|---|---|
| Between Groups | 111.50 | (k-1)= 4-1 = 3 | 111.50/3= 37.17 | MSB/MSE= 37.17/4.04= 9.196 | $F_{0.05}$ = 3.49 for (3,12) d.f |
| Within Groups (Errors) | 48.50 | (N-K)= 16-4= 12 | 48.50/12= 4.04 | | |
| Total | 160 | 16-1 = 15 | | | |

Since the observed value of F (i.e., 9.196) exceeds the 5% tabulated value (i.e., 3.49) for (3,12) d.f., we reject the null hypothesis of equality of population means, and conclude that the retail prices of the commodity in the four cities are not equal.

In order to test which of the cities differ in prices, we calculate the critical difference (C.D.).

C.D. = $s\sqrt{2n}\ t_{0.025}$ (for 12d.f.)$s^2$ = MSE

$s = \sqrt{MSE} = \sqrt{4.04}$

C.D. = √4.04* √2*4 * 2.18 = 12.39

The sample totals (of the reduced observations) are $T_1$= -20, $T_2$= -30, $T_3$= -41, $T_4$= -13

We have

$| T_1 - T_2| = 10$

$| T_1 - T_3| = 21$

$| T_1 - T_4| = 7$

$| T_2 - T_3| = 11$

$| T_2 - T_4| = 17$

$| T_3 - T_4| = 28$

Comparing these figures with the C.D. (i.e., 12.39) we find that the cities A and C, B and D,C and D differ in prices. Cities A and B and A and D may be taken to be having same prices.

**Example of Two Way ANOVA**

**Illustration: 9**

The following table gives the estimates of acreage of cultivable land but not cultivated land out of 100 acres of total land, as obtained by three investigators in each of three districts. Perform an analysis of variance to test whether there are significant differencesbetween investigators and districts. [Given $F_{0.05}$ = 6.94 for d.f. (2, 4)]

| **Investigator** | **District** | | |
|---|---|---|---|
| | **I** | **II** | **III** |
| **A** | 23 | 28 | 26 |
| **B** | 24 | 25 | 27 |
| **C** | 24 | 22 | 26 |

**Solution**

Each observation is reduced by 24, and shown below: Calculation for Analysis of Variance

| Investigator | District | | | Total |
|---|---|---|---|---|
| | I | II | III | |
| A | -1 | 4 | 2 | 5 |
| B | 0 | 1 | 3 | 4 |
| C | 0 | -2 | 2 | 0 |
| Total ($T_i$) | -1 | 3 | 7 | T= 9 |
| sample size | $n_1$= 3 | $n_2$= 3 | $n_3$= 3 | N = 9 |

Total of the squares of all figures

$\sum\sum xij^2 = (-1)^2 + (4)^2 + (2)^2 + (0)^2 + (1)^2 + (3)^2 + (0)^2 + (-2)^2 + (2)^2 = 39$

Correction Factor (C.F.) = $T^2/N = (9)^2/9 = 81/9 = 9$

Total Sum of Squares (SS) = $\sum\sum xij^2$ - C.F. = 39-9 = 30

Sum of Squares (SS) between Investigators = $(T_1^2 + T_1^2 + T_1^2)/3$ – C.F

= $(5^2 + 4^2 + 0^2)/3 - 9$

= 41/3 – 9

=4.67

Sum of Squares (SS) between Districts = $(T_1^{'2} + T_1^{'2} + T_1^{'2})/3$ – C.F.

= $[(-1)^2 + (3)^2 + (7)^2]/3 - 9$

= 59/3 – 9

= 10.67

SS due to Error = Total SS – (SS between investigators) - (SS between districts)

= 30 – 4.67 – 10.67

= 14.66

**Analysis of Variance Table**

| **Source of Variation** | **S.S** | **d.f** | **Mean Squares(M.S)** | **Observed F Value** | **Tabulated F Value** |
|---|---|---|---|---|---|
| Between Investigators | 4.67 | 3-1 = 2 | 4.67/2 = 2.34 | 2.34/3.67 = 0.64 | F0.05 = 6.94 for (2,4) d.f |
| Between Districts | 10.67 | 3-1 = 2 | 10.67/2 = 5.34 | 5.34/3.67 = 1.46 | F0.05 = 6.94 for (2,4) d.f |
| Within Groups (Errors) | 14.66 | 4 | 14.66/4 = 3.67 | | |
| Total | 30 | 9-1 = 8 | | | |

Since the observed value of F for experimenters (i.e., 0.64) is less than the corresponding tabulated value (i.e., 6.94) for d.f. (2, 4), it is not significant at 5% level. We conclude that the mean acreage of cultivable land in the three districts as determined by the three investigators may not be different from one another, i.e., there are no significant differences between investigators.

Since the observed value of F for districts (i.e., 1.46) is less than the corresponding tabulated value (i.e., 6.94) for d.f (2,4), it is not significant at 5% level. We conclude that the estimates of acreage of cultivable land in the three districts may not be different from one another, i.e., there are no significant differences between districts.

## Important questions

**Choose the correct answer**

1. Which of these distributions is used for a testing hypothesis?

a) Normal Distribution **b) Chi-Squared Distribution**

c) Gamma Distribution d) Poisson Distribution

2. What is the chi square test used for statistics?

a) comparing means of two groups b) testing the significance of correlation

**c) comparing proportions of categorical variables** d) Analyzing the variables of continuous data

3. What type of data suitable for chi-square test?

**a) Categorical data** b) Interval data c) Ordinal data d) Continuous data

4. A t-test is a significance test that assesses

**a) The means of two independent groups** b) The medians of two dependent groups

c) The modes of two independent variables d) The deviation of three independent variables

5. To use a t-test, the dependent variable must have _______ data.

a) Nominal or interval b) Ordinal or ratio **c) Interval or ratio** d) Ordinal or interval

6. Which below one is not a a type of t-tests?

a) One-sample t-tests **b) Null Hypothesis t-tests**

c) Independent sample t-tests d) Paired samples t-tests

7. The f test is used to compare

a) Two means from independent samples **b) Two variances from independent samples**

c) Two means from dependent samples d) Two variances from dependent samples

8. The f test is based on the ratio of

a) sample mean and population mean    **b) two sample variances**

c) two sample means    d) two population variances

9. In a one way ANOVA, the f test is used to compare the variability between:

**a) two or more sample means**    b) two or more dependent variables

c) two or more independent variables    d) two or more populations

**Questions:**

1. What is hypothesis?
2. Illustrate the points to be considered while formulating Hypothesis
3. Explain the procedure for hypothesis testing.
4. Distinguish between Type I error and Type II error.
5. What is **'t'** test and its types?
6. Give brief note on F test?
7. What is ANOVA?
8. Explain about Chi-square test and its uses.

**Exercises:**

1. The table given below the data obtained during outbreak of small –pox:

| | Attacked | Not attacked | Total |
|---|---|---|---|
| Vaccinated | 31 | 469 | 500 |
| Not vaccinated | 185 | 1315 | 1500 |
| Total | 16 | 1784 | 2000 |

Test the effectiveness of vaccination in preventing the attack from smallpox. Test your result with the help of $\chi^2$ at5% level of significance. (Given: For ν =1, $\chi^2_{0.05}$= 3.84).

**Answer: calculated value = 14.642. Null hypothesis is rejected**

2. In an anti malaria campaign in a certain area, quinine was administered to 812 persons out of a total population of 3248. The number of fever cases is shown below:

| Treatment | Fever | No fever | Total |
|---|---|---|---|
| Quinine | 140 | 30 | 170 |
| No Quinine | 60 | 20 | 80 |
| Total | 200 | 50 | 250 |

Discuss the usefulness of quinine in checking malaria. (Given: For $\nu$ =1, $\chi^2_{0.05}$= 3.84)

**Answer: calculated value = 1.839. Null hypothesis is accepted**

3. To test the significance of variation in the retail prices of a commodity in three principal cities, Mumbai, Kolkata, and Delhi, four shops were chosen at random in each city and the prices that lack confidence in their mathematical ability observed in rupees were as follows:

| | | | | |
|---|---|---|---|---|
| Kanpur | 15 | 7 | 11 | 13 |
| Lucknow | 14 | 10 | 10 | 6 |
| Delhi | 4 | 10 | 8 | 8 |

Do the data indicate that the prices in the three cities are significantly different? (Given: For $V_1$ = 2,$V_2$ = 9, and $F_{0.05}$ = 4.26).

**Answer: calculated Value F = 1.71; Null hypothesis is accepted.**

4. The following table gives the number of refrigerators sold by 4 salesmen in threemonths May, June and July:

| Month | Salesman | | | |
|---|---|---|---|---|
| | A | B | C | D |
| March | 50 | 40 | 48 | 39 |
| April | 46 | 48 | 50 | 45 |
| May | 39 | 44 | 40 | 39 |

Is there a significant difference in the sales made by the four salesmen? Is there a significantdifference in the sales made during different months? (Given: For $V_1$ = 3,$V_2$ = 6, and $F1_{0.05}$ = 4.75 & Given: For $V_1$ = 2,$V_2$ = 6, and $F2_{0.05}$ = 5.14)

**Answer: calculated Value $F_1$ = 1.018; Null hypothesis is accepted &**
**Calculated Value $F_2$ = 3.327; Null hypothesis is accepted.**

# References

## Books

Hazarika Padmalochan,A textbook of Business Statistics , S.Chand Publications

- Vohra ND, Business Statistics: Text and Problems – With Introduction to Business Analytics, Mc Graw Hill ,2021
- J.K. Sharma, Business Statistics, Pearson Education, New Delhi,2007.
- P.R. Vittal, Business Mathematics and Statistics, Margham Publications, Chennai, 2004.
- S.P. Gupta, Statistical Methods, Sultan Chand &Sons, NewDelhi,2007.
- Alexander Holmes, Barbara Illowsky and Susan Dean, Introductory Business Statistics , 12th Media Services, 2017
- Business Statistics & OR - Dr. S. P. Rajagopalan, Tata McGraw-Hill
- David M.Levine, David F.Stephan etal. Business Statistics : A first Course, 7th edition
- Dina Nath Pandit, Statistics: A Modern Approach , Hindustan Publishing corporation
- S.P. Gupta, Elements of Business Statistics, Sultan Chand & Sons, NewDelhi,2007.

## Websites

- http://www.statisticshowto.com
- https://theintactone.com/2019/09/01/ccsubba-204-business-statistics/
- https://statisticsbyjim.com/basics/measures-central-tendency-mean-median-mode/
- https://ug.its.edu.in/sites/default/files/Business%20Statistics.pdf