PRACTICAL MANUAL
OF
ANIMAL GENETICS AND BREEDING

(Biostatistics and Computer Application)

(**UNIT-1**)

Compiled By

Prof. J.S. Poonia

Dr. Vijay B. Sharma
(Associate Professor)

Dr. Yamini
(Assistant Professor)

***NAME*_________________________________**

***ROLL No.*_______________________________**

***CLASS*______________*BATCH*_____________**

Department of Animal Genetics and Breeding

Mahatma Jyotiba Fule College of Veterinary and Animal Science
Chomu, Jaipur (Rajsthan)

FORWARD

*I am very happy to go though the practical manual entitled, "***ANIMAL GENETICS AND BREEDING***" (***Biostatistics and Computer Application***) Unit-1 Prepared by* ***Prof. J.S. Poonia, Dr. Vijay B. Sharma****, Associate professor and* ***Dr. Yamini,*** *Assistant Professor,* ***Dept. of Animal Genetics & Breeding****, Mahatma Jyotiba Fule College of Veterinary & Animal science.*

It is appreciable to note that the manual covers the whole practical syllabus of B.V.Sc. & A.H. course as per the standards laid down by Veterinary Council of India.

The authors have devoted keenly to prepare this manual with their excellent knowledge and expertise in the field of biostatistics. Definitely this manual will be helpful for smooth and effective conduction of practical exercises and ensure a handbook for students for entire life in the profession.

I congratulate Prof. J.S. Poonia, Dr. Vijay B. Sharma and Dr. Yamini for the efforts put in bringingout this practical manual.

DEAN
Mahatma Jyotiba Fule College of Veterinary &
Animal science, Chomu, Jaipur

PREFACE

This Manual has been compiled for the undergraduate students of B.V.Sc. & A.H. in accordance with the syllabus designed by the Veterinary Council of India. The manual is an effort to outline the statistics and computer science in context of veterinary science. It highlights the basics applications of statistics and computer science as per the VCI syllabus for under graduates.

Author of this manual extend his thanks to the distinguished authors of various standard text books on statistics and computer science.

We hope this manual will serve very useful tool to the undergraduate and graduate students of Veterinary Science who are undergoing courses in veterinary Pathology.

It is our pleasure to thank Dean Sir, M.J.F College of veterinary and Animal Sciences, Chomu, Jaipur for providing necessary facilities and rendering all helps in preparing this course manual.

We happily acknowledge the healthy and timely assistance of computer operator and typist *Mr. Ashutosh Sharma* for very existence of this manual.

Date ………..   Course Teacher
Place…………………….   Dept. of Animal Genetics and Breeding

INDEX

(Unit-1: Biostatistics and Computer Application)

EXERCISE No. 1

INTRODUCTION TO BIOSTATISTICS

What is Biostatistics?

Biostatistics is defined as the applications of statistical methods to the problems of biology, medicine, and public health. It is also called as ***Biometry.***

The word 'Statistics' is derived from the Latin word 'Statis' which means a "political state." Clearly, statistics is closely linked with the administrative affairs of a state such as facts and figures regarding defence force, population, housing, food, financial resources etc.

The word statistics has several meanings. In the first place, it is a **plural noun** which describes a collection of numerical data such as employment statistics, accident statistics, population statistics, birth and death, income and expenditure, of exports and imports etc. It is in this sense that the word 'statistics' is used by a layman or a newspaper.

Secondly the word statistics as a **singular noun** is used to describe a branch of applied mathematics, whose purpose is to provide methods of dealing with collections of data and extracting information from them in compact form by tabulating, summarizing and analyzing the numerical data or a set of observations.

The various methods used are termed as statistical methods and the person using them is known as a statistician. In a specialized sense **statistics** describes various numerical items which are produced by using statistics (in the second sense) to statistics (in the first sense). Averages, standard deviation etc. are all statistics in this specialized third sense.

The word **'statistics'** in the second sense is defined by **Croxton and Cowden** as follows:" The collection, presentation, analysis and interpretation of the numerical data."

This definition clearly points out four stages in a statistical investigation, namely:

1) **Collection of data**
2) **Presentation of data**
3) **Analysis of data**
4) **Interpretation of data.**

Kinds or Branches of Statistics: Statistics may be divided into two main branches

1. Descriptive Statistics

2. Inferential Statistics

Descriptive Statistics: Descriptive statistics deals with the collection of data, its presentation in various forms, such as tables, graphs and diagrams and finding averages and other measures which would describe the data.

Inferential Statistics: Inferential statistics deals with techniques used for the analysis of data, making estimates and drawing conclusions from limited information obtained through sampling and testing the reliability of the estimates.

Uses:

1. To present the data in a concise and definite form: Statistics helps in classifying and tabulating raw data for processing and further tabulation for end users.
2. To make it easy to understand complex and large data: This is done by presenting the data in the form of tables, graphs, diagrams etc., or by condensing the data with the help of means, dispersion etc.
3. For comparison: Tables, measures of means and dispersion can help in comparing different sets of data.
4. In forming policies: It helps in forming policies like a production schedule, based on the relevant sales figures. It is used in forecasting future demands.
5. Enlarging individual experiences: Complex problems can be well understood by statistics, as the conclusions drawn by an individual are more definite and precise than mere statements on facts.
6. In measuring the magnitude of a phenomenon: Statistics has made it possible to count the population of a country, the industrial growth, the agricultural growth, the educational level (of course in numbers).

Limitations:

1. **Statistics does not deal with individual measurements.** Since statistics deals with aggregates of facts, it can't be used to study the changes that have taken place in individual cases. For example, the wages earned by a single industry worker at any time, taken by itself is not a statistical datum. But the wages of workers of that industry can be used statistically. Similarly the marks obtained by John of your class or the height of Beena (also of your class) are not the subject matter of statistical study. But the average marks or the average height of your class has statistical relevance.

2. **Statistics cannot be used to study qualitative phenomenon** like morality, intelligence, beauty etc. as these can‘t be quantified. However, it may be possible to analyze such problems statistically by expressing them numerically. For example we may study the intelligence of boys on the basis of the marks obtained by them in an examination.

3. **Statistical results are true only on an average**: The conclusions obtained statistically are not universal truths. They are true only under certain conditions. This is because statistics as a science is less exact as compared to the natural science.

4. **Statistical data, being approximations, are mathematically incorrect**. Therefore, they can be used only if mathematical accuracy is not needed.

5. **Statistics, being dependent on figures, can be manipulated** and therefore can be used only when the authenticity of the figures has been proved beyond doubt.

Question1. Define Biostatistics. Write the uses and limitations of statistics.

EXERCISE No. 2

COLLECTION OF DATA

In any statistical investigation, the collection of the numerical data is the first and the most important matter to be attended. Often a person investigating will have to collect the data from the actual field of inquiry. For this he may issue suitable questionnaires to get necessary information or he may take actual interviews. Personal interviews are more effective than questionnaires, which may not evoke an adequate response.

Another method of collecting data may be available in publications of Government bodies or other public or private organizations. Sometimes the data may be available in publications of Government bodies or other public or private organizations.

Data: Statistical data is a set of numbers or series of numerical data. Statistical data help in the process of decision making.

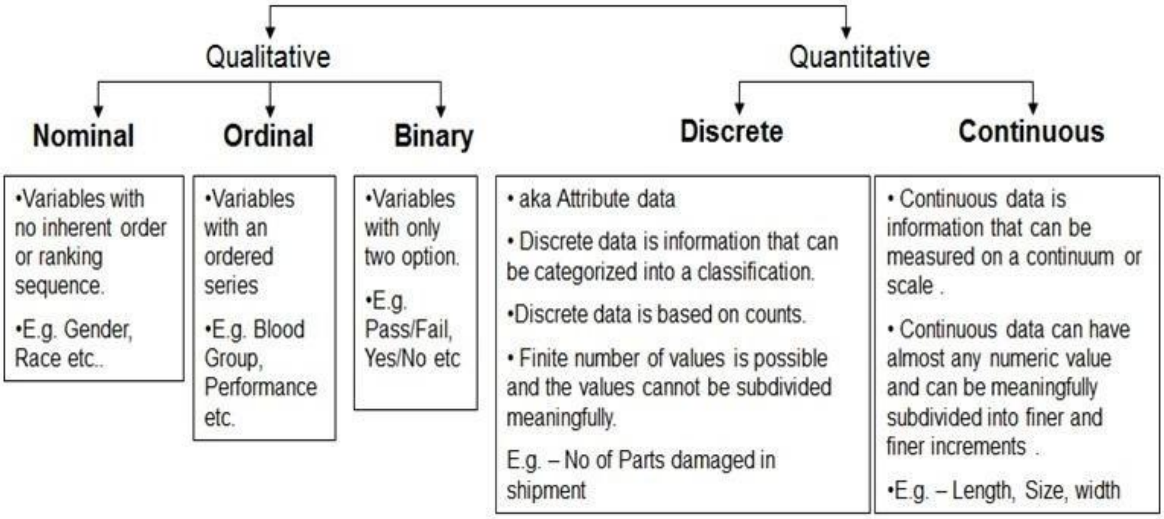
Types of Data: There are 2 types of data:

1. Primary Data
2. Secondary Data

1. **Primary Data** – primary data are original and first hand information and refers to the data that the investigator collects for the very first time. This type of data has not been collected either by this or any other investigator before. A primary data will provide the investigator with the most reliable first-hand information about the respondents. The investigator would have a clear idea about the terminologies uses, the statistical units employed, the research methodology and the size of the sample. Primary data may either be internal or external to the organization.

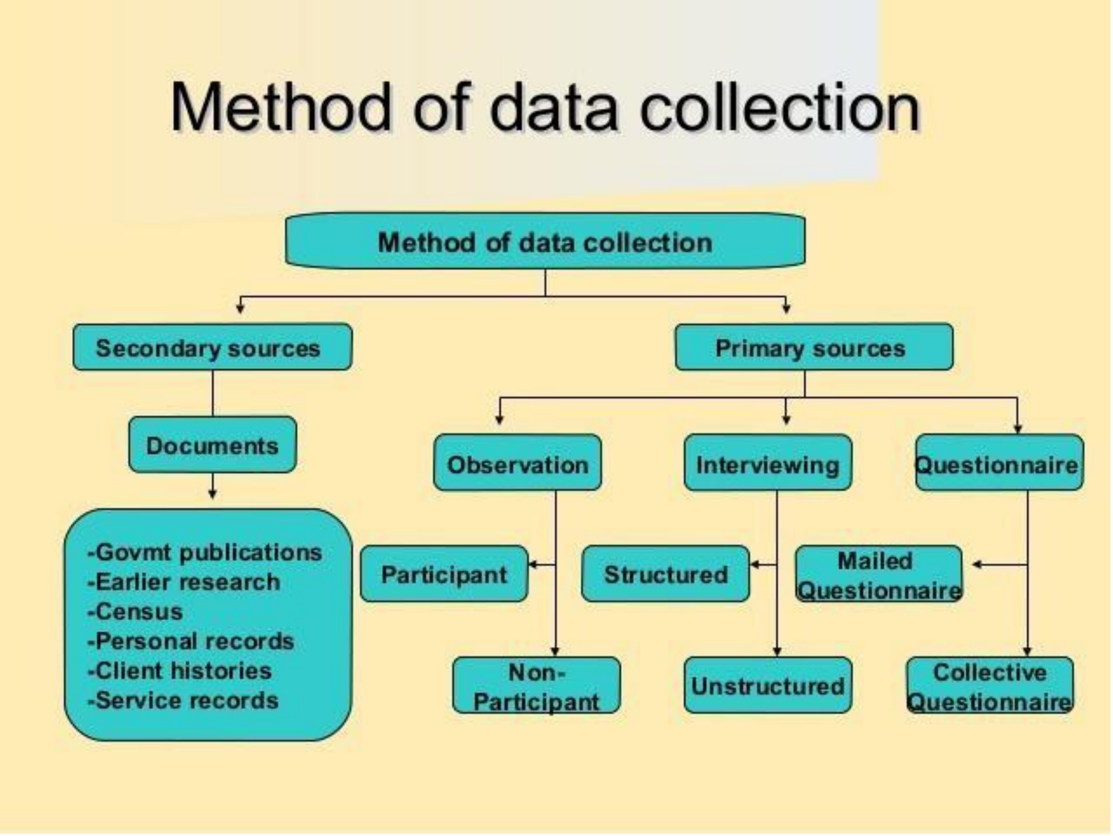
2. **Secondary Data** – refers to the data that the investigator collects from another source. Past investigators or agents collect data required for their study. The investigator is the first researcher or statistician to collect this data. Moreover, the investigator does not have a clear idea about the intricacies of the data. There may be ambiguity in terms of the sample size and sample technique. There may also be unreliability with respect to the accuracy of the data.

It is useful to distinguish between two broad **types** of variables: qualitative and quantitative (or numeric). Each is broken down into two sub-**types**: qualitative **data** can be ordinal or nominal, and numeric **data** can be discrete (often, integer) or continuous.


Qualitative
Quantitative
Nominal
Ordinal
Binary
Discrete
Continuous
•Variables with no inherent order or ranking sequence.
•E.g. Gender, Race etc..
•Variables with an ordered series
•E.g. Blood Group, Performance etc.
•Variables with only two option.
•E.g. Pass/Fail, Yes/No etc
• aka Attribute data
• Discrete data is information that can be categorized into a classification.
•Discrete data is based on counts.
• Finite number of values is possible and the values cannot be subdivided meaningfully.
E.g. – No of Parts damaged in shipment
• Continuous data is information that can be measured on a continuum or scale .
• Continuous data can have almost any numeric value and can be meaningfully subdivided into finer and finer increments .
•E.g. – Length, Size, width

Methods of Data Collection:

Data collection is the process of gathering and measuring data or information of any variables of interest in a standardized and established manner that enables the collector to answer or test hypothesis and evaluate outcomes of the particular collection.



Primary Data-

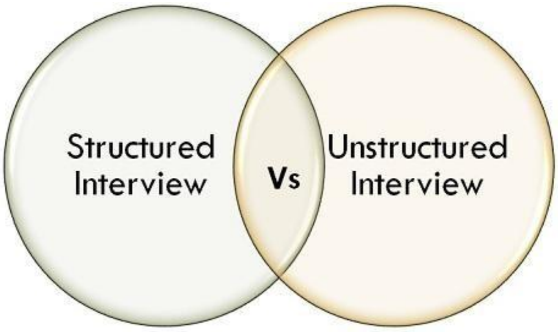
Discussed below are 5 broad classifications of the methods of collecting primary data.

1. **Direct Personal Investigation**

It consists of the collection of data by the investigator in a direct manner. The investigator (or researcher) is responsible for personally approaching a respondent and investigating the research and gather appropriate information. In other words, the researcher himself enters the field and solicits data that he requires to take the research forward. Thus, this method of data collection ensures first-hand information. This data is all the more reliable for an intensive research. But in an extensive research, this data is inadequate and proves to be unreliable. This method of collection of data is time-consuming. Hence, it tends to get handicapped when there is lack of time resource. However, the greatest demerit is that this method is very subjective in nature and is not suitable for objective based extensive researches.

2. **Interview**

Interview is described as an in-depth conversation between two or more persons, in a formal way, so as to figure out candidate‗s acceptability for the job. It is one of the most effective tools for data collection and selection. It is one to one communication between the interviewer and interviewee; wherein both the parties get a chance to learn about each other. Interviews can be structured interview or unstructured interview.



The **structured interview** uses preset questions, which are asked to all the candidates. On the other extreme, in an **unstructured interview**, the questions which are asked are not determined in advance, rather they are spontaneous.

Interview consists of the collection of data by the investigator in an indirect manner. The investigator (or enumerator) approaches (either by telephonic interviews) an indirect respondent who possesses the appropriate information for the research. Thus, this method of data collection ensures first-hand information because the interviewers can cross-question for the right and appropriate information.

3. **Mailed Questionnaire**

It consists of mailing a set or series of questions related to the research. The respondent answers the questionnaire and forwards it back to the investigator after marking his/her responses. This method of collection of data has proven to be time-saving. It is also a very cost-efficient manner of collecting the required data. An investigator who has the access to the internet and an email account can undertake this method of data collection. The researcher can only investigate those respondents who also have access to the internet and an email account. This remains the only major restriction of this method.

4. **Schedules**

Scheduling involves a face to face situation with the respondents. In this method of collecting data, the interviewer questions the respondent according to the questions mentioned in a form. This form is known as a schedule. This is different than a questionnaire. A questionnaire is personally filled by the respondents and the interviewer may or may not be physically present. Whereas, the schedule is filled by the enumerator or interviewer after asking the respondent his/her answer to a specific question. And in scheduling method of collecting data, the interviewer or enumerator is physically present.

5. **Local agencies**

In this method, the information is not directly or indirectly collected by either the interviewer of the enumerator. Instead, the interviewer hires or employs a local agency to work for him/her and help in gathering appropriate information. These local agents are often known as correspondents as well. Correspondents are only responsible for gathering accurate and reliable information. They work according to their preference and adopt different methods to do so.

Sources of Secondary Data: Discussed below are 2 broad classifications of the sources of secondary data.

1. **Published Sources**

There are many national organizations, international agencies and official publications that collect various statistical data. They collect data related to business, commerce, trade, prices, economy, productions, services, industries, currency and foreign affairs. They also collect information related to various (internal and external) socio-economic phenomena and publish them. These publications contain statistical reports of various kinds. Central Government Official Publication, Publications of Research Institutions, Committee Reports and International Publications are some published sources of secondary data.

2. **Unpublished Sources**

Some statistical data are not always a part of publications. Such data are stored by institutions and private firms. Researchers often make use of these unpublished data in order to make their researches all the more original.

Question1: Define data. What are the different types of data?

Question 2: Describe the different methods and sources of data collection.

EXERCISE No. 3

CLASSIFICATION AND TABULATION OF DATA

Therefore it becomes, very necessary to tabulate and summarize the data to an easily manageable form. In doing so we may overlook its details. But this is not a serious loss because Statistics is not interested in an individual but in the properties of aggregates. For a layman, presentation of the raw data in the form of tables or diagrams is always more effective.

Classification

Classification is the process of **dividing or arranging the data** into different groups (viz. classes) according to their resemblance. Groups are homogeneous within but heterogeneous between them. It helps in understanding the salient features of the data and also the comparison with similar data.

Classification is the process of arranging data into sequences and groups according to their common characteristics or separating them into different but related parts. **- Secrist**

Objectives / Purposes of Classification:

1. To simplify and condense the large data.
2. To present the facts to easily in understandable form.
3. To allow comparisons.
4. To help to draw valid inferences.
5. To relate the variables among the data.
6. To help further analysis.
7. To eliminate unwanted data.
8. To prepare tabulation

Modes/ Types of classification/ Methods of Classification:

This refers to the class categories into which the data could be sorted out and tabulated. These categories depend on the nature of data and purpose for which data is being sought.

Types of Classification:

The data is classified in the following ways:

a. **Geographical** (on the basis of area or region or place wise)
b. **Chronological** (On the basis of Temporal / Historical, i.e. with respect to time)
c. **Qualitative** (on the basis of character / attributes)
d. **Numerical, quantitative** (on the basis of magnitude)

Geographical Classification In geographical classification, the classification is based on the geographical regions or areas.

Example:

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

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

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

Example:

Sales reported by a departmental store

Month	**Sales (Rs.) in lakhs**
January	22
February	26
March	32
April	24
May	17
June	30

Qualitative Classification: When facts are grouped according to the qualities (attributes) like religion, literacy, business etc., the classification is called as 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.

Example:

Grouping or classification of a population on the basis of , Sex, Literacy, Education, Class grade etc.

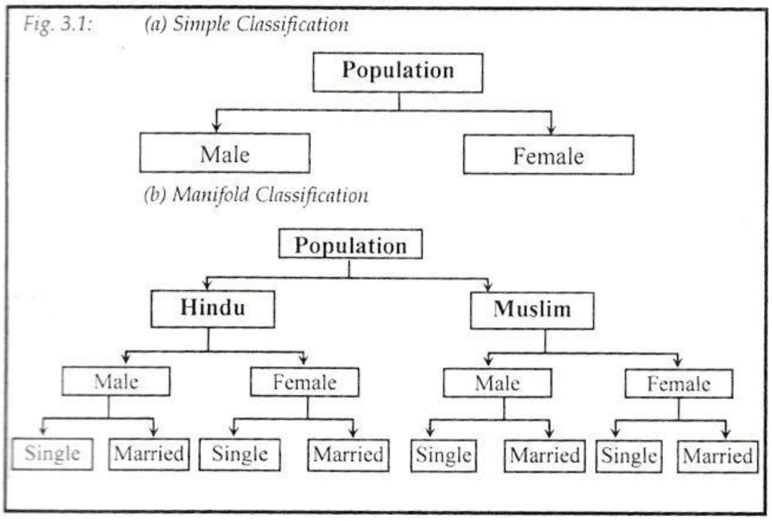
Further, it may be classified as : a) Simple classification b) Manifold classification

a). Simple Classification: If the classification is done into only two classes then classification is known as simple classification. It is also known as classification according to Dichotomy. In this data (facts) are divided into groups according to their qualities, the classification is called as 'Simple Classification'.

Ex: a) Population in to Male / Female b) Population into Educated / Uneducated

b). Manifold or Multiple Classifications: In this classification, the classification is based on more than one attribute at a time.

In this method data is classified using one or more qualities. First, the data is divided into two groups (classes) using one of the qualities. Then using the remaining qualities, the data is divided into different subgroups. For example, the population of a country is classified using three attributes: sex, literacy and business.



Quantitative Classification or Frequency Distribution (Grouped Data):

A frequency distribution is a statistical table which shows the set of all distinct values of the variable arranged in order of magnitude, either individually or in groups with their corresponding frequencies. - **Croxton and Cowden**

Frequency distribution is a table used to organize the data. The left column (called classes or groups) includes numerical intervals on a variable under study. The right column contains the list of frequencies, or number of occurrences of each class/group. Intervals are normally of equal size covering the sample observations range. It is simply a table in which the gathered data are grouped into classes and the number of occurrences, which fall in each class, is recorded.

A frequency distribution can be classified as

1. Series of individual observation
2. Discrete frequency distribution
3. Continuous frequency distribution

1. **Series of Individual**:

Observation Series of individual observation is a series where the items are listed one after the each observation.

For statistical calculations, these observations could be arranged is either ascending or descending order. This is called as array.

Roll No.	**Marks obtained in statistics paper**
1	83
2	80
3	72
4	92
5	65

The above data list is a raw data. The presentation of data in above form doesn"t reveals any information. If the data is arranged in ascending / descending in the order of their magnitude, which gives better presentation then, it is called arraying of data.

2. **Discrete (ungrouped) Frequency Distribution**:

If the data series are presented in such a way that indicating its exact measurement of units, then it is called as discrete frequency distribution. Discrete variable is one where the variants differ from each other by definite amounts.

Example:

Assume that a survey has been made to know number of post-graduates in 10 families at random; the resulted raw data could be as follows. 0, 1, 3, 1, 0, 2, 2, 2, 2, 4

This data can be classified into an ungrouped frequency distribution.

Number of post-graduates becomes variable (x) for which we can list the frequency of occurrence (f) in a tabular from as follows;

Number of post graduates (x)	**Frequency (f)**
0	2
1	2
2	4
3	1
4	1

The above example shows a discrete frequency distribution, where the variable has discrete numerical values.

3. **Continuous Frequency Distribution (grouped frequency distribution)**:

Continuous data series is one where the measurements are only approximations and are expressed in class intervals within certain limits. In continuous frequency distribution the class interval theoretically continuous from the starting of the frequency distribution tills the end without break.

According to **Boddington**, the variable which can take very intermediate value between the smallest and largest value in the distribution is a continuous frequency distribution.

Example:

Marks obtained by 20 students in students‟ exam for 50 marks are as given below –

Convert the data into continuous frequency distribution form.

18	23	28	29	44	28	48	33	32	43
24	29	32	39	49	42	27	33	28	29

By grouping the marks into class interval of 10 following frequency distribution tables can be formed.

Marks	**No. of students**
0-5	0
5-10	0
10-15	0
15-20	1
20-25	2
25-30	7

30-35	4
35-40	1
40-45	3
45-50	2

Tabulation

It is the process of condensing or summarizing classified or grouped data in the form of table for convenience, in statistical processing, presentation and interpretation of the information.

Major Objectives of Tabulation:

1. To simplify the complex data
2. To facilitate comparison
3. To economise the space
4. To draw valid inference / conclusions
5. To help for further analysis

Classification of Tables: Classification is done based on

1. Coverage (Simple and complex table)
2. Objective / purpose (General purpose / Reference table / Special table or summary table)
3. Nature of inquiry (primary and derived table)

Ex: a) **Simple table:** 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**

b) **Two-Way Table:** Classification is based on two characteristics

Distribution of marks Class on the basis of gender and marks

Class Marks	**No. of students**		
	Boys	**Girls**	**Total**
30-40	10	10	20
40-50	15	5	20
50-60	3	7	10
Total	28	22	50

Comparison Chart

BASIS FOR COMPARISON	**CLASSIFICATION**	**TABULATION**
Meaning	Classification is the process of grouping data into different categories, on the basis of nature, behaviour, or common characteristics.	Tabulation is a process of summarizing data and presenting it in a compact form, by putting data into statistical table.
Order	After data collection	After classification
Arrangement	Attributes and variables	Columns and rows
Purpose	To analyse data	To present data
Bifurcates data into	Categories and sub-categories	Headings and sub-headings

Question 1: What do you mean by Tabulation and Classification?

Question 2: Define Frequency Distribution

Question 3: Write down the different types /methods of Tabulation and Classification

EXERCISE No. 4

DIAGRAMMATIC AND GRAPHICAL PRESENTATION OF DATA

An attractive representation of statistical data is provided by charts, diagrams and pictures. Diagrammatic representation can be used for both the educated section and uneducated section of the society. Furthermore, any hidden trend present in the given data can be noticed only in this mode of representation. However, compared to tabulation, this is less accurate. So if there is a priority for accuracy, we have to recommend tabulation.

Diagrammatic Representation:

We are going to consider the following types of diagrams:

1. **Line diagram**
2. **Bar diagram**
3. **Pie chart**

Let us discuss the above four diagrammatic representations of data in detail.

Line diagram

When the time series exhibit a wide range of fluctuations, we may think of logarithmic or ratio chart where "Log y" and not "y" is plotted against "t".

We use multiple line charts for representing two or more related time series data expressed in the same unit and multiple – axis chart in somewhat similar situations, if the variables are expressed in different units.

Line diagram - Example

Example:

The profits in thousands of dollars of an industrial house for 2002, 2003, 2004, 2005, 2006, 2007 and 2008 are 5, 8, 9, 6, 12, 15 and 24 respectively.

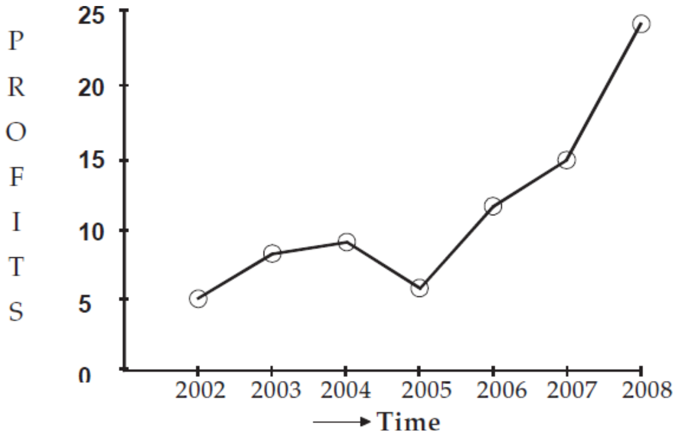
Represent these data using a suitable diagram during 2002 to 2008.

Solution:

We can represent the profits for 7 consecutive years by drawing either a line diagram as given below. Let us consider years on horizontal axis and profits on vertical axis.

For the year 2002, the profit is 5 thousand dollars. It can be written as a point (2002, 5). In the same manner, we can write the following points for the succeeding years as (2003, 8), (2004, 9), (2005, 6), (2006, 12), (2007, 15) and (2008, 24)

Now, plotting all these point and joining them using ruler, we can get the line diagram.



Bar Diagram:

Simple Bar:

A simple bar chart is used to represent data involving only one variable classified on a spatial, quantitative or temporal basis. In a simple bar chart, we make bars of equal width but variable length, i.e. the magnitude of a quantity is represented by the height or length of the bars.

The following steps are used to draw a simple bar diagram:

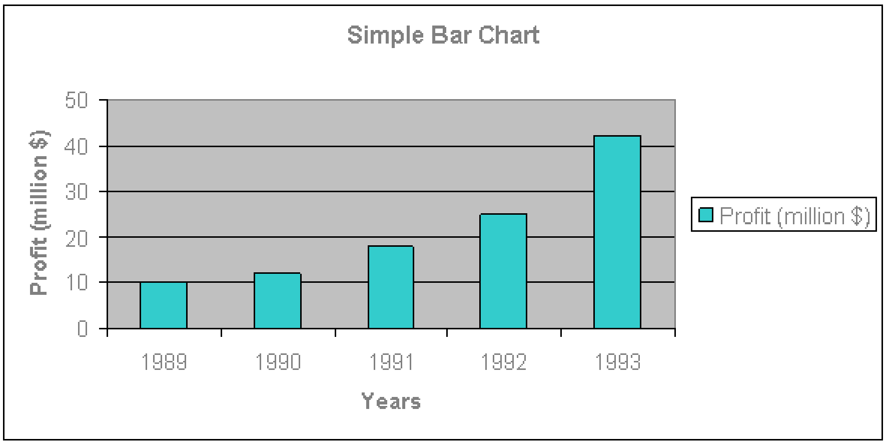
- Draw two perpendicular lines, one horizontally and the other vertically, at an appropriate place on the paper.
- Take the basis of classification along the horizontal line (X−X− axis) and the observed variable along the vertical line (Y−Y− axis), or vice versa.
- Mark signs of equal breadth for each class and leave equal or not less than half a breadth between two classes.
- Finally mark the values of the given variable to prepare required bars.

Example:

Draw simple bar diagram to represent the profits of a bank for 55 years.

Years	1989	1990	1991	1992	1993
Profits (million $$)	10	12	18	25	42

Solution:



We consider Multiple to compare related series. Component or sub-divided Bar diagrams are applied for representing data divided into a number of components.

Bar diagrams for comparing different components of a variable and also the relating of the components to the whole. For this situation, we may also use Pie chart or Pie diagram or circle diagram.

Multiple Bar Chart

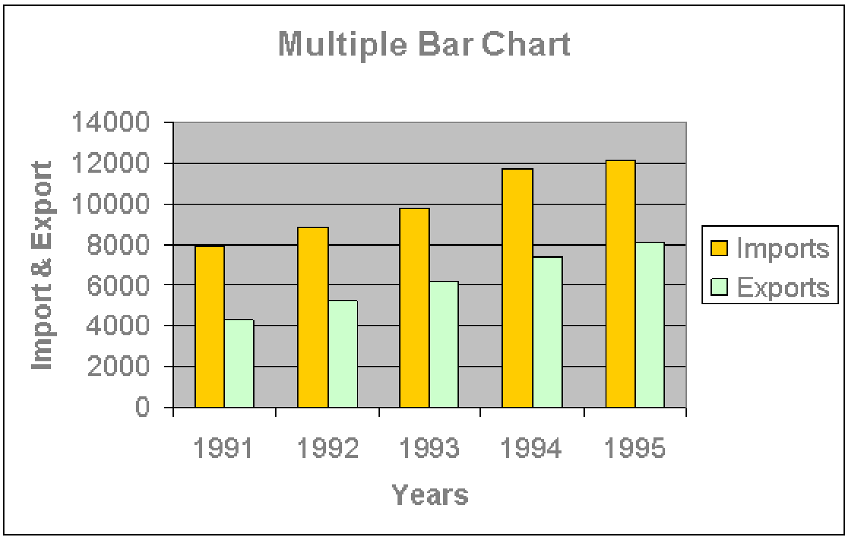
In a multiple bars diagram two or more sets of inter-related data are represented (multiple bar diagram faciliates comparison between more than one phenomenon). The technique of making a simple bar chart is used to draw this diagram but the difference is that we use different shades, colours, or dots to distinguish between different phenomena.

Example:

Draw a multiple bar chart to represent the imports and exports of Canada (values in $) for the years 1991 to 1995.

Years	Imports	Exports
1991	7930	4260
1992	8850	5225
1993	9780	6150
1994	11720	7340
1995	12150	8145

Solution:



Component Bar Chart:

A sub-divided or component bar chart is used to represent data in which the total magnitude is divided into different or components.

In this diagram, first we make simple bars for each class taking the total magnitude in that class and then divide these simple bars into parts in the ratio of various components. This type of diagram shows the variation in different components within each class as well as between different classes. A sub-divided bar diagram is also known as a component bar chart or stacked chart.

Example:

The table below shows the quantity in hundred kgs of wheat, barley and oats produced in a certain form during the years 1991 to 1994.

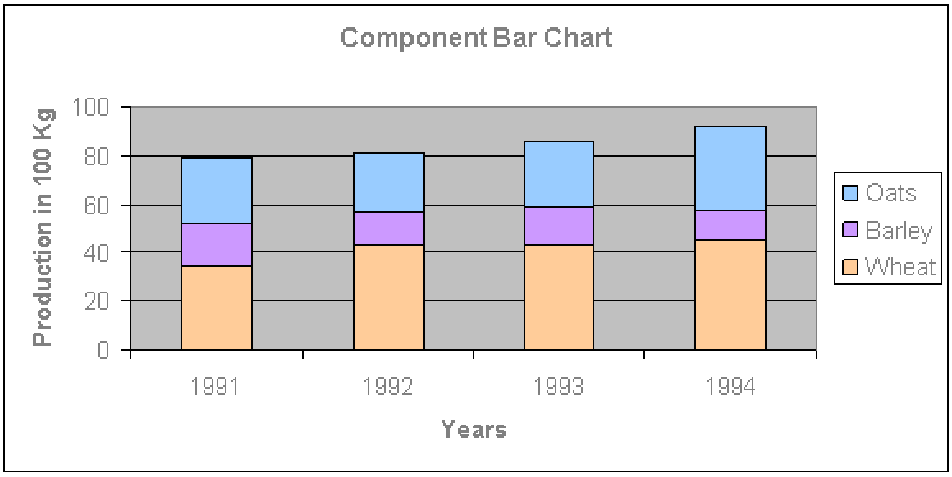
Years	**Wheat**	**Barley**	**Oats**
1991	34	18	27
1992	43	14	24
1993	43	16	27
1994	45	13	34

Construct a component bar chart to illustrate this data.

Solution:

To make the component bar chart, first of all we have to take a year-wise total production.

Years	**Wheat**	**Barley**	**Oats**	**Total**
1991	34	18	27	79
1992	43	14	24	81
1993	43	16	27	86
1994	45	13	34	92



Pie chart

In a pie chart, the various observations or components are represented by the sectors of a circle and the whole circle represents the sum of the value of all the components .Clearly, the total angle of 360° at the centre of the circle is divided according to the values of the components.

The central angle of a component is = [Value of the component / Total value] x 360°

Sometimes, the value of the components is expressed in percentages. In such cases,

The central angle of a component is = [Percentage value of the component / 100] x 360°

Example:

The number of hours spent by a school student on various activities on a working day, is given below.

Construct a pie chart using the angle measurement.

Activity	Sleep	School	Play	Homework	Others
Number of hours	8	6	3	3	4

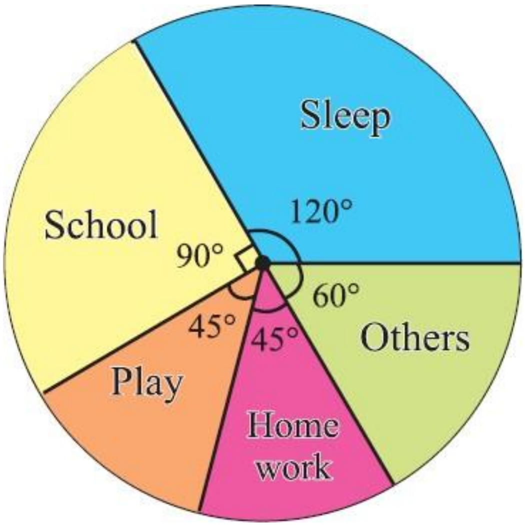
Solution:

The central angle of a component is = [Value of the component / Total value] x 360°

We may calculate the central angles for various components as follows:

Activity	**Duration in hours**	**Central angle**
Sleep	8	$\frac{8}{24} \times 360^{0} = 120^{0}$
School	6	$\frac{6}{24} \times 360^{0} = 90^{0}$
Play	3	$\frac{3}{24} \times 360^{0} = 45^{0}$
Homework	3	$\frac{3}{24} \times 360^{0} = 45^{0}$
Others	4	$\frac{4}{24} \times 360^{0} = 60^{0}$
Total	24	360^{0}

From the above table, clearly, we obtain the required pie chart as shown below.



Graphical Presentation:

Graphical presentation of frequency distribution is done by

1. Histogram
2. Frequency polygon
3. Frequency curve
4. Cumulative Frequency Polygon or Ogive

1. Histogram:

A graph formed by a series of adjacent rectangles, whose height is proportional to the class frequency, and the width is proportional to the width of the class, such that the area is proportional to the total frequency is called a histogram. A two dimensional graphical representation of a continuous frequency distribution is called a histogram.

In histogram, the bars are placed continuously side by side with no gap between adjacent bars. That is, in histogram rectangles are erected on the class intervals of the distribution. The areas of rectangle are proportional to the frequencies.

Points to be taken care of while constructing a histogram:

a) Inclusive to be converted to exclusive

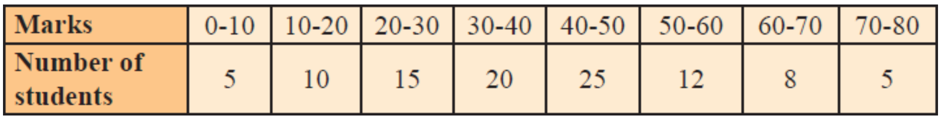
b) If class intervals are of unequal width, then histogram is constructed for frequency density against the width of the class interval.

c) Class limits along the x axis

d) Frequencies along the y axis

Example:

Draw a histogram for the following table which represent the marks obtained by 100 students in an examination:



Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Number of students	5	10	15	20	25	12	8	5

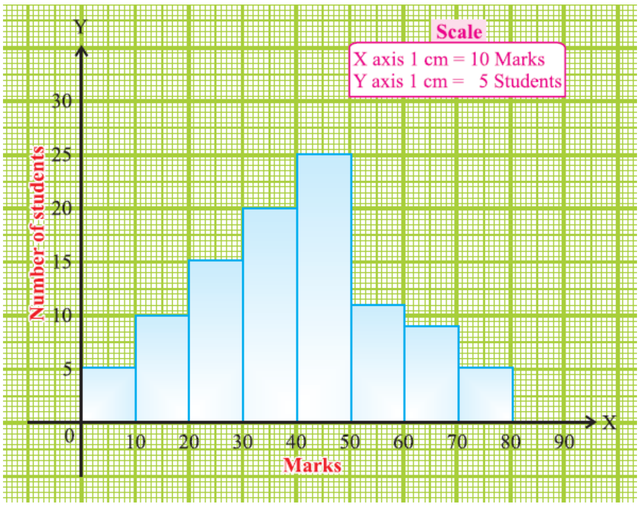
Solution:

The class intervals are all equal with length of 10 marks.

Let us denote these class intervals along the X-axis.

Denote the number of students along the Y-axis, with appropriate scale.

The histogram is given below.



2. **Frequency Polygon:**

It is an important type of graph used to represent either continuous or discrete frequency distribution. so called because it resembles a plane geometrical figure polygon representing frequency distribution. Hence the total area under the figure is proportional to the total frequency**. Frequency polygons are very similar to histograms**, except histograms have bars and frequency polygons have dots and lines connecting the frequencies of each class interval.

Steps to Draw a Frequency Polygon

1. Mark the class intervals for each class on the horizontal axis. We will plot the frequency on the vertical axis.
2. Calculate the classmark for each class interval. The formula for class mark is:

   Classmark = (Upper limit + Lower limit) / 2

3. Mark all the class marks on the horizontal axis. It is also known as the mid-value of every class.
4. Corresponding to each class mark, plot the frequency as given to you. The height always depicts the frequency. Make sure that the frequency is plotted against the class mark and not the upper or lower limit of any class.
5. Join all the plotted points using a line segment. The curve obtained will be kinked.
6. This resulting curve is called the frequency polygon.

Note that the above method is used to draw a frequency polygon without drawing a histogram. You can also draw a histogram first by drawing rectangular bars against the given class intervals. After this, you must join the midpoints of the bars to obtain the frequency polygon. Remember that the bars will have no spaces between them in a histogram.

Example: In a batch of 400 students, the height of students is given in the following table. Represent it through a frequency polygon.

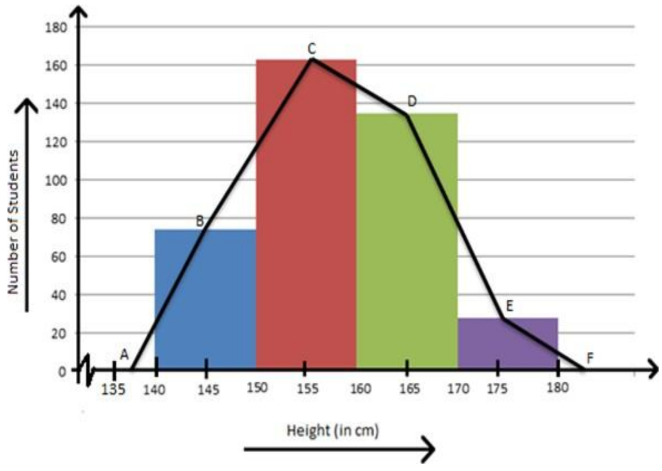
Height (in cm)	Number of Students(Frequency)
140 – 150	74
150 – 160	163
160 – 170	135
170 – 180	28
Total	400

Solution: Following steps are to be followed to construct a histogram from the given data:

a. The heights are represented on the horizontal axes on a suitable scale as shown.

b. The number of students is represented on the vertical axes on a suitable scale as shown.

c. Now rectangular bars of widths equal to the class- size and the length of the bars corresponding to a frequency of the class interval is drawn.

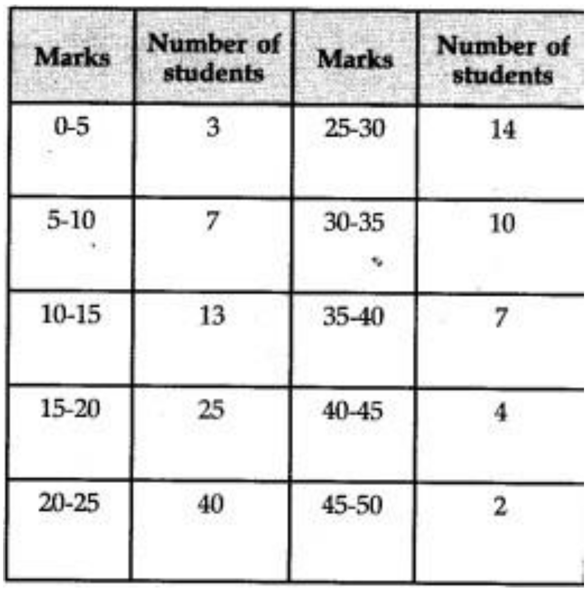


3. Frequency Curve:

It is a smooth free hand curve drawn along the frequency polygon by eliminating random or erratic fluctuations to show the regular movement of frequencies. Total area is proportional to the total frequency.

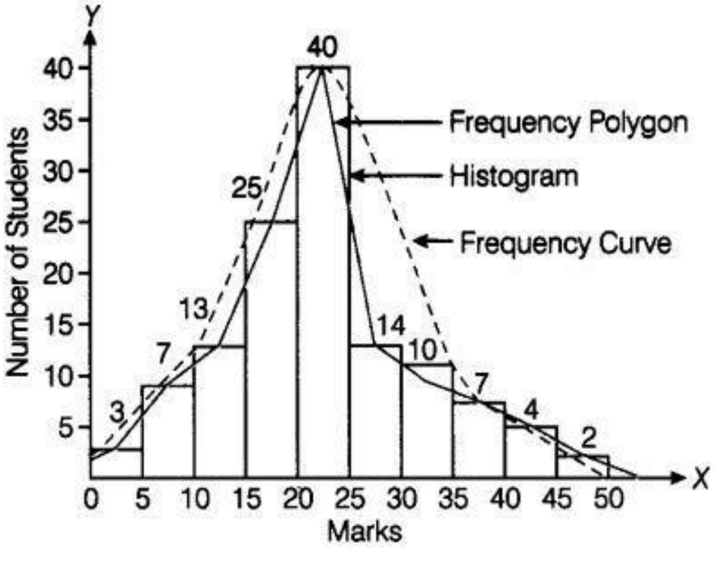
Frequency distribution curves are like frequency polygons. In frequency distribution, instead of using straight line segments, a smooth curve is used to connect the points.

Example: Make a frequency curve of the given frequency data



Marks	Number of students	Marks	Number of students
0-5	3	25-30	14
5-10	7	30-35	10
10-15	13	35-40	7
15-20	25	40-45	4
20-25	40	45-50	2

The frequency curve for the given data is shown as:



4. **Cumulative Frequency Polygon or Ogive:**

The curves drawn for cumulative frequencies against class limits are called ogives. Ogive is another statistical tool primarily used for finding out different quartiles in a distribution. From such a device we can also identify the number of observations lying above or below a certain value of the concerned variable.

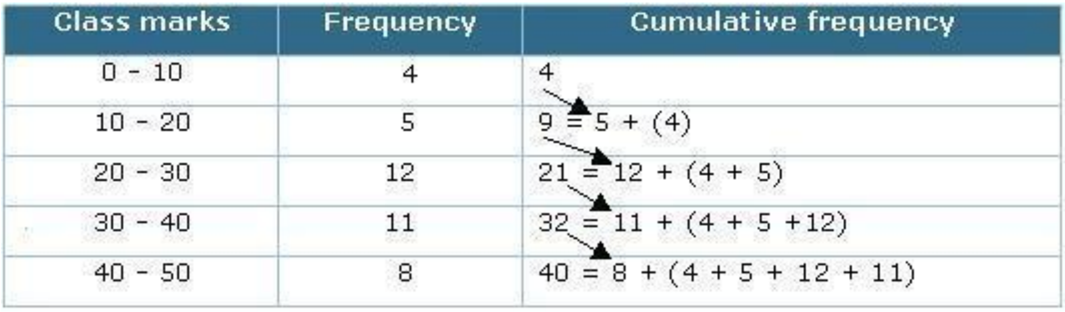
This kind of a diagram is drawn for a frequency distribution of a continuous variable in terms of cumulative frequencies of both the types (more than or less than type). While drawing this diagram we consider the given values of the variable horizontally and the corresponding cumulative frequencies (of either type) vertically.

Two Types of Ogives:

1. Less than ogive
2. More than ogive

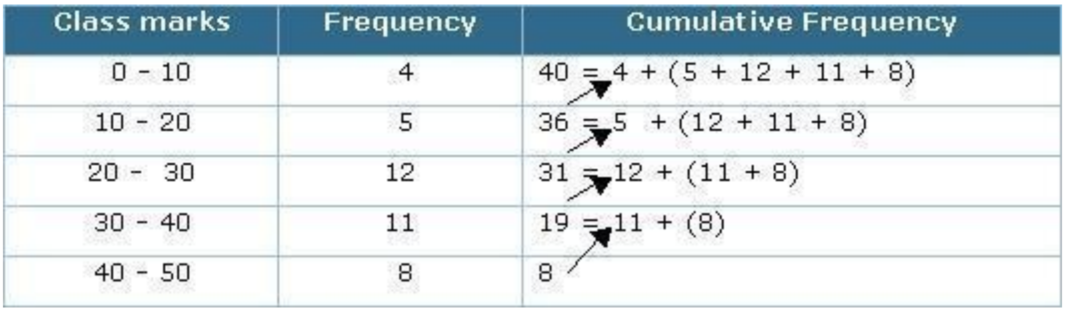
Cumulative frequency of less than type is zero for the lowest given value of the variable and similarly cumulative frequency of greater than type is zero for the highest value of the variable considered.

Less than type cumulative Frequency



Class marks	Frequency	Cumulative frequency
0 - 10	4	4
10 - 20	5	9 = 5 + (4)
20 - 30	12	21 = 12 + (4 + 5)
30 - 40	11	32 = 11 + (4 + 5 +12)
40 - 50	8	40 = 8 + (4 + 5 + 12 + 11)

More than type cumulative Frequency



Class marks	Frequency	Cumulative Frequency
0 - 10	4	40 = 4 + (5 + 12 + 11 + 8)
10 - 20	5	36 = 5 + (12 + 11 + 8)
20 - 30	12	31 = 12 + (11 + 8)
30 - 40	11	19 = 11 + (8)
40 - 50	8	8

Using the data available from a production organisation, Ogives of both the types are drawn below for our ready reference.

With the help of less than ogive one can find the frequencies less than the value of the given variable and the point of the variable for the given frequency with the help of more than ogive.

One can find the frequencies more than the value of the given variable and the point of the variable for the given frequency hence it is used to locate partition values.

How to Plot a Less Than Type Ogive:

Here we use the upper limit of the classes to plot the curve.

1. In the graph, put the upper limit on the x-axis
2. Mark the cumulative frequency on the y-axis.
3. Plot the points (x,y) using upper limits (x) and their corresponding cumulative frequency (y)
4. Join the points by a smooth freehand curve. It looks like an elongated *S.*

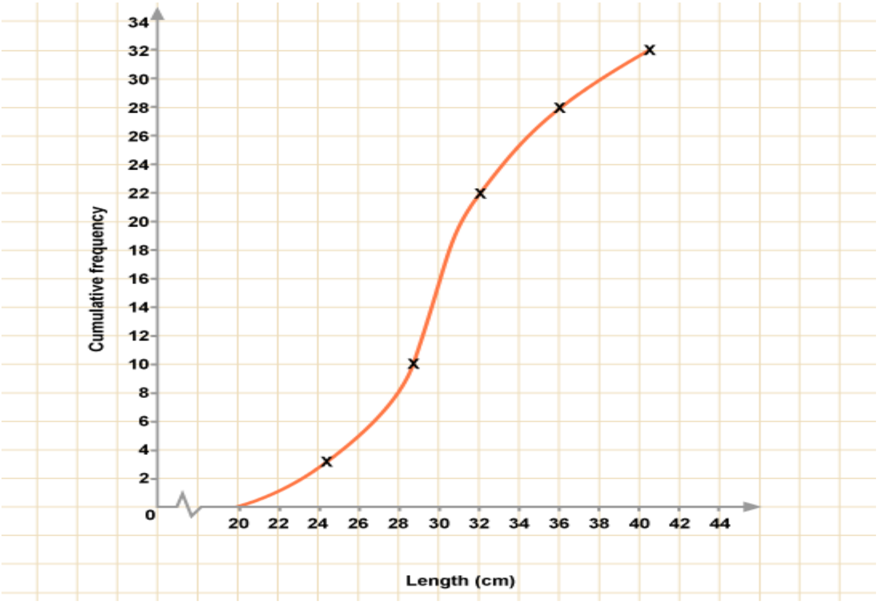
Cumulative Graphs can also be used to calculate the Median of given data. If you draw both the curves on the same graph, the point at which they intersect, the corresponding value on the x-axis, represents the Median of the given data set.

Example: The table shows the lengths (in cm) of 32 cucumbers.

Length	Frequency	Cumulative Frequency(Less than type)
21-24	3	3
25-28	7	10 (= 3 + 7)
29-32	12	22 (= 3 + 7 + 12)
33-36	6	28 (= 3 + 7 + 12 + 6)
37-40	4	32 (= 3 + 7 + 12 + 6 + 4)

Before drawing the cumulative frequency diagram, we need to work out the cumulative frequencies. This is done by adding the frequencies in turn.

The points are plotted at the upper class boundary. In this example, the upper class boundaries are 24.5, 28.5, 32.5, 36.5 and 40.5. Cumulative frequency is plotted on the vertical axis.



How to Plot a More Than Type Ogive:

Here we use the lower limit of the classes to plot the curve.

1. In the graph, put the lower limit on the x-axis
2. Mark the cumulative frequency on the y-axis.
3. Plot the points (x,y) using lower limits (x) and their corresponding Cumulative frequency (y)
4. Join the points by a smooth freehand curve. It looks like an upside down *S.*

Example: Finding median using Ogive

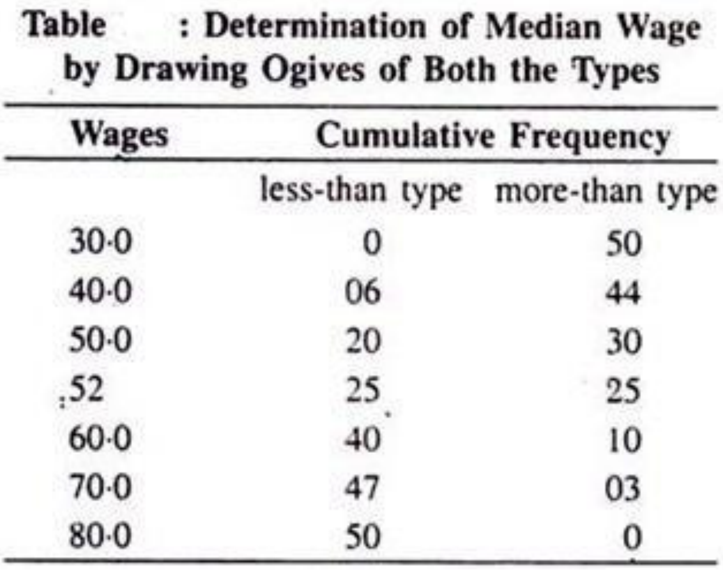


Table : Determination of Median Wage by Drawing Ogives of Both the Types

Wages	Cumulative Frequency	
	less-than type	more-than type
30·0	0	50
40·0	06	44
50·0	20	30
52	25	25
60·0	40	10
70·0	47	03
80·0	50	0

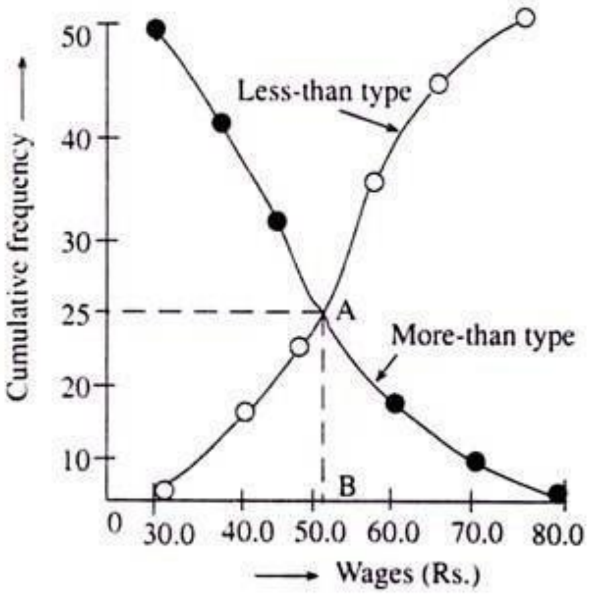


Fig. : Diagrammatic Determination of Median Wage

Question1: What are the Uses/importance of Graph and Diagram?

Question2: What are the differences between Graph and diagram?

Question3: Differentiate frequency curve and frequency polygon.

Question4: Plot the Ogive of the following given data:

Class Interval	**Frequency**
10-20	5
20-30	7
30-40	12
40-50	10
50-60	6

Question 5: Construct the histogram of following frequency data

Height of Trees (ft)	60-65,	65-70,	70-75,	75-80,	80-85,	85-90
No. of trees:	3	3	8	10	5	2

Question 6: The favourite flavours of ice-cream for the children in a locality are given in percentage as follow. Draw the pie chart to represent the given information

Flavours	% of Students Prefer the Flavours
Vanilla	25 %
Strawberry	15 %
Chocolate	10 %
Kesar-Pista	30 %
Mango Zap	20 %

EXERCISE No.-5

MEASURES OF CENTRAL TENDENCY

In the previous exercise, we have studied how to collect raw data, its classification and tabulation in a useful form, which contributes in solving many problems of statistical concern.

Yet this is not sufficient, for in practical purposes, there is need for further condensation, particularly when we want to compare two or more different distributions. We may reduce the entire distribution to one number which represents the distribution.

In a general sense the average is understood as the centre of a distribution or the most typical case of the observations.

A numerical measure of this central characteristic is known as a measure of central tendency. The mean is defined as a simple average of the numerical data.

The fundamental measures of central tendencies are:

1. Arithmetic mean
2. Median
3. Mode
4. Geometric mean
5. Harmonic mean

However, the most common measures of central tendencies or locations are arithmetic mean, median and mode.

Arithmetic Mean

This is the most commonly used average which you have also studied and used in lower grades.

The **mean**, often called the average, of a numerical set of data, is simply the sum of the data values divided by the number of values. This is also referred to as the arithmetic mean. The mean is the balance or middle most point of a distribution.

Methods of Calculating Mean

Data can be divided in three types:

1. Individual Series
2. Discrete data
3. Continuous or Class Interval series or Grouped data

There are two methods for calculating mean:

a. Direct Method and b. Short-cut Method

a.Direct Method

Individual Series

If x_1, x_2,…………x_n are the n values of x variable then,

A.M. = $A.M.=(\bar{x})=\frac{x_1+x_2+.........+x_n}{n}$ or $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

H.M. = $A.M.=(\bar{x})=\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+.........+\frac{1}{x_n}}$ or $\bar{x}=\frac{n}{\sum\frac{1}{x_i}}$

G.M. = $G.M.=(\bar{x})=(x_1\times x_2\times.........\times x_n)^{1/n}=\log G=\frac{1}{n}\sum_{i=1}^{n}\log x_i=anti\log\frac{1}{n}\sum_{i=1}^{n}\log x_i$

Discrete Data (Direct Method)

If x_1, x_2,…………x_n are the n values of x variable with frequencies f_1,f_2,…..f_n then,

A.M. = $A.M.=(\bar{x})=\frac{x_1f_1+x_2f_2+.........+x_nf_n}{n}$ or $\bar{x}=\frac{1}{N}\sum_{i=1}^{n}x_if_i$

H.M. = $H.M.=(\bar{x})=\frac{N}{\sum\left(\frac{f_i}{x_i}\right)}$

G.M. = $G.M.=(\bar{x})=anti\log\left(\frac{1}{N}\sum_{i=1}^{n}f_i.\log x_i\right)$

where, N= $\sum f_i$

Continuous or Class Interval series or Grouped data

If $x_1, x_2, \dots\dots\dots\dots x_n$ are the n values of x variable given in the form of class interval like 10-20, 20-30, ………90-100 with frequencies $f_1, f_2, \dots\dots f_n$ then, find out the mid value of each class interval and use mid value as x and calculate the mean in usual manner.

A.M. = $A.M. = (\bar{x}) = \frac{x_1 f_1 + x_2 f_2 + \dots\dots\dots + x_n f_n}{n}$ $or\ \bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i f_i$

H.M. = $H.M. = (\bar{x}) = \frac{N}{\sum\left(\frac{f_i}{x_i}\right)}$

G.M. = $G.M. = (\bar{x}) = anti\log\left(\frac{1}{N}\sum_{i=1}^{n} f_i .\log x_i\right)$

where, N= $\sum f_i$

b. Short-cut Method

For Discrete and Grouped Data

1. A suitable variable as assumed mean, say a and find the deviation $d_i = x_i$-a
2. Multiply respective value of f and d to get product fx and apply formula

$$Mean = a + \frac{\sum f_i d_i}{N}$$

Example 1. **Compute the arithmetic mean of the following by direct and short -cut methods both:**

Class	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**
Freqyebcy	**8**	**26**	**30**	**20**	**16**

Solution.

Class	Mid Value x	f	fx	d= x-A A = 45	f d
20-30	25	8	200	-20	-160
30-40	35	26	910	-10	-260
40-50	45	30	1350	0	0
50-60	55	20	1100	10	200
6070	65	16	1040	20	320
Total		N = 100	∑ fx = 4600		∑f d = 100

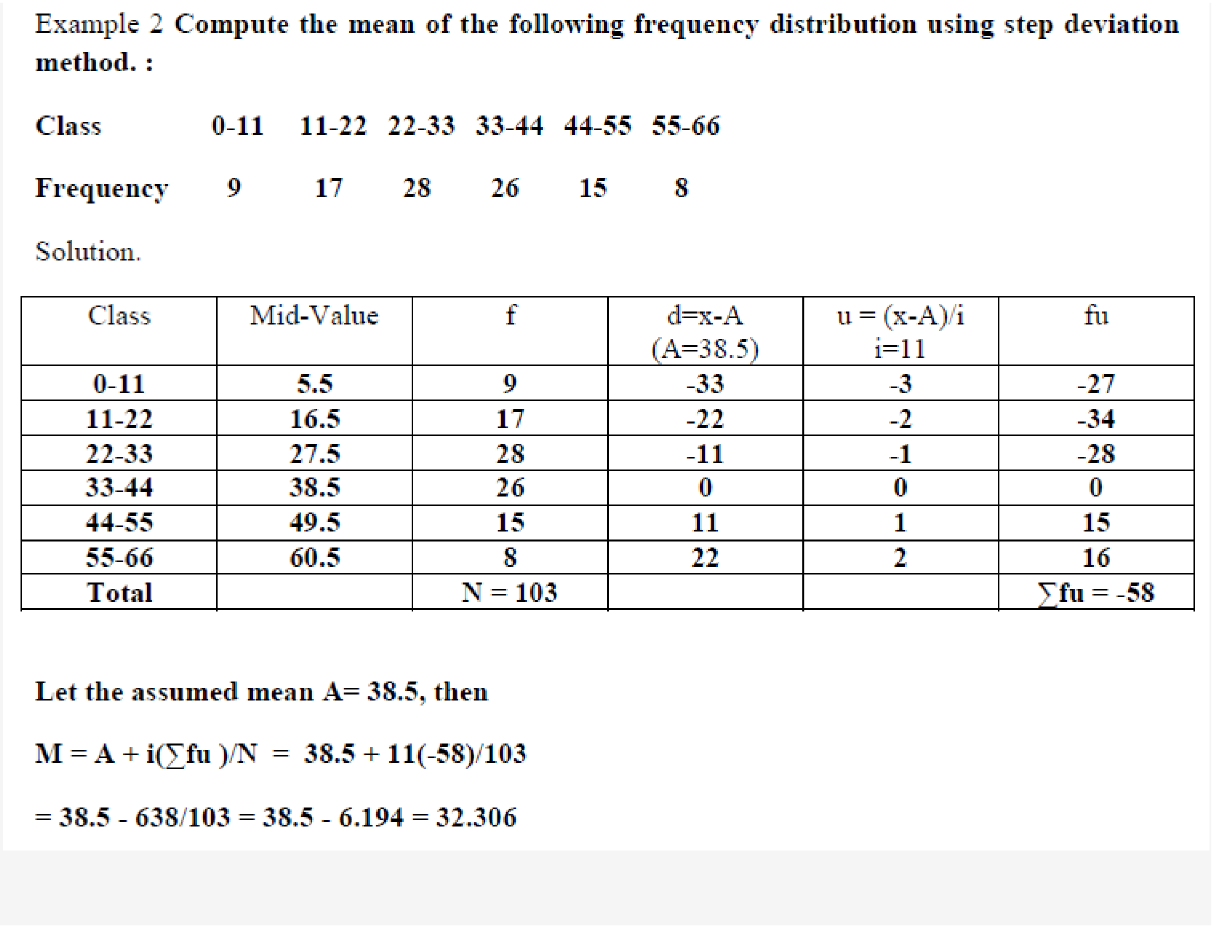
By direct method

$$M = (\sum fx)/N = 4600/100 = 46.$$

By short cut method.

Let assumed mean A= 45.

$$M = A + (\sum fd)/N = 45+100/100 = 46.$$



Example 2 **Compute the mean of the following frequency distribution using step deviation method. :**

Class	**0-11**	**11-22**	**22-33**	**33-44**	**44-55**	**55-66**
Frequency	**9**	**17**	**28**	**26**	**15**	**8**

Solution.

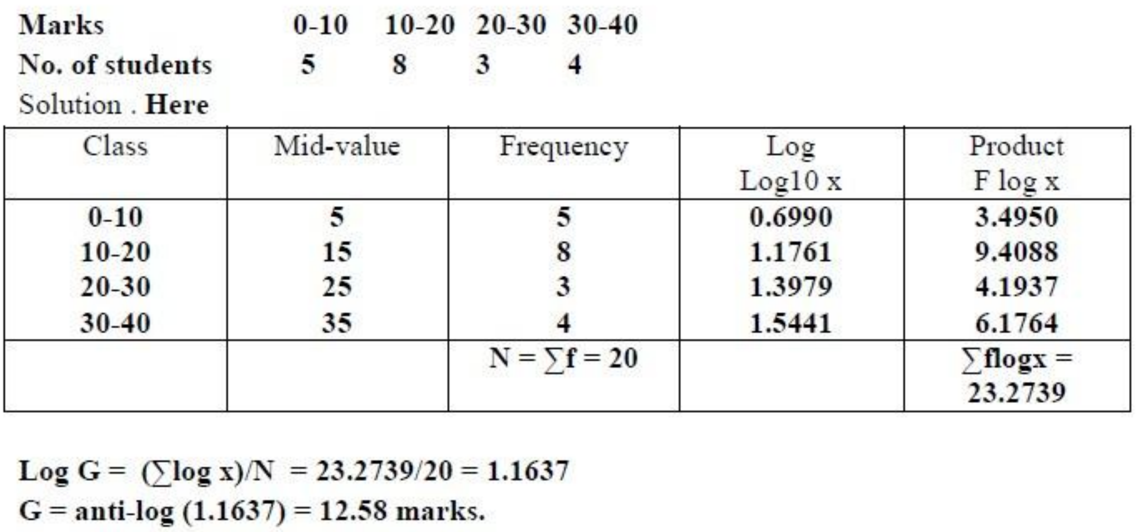
Class	Mid-Value	f	d=x-A (A=38.5)	u = (x-A)/i i=11	fu
0-11	5.5	9	-33	-3	-27
11-22	16.5	17	-22	-2	-34
22-33	27.5	28	-11	-1	-28
33-44	38.5	26	0	0	0
44-55	49.5	15	11	1	15
55-66	60.5	8	22	2	16
Total		N = 103			∑fu = -58

Let the assumed mean A= 38.5, then

$M = A + i(\sum fu)/N = 38.5 + 11(-58)/103$

$= 38.5 - 638/103 = 38.5 - 6.194 = 32.306$

Example: 3 Compute the Geometric mean of following distribution



Marks	0-10	10-20	20-30	30-40
No. of students	5	8	3	4

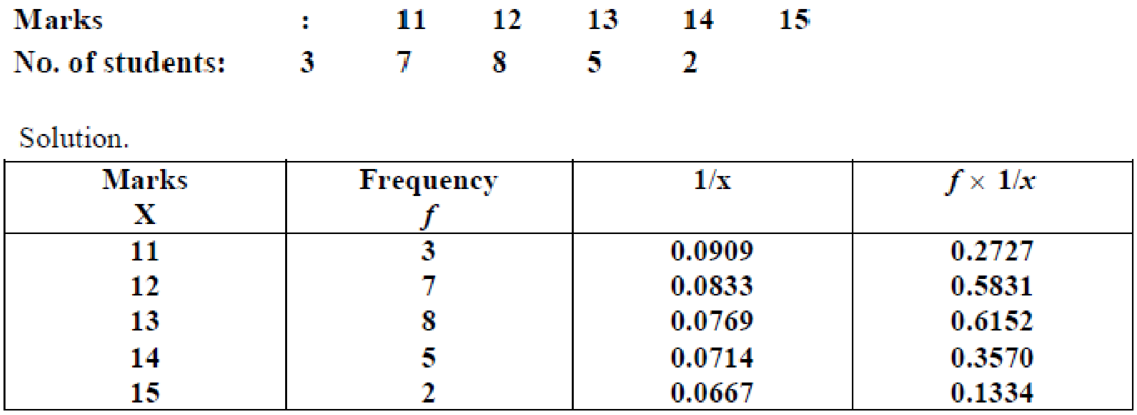
Solution . **Here**

Class	Mid-value	Frequency	Log Log10 x	Product F log x
0-10	5	5	0.6990	3.4950
10-20	15	8	1.1761	9.4088
20-30	25	3	1.3979	4.1937
30-40	35	4	1.5441	6.1764
		N = ∑f = 20		∑flogx = 23.2739

Log G = (∑log x)/N = 23.2739/20 = 1.1637

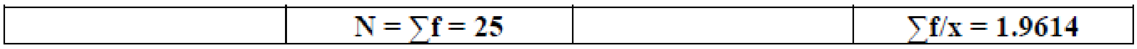
G = anti-log (1.1637) = 12.58 marks.

Example.4: Find the harmonic mean of given data

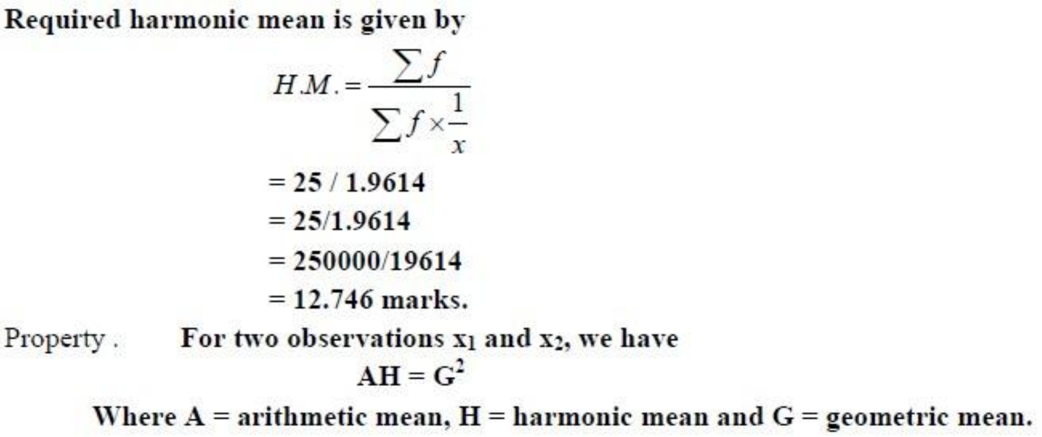


Marks	:	11	12	13	14	15
No. of students:	3	7	8	5	2	

Solution.



Marks X	Frequency f	1/x	$f \times 1/x$
11	3	0.0909	0.2727
12	7	0.0833	0.5831
13	8	0.0769	0.6152
14	5	0.0714	0.3570
15	2	0.0667	0.1334
	N = ∑f = 25		∑f/x = 1.9614



Required harmonic mean is given by

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

= 25 / 1.9614

= 25/1.9614

= 250000/19614

= 12.746 marks.

Property . **For two observations x_1 and x_2, we have**

$$AH = G^2$$

Where A = arithmetic mean, H = harmonic mean and G = geometric mean.

The Median

The **median** is the number that is in the middle value if the data has been organized either in descending order or in ascending order. Organized data means the numbers are arranged from smallest to largest or from largest to smallest.

The median for an odd number of data values is the value that divides the data into two halves.

Methods for Calculating Median

1. **Individual Series:**

If N represents the number of data values and N is an odd number, then the median will be

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

If N is an even number, then the median will be:

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

2. **Discrete Series:**
   1. Arrange the values x of given series in ascending order in first column
   2. Find cumulative frequency c.f. of given respective frequency in second column
   3. Find the median item, $\left(\frac{N+1}{2}\right)^{th} term$, where N is total of frequency
   4. Check the cumulative frequency in which median item is included
   5. The value of x corresponding to this c.f. is median

3. **Grouped data or Class Interval data:**
   1. Find cumulative frequency c.f. of given respective frequency in second column
   2. Find median class as $\left(\frac{N}{2}\right)^{th} term$
   3. Find the median by $Md = l_1 \frac{\frac{N}{2} - C}{f} \times i$

Where l_1 = lower limit of median class
C = cumulative frequency preceding median class
f = frequency of median class
i = class interval

Example 1 – **According to the census of 1991, following are the population figure, in thousands, of 10 cities :**

1400, 1250, 1670, 1800, 700, 650, 570, 488, 2100, 1700.

Find the median.

Solution. **Arranging the terms in ascending order.**

488, 570, 650, 700, 1250, 1400, 1670, 1800, 2100.

Here n=10, therefore the median is the mean of the measure of the 5th and 6th terms.

Here 5th term is 1250 and 6th term is 1400.

Median (Md) = (1250+14000)/2 Thousands

= 1325 Thousands

Examples 2. **Find the median for the following distribution:**

Wages in Rs.	**0-10**	**10-20**	**20-30**	**30-40**	**40-50**
No. of workers	**22**	**38**	**46**	**35**	**20**

Solution . **We shall calculate the cumulative frequencies.**

Wages in Rs.	No. of Workers f	Cumulative Frequencies (c.f.)
0-10	**22**	**22**
10-20	**38**	**60**
20-30	**46**	**106**
30-40	**35**	**141**
40-50	**20**	**161**

Here N = 161. Therefore median is the measure of (N + 1)/2th term i.e 81st term. Clearly 81st term is situated in the class 20-30. Thus 20-30 is the median class. Consequently.

Median $M_d = l + \frac{\frac{n}{2} - cf}{f} \times i$

= 20 + (½ × 161 – 60) / 46 × 10

= 20 + 205/46 = 20 + 4.46 = 24.46.

Example 3. **Find the median of the following frequency distribution:**

Marks	No. of students	Marks	No. of students
Less than 10	15	Less than 50	106
Less than 20	35	Less than 60	120
Less than 30	60	Less than 70	125
Less than 40	84		

Solution . **The cumulative frequency distribution table :**

Class (Marks)	Frequency f (No. of students)	Cumulative Frequency (C. F.)
0-10	15	15
10-20	20	35
20-30	25	60
30-40	24	84
40-50	22	106
50-60	14	120
60-70	5	125
Total	N = 125	

Median = measure of $\left(\frac{125 + 1}{2}\right)^{th}$ **term**

= 63rd term.

Clearly 63rd term is situated in the class 30-40.

Thus median class = 30 - 40

Median $M_d = l + \frac{\frac{n}{2} - cf}{f} \times i$

=30 + (125/2 – 60) / 24 × 10

= 30 + 25/24

= 30+1.04 = 31.04

Example 4: Find the median of the given data

Value **X**	**Frequency** **f**
20	2
29	4
30	4
39	3
44	2

Solution:

Value **x**	**Frequency** **f**	**Cumulative** **frequency**
20	2	2
29	4	2 + 4 = 6
30	4	6 + 4 = **10**
39	3	10 + 3 = 13
44	2	13 + 2 = 15
	$\sum f = 15$	$\sum fx = 481$

$\sum f = 15$ items,

The median item= $\left(\frac{N+1}{2}\right)^{th}$ $term$ = 16/2 = 8

The 8th item in the ordered data array will be the median. The 8 item will be included in the cumulative frequency 10.

Hence the median of the distribution is the x value corresponding to cumulative frequency 10 which reads as 30.

Median of the data = 30.

Mode:

The most common value in a given series is mode. It is the most frequent value in a given data or series or the value whose frequency is highest in a series.

Methods for Calculating Mode

1. **Individual Series:**

If $x_1, x_2, \ldots\ldots\ldots\ldots x_n$ are the n values of x variable then, mode in a series is found out by observing the value which is most frequent.

2. **Discrete Series:**

If $x_1, x_2, \ldots\ldots\ldots\ldots x_n$ are the n values of x variable with frequencies $f_1, f_2, \ldots.. f_n$ then, the value whose frequency is maximum is mode value.

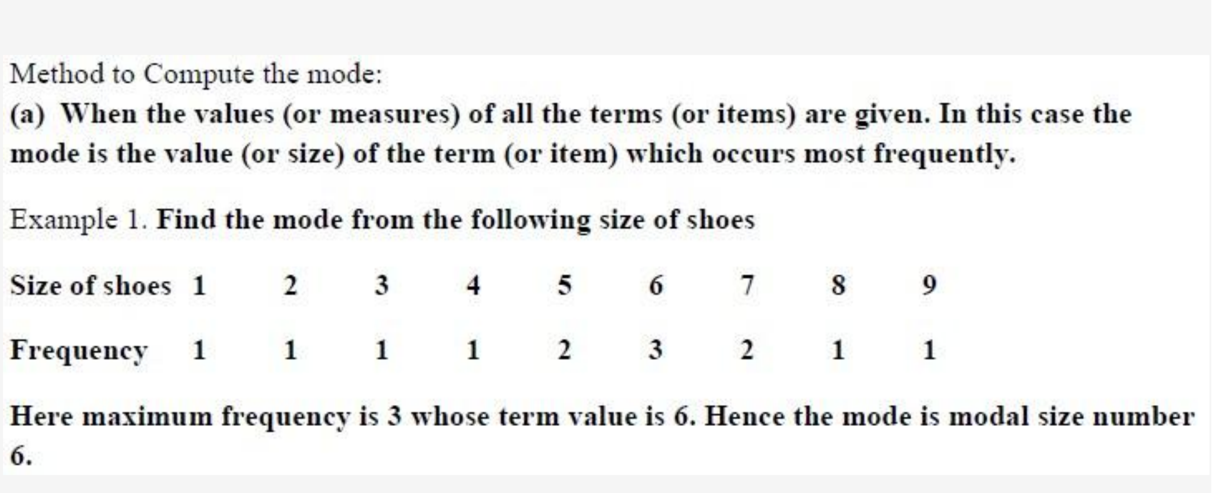
3. **Grouped data or Class Interval data:**

If $x_1, x_2, \ldots\ldots\ldots\ldots x_n$ are the n values of x variable given in the form of class interval like 10-20, 20-30, ………90-100 with frequencies $f_1, f_2, \ldots.. f_n$ then, value of mode is determined by the following formula:

$$Mode = l_1 + \frac{f_0 - f_1}{2f_0 - f_1 - f_2} \times i$$

Where l_1 = lower limit of modal class
f_0 = frequency of modal class
f_1 = frequency preceding modal class
f_2 = frequency proceeding modal class
i = class interval

Modal class is the class interval with highest frequency



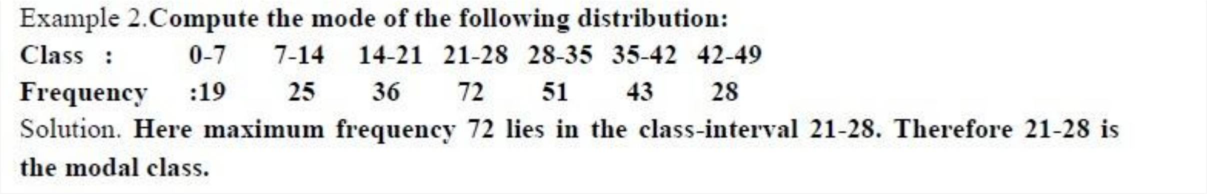
Method to Compute the mode:

(a) When the values (or measures) of all the terms (or items) are given. In this case the mode is the value (or size) of the term (or item) which occurs most frequently.

Example 1. **Find the mode from the following size of shoes**

Size of shoes	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Frequency	**1**	**1**	**1**	**1**	**2**	**3**	**2**	**1**	**1**

Here maximum frequency is 3 whose term value is 6. Hence the mode is modal size number 6.



Example 2.**Compute the mode of the following distribution:**

Class :	**0-7**	**7-14**	**14-21**	**21-28**	**28-35**	**35-42**	**42-49**
Frequency	**:19**	**25**	**36**	**72**	**51**	**43**	**28**

Solution. **Here maximum frequency 72 lies in the class-interval 21-28. Therefore 21-28 is the modal class.**

Where l_1 = lower limit of modal class =21
f_0 = frequency of modal class = 72
f_1 = frequency preceding modal class = 36
f_2 = frequency proceeding modal class = 51
i = class interval = 7

$$Mode = l_1 + \frac{f_0 - f_1}{2f_0 - f_1 - f_2} \times i$$

= 21+72-36/(2*72 - 36 -51)*7

=21+252/57

=21+4.421

=25.421

Question 1: Differentiate the Mean, Mode and Median

Question 2: Calculate the mean, mode and median of following data

4, 6, 2, 2, 2, 6, 8, 7, 12, 6, 1, 3 and 3

Question 3: Calculate the mode of the following distribution

Class	0-7	7-14	14-21	21-28	28-35	35-42	42-49
Frequency	19	25	36	72	51	43	28

Question 4: Calculate the mean, mode and median of the following distribution

Class	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	4	6	8	5	3	2

EXERCISE No. 6

MEASURES OF DISPERSION

An average is an attempt to summarize a set of data using just one number. As some of our examples have shown, an average taken by itself may not always be very meaningful. We need a statistical cross-reference that measures the spread of the data.

There are following measures of dispersion:

1. Range
2. Variance
3. Standard Deviation
4. Coefficient of Variance
5. Standard error

Range:

The range is the difference between the largest and smallest values of a data distribution**.**

A large bakery regularly orders cartons of Maine blueberries. The average weight of the cartons is supposed to be 22 ounces. Random samples of cartons from two suppliers were weighed.

The weights in ounces of the cartons were

Supplier I: 17 22 22 22 27

Supplier II: 17 19 20 27 27

(a) Compute the range of carton weights from each supplier.

Range = Largest value _ Smallest value

Supplier I range = 27 - 17 =10 ounces

Supplier II range = 27 -17 =10 ounces

(b) **Compute the mean weight of cartons from each supplier**.

In both cases the mean is 22 ounces.

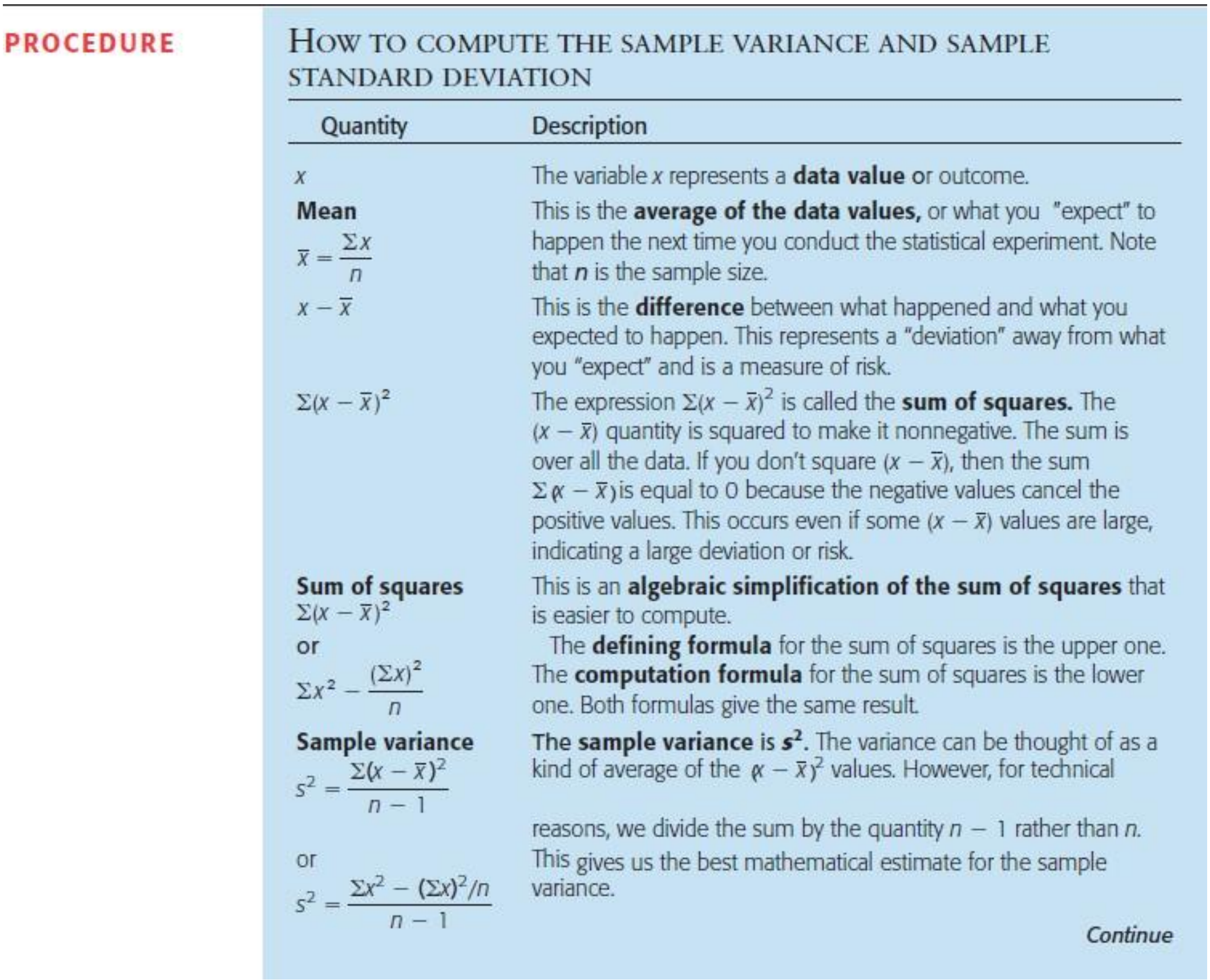
(c) Look at the two samples again. **The samples have the same range and mean. How do they differ?** The bakery uses one carton of blueberries in each blueberry muffin recipe. It is important that the cartons be of consistent weight so that the muffins turn out right.

Supplier I provides more cartons that have weights closer to the mean. Or, put another way, the weights of cartons from Supplier I are more clustered around the mean. The bakery might find Supplier I more satisfactory. As we see in Example 5, although the range tells the difference between the largest and smallest values in a distribution, it does not tell us how much other values vary from one another or from the mean.

Variance and Standard Deviation:

We need a measure of the distribution or spread of data around an expected value (either or m). The ***variance*** and ***standard deviation*** provide such measures. Formulas and rationale for these measures are described in the next Procedure display. Then, examples and guided exercises show how to compute and interpret these measures.

As we will see later, the formulas for variance and standard deviation differ slightly depending on whether we are using a sample or the entire population.



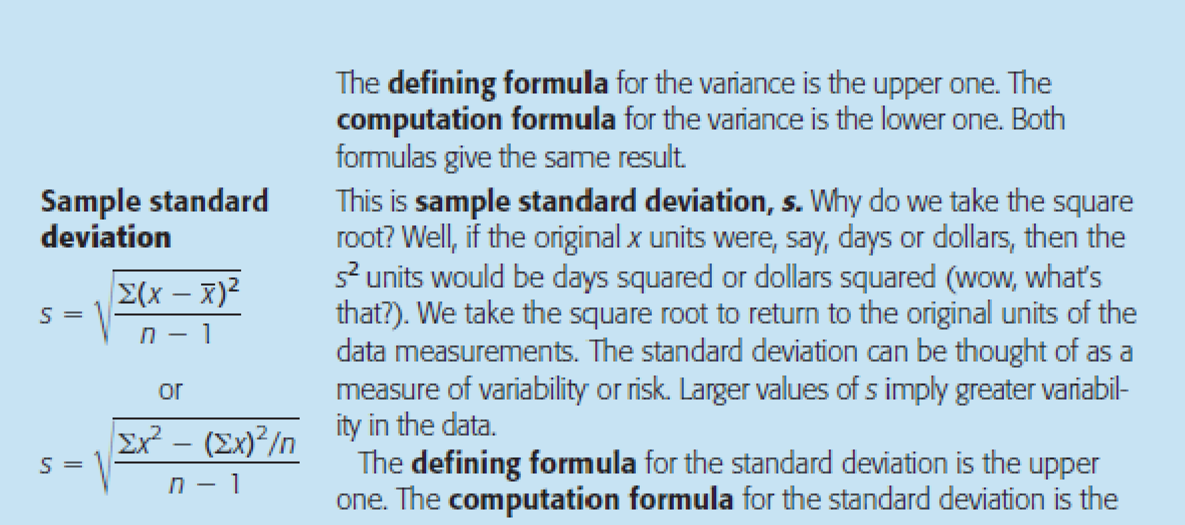
PROCEDURE

HOW TO COMPUTE THE SAMPLE VARIANCE AND SAMPLE STANDARD DEVIATION

Quantity	Description
x	The variable x represents a **data value** or outcome.
Mean $\bar{x} = \frac{\Sigma x}{n}$	This is the **average of the data values,** or what you "expect" to happen the next time you conduct the statistical experiment. Note that ***n*** is the sample size.
$x - \bar{x}$	This is the **difference** between what happened and what you expected to happen. This represents a "deviation" away from what you "expect" and is a measure of risk.
$\Sigma(x - \bar{x})^2$	The expression $\Sigma(x - \bar{x})^2$ is called the **sum of squares.** The $(x - \bar{x})$ quantity is squared to make it nonnegative. The sum is over all the data. If you don't square $(x - \bar{x})$, then the sum $\Sigma(x - \bar{x})$ is equal to 0 because the negative values cancel the positive values. This occurs even if some $(x - \bar{x})$ values are large, indicating a large deviation or risk.
Sum of squares $\Sigma(x - \bar{x})^2$ or $\Sigma x^2 - \frac{(\Sigma x)^2}{n}$	This is an **algebraic simplification of the sum of squares** that is easier to compute. The **defining formula** for the sum of squares is the upper one. The **computation formula** for the sum of squares is the lower one. Both formulas give the same result.
Sample variance $s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$ or $s^2 = \frac{\Sigma x^2 - (\Sigma x)^2/n}{n - 1}$	**The sample variance is s^2.** The variance can be thought of as a kind of average of the $(x - \bar{x})^2$ values. However, for technical reasons, we divide the sum by the quantity $n - 1$ rather than n. This gives us the best mathematical estimate for the sample variance.

Continue

PROCEDURE *continued*



The **defining formula** for the variance is the upper one. The **computation formula** for the variance is the lower one. Both formulas give the same result.

Sample standard deviation

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

or

$$s = \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2/n}{n - 1}}$$

This is **sample standard deviation, *s*.** Why do we take the square root? Well, if the original *x* units were, say, days or dollars, then the s^2 units would be days squared or dollars squared (wow, what's that?). We take the square root to return to the original units of the data measurements. The standard deviation can be thought of as a measure of variability or risk. Larger values of *s* imply greater variability in the data.

The **defining formula** for the standard deviation is the upper one. The **computation formula** for the standard deviation is the

Defining formulas (sample statistics)

$$\text{Sample variance} = s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} \qquad (1)$$

$$\text{Sample standard deviation} = s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}} \qquad (2)$$

where x is a member of the data set, $\bar{x}$ is the mean, and n is the number of data values. The sum is taken over all data values.

Computation formulas (sample statistics)

$$\text{Sample variance} = s^2 = \frac{\Sigma x^2 - (\Sigma x)^2/n}{n - 1} \qquad (3)$$

$$\text{Sample standard deviation} = s = \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2/n}{n - 1}} \qquad (4)$$

where x is a member of the data set, $\bar{x}$ is the mean, and n is the number of data values. The sum is taken over all data values.

COMMENT The computation formula for the population standard deviation is

$$\sigma = \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2/N}{N}}$$

Population Parameters

$$\text{Population mean} = \mu = \frac{\Sigma x}{N}$$

$$\text{Population variance} = \sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$$

$$\text{Population standard deviation} = \sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}$$

where N is the number of data values in the population and x represents the individual data values of the population.

Coefficient of Variation

A disadvantage of the standard deviation as a comparative measure of variation is that it depends on the units of measurement. This means that it is difficult to use the standard deviation to compare measurements from different populations. For this reason, statisticians have defined the *coefficient of variation,* which expresses the standard deviation as a percentage of the sample or population mean.

If $\bar{x}$ and s represent the sample mean and sample standard deviation, respectively, then the sample coefficient of variation CV is defined to be

$$CV = \frac{s}{\bar{x}} \cdot 100$$

If μ and σ represent the population mean and population standard deviation, respectively, then the population coefficient of variation CV is defined to be

$$CV = \frac{\sigma}{\mu} \cdot 100$$

Notice that the numerator and denominator in the definition of CV have the same units, so CV itself has no units of measurement. This gives us the advantage of being able to directly compare the variability of two different populations using the coefficient of variation.

In the next example and guided exercise, we will compute the CV of a population and of a sample and then compare the results.

Sample standard deviation (defining formula)

Big Blossom Greenhouse was commissioned to develop an extra large rose for the Rose Bowl Parade. A random sample of blossoms from Hybrid A bushes yielded the following diameters (in inches) for mature peak blooms.

2 3 3 8 10 10

Find the sample variance and standard deviation.

SOLUTION: Several steps are involved in computing the variance and standard deviation. A table will be helpful (see Table 3-1 on the next page). Since $n = 6$, we take the sum of the entries in the first column of Table 3-1 and divide by 6 to find the mean $\bar{x}$.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{36}{6} = 6.0 \text{ inches}$$

TABLE 3-1 Diameters of Rose Blossoms (in inches)

Column I	Column II	Column III
x	$x - \bar{x}$	$(x - \bar{x})^2$
2	$2 - 6 = -4$	$(-4)^2 = 16$
3	$3 - 6 = -3$	$(-3)^2 = 9$
3	$3 - 6 = -3$	$(-3)^2 = 9$
8	$8 - 6 = 2$	$(2)^2 = 4$
10	$10 - 6 = 4$	$(4)^2 = 16$
10	$10 - 6 = 4$	$(4)^2 = 16$
$\Sigma x = 36$		$\Sigma(x - \bar{x})^2 = 70$

Using this value for $\bar{x}$, we obtain Column II. Square each value in the second column to obtain Column III, and then add the values in Column III. To get the sample variance, divide the sum of Column III by $n - 1$. Since $n = 6$, $n - 1 = 5$.

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} = \frac{70}{5} = 14$$

Now obtain the sample standard deviation by taking the square root of the variance.

$$s = \sqrt{s^2} = \sqrt{14} \approx 3.74$$

(Use a calculator to compute the square root. Because of rounding, we use the approximately equal symbol, ≈.)

Example1. **Calculate the S.D. and coefficient of variation (C.V.) for the following table :**

Class	:	**0-10**	**10-20**	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**	**70-80**
Frequency	:	**5**	**10**	**20**	**40**	**30**	**20**	**10**	**5**

Solution. **We prepare the following table for the computation of S.D.**

Class	Mid-value x	f	$u = \frac{x-35}{10}$	fu	fu^2
0-10	5	5	-3	-15	45
10-20	15	10	-2	-20	40
20-30	25	20	-1	-20	20
30-40	35	40	0	0	0
40-50	45	30	1	30	30
50-60	55	20	2	40	80
60-70	65	10	3	30	90
70-80	75	5	4	20	80
		$N=\sum f = 140$		$\sum fu = 65$	$\sum fu^2 = 385$

Let assumed mean = 35 = A (say) and h = 10

A.M. , M $= A + h\,(\sum fu)/N$ **= 35 + 10 (65/140)**

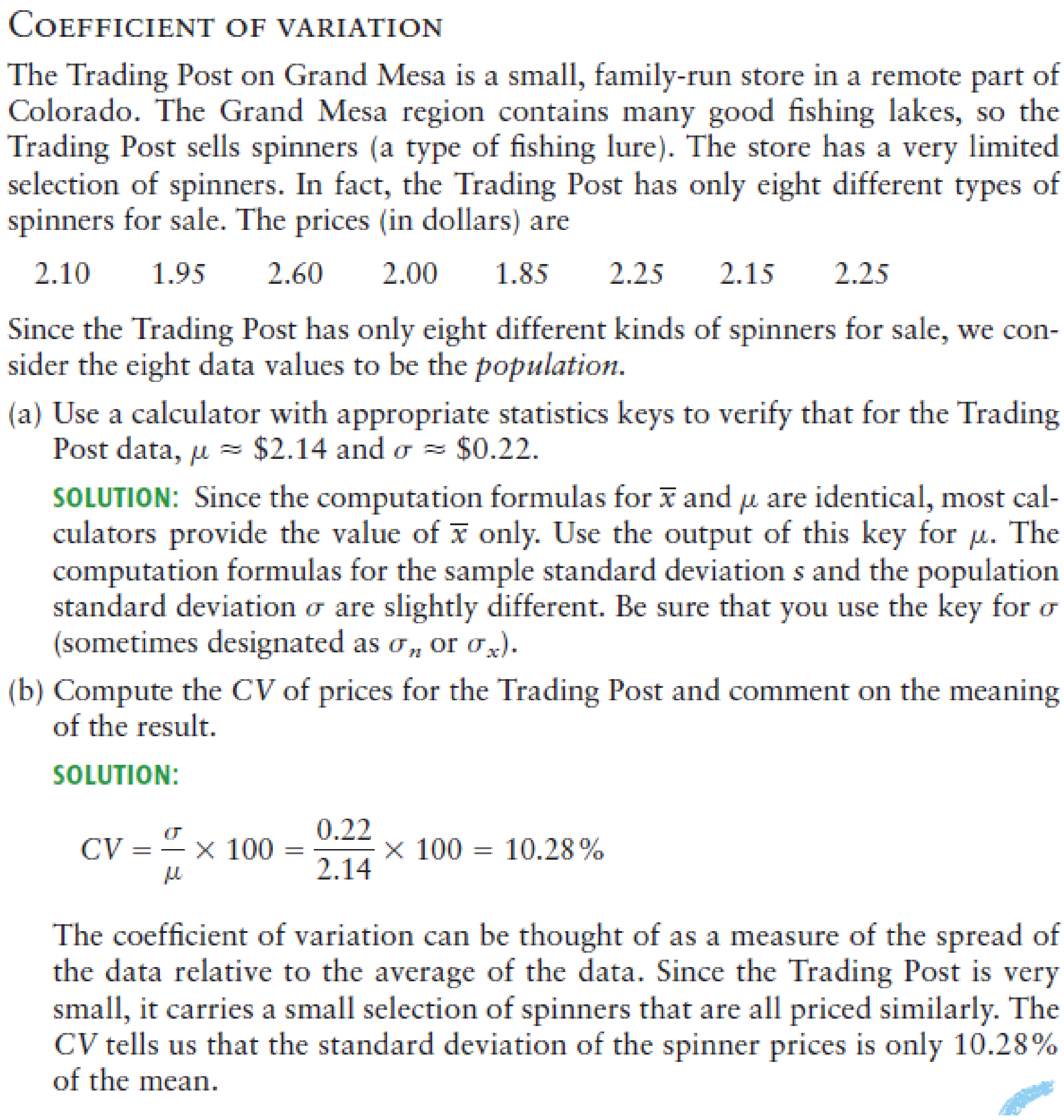
= 35 + 4.64 = 39.64

S.D., σ $= h\sqrt{[\frac{\sum fu^2}{N} - (\frac{\sum fu}{N})^2]}$

$$= 10\sqrt{[\frac{385}{140} - (.464)^2]}$$

$$= 10\sqrt{[2.75 - .215]} \;=\; 10\sqrt{(2.535)} \;=\; \mathbf{10 \text{ x } 1.59 = 15.9}$$

C.V. = σ/M x100 = 15.9/39.64 x 100 = 40.11%.



Coefficient of variation

The Trading Post on Grand Mesa is a small, family-run store in a remote part of Colorado. The Grand Mesa region contains many good fishing lakes, so the Trading Post sells spinners (a type of fishing lure). The store has a very limited selection of spinners. In fact, the Trading Post has only eight different types of spinners for sale. The prices (in dollars) are

2.10 1.95 2.60 2.00 1.85 2.25 2.15 2.25

Since the Trading Post has only eight different kinds of spinners for sale, we consider the eight data values to be the *population*.

(a) Use a calculator with appropriate statistics keys to verify that for the Trading Post data, $\mu \approx \$2.14$ and $\sigma \approx \$0.22$.

SOLUTION: Since the computation formulas for $\overline{x}$ and μ are identical, most calculators provide the value of $\overline{x}$ only. Use the output of this key for μ. The computation formulas for the sample standard deviation s and the population standard deviation σ are slightly different. Be sure that you use the key for σ (sometimes designated as σ_n or σ_x).

(b) Compute the *CV* of prices for the Trading Post and comment on the meaning of the result.

SOLUTION:

$$CV = \frac{\sigma}{\mu} \times 100 = \frac{0.22}{2.14} \times 100 = 10.28\%$$

The coefficient of variation can be thought of as a measure of the spread of the data relative to the average of the data. Since the Trading Post is very small, it carries a small selection of spinners that are all priced similarly. The *CV* tells us that the standard deviation of the spinner prices is only 10.28% of the mean.

Standard Error:

It is the standard deviation of the sampling distribution of a statistic. It can be abbreviated as S.E. The standard error of the sample mean depends on both the standard deviation and the sample size, by the simple relation SE = SD/ √(sample size).

The magnitude of the standard error gives an index of the precision of the estimate of the parameter. The reciprocal is generally taken as the measure of the reliability or the precision of the statistic. In other words, it is inversely proportional to the sample size. This means that the greater the standard error, the smaller the size of the sample.

Question1: Calculate the range of following data 24, 52, 36, 89, 100, 12, 40, 60, 15

Question2: Calculate the sample standard deviation of following data 2, 2, 5, 7

Question 3: Find the mean deviation about the mean of following data:

12, 3, 18, 17, 4, 9, 19, 17, 20, 15, 8, 17, 2, 3, 16, 11, 3, 1, 0, 5

Question 4: Calculate the variance and standard deviation of following frequency distribution

Marks obtained in Physics:	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Number of students:	3	7	12	15	8	3	2

Question 5: Following values are calculated in respect of height and weight of students

	Height	**Weight**
Mean	162.6 cm	52.36 kg
Variance	127.69 cm^2	23.13 kg^2

Can we say that weights show greater variation than the height?

EXERCISE No.-7

CORRELATION AND REGRESSION ANALYSIS

The statistical methods discussed so far are used to analyze data involving only one variable. Often an analysis of data concerning two or more variables is needed to look for any statistical relationship or association between them.

There are **two different techniques** which are used for the study of two or more variables:

1. **Regression 2. Correlation**

Both study the behaviour of the variables but they differ in their end results. Regression studies the relationship where dependence is necessarily involved. One variable is dependent on a certain number of variables. Regression can be used for predicting the values of a variable which depends upon other variables.

Correlation attempts to study the strength of the mutual relationship between two variables. In correlation we assume that the variables are random and dependence of any nature is not involved.

Correlation

Correlation is a technique which measures the strength of association between two variables. Both the variables X and Y may be random and it may be that one variable is independent (non-random) and the other is correlated or dependent. When the changes in one variable appear to be linked with the changes in the other variable, the two variables are said to be correlated.

Types of Correlation:

a. **Positive and Negative Correlation**

Positive Correlation

A correlation in the same direction is called a positive correlation. If one variable increases the other also increases and when one variable decreases the other also decreases.

For example, the length of an iron bar will increase as the temperature increases.

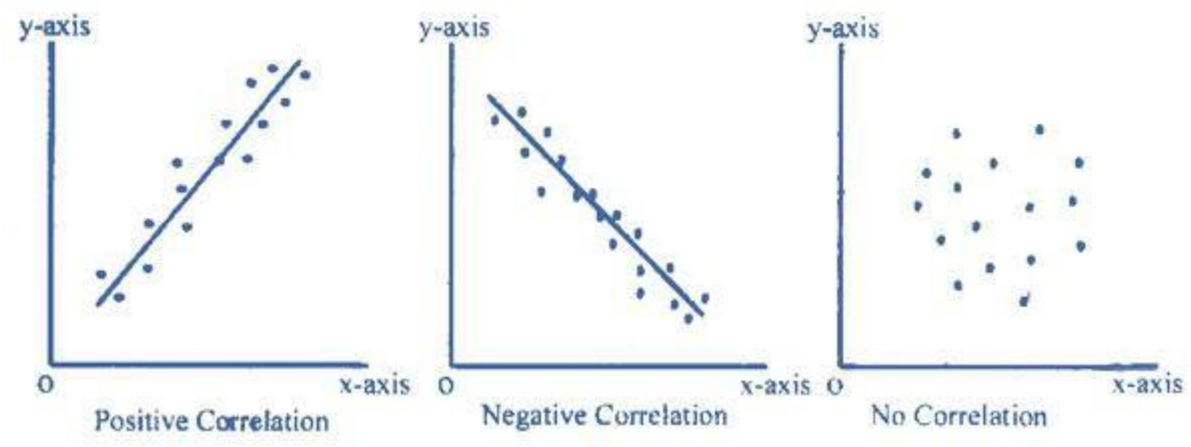
Negative Correlation

Correlation in the opposite direction is called a negative correlation. Here if one variable increases the other decreases and vice versa.

For example, the volume of gas will decrease as the pressure increases, or the demand for a particular commodity increases as the price of such commodity decreases.

No Correlation or Zero Correlation

If there is no relationship between the two variables such that the value of one variable changes and the other variable remains constant, it is called no or zero correlation.



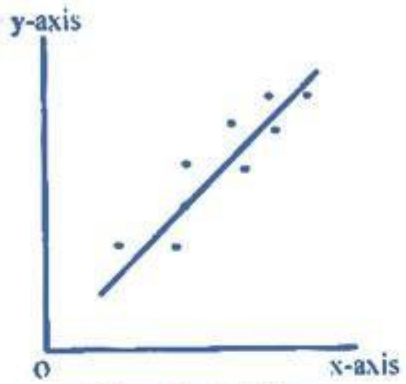
b. **Linear and Non Linear Correlation**

Linear Correlation

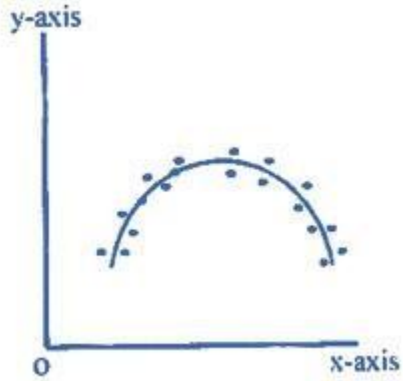
Correlation is said to be linear if the ratio of change is constant. When the amount of output in a factory is doubled by doubling the number of workers, this is an example of linear correlation. **In other words**, when all the points on the scatter diagram tend to lie near a line which looks like a straight line, the correlation is said to be linear. This is shown in the figure on the left below.

Non Linear (Curvilinear) Correlation

Correlation is said to be non linear if the ratio of change is not constant. In other words, when all the points on the scatter diagram tend to lie near a smooth curve, the correlation is said to be non linear (curvilinear). This is shown in the figure on the right below.



Linear Correlation



Non Linear Correlation

Methods of Study of Correlation:

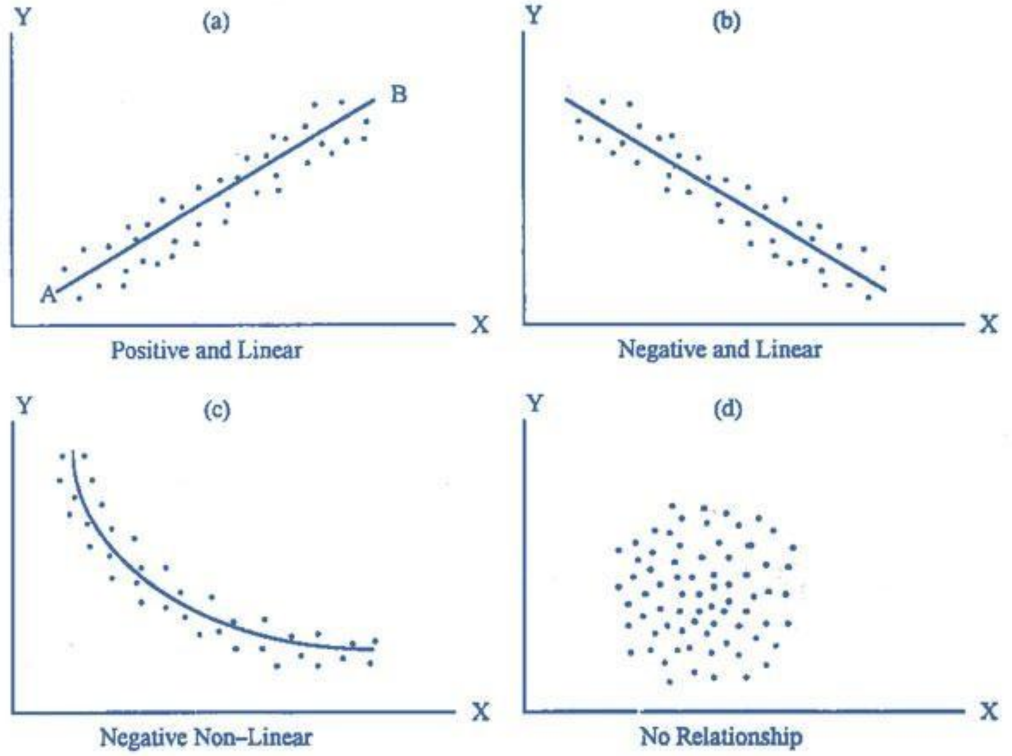
1. Scatter Diagram method
2. Karl Pearson‘s Method by finding correlation coefficient

1. **The Scatter Diagram:** A scatter diagram is a graphic picture of the sample data.

Suppose a random sample of n pairs of observations has the values (X1,Y1), (X2,Y2), (X3,Y3) ,…, (Xn,Yn). These points are plotted on a rectangular co-ordinate system putting the independent variable on the X-axis and the dependent variable on the Y-axis. No matter what the independent variable is, it must be placed on the X-axis.

Suppose the plotted points are as shown in figure (a). Such a diagram is called a scatter diagram. In this figure, we see that when X has a small value Y is also small, and when X has a large value Y also has a large value. This is called a direct or positive relationship between X and Y. The plotted points cluster around a straight line. It appears that if a straight line is drawn passing through the points, the line will be a good approximation to represent the original data.

Suppose we draw a line AB to represent the scattered points. The line AB rises from left to right and has a positive slope. This line can be used to establish an approximate relation between the random variable Y and the independent variable X. It is a nonmathematical method in the sense that different people may draw different lines. This line is called the regression line obtained by inspection or judgment.



Making a scatter diagram and drawing a line or curve is the primary investigation to assess the type of relationship between the variables. The knowledge gained from the scatter diagram can be used for further analysis of the data. In most of the cases, the diagrams are not as simple as in figure (a). There are quite complicated diagrams and it is difficult to choose a proper mathematical model to represent the original data. The scatter diagram gives an indication of the appropriate model which should be used for further analysis with the help of the method of least squares.

Figure (b) shows that the points in the scatter diagram are falling from the top left corner to the right. This is a relation called inverse or indirect. The points are in the neighborhood of a certain line called the regression line.

As long as the scattered points show closeness to a straight line in some direction, we draw a straight line to represent the sample data. But when the points do not lie around a straight line, we do not draw the regression line. Figure (c) shows that the plotted points have a tendency to fall from left to right in the form of a curve. This is a relation called non-linear or curvilinear. Figure (d) shows points which apparently do not follow any pattern. If X has a small value, Y may have a small or large value.

There seems to be no relationship between X and Y. Such a diagram suggests that there is no relationship between the two variables.

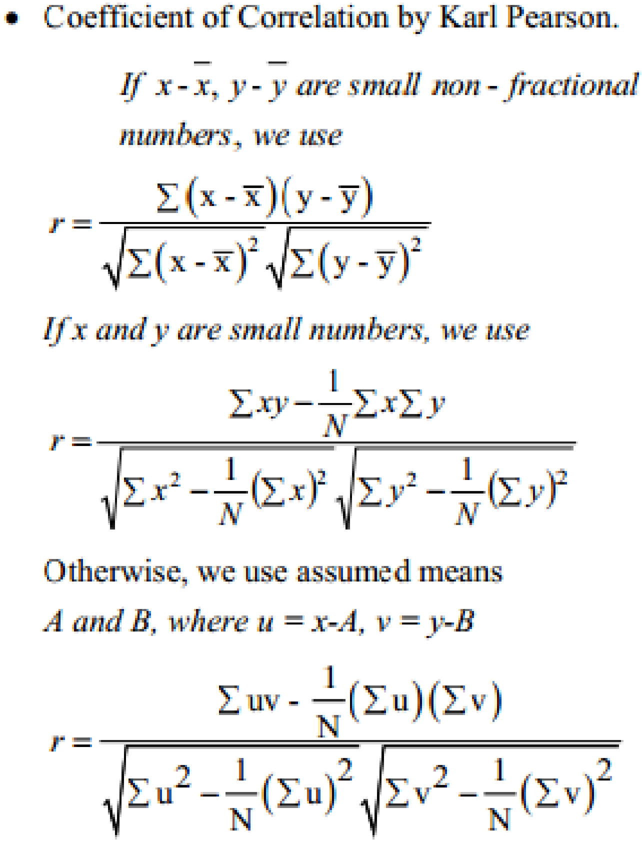
2. Coefficient of Correlation:

If X and Y are two variable then correlation coefficient Cov(X,Y) is given as:

$$r = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X).\operatorname{var}(Y)}}$$

$$r = \frac{\operatorname{cov}(X,Y)}{\sqrt{\sigma_X.\sigma_Y}}$$

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right]}}$$



- Coefficient of Correlation by Karl Pearson.

If $x - \bar{x}$, $y - \bar{y}$ are small non - fractional numbers, we use

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2}\sqrt{\Sigma(y - \bar{y})^2}}$$

If x and y are small numbers, we use

$$r = \frac{\Sigma xy - \frac{1}{N}\Sigma x \Sigma y}{\sqrt{\Sigma x^2 - \frac{1}{N}(\Sigma x)^2}\sqrt{\Sigma y^2 - \frac{1}{N}(\Sigma y)^2}}$$

Otherwise, we use assumed means

A and B, where u = x-A, v = y-B

$$r = \frac{\Sigma uv - \frac{1}{N}(\Sigma u)(\Sigma v)}{\sqrt{\Sigma u^2 - \frac{1}{N}(\Sigma u)^2}\sqrt{\Sigma v^2 - \frac{1}{N}(\Sigma v)^2}}$$

Examples of Correlation

Calculate and analyze the correlation coefficient between the number of study hours and the number of sleeping hours of different students.

Number of Study Hours	**2**	**4**	**6**	**8**	**10**
Number of Sleeping Hours	**10**	**9**	**8**	**7**	**6**

Solution:

The necessary calculations are given below:

X	**Y**	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X}) \cdot (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
2	**10**	**-4**	**+2**	**-8**	**16**	**4**
4	**9**	**-2**	**+1**	**-2**	**4**	**1**
6	**8**	**0**	**0**	**0**	**0**	**0**
8	**7**	**+2**	**-1**	**-2**	**4**	**1**
10	**6**	**+4**	**-2**	**-8**	**16**	**1**
$\sum X=30$, $\sum Y=40$ $\sum(X-\bar{X})=0$, $\sum(Y-\bar{Y})=0$ $\sum(X-\bar{X})(Y-\bar{Y})=-20$, $\sum(X-\bar{X})^2=40$, $\sum(Y-\bar{Y})^2=10$						

$\bar{X}=\sum X/n=30/5=6$

$\bar{Y}=\sum Y/n=40/5=8$

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right]\left[\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right]}} = -\frac{20}{20} = -1$$

There is a perfect negative correlation between the number of study hours and the number of sleeping hours.

<u>Example</u>:

From the following data, compute the coefficient of correlation between X and Y:

	X Series	Y Series
Number of Items	**15**	**15**
Arithmetic Mean	**25**	**18**
Sum of Square Deviations	**136**	**138**

The summation of the products of the deviations of X and X series from their arithmetic means = **122**.

Solution:

Here $n=15$, $\bar{X}=25$, $\bar{Y}=18$, $\sum(X-\bar{X})^2=\sum(Y-\bar{Y})^2=138$ and $\sum(X-\bar{X})^2(Y-\bar{Y})^2= 122$

Hence,

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right]\left[\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right]}}=\frac{122}{\sqrt{(136)(138)}}=\frac{122}{137}=0.89$$

Regression Analysis

The **Regression Analysis** is a statistical tool used to determine the probable change in one variable for the given amount of change in another. This means, the value of the unknown variable can be estimated from the known value of another variable.

Regression analysis is a form of predictive modelling technique which investigates the relationship between a **dependent** (target) and **independent variable (s)** (predictor).

This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

Linear Regression

In Linear regression the dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.

Linear Regression establishes a relationship between **dependent variable (Y)** and one or more **independent variables (X)** using a **best fit straight line** (also known as regression line). It is represented by an equation **Y=a+b*X + e**, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).

The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable.

Methods of Regression Analysis: We can study regression by the following methods:

1. Graphic method (regression lines)
2. Algebraic method

Regression Lines: There are two regression line:

The regression line of x on y → $x - \overline{x} = b_{xy}\,(y - \overline{y})$

The regression line of y on x → $y - \overline{y} = b_{yx}\,(x - \overline{x})$

Algebraic Method: The algebraic method for simple linear regression can be understood by two methods:

1. Regression Equations
2. Regression Coefficients

Regression Equations: These equations are known as estimating equations. Regression equations are algebraic expressions of the regression lines.

As there are two regression lines, there are two regression equations :

(i) x on y is used to describe the variations in the values of x for given changes in y.

(ii) y on x is used to describe the variations in the values of y for given changes in x.

The regression equations **of y on x** is expressed as Y = a + bX

The regression equations **of x on y** is expressed as X = a + bY

In these equations a and b are constants which determine the position of the line completely.

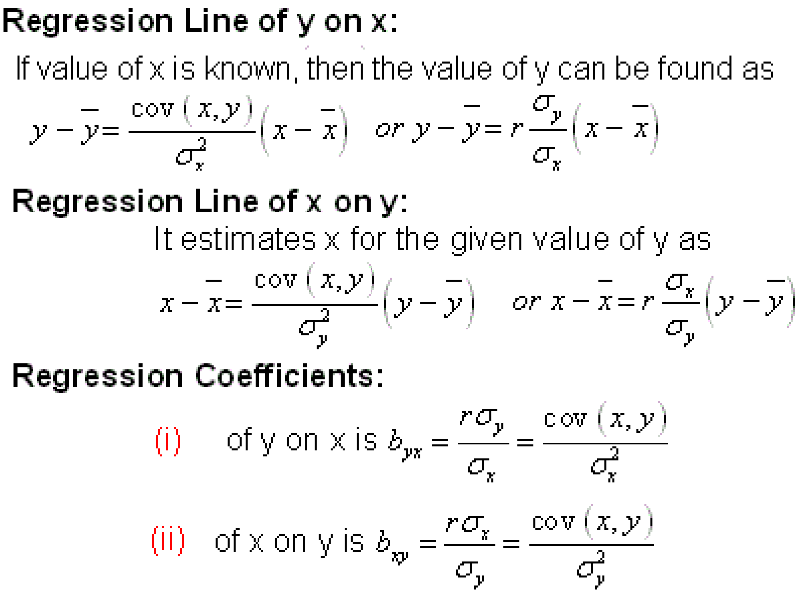
Regression Coefficient:

The **Regression Coefficient** is the constant _b' in the regression equation that tells about the change in the value of dependent variable corresponding to the unit change in the independent variable. There are two regression coefficients

If there are two regression equations, then there will be two regression coefficients b_{xy} **and** b_{yx}

a) **Regression Coefficient of X on Y:** The regression coefficient of X on Y is represented by the symbol $\mathbf{b_{xy}}$ that measures the change in X for the unit change in Y.

b) **Regression Coefficient of Y on X:** The symbol b_{yx} is used that measures the change in Y corresponding to the unit change in X.



Regression Line of y on x:

If value of x is known, then the value of y can be found as

$$y-\bar{y}=\frac{\operatorname{cov}(x,y)}{\sigma_x^2}(x-\bar{x}) \quad or\ y-\bar{y}=r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

Regression Line of x on y:

It estimates x for the given value of y as

$$x-\bar{x}=\frac{\operatorname{cov}(x,y)}{\sigma_y^2}(y-\bar{y}) \quad or\ x-\bar{x}=r\frac{\sigma_x}{\sigma_y}(y-\bar{y})$$

Regression Coefficients:

(i) of y on x is $b_{yx}=\frac{r\sigma_y}{\sigma_x}=\frac{\operatorname{cov}(x,y)}{\sigma_x^2}$

(ii) of x on y is $b_{xy}=\frac{r\sigma_x}{\sigma_y}=\frac{\operatorname{cov}(x,y)}{\sigma_y^2}$

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$a = \frac{\sum y - b\sum x}{n}$$

How to find regression equation and relationship between two variables using the slope and y-intercept?

A regression is a statistical analysis assessing the association between two variables. In simple linear regression, a single independent variable is used to predict the value of a dependent variable.

Regression Formula:

A. **Regression Equation (y)** = a + bx
B. **Slope (b) =** (NΣXY - (ΣX)(ΣY)) / (NΣX^2 - $(ΣX)^2$)
C. **Intercept (a) =** (ΣY - b(ΣX)) / N

Where**,**

x and y are the variables.

b = The slope of the regression line/ Regression coefficient

a = The intercept point of the regression line and the y axis.

N = Number of values or elements

X = First Score

Y = Second Score

ΣXY = Sum of the product of first and Second Scores

ΣX = Sum of First Scores

ΣY = Sum of Second Scores

ΣX^2 = Sum of square First Scores

Example:

To find the Simple/Linear Regression of the given values

X –Values	Y- Values
60	3.1
61	3.6
62	3.8
63	4
65	4.1

To find regression equation, we will first find slope, intercept and use it to form regression equation.

Step 1:

Count the number of values. N = 5

Step 2:

Find XY, X^2

See the below table

X Value	Y Value	X*Y	X*X
60	3.1	60 * 3.1 =186	60 * 60 = 3600
61	3.6	61 * 3.6 = 219.6	61 * 61 = 3721
62	3.8	62 * 3.8 = 235.6	62 * 62 = 3844
63	4	63 * 4 = 252	63 * 63 = 3969
65	4.1	65 * 4.1 = 266.5	65 * 65 = 4225

Step 3:

Find ΣX, ΣY, ΣXY, ΣX^2.

ΣX = 311 ΣY = 18.6 ΣXY = 1159.7 ΣX^2 = 19359

Step 4:

Substitute in the above slope formula given. Slope/ Regression coefficient (b)

(b) = (NΣXY - (ΣX)(ΣY)) / (NΣX^2 - (ΣX)^2)

= ((5)*(1159.7)-(311)*(18.6))/((5)*(19359)-(311)^2)

= (5798.5 - 5784.6)/(96795 - 96721) = 13.9/74 = 0.18784

Step 5:

Now, again substitute in the above intercept formula given.

Intercept (a) = (ΣY – b(ΣX)) / N

= (18.6 - 0.18784(311))/5 = (18.6 - 58.41824)/5 = -39.81824/5 = -7.964

Step 6:

Then substitute these values in regression equation formula Regression Equation

(y) = a + bx

=-7.964+0.188x.

Suppose if we want to know the approximate y value for the variable x = 64. Then we can substitute the value in the above equation.

Regression Equation(y) = a + bx = -7.964+0.188(64). = -7.964+12.032. = 4.068

This example will guide you to find the relationship between two variables by calculating the Regression from the above steps.

Question1: Calculate the correlation coefficient between height of father and height of son from the given data:

Height of father (in inches)	64	65	66	67	68	69	70
Height of son (in inches)	66	67	65	68	70	68	72

Question 2: The coefficient of correlation between two variates X and Y is 0.8 and their covariance is 20. If variance of X series is 16, find the standard deviation of Y series

Question 3: Write the relation between correlation and regression.

Question 4: Compute the two regression coefficients from the data given below and find the correlation coefficient by using regression coefficients.

X	6	7	4	8	5
Y	8	7	5	9	2

EXERCISE No.-8

SIMPLE PROBABILITY

Probability is the measure of the likelihood that an event will occur. Probability of an event is the ratio of the **number of observations of the event** to the **total numbers of the observations**.

Important Points to be remember:

1. All possible outcomes in a trial , are called exhaustive events. For example, if an unbiased die is rolled, then we may obtain any one of the six numbers1,2,3,4,5 and 6. Hence there are six exhaustive events in this trial.
2. The total number of favorable outcomes (or ways) in a trail, to happen an event, are called favorable events. For example, If a pair of fair dice is tossed then the favorable events to get the sum 7 are six : (1,6), (2,5), (3,4), (4,3), (5,2), (6,1,).
3. When we throw a coin, then either a Head (H) or a Tail (T) appears.
4. A dice is a solid cube, having 6 faces, marked 1, 2, 3, 4, 5, 6 respectively. When we throw a die, the outcome is the number that appears on its upper face.
5. A pack of cards has 52 cards.

   It has 13 cards of each suit, name **Spades, Clubs, Hearts and Diamonds**.

   Cards of spades and clubs are **black cards**.

   Cards of hearts and diamonds are **red cards**.

   There are 4 honours of each unit.

   There are **Kings, Queens and Jacks**. These are all called **face cards**.

Probability of Event:

Probability of occurrence of an event E is calculated as,

$$P(E) = \frac{\text{Number of outcomesfavourable tooccuranceof event E}}{\text{Totalnumber of allpossibleoutcomesS}}$$

Example:

Tickets numbered 1 to 20 are mixed up and then a ticket is drawn at random. What is the probability that the ticket drawn has a number which is a multiple of 3 or 5?

Solution:

Here, S = {1, 2, 3, 4, ..., 19, 20}.

Let E = event of getting a multiple of 3 or 5 = {3, 6 , 9, 12, 15, 18, 5, 10, 20}.

$$\therefore\ P(E) = \frac{n(E)}{n(S)} = \frac{9}{20}.$$

Example:

Two dice are thrown simultaneously. What is the probability of getting two numbers whose product is even?

Solution:

In a simultaneous throw of two dice, we have *n*(S) = (6 x 6) = 36.

Then, E = {(1, 2), (1, 4), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 2), (3, 4),
(3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (5, 2), (5, 4), (5, 6), (6, 1),
(6, 2), (6, 3), (6, 4), (6, 5), (6, 6)}

$\therefore$ *n*(E) = 27.

$$\therefore\ P(E) = \frac{n(E)}{n(S)} = \frac{27}{36} = \frac{3}{4}.$$

Example:

From a pack of 52 cards, two cards are drawn together at random. What is the probability of both the cards being kings?

Solution:

Let S be the sample space.

$$\text{Then, } n(S) = {}^{52}C_2 = \frac{(52 \text{ x } 51)}{(2 \text{ x } 1)} = 1326.$$

Let E = event of getting 2 kings out of 4.

$$\therefore\ n(E) = {}^4C_2 = \frac{(4 \text{ x } 3)}{(2 \text{ x } 1)} = 6.$$

$$\therefore\ P(E) = \frac{n(E)}{n(S)} = \frac{6}{1326} = \frac{1}{221}.$$

Addition Rule:

If A and B are events, the probability of obtaining either of them. The probability of occurrence of two mutually exclusive events either A or B is the sum of the probabilities of the events occurring separately

If two events A and B are mutually exclusive then A ∩B =0,

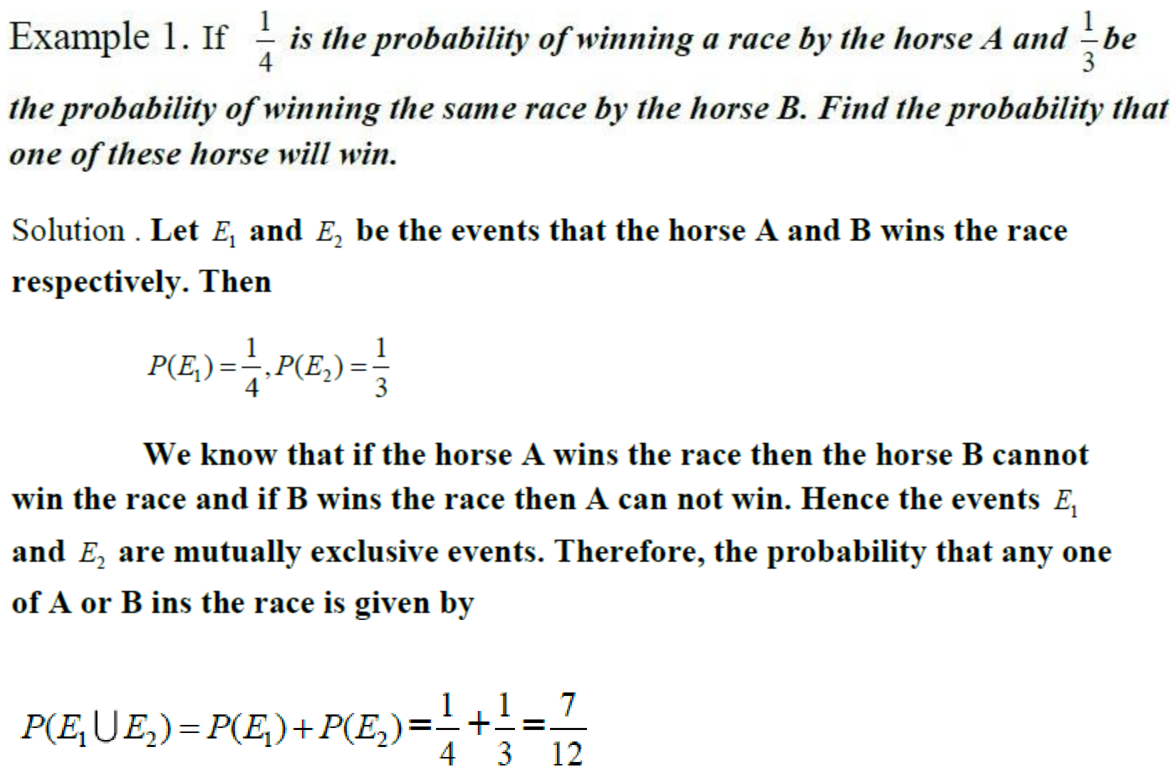
$$P\ (A \cup B) = P\ (A) + P\ (B)$$

The above formula can be expanded to consider more than two exclusive events:

$$P\ (A \text{ or } B \text{ or } C \text{ or } D\ldots \text{ or } Z) = P\ (A) + P\ (B) + P(C) + \ldots + P\ (Z)$$

When A and B are not mutually exclusive, then A ∩B ≠0,

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



Example 1. If $\frac{1}{4}$ ***is the probability of winning a race by the horse A and*** $\frac{1}{3}$ ***be the probability of winning the same race by the horse B. Find the probability that one of these horse will win.***

Solution . **Let** E_1 **and** E_2 **be the events that the horse A and B wins the race respectively. Then**

$$P(E_1) = \frac{1}{4}, P(E_2) = \frac{1}{3}$$

We know that if the horse A wins the race then the horse B cannot win the race and if B wins the race then A can not win. Hence the events E_1 **and** E_2 **are mutually exclusive events. Therefore, the probability that any one of A or B ins the race is given by**

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) = \frac{1}{4} + \frac{1}{3} = \frac{7}{12}$$

Multiplication Rule:

The probability of independent events A and B occurring is the product of the probabilities of the events occurring separately

$$P\ (A \text{ and } B) = P\ (A).\ P\ (B)$$

The above formula can be expanded. If A, B, C... Z are independent events, then:

P (A and B and C and ... and Z) = P (A). P (B). P(C) ... P (Z)

When events are **dependent**, each possible outcome is related to the other. Given two events A and B, the probability of obtaining both A and B is the product of the probability of obtaining one of the events times the conditional probability of obtaining the other event, given the first event has occurred.

P (A and B) = P (A). P (B|A)

The multiplication rule for several dependent events can be extended to several dependent events:

$$P\ (A \text{ and } B \text{ and } C) = P\ (A).\ P\ (B|A).\ P\ (C|A \text{ and } B)$$

- ➢ Two events A and B are mutually exclusive or disjoint if:
  - ✓ $P\ (A\cup B) = P\ (A) + P\ (B)$
  - ✓ $P\ (A\cap B) = 0$
- ➢ Two events A and B are independent if:
  - ✓ $P\ (A\cap B) = P\ (A) \cdot P\ (B)$

Example: Consider another example where a pack contains 4 blue, 2 red and 3 black pens. If a pen is drawn at random from the pack, replaced and the process repeated 2 more times, what is the probability of drawing 2 blue pens and 1 black pen?

Solution

Here, total number of pens = 9

Probability of drawing 1 blue pen = 4/9

Probability of drawing another blue pen = 4/9

Probability of drawing 1 black pen = 3/9

Probability of drawing 2 blue pens and 1 black pen = 4/9 * 4/9 * 3/9 = 48/729 = 16/243

Conditional Probability

When probability of an event A is given and we have to find the probability of other event B based on that event A, then the probability obtained is called conditional probability of B given A. It is denoted by P (B/A).

If A and B are two events such that given event B has happened and we have to find the probability of event A, then probability of A given B, denoted by P(A/B) is given by

$$P(A/B) = P(A \cap B) / P(B)$$

Equivalently $$P(B|A) = P(A \cap B)/P(A)$$

$$P(A \cap B) = P(B|A)\,P(A) \text{ or } P(A \cap B) = P(B/A) \times P(A)$$

Example1. ***if A and B are two events, where***

$P(A)=\frac{1}{2}, P(B)=\frac{1}{3}, and\ P(A\cap B)=\frac{1}{4}$***,then evaluate the following:***

a. $P(A/B)$

b. $P(B/A)$

c. $P(A\cup B)$

Solution.

a. $$P(A\cap B)=P(B).P(A/B)$$

$$P(A/B)=\frac{P(A\cap B)}{P(B)}=\frac{\frac{1}{4}}{\frac{1}{3}}=\frac{3}{4}$$

b. $$P(B/A)=\frac{P(B\cap A)}{P(A)}=\frac{P(A\cap B)}{P(A)}=\frac{\frac{1}{4}}{\frac{1}{2}}=\frac{2}{4}=\frac{1}{2}$$

c. $$P(A\cup B)=P(A)+P(B)-P(A\cap B)=\frac{1}{2}+\frac{1}{3}-\frac{1}{4}=\frac{7}{12}$$

Question 1: A dice is thrown 3 times .what is the probability that at least one head is obtained?

Question 2: 1 card is drawn at random from the pack of 52 cards.
(i) Find the Probability that it is an honor card.
(ii) It is a face card.

Question 3: Three dice are rolled together. What is the probability as getting at least one '4'?

Question 4: A problem is given to three persons P, Q, R whose respective chances of solving it are 2/7, 4/7, 4/9 respectively. What is the probability that the problem is solved?

Question 5: In a class, 40% of the students study math and science. 60% of the students study math. What is the probability of a student studying science given he/she is already studying math?

Question 6: What is the probability of getting a 2 or a 5 when a die is rolled?

Question 7: Write-down the laws of probability.

EXERCISE No.-9

TEST OF SIGNIFICANCE (HYPOTHESIS TESTING)

Once sample data has been gathered through an observational study or experiment, statistical inference allows analysts to assess evidence in favour or some claim about the population from which the sample has been drawn. The methods of inference used to support or reject claims based on sample data are known as ***tests of significance***. **Significance testing** aims to quantify evidence against a particular hypothesis being true. **Hypothesis testing** rather looks at evidence for a particular hypothesis being true.

The usual process of hypothesis testing consists following steps:

1. State the null hypothesis H_0 and the alternative hypothesis H_1 about the population parameter.
2. Choose the level of significance. (The acceptable risk of a Type 1 error).
3. Choose the appropriate test statistic to assess the truth of the null hypothesis.
4. Compute the value of used test statistic.
5. Calculate the p-value of the test statistic.
6. Compare this p-value to the original α.
    - If $p < \alpha$, reject H_0. Conclude that H_1 is plausible.
    - If $p \geq \alpha$ do not reject H_0. You do not conclude that H_1 is plausible.

Test statistics generally which are used in testing of statistical hypothesis:

There are following test statistics generally that are used in testing of hypothesis

1. Student‘s t-test
2. Chi-square test
3. F-test
4. Z-test or normal deviate test

Applications of t- statistics:

The t- test has a large number of applications in statistics which are given as below:

1. To test if sample mean differs significantly from hypothetical value of population mean (μ_0).
2. To test the equality of two population means based on sample means.
3. To test the significance of sample correlation coefficient.
4. To test the significance of sample regression coefficient.
5. To test the significance of sample partial correlation coefficient

One-sample *t*-test.:

Ans. Suppose we want to test the null hypothesis that the population mean is equal to a specified value μ_0 , (H0: $\mu = \mu_0$ v/s H1 : $\mu \neq \mu_0$) on the basis of random sample x1, x2, x3…xn of size n <30 drawn from a normal population which variance is unknown.

One uses the following t-statistic

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}, \quad \text{where } s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{(n-1)}}$$

Where $\bar{x}$ the sample mean, s is is the standard deviation of the sample and n is the sample size. The degrees of freedom used in this test are $n - 1$.

Test statistic for testing the equality of two populations mean:

For testing the equality of two populations mean on basis of two samples taken from these normal populations, the independent two-sample *t*-test is used. A two-sample t-test examines whether two samples are different and is commonly used when the variances of two normal distributions are unknown and small sample size is small.

There are following two different conditions:

1. **When two sample sizes (n) are equal and population variances are same and unknown:**

The *t* statistic to test whether the means of two populations are different can be calculated as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where,

$$s_p = \sqrt{\frac{\sum (x_{1i} - \bar{x}_1)^2 + \sum (x_{2i} - \bar{x}_2)^2}{(n_1 + n_2 - 2)}} \qquad \text{or} \qquad s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Where sp is grand or pooled standard deviation of two samples.

For significance testing, the degree of freedom for this test is $n_1 + n_2 - 2$ where n is the number of participants in each group.

2. **When two sample sizes (n) are equal or unequal and population variances are unequal and unknown:**

This test, also known as Welch's *t*-test is used only when the two population variances are not assumed to be equal (the two sample sizes may or may not be equal). The *t* statistic to test whether the population means are different is calculated as:

$$t=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \qquad , s=\sqrt{\frac{\sum(x_i-\bar{x})^2}{(n-1)}}$$

Here s_1^2 and s_2^2 are the separate sample variances of two samples not a pooled variance.

Dependent *t*-test or paired t-test for paired samples:

When the observations are taken on the same item or experimental material or items are paired before taking the observation or when the samples are dependent i.e. when there is only one sample that has been tested twice (repeated measures) or when there are two samples that have been matched or paired then to test the differences between n pairs of samples paired t-test is used.

$$t=\frac{\bar{d}}{\frac{s_d}{\sqrt{n}}},$$

Where, $\bar{d}=\frac{1}{n}\sum_{i=1}^{n}d_i$

d_i is the difference between observations of n pairs of samples $d_i=x_{1i}-x_{2i}$

$$s_d=\sqrt{\frac{\sum(d_i-\bar{d})^2}{(n-1)}}$$

The degree of freedom used is $n-1$.

Unlike the two sample t-test, the two samples are not independent but are correlated, paired and having equal number of observations. Paired samples are also called **matched samples** or **repeated measures**.

t- Test for testing the significance of observed sample correlation coefficient.

If r is sample correlation coefficient in a sample of n pairs of observations from bivariate normal population. Here we have to test,

H_0 : r =0 , i.e. correlation coefficient is not significant

H_1 : r$\neq$ 0 i.e. correlation coefficient is significant

The t-statistic is given as follow:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Here d.f. is n-2.

Chi-square (χ^2) statistic:

There are following uses of chi-square test statistic

1. To test the whether a hypothetical value of population variance (σ_0^2) is true or not?

   Or

   To test if the given sample has been drawn from a population with specific variance σ0.
2. To decide whether there is any difference between the observed (experimental) value and the expected (theoretical) value (**Test of goodness of fit**).
3. To test whether the two characteristics are independent or not (**Test of independence of two attributes, Contingency table**).
4. To test the validity of hypothetical ratio.
5. To test homogeneity of several population variance (Bartlett‘s test).
6. To test equality of several population correlation coefficients.
7. To test equality of more than two population proportions.

Test statistics to test if the given sample has been drawn from a population with specific variance σ$_0$:

Let there are n random sample (n<30) $x_{1,}x_2,..........x_n$ from a normal population then the hypothesis $H_0 : \sigma^2 = \sigma_0^2$ v/s $H_1 : \sigma^2 \neq \sigma_0^2$ can be tested by χ^2 -test.

The test statistics is

$$\chi^2 = \frac{(n-1)\, s^2}{\sigma_0^2}$$

Where, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})$, n is sample size and σ_0^2 is hypothetical value of population variance.

Here χ^2 has (n-1) d.f.

Test of Goodness of Fit:

The chi-square test is used to test if there is discrepancy between observed or experimental frequency and theoretically determined frequency from the assumed distribution of event. It is used to determine whether sample data are consistent with a hypothesized distribution. If the discrepancy is not large we can assume that assumption about the distribution of that event is correct. The test is applied when there is one categorical variable from a single population.

Here we test, H_0: The data are consistent with a specified distribution

H_1: The data are not consistent with a specified distribution

The chi-square is the sum of the squared difference between observed (*o*) and the expected (e) data (frequencies), divided by the expected data in all possible categories.

The chi-square test statistic is given as:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Where O_i are the observed frequencies and E_i is expected frequencies

Here χ^2 has (n-1) d.f.

Steps to calculate the chi square statistic x^2 in testing of goodness of fit:

1. Calculate the expected value for each observed value in the table, if not given.
2. For each observed number in the table subtract the corresponding expected number (O — E).
3. Square the difference [(O —E) ^{2}].

4. Divide the squares obtained for each cell in the table by the expected number for that cell [(O - E) ^{2} / E].
5. Sum all the values for (O - E) ^{2} / E. This is the value of chi square statistic.

Suppose anyone preformed a simple monohybrid cross between two individuals that were heterozygous for the trait of interest (Aa x Aa) In this test he has the phenotypic ratio 85 of the A type and 15 of the a-type (homozygous recessive). However, in a monohybrid cross between two heterozygotes he has predicted a 3:1 ratio of phenotypes. In other words, he has expected to get 75 A-type and 25 a-type. Test where the given data is in correspond with hypothetical ratio?

Ans. Here we test, H_0: The given data are consistent with the hypothetical ratio

H_1: The data are not consistent with the hypothetical ratio

The observed frequencies are O_1 =85 and O_2= 15, expected frequencies E1 = 75 and E2 = 25

As $$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Now, $$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} = \frac{(85-75)^2}{75} + \frac{(15-25)^2}{25} = 1.33+4.0 = 5.33$$

Thus the calculated χ^2 value is 5.33. As the Calculated value of χ^2 is greater than tabulated χ^2 value at 0.05 alpha level of significance and 1 degrees of freedom (df =1), ($\chi^2_{calculated}$ 5.33> $\chi^2_{tabulated}$ 3.841). So the null hypothesis that given data are consistent with the hypothetical ratio or the two distributions are the same is rejected.

Testing of Independence of Two Attributes:

A **test of independence** assesses whether unpaired observations on two variables, expressed in a contingency table, are independent of each other. There is simplest set of the 2 x 2 table to the general notation shown below in Table using the letters a, b, c, and d to denote the contents of the cells, then we would have the following table:

Variable-1 (Row)	**Variable2 (Column)**		**Totals**
	Data type-1	Data type-2	
Category-1	a	b	**a+b**
Category-2	c	d	**c+d**
Total	**a+c**	**b+d**	**a+b+c+d = N**

Here we have to test, H_0 :The two categorical variables are independent.

H_1 : The two categorical variables are related.

For a 2 x 2 contingency table the Chi Square statistic is calculated by the following formula:

$$\chi^2 = \frac{N\left(ad - bc\right)^2}{\left(a+b\right)\left(c+d\right)\left(a+c\right)\left(b+d\right)}$$

Where, the four components of the denominator are the four totals from the table columns and rows.

Here the d.f. is (number of rows minus one) x (number of columns minus one) i.e. (r-1) x (c-1)

The following table gives data regarding a clinic trial in which animals are treated with a specific drug and increase in heart rate when animals are treated and not treated, are recorded.

	Heart rate increased	No heart rate increased	Total
Treated	30	14	50
Not treated	30	25	55
Total	66	39	105

Test that the animals receiving the drug would show increased heart rates compared to those that did not receive the drug.

Solution: Here we are to test ,H0: The proportion of animals whose heart rate increased is independent of drug treatment.

H1: The proportion of animals whose heart rate increased is associated with drug treatment.

Applying the chi-square test $\chi^2 = \dfrac{N\left(ad-bc\right)^2}{\left(a+b\right)\left(c+d\right)\left(a+c\right)\left(b+d\right)}$

$$\chi^2 = \frac{105\left[(36\times 25)-(14\times 30)\right]^2}{\left(50\times 55\times 39\times 66\right)} = \frac{105\left[(900)-(420)\right]^2}{\left(7078500\right)} = \frac{24192000}{7078500} = 3.41$$

Here the degrees of freedom is (number of rows minus one) x (number of columns minus one) for given data d.f is (2-1) x (2-1) = 1.

As calculated chi square statistic at the 0.05 level of significance, and df = 1 is χ^2 = 3.41 is less than tabulated value of chi-square (2.706). So we accept the null hypothesis. In other words, there is no statistically significant difference in the proportion of animals whose heart rate increased.

Chi-square Test of Homogeneity:

The test is applied to a single categorical variable from two different populations. It is used to determine whether two or more independent samples are homogeneous or drawn from same population or from different populations. For instance, in a survey of TV viewing preferences, we might ask respondents to identify their favourite program. We might ask the same question of two different populations, such as boys and girls. We could use a chi-square test for homogeneity to determine whether boys viewing preferences differed significantly from girls viewing preferences.

This application of χ^2 test can be regarded as an extension of the χ^2 test of independence of attributes.

The chi-square test statistic is given as:

$$\chi^2 = \sum_i \frac{\left(O_i - E_i\right)^2}{E_i}$$

Where O_i are the observed frequencies and E_i is expected frequencies

The expected frequencies are calculated by using following formula:

Expected frequency (E) = $\frac{(Row\ total)(Column\ total)}{Total\ number\ of\ observations}$

Here χ^2 has the d.f. is (number of columns minus one) x (number of rows minus one) i.e.(r-1) x (c-1)

Test Statistic to Test the Validity of Hypothetical Ratio r:

Ans. Let there are two classes C_1 and C_2 with frequencies a and b. To test the hypothesis that the given frequencies a and b are in the ratio r_1 and r_2, the Chi-square (χ^2) test is used.

To test the validity of hypothetical ratio r_1: r_2 the Chi-square (χ^2) statistics is

$$\chi^2 = \frac{(a-rb)^2}{r(a+b)}, \text{ where r = } \frac{r_1}{r_2}$$

Here Chi-square (χ^2) has 1 d.f.

F- Test Statistic:

This test is also known as Fisher‘s F- test and is based on f-distribution. F is defined as the ratio of two independent chi-square variates devided by their respective degree of freedom. If X and Y are two independent chi-square variates with v_1 and v_2 d.f. respectively, then F-statistic is defined as

$$F = \frac{X / v_1}{Y / v_2}.$$

Applications of F- test:

There are following use of F-test:

1. To test the equality of several population means (ANOVA).
2. To test equality of variance of two normal populations, when the sample size is small i.e. n < 30.
3. To test **equality of several regression** coefficients.
4. To test the significance of multiple correlation coefficient.
5. To test the significance of sample correlation ratio.
6. To test the linearity of regression (ANOVA)

Normal deviate test:

z -test for single mean is used to test a hypothesis on a specific value of the population mean if sample size n ≥ 30 and population standard deviation is known. When the sample size is large it is supposed that the sample variance is almost equal to population variance. So in testing of H0: $\mu = \mu_0$ v/s H1 : $\mu \neq \mu_0$, we use population standard deviation instead of sample standard deviation.

If the standard deviation of the **population is known** the z-statistic is

$$z = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Where σ is population standard deviation and n is sample size.

If the standard deviation of the **population is not known** and the sample size is 30 or above 30, then the σ is replaced by estimated value of s, sample standard deviation.

***z test* to test a hypothesis on a specific value of the population standard deviation if sample size n ≥ 30:**

Unlike the chi square test for single variance, this test is used if n ≥ 30. Let we have a large sample of size n taken from a normal population with unknown standard deviation σ Here we test the following hypothesis $H_0 : \sigma = \sigma_0$ v/s $H_1 : \sigma \neq \sigma_0$

The z-test statistic is

$$z = \frac{|s - \sigma|}{s/\sqrt{2n}} \square N(0,1), \text{ where n is large.}$$

$s/\sqrt{2n}$ is standard error of standard deviation of sample of size n.

z- Test Statistic for Testing the Equality of Two Populations Mean for Large Sample:

For testing the equality of two populations mean on basis of two samples taken from these normal populations and if the variances of two normal populations are known and sample size is large, n ≥ 30.

Let two samples of size n1 and n2 are drawn from two different normal populations having mean μ_1 & μ_2 and variances σ_2 & σ_1 respectively.

The *z*-statistic to test whether the means of two populations are different can be calculated as follows: $z = \dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \square N(0,1)$

If two samples are drawn from the population having same variances i.e. $\sigma_1 = \sigma_2 = \sigma$. The *z*-statistic

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sigma \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \square N(0,1)$$

Test statistic for testing the equality of variances of two populations when the sample size of each sample is 30 or larger:

Ans. Suppose there are two large sample of size n1 and n2 with standard deviations s1 and s2 taken from two normal populations with mean μ_1 & μ_2 and variances σ_2 & σ_1 respectively.

Here we test the $H_0 : \sigma_1 - \sigma_2 = 0 \, vs \, H_1 : \sigma_1 - \sigma_2 \neq 0$

$$z = \frac{|s_1 - s_2|}{\sqrt{\dfrac{s_1^2}{2n_1} + \dfrac{s_2^2}{2n_2}}} \square N(0,1)$$

Where, s_1^2 and s_2^2 are sample variances of large samples of size n1 and n2 respectively.

The z test statistic for testing the significance of **specific value of** ***single*** **population proportion:**

Here we test the $H_0 : p = p_0$ against the $H_1 : p \neq p_0$ where p is the population proportion and p_0 is a specific value of the population proportion.

The test statistic is a z –test is given by the following equation.

$$z = \frac{(p - P)}{\sigma_{prop}}$$

Where σ is the standard error of the sampling distribution of population proportion and is calculated by using following formula $\sigma_{prop} = \sqrt{\dfrac{P(1-P)}{n}}$

Where P is the hypothesized value of population proportion in the null hypothesis, p is the sample proportion, and σ is the standard error of the sampling distribution and n is the sample size.

In a sample of 300 bulbs manufactured by a company 20 bulbs were found defected. But the company holder claimed that only 35 of their product is defective. Test the claim of company holder at the 5% level of significance.

Ans. Here n= 300, the hypothesized value of population proportion P = 3/100 = 0.03, q = 1-0.03 = 0.97

Here hypothesis is $H_0 : p = 0.03 \quad H_1 : P \neq 0.03$

$$SE_{\mathrm{Prop}} = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{0.03 \times 0.97}{300}} \, 0.00984$$

Sample proportion p = 20/300= 0.066

Z-statistics = $z = \frac{(p-P)}{\sigma_{prop}} = = \frac{(0.066-0.03)}{0.00984} = 3.719$

As the 5% level of significance the table value of z is 1.64. The calculated value of z is more than tabulated value of z so $H_0 : p = 0.03$ is rejected i.e. the claim of company holder is not tenable.

Z test for testing the significance of difference between proportions of two populations:

Let two large samples of size n_1 and n_2 are taken from two different normal populations and we have to test the $H_0 : P_1 = P_2$ i.e., there is no difference between the two population proportions against the $H_1 : P_1 \neq P_2$ i.e., there is difference between the two population proportions

The test statistic is a z-score (z) defined by the following equation.

$$z = \frac{| p_1 - p_2 |}{SE_{prop\,diff}} \square N(0,1)$$

where p_1 and p_2 are the proportion from sample 1 and sample 2 respectively and SE is the standard error of the sampling distribution difference between two proportions.

Where standard error, $SE_{propdiff} = \sqrt{p_0\left(1-p_0\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

Where, p_0 is the pooled sample proportion.

The pooled sample proportion p_0 used to compute the standard error of the sampling distribution is calculated as follow:

$$p_0 = \frac{p_1.n_1 + p_2.n_2}{(n_1+n_2)}$$

Where p_1 is the sample proportion from population 1, p_2 is the sample proportion from population 2, n_1 is the size of sample 1, and n_2 is the size of sample 2.

Example: In a random sample of 500 students from university A 200 are found to be tobacco. And in sample of 400 students from university B 200 are found to be consumer of tobacco. Test whether there is difference between proportion of students of tobacco consumer in two university.

Solution: Here we have to test, $H_0 : P_1 = P_2$ against the $H_1 : P_1 \neq P_2$.

$$p_1 = \frac{200}{500} = 0.4 \text{ and } p_2 = \frac{200}{400} = 0.5$$

Pooled proportion $p_0 = \frac{p_1.n_1 + p_2.n_2}{(n_1+n_2)} = = \frac{0.4\times 500 + 0.5\times 400}{500+400} = 0.444$,

$$1\text{-}\ p_0 = 1\text{-}0.444 = 0.556$$

$$SE_{propdiff} = \sqrt{p_0\left(1-p_0\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \sqrt{0.444\left(1-0.444\right)\left(\frac{1}{500}+\frac{1}{400}\right)} = \sqrt{.0246\times 0.045} = 0.333$$

z- statistic $z = \frac{|\,p_1 - p_2\,|}{SE_{propdiff}} = = \frac{|\,0.4 - 0.5\,|}{0.333} = 3.0$

As z is exactly equal to 3SE, in general this difference between proportion in not significant but at the 5% or 1% level of significance may be significant.

Question 1: A manufacturer of dry cells claimed that the life of their cells is 24.0 hrs. A sample of 10 cells had a mean life of 22.5 hrs with standard deviation of 3.0 hrs. On the basis of available information, tests whether the claim of manufacturer is correct.

(Given: t0.05,9= 2.2623)

Question 2: In a breeding experiment, the ratio of springs in four classes was expected to be 1:3:3:9. The experiment yield the data as follow:

Class	AA	Aa	Aa	aa
No. Of Springs	8	29	37	102

Test whether the given data is in agreement with the hypothetical ratio.

Question 3: What should be the sample size for using the t-test and Z-test?

Question 4: Give the conditions for using the paired t test.

Question 5: Give the skeleton of ANAOVA table

EXERCISE No. 10

Design of Experiments

(CRD & RBD)

Design of experiment means how to design an experiment in the sense that how the observations or measurements should be obtained to answer a query in a valid, efficient and economical way. The designing of experiment and the analysis of obtained data are inseparable. If the experiment is designed properly keeping in mind the question, then the data generated is valid and proper analysis of data provides the valid statistical inferences. If the experiment is not well designed, the validity of the statistical inferences is questionable and may be invalid.

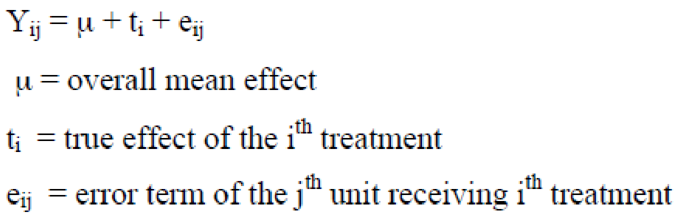
1. Completely Randomized Design (CRD):

CRD is the basic single factor design. In this design the treatments are assigned completely at random so that each experimental unit has the same chance of receiving any one treatment. But CRD is appropriate only when the experimental material is homogeneous. As there is generally large variation among experimental plots due to many factors CRD is not preferred in field experiments.

In laboratory experiments and greenhouse studies it is easy to achieve homogeneity of experimental materials and therefore CRD is most useful in such experiments.

Layout of a CRD Completely randomized Design is the one in which all the experimental units are taken in a single group which are homogeneous as far as possible.

The statistical model for CRD with one observation per unit



$Y_{ij} = \mu + t_i + e_{ij}$

μ = overall mean effect

t_i = true effect of the i^{th} treatment

e_{ij} = error term of the j^{th} unit receiving i^{th} treatment

The arrangement of data in CRD is as follows:

	Treatments				
	T_1	T_2	T_i	T_K	
	y_{11}	y_{21}	y_{i1}	Y_{K1}	
	y_{12}	y_{22}	y_{i2}	Y_{K2}	
	y_{1r1}	y_{2r2}	y_{iri}	$Y_{k\,rk}$	
Total	Y_1	Y_2	Y_i	T_k	GT

(GT – Grand total)

The null hypothesis will be

$H_o : \mu_1 = \mu_2 = \ldots\ldots\ldots\ldots = \mu_k$ or There is no significant difference between the treatments

And the alternative hypothesis is

$H_1 : \mu_1 \neq \mu_2 \neq \ldots\ldots\ldots\ldots \neq \mu_k$. There is significant difference between the treatments

The different steps in forming the analysis of variance table for a CRD are:

1. $$\text{C.F} = \frac{(GT)^2}{n}$$

   n= Total number of observations

2. $$\text{Total SS} = \text{TSS} = \sum_{i=1}^{k}\sum_{j=1}^{v} y_{ij}^{\;\;2} - C.F$$

3. $$\text{Treatment SS} = \text{TrSS} = \frac{Y_1^2}{r_1} + \frac{Y_2^2}{r_2} + \ldots\ldots\ldots + \frac{Y_k^2}{r_k} - C.F$$

   $$= \sum_{i=1}^{k} \frac{Y^2{}_i}{r_i} - C.F$$

4. $$\text{Error SS} = \text{ESS} = \sum_{i=1}^{k}\sum_{j=1}^{r} y_{ij}^{\;\;2} - \sum_{i=1}^{k}\frac{Y_i^2}{r_i}$$

   $$= \text{TSS} - \text{TrSS}$$

5. Form the following ANOVA table and calculate F value.

Source of variation	**d.f.**	**SS**	**MS**	**F**
Treatments	t-1	TrSS	TrMS= $\frac{TrSS}{t-1}$	$\frac{TrMS}{EMS}$
Error	n-t	ESS	EMS= $\frac{ESS}{n-t}$	
Total	n-1	TSS		

6. Compare the calculated F with the critical value of F corresponding to treatment degrees of freedom and error degrees of freedom so that acceptance or rejection of the null hypothesis can be determined.

7. If null hypothesis is rejected that indicates there is significant differences between the different treatments.

8. Calculate C D value.

C.D. = SE(d). t

$$\text{where S.E(d)} = \sqrt{EMS(\frac{1}{r_i} + \frac{1}{r_j})}$$

r_i = number of replications for treatment i

r_j = number of replications for treatment j and

t is the critical t value for error degrees of freedom at specified level of significance, either 5% or 1%.

1-58-18 D(0.9)	7-96-29 F(41.0)	13-64-21 E(39.2)	19-20-07 B(37.5)	25-25-08 B(38.4)
2-97-30 F(40.6)	8-51-15 C(39.5)	14-52-16 D(41.7)	20-73-23 E(38.7)	26-60-19 D(40.1)
3-42-11 C(40.9)	9-74-24 E(39.7)	15-62-20 D(39.4)	21-44-12 C(39.8)	27-95-28 F(39.8)
4-07-02 A(31.3)	10-79-25 E(40.6)	16-28-09 B(38.8)	22-01-01 A(32.2)	28-15-04 A(33.9)
5-49-14 D(39.2)	11-13-03 A(29.2)	17-92-27 F(41.1)	23-31-10 B(37.4)	29-53-17 D(40.0)
6-14-05 A(33.4)	12-85-26 F(41.5)	18-45-13 C(38.6)	24-17-06 B(35.8)	30-65-22 E(41.9)

Table 7-1. Root yields (tons/acre) of plots fertilized with six levels of nitrogen.

Treatment			Replications			Total	Mean
(lb. /acre)						$(Y_{i.})$	$(\bar{Y}_{i.})$
A(0)	31.3	33.4	29.2	32.2	33.9	160.0	32.00
B(50)	38.8	37.5	37.4	35.8	38.4	187.9	37.58
C(100)	40.9	39.2	39.5	38.6	39.8	198.0	39.60
D(150)	40.9	41.7	39.4	40.1	40.0	202.1	40.42
E(200)	39.7	40.6	39.2	38.7	41.9	200.1	40.02
F(250)	40.6	41.0	41.5	41.1	39.8	204.0	40.80
Overall						1152.1	38.40

Source	df	SS	MS	F
Total	29	311.13		
Nitrogen treatments	5	277.69	55.54	39.95
Experimental error	24	33.44	1.39	

The procedure involved in constructing such an AOV table is illustrated by the following steps.

Step 1: Outline the AOV table and list the sources of variation and degrees of freedom. There are two sources of variation, between and within treatments. Degrees of freedom are one less than the number of observations in each source of variation. There are 6 treatments, therefore there are 5 degrees of freedom for the between treatment sum of squares (SST). There are 5 replications per treatment, therefore there are 4 degrees of freedom for each treatment times 6 treatments, which gives 24 degrees of freedom for the within treatment sum of squares (SSE). The degrees of freedom associated with the total variation in the experiment is one less than the total number of experimental units: 30 - 1 = 29. Note that the degrees of freedom associated with the sources of variation are additive, 5 + 25 = 29.

Step 2: Calculate the correction term (C).

$$C = Y^2../kr = (1152.1)^2/6(5) = 44244.48$$

This is actually the sum of squares due to the mean.

Step 3: Calculate the total sum of squares (TSS).

$$\begin{aligned} TSS &= \Sigma\Sigma(Y_{ij} - \overline{Y}..)^2 \\ &= \Sigma\Sigma_{ij}^2 - C \\ &= 31.3^2 + 38.8^2 + ... \end{aligned}$$

The correction term is used so that the sum of squares is calculated about the general mean $\overline{Y}..$ not about 0.

Step 4: Calculate the sum of squares and mean square for treatments.

$$\begin{aligned} SST &= r\,\Sigma(\overline{Y}_{i.} - \overline{Y}..)^2 \\ &= \Sigma Y_{i}^{2}./r - C \\ &= (160.0^2 + 187.9^2 + ... + 204.0^2)/5 - C \\ &= 44522.17 - 44244.48 = 277.69 \end{aligned}$$

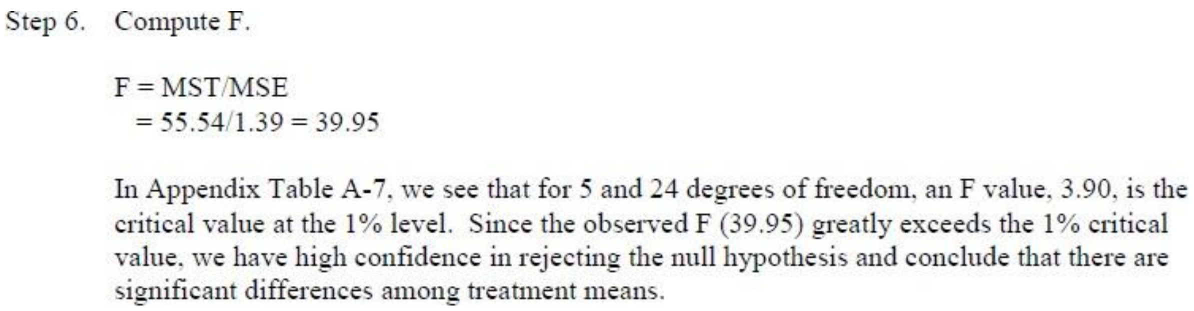
A mean square is calculated by dividing the sum of squares by its degrees of freedom.

$$\begin{aligned} MST &= SST/(k\text{-}1) \\ &= 277.69/(6\text{-}1) = 55.54 \end{aligned}$$

Step 5: Calculate the sum of squares and mean square for error.

$$\begin{aligned} SSE &= TSS - SST \\ &= 311.13 - 277.69 = 33.44 \end{aligned}$$

$$\begin{aligned} MSE &= SSE/k(r\text{-}1) \\ &= (33.44)/24 = 1.39 \end{aligned}$$



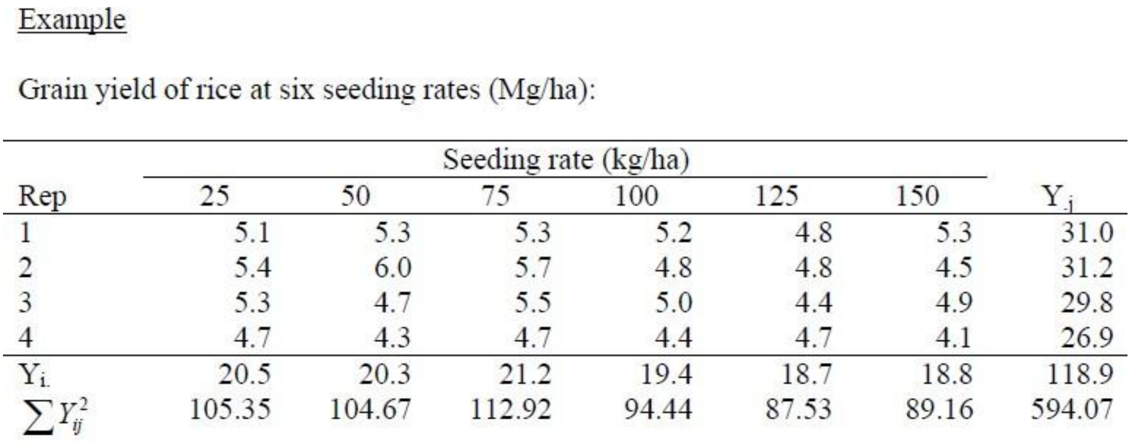
Step 6. Compute F.

F = MST/MSE
= 55.54/1.39 = 39.95

In Appendix Table A-7, we see that for 5 and 24 degrees of freedom, an F value, 3.90, is the critical value at the 1% level. Since the observed F (39.95) greatly exceeds the 1% critical value, we have high confidence in rejecting the null hypothesis and conclude that there are significant differences among treatment means.

2. **Randomized Block Design (RBD)**

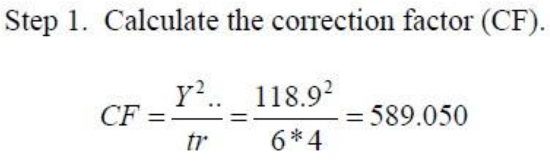
It is perhaps the most commonly encountered design that can be analyzed as a two-way AOV. If large numbers of treatments are to be compared then large numbers of experimental units are required. This will increase the variation among the responses and CRD may not be appropriate to use. In such a case when the experimental material is not homogeneous and there are v treatments to be compared, then it may be possible to • Group the experimental material into blocks of sizes v units. • Blocks are constructed such that the experimental units within a block are relatively homogeneous and resemble to each other more closely than the units in the different blocks. • If there are b such blocks, we say that the blocks are at b levels. Similarly if there are v treatments, we say that the treatments are at v levels. The responses from the b levels of blocks and v levels of treatments can be arranged in a two-way layout.



Example

Grain yield of rice at six seeding rates (Mg/ha):

	Seeding rate (kg/ha)						
Rep	25	50	75	100	125	150	$Y_{.j}$
1	5.1	5.3	5.3	5.2	4.8	5.3	31.0
2	5.4	6.0	5.7	4.8	4.8	4.5	31.2
3	5.3	4.7	5.5	5.0	4.4	4.9	29.8
4	4.7	4.3	4.7	4.4	4.7	4.1	26.9
$Y_{i.}$	20.5	20.3	21.2	19.4	18.7	18.8	118.9
$\sum Y_{ij}^2$	105.35	104.67	112.92	94.44	87.53	89.16	594.07



Step 1. Calculate the correction factor (CF).

$$CF = \frac{Y^2..}{tr} = \frac{118.9^2}{6*4} = 589.050$$

Step 2. Calculate the Total SS.

$$Total\ SS = \sum Y_{ij}^2 - CF$$

$$= (5.1^2 + 5.4^2 + 5.3^2 + ... + 4.1^2) - CF$$

$$= 5.02$$

Step 3. Calculate the Replicate SS (Rep SS)

$$\text{Rep}\ SS = \sum \frac{Y_{.j}^2}{t} - CF$$

$$= \frac{\left(31.0^2 + 31.2^2 + 29.8^2 + 26.9^2\right)}{6} - CF$$

$$= 1.965$$

Step 4. Calculate the Treatment SS (Trt SS)

$$Trt\ SS = \sum \frac{Y_{i.}^2}{r} - CF$$

$$= \frac{\left(20.5^2 + 20.3^2 + 21.2^2 + 19.4^2 + 18.7^2 + 18.8^2\right)}{4} - CF$$

$$= 1.2675$$

Step 5. Calculate the Error SS

Error SS = Total SS – Rep SS – Trt SS

= 1.7875

Step 6. Complete the ANOVA Table

SOV	Df	SS	MS	F
Rep	r-1 = 3	1.9650	0.6550	Rep MS/Error MS = 5.495^{**}
Trt	t-1 = 5	1.2675	0.2535	Trt MS/Error MS = 2.127^{ns}
Error	(r-1)(t-1) = 15	1.7875	0.1192	
Total	tr-1 = 23	5.0200		

Step 7. Look up Table F-values for Rep and Trt:

Rep	Trt
$F_{.05;3,15} = 3.29$	$F_{.05;5,15} = 2.90$
$F_{.01;3,15} = 5.42$	$F_{.01;5,15} = 4.56$

Step 8. Make conclusions.

Rep: Since $F_{calc.}(5.495) > F_{Tab.}$ at the 95 and 99% levels of confidence, we reject H_o: All replicate means are equal.

TRT: Since $F_{calc.}(2.127) < F_{Tab.}$ at the 95 and 99% levels of confidence, we fail to reject H_o: All treatment means are equal.

Question 1: A dataset of gain in body weight involving four feed stuff A, B, C, D tried on 20 chicks is given below. All twenty chicks are alike in all respect expect feeding and each feed is given to 5 chicks. Analyse the data. (Given F $_{0.05(3,6)}$ = 3.06)

Feed	**Body Wt. Gain**				
A	55	49	42	21	52
B	61	112	30	89	63
C	42	97	81	95	92
D	169	137	169	85	154

Question 2: The result (yield) of experiment involving six treatments in four blocks are given below and treatments are indicated by numbers within parentheses.

Test whether the treatments differ significantly? (Given $F_{0.05, (3,15)}$=5.42, $F_{0.05\ (5,15)}$ =4.5)

Blocks	**Result (yield) of experiment involving six treatments**					
1	**(1)** **24.7**	**(2)** **27.7**	**(3)** **20.6**	**(4)** **16.2**	**(5)** **16.2**	**(6)** **24.9**
2	**(3)** **22.7**	**(2)** **28.8**	**(1)** **27.3**	**(4)** **15.0**	**(6)** **22.5**	**(5)** **17.0**
3	**(6)** **26.3**	**(4)** **19.6**	**(1)** **38.5**	**(3)** **36.8**	**(2)** **39.5**	**(3)** **15.4**
4	**(5)** **17.7**	**(2)** **31.0**	**(1)** **28.5**	**(4)** **14.1**	**(3)** **34.9**	**(6)** **22.6**

EXERCISE No.-11

COMPUTER BASICS AND COMPONENTS OF COMPUTER

Computer is an electronic device which is used to store the data, as per given instructions it gives results quickly and accurately. Computer itself is a combination of different type of separate electronic device. i.e. computer only will be computer if it has INPUT DEVICE, PROCESS UNIT, and OUTPUT DEVICE.

Generations of the Computer

Charlse Babbase is known as father of computer he has invented first analytical computer in year 1822

- First Generation (1940 – 1955) : Electronic Numerical Integrator and Computer (ENIAC) , EDVAC
- Second Generation (1956 – 1965) : IBM 1401
- Third Generation (1966 – 1975) : IBM System/360
- Fourth Generation (1976 – 1985) : Macintosh 128k
- Fifth Generation (1986 -till date) : Super computer

Number System

- Binary Number System : It has only base 2 i.e 0 and 1
- Octal Number System : Base of octal is 8 i.e. 0, 1, 2, 3, 4, 5, 6, 7
- Decimal Number System : Base of Decimal is 10 i.e. 0 1 2 3 4 5 6 7 8 9
- Hexadecimal Number System : Base of this number system is 16 i.e. 0 1 2 3 4 5 6 7 8 9 A B C D E F

Storage Capacity

Today‘s computers can store large volumes of data. A piece of information once recorded (or stored) in the computer can never be forgotten and can be retrieved almost instantaneously.

- ✓ 1 KB = 1024 BYTE
- ✓ 1 MB = 1024 KB
- ✓ 1 GB = 1024 MB
- ✓ 1 TB = 1024 PB
- ✓ 1 PB = 1024 EB

- **Data**: Data is a raw material of information.
- **Information**: Proper collection of the data is called information.
- **Universal Serial Bus (USB)**: This is used to connect the external device to the computer.

Components of Computer:

A computer is a programmable machine designed to perform arithmetic and logical operations automatically and sequentially on the input given by the user and gives the desired output after processing. Computer components are divided into two major categories namely **hardware and software**. Hardware is the machine itself and its connected devices such as monitor, keyboard, mouse etc. Software is the set of programs that make use of hardware for performing various functions.

A computer system consists of mainly four basic units; namely input unit, storage unit, central processing unit and output unit. Central Processing unit further includes Arithmetic logic unit and control unit, as shown in the figure:

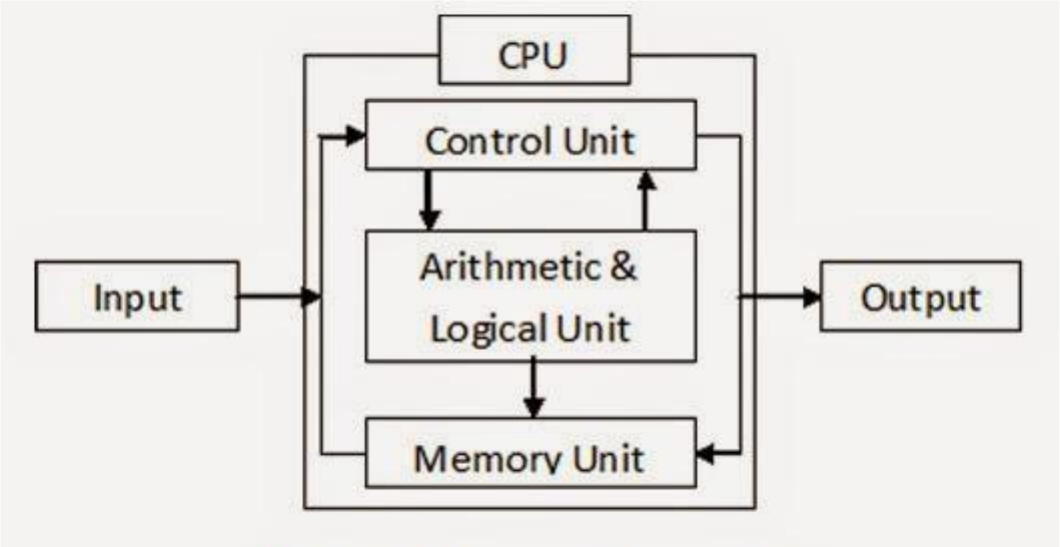


Fig. Block Di agram of Computer

A computer performs five major operations or functions irrespective of its size and make. These are

- ✓ It accepts data or instructions as input
- ✓ It stores data and instruction
- ✓ It processes data as per the instructions
- ✓ It controls all operations inside a computer
- ✓ It gives results in the form of output

Functional Units:

Input Unit: This unit is used for entering data and programs into the computer system by the use for processing.

Storage Unit: The storage unit is used for storing data and instructions before and after processing.

Output Unit: The output unit is used for storing the result as output produced by the computer after processing.

Processing Unit: The task of performing operations like arithmetic and logical operations is called processing. The Central Processing Unit (CPU) takes data and instructions from the storage unit and makes all sorts of calculations based on the instructions given and the type of data provided. It is then sent back to the storage unit.

CPU includes Arithmetic logic unit (ALU) and control unit (CU)

Arithmetic Logic Unit: All calculations and comparisons, based on the instructions provided, are carried out within the ALU. It performs arithmetic functions like addition, subtraction, multiplication, division and also logical operations like greater than, less than and equal to etc.

Control Unit: Controlling of all operations like input, processing and output are performed by control unit. It takes care of step by step processing of all operations inside the computer.

Memory: Computer‗s memory can be classified into two types; **primary memory** and **secondary memory**.

Primary Memory can be further classified as **RAM and ROM**.

RAM or Random Access Memory is the unit in a computer system. It is the place in a computer where the operating system, application programs and the data in current use are kept temporarily so that they can be accessed by the computer‗s processor. It is said to be ‗volatile‗ since its contents are accessible only as long as the computer is on. The contents of RAM are no more available once the computer is turned off.

ROM or Read Only Memory is a special type of memory which can only be read and contents of which are not lost even when the computer is switched off. It typically contains manufacturer‗s instructions. Among other things, ROM also stores an initial program called the _bootstrap loader‘ whose function is to start the operation of computer system once the power is turned on.

Secondary Memory:

RAM is volatile memory having a limited storage capacity. Secondary/auxiliary memory is storage other than the RAM. These include devices that are peripheral and are connected and controlled by the computer to enable permanent storage of programs and data.

1. **CD ROM**

Secondary storage devices are of two types; magnetic and optical. Magnetic devices include hard disks and optical storage devices are CDs, DVDs, Pen drive, Zip drive etc.

2. **Hard Disk**

Hard disks are made up of rigid material and are usually a stack of metal disks sealed in a box. The hard disk and the hard disk drive exist together as a unit and is a permanent part of the computer where data and programs are saved. These disks have storage capacities ranging from 1GB to 80 GB and more. Hard disks are rewritable.

3. **Compact Disk**

Compact Disk (CD) is portable disk having data storage capacity between 650-700 MB. It can hold large amount of information such as music, full-motion videos, and text etc. CDs can be either read only or read write type.

4. **Digital Video Disk**

Digital Video Disk (DVD) is similar to a CD but has larger storage capacity and enormous clarity. Depending upon the disk type it can store several Gigabytes of data. DVDs are primarily used to store music or movies and can be played back on your television or the computer too. These are not rewritable.

Input / Output Devices:

These devices are used to enter information and instructions into a computer for storage or processing and to deliver the processed data to a user. Input/Output devices are required for users to communicate with the computer. In simple terms, input devices bring information INTO the computer and output devices bring information OUT of a computer system. These input/output devices are also known as peripherals since they surround the CPU and memory of a computer system.

Input Devices

An input device is any device that provides input to a computer. There are many input devices, but the two most common ones are a keyboard and mouse.

Keyboard: The keyboard is very much like a standard typewriter keyboard with a few additional keys. The basic QWERTY layout of characters is maintained to make it easy to use the system. The additional keys are included to perform certain special functions. These are known as function keys that vary in number from keyboard to keyboard.

Mouse: A device that controls the movement of the cursor or pointer on a display screen. A mouse is a small object you can roll along a hard and flat surface. Its name is derived from its shape, which looks a bit like a mouse. As you move the mouse, the pointer on the display screen moves in the same direction.

Trackball: A trackball is an input device used to enter motion data into computers or other electronic devices. It serves the same purpose as a mouse, but is designed with a moveable ball on the top, which can be rolled in any direction.

Touchpad: A touch pad is a device for pointing (controlling input positioning) on a computer display screen. It is an alternative to the mouse. Originally incorporated in laptop computers, touch pads are also being made for use with desktop computers. A touch pad works by sensing the user‛s finger movement and downward pressure.

Touch Screen: It allows the user to operate/make selections by simply touching the display screen. A display screen that is sensitive to the touch of a finger or stylus. Widely used on ATM machines, retail point-of-sale terminals, car navigation systems, medical monitors and industrial control panels.

Light Pen: Light pen is an input device that utilizes a light-sensitive detector to select objects on a display screen.

Magnetic ink character recognition (MICR): MICR can identify character printed with a special ink that contains particles of magnetic material. This device particularly finds applications in banking industry.

Optical mark recognition (OMR): Optical mark recognition, also called mark sense reader is a technology where an OMR device senses the presence or absence of a mark, such as pencil mark. OMR is widely used in tests such as aptitude test.

Bar code reader: Bar-code readers are photoelectric scanners that read the bar codes or vertical zebra strips marks, printed on product containers. These devices are generally used in super markets, bookshops etc.

Scanner: Scanner is an input device that can read text or illustration printed on paper and translates the information into a form that the computer can use. A scanner works by digitizing an image.

Output Devices:

The output is usually produced in one of the two ways – on the display device, or on paper (hard copy).

Monitor: is often used synonymously with ―computer screen‖ or ―display.‖ Monitor is an output device that resembles the television screen. It may use a Cathode Ray Tube (CRT) to display information. The monitor is associated with a keyboard for manual input of characters and displays the information as it is keyed in. It also displays the program or application output. Like the television, monitors are also available in different sizes

Printer: Printers are used to produce paper (commonly known as hard copy) output. Based on the technology used, they can be classified as **Impact or Non-impact printers.**

Impact printers use the typewriting printing mechanism wherein a hammer strikes the paper through a ribbon in order to produce output. Dot-matrix and Character printers fall under this category.

Non-impact printers do not touch the paper while printing. They use chemical, heat or electrical signals to etch the symbols on paper. Inkjet, Deskjet, Laser, Thermal printers fall under this category of printers.

Plotter: Plotters are used to print graphical output on paper. It interprets computer commands and makes line drawings on paper using multi colored automated pens. It is capable of producing graphs, drawings, charts, maps etc. • **Facsimile (FAX)**: Facsimile machine, a device that can send or receive pictures and text over a telephone line. Fax machines work by digitizing an image.

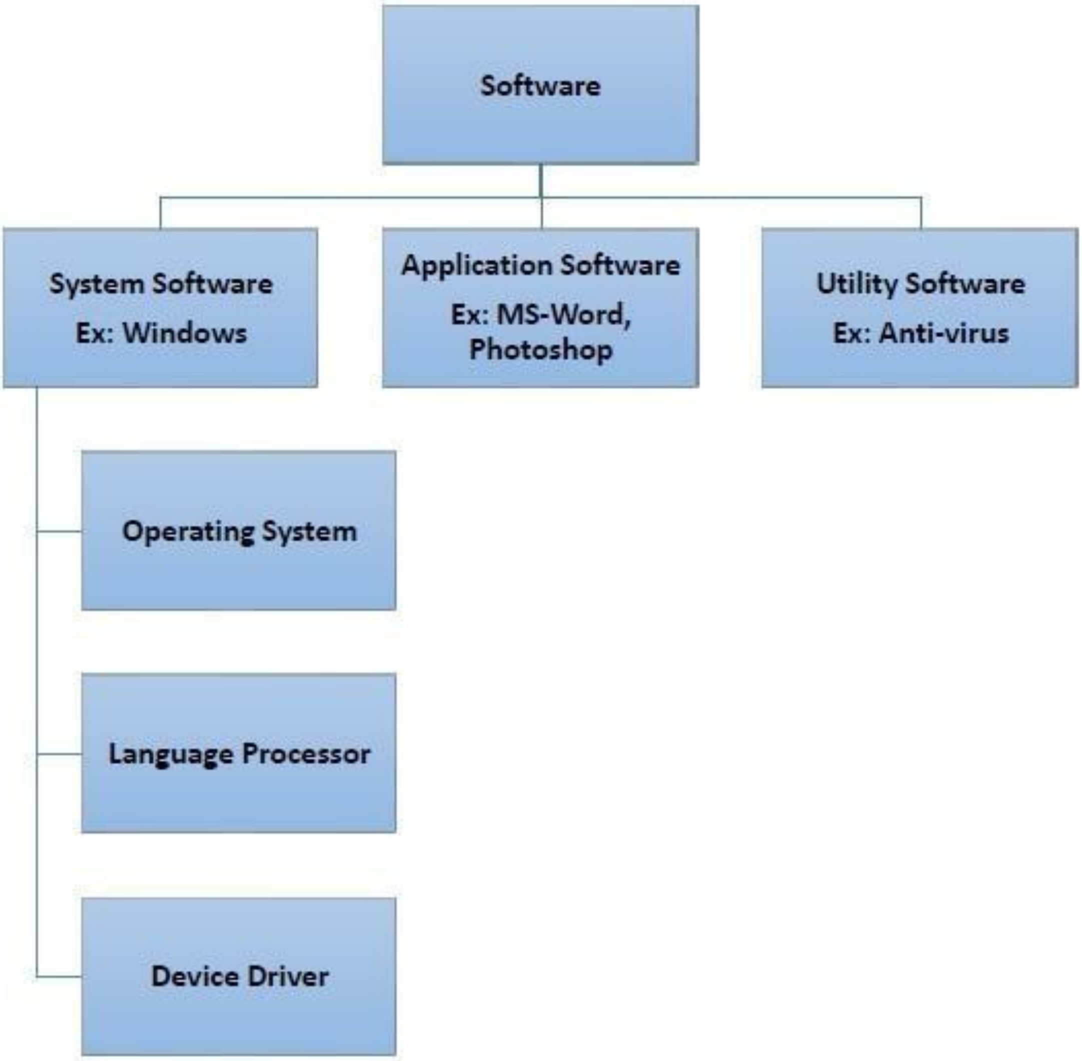
Sound cards and Speaker(s): An expansion board that enables a computer to manipulate and output sounds. Sound cards are necessary for nearly all CD-ROMs and have become commonplace on modern personal computers. Sound cards enable the computer to output sound through speakers connected to the board, to record sound input from a microphone connected to the computer, and manipulate sound stored on a disk.

Computer Software

A set of instructions that achieve a single outcome are called program or procedure. Many programs functioning together to do tasks make software.

There are three categories of software −

- System Software
- Application Software
- Utility Software



SystemSoftware:

Software required to run the hardware parts of the computer and other application software are called system software. System software acts as interface between hardware and user applications.

Based on its function, system software is of four types −

- Operating System
- Language Processor
- Device Drivers

Operating System

System software that is responsible for functioning of all hardware parts and their interoperability to carry out tasks successfully is called **operating system (OS)**. OS is the first software to be loaded into computer memory when the computer is switched on and this is called **booting**. OS manages a computer‘s basic functions like storing data in memory, retrieving files from storage devices, scheduling tasks based on priority, etc.

Language Processor

As discussed earlier, an important function of system software is to convert all user instructions into machine understandable language. When we talk of human machine interactions, languages are of three types −

- **Machine-level language** − This language is nothing but a string of 0s and 1s that the machines can understand. It is completely machine dependent.
- **Assembly-level language** − This language introduces a layer of abstraction by defining **mnemonics**. **Mnemonics** are English like words or symbols used to denote a long string of 0s and 1s. For example, the word –READ‖ can be defined to mean that computer has to retrieve data from the memory. The complete **instruction** will also tell the memory address. Assembly level language is **machine dependent**.
- **High level language** − This language uses English like statements and is completely independent of machines. Programs written using high level languages are easy to create, read and understand. BASIC (Beginners All Purpose Symbolic Instruction Code)

✓ FORTRAN (Formula Translation)

- ✓ PL/I (Programming Language, Version 1)
- ✓ LISP (List Processing)
- ✓ C++
- ✓ Java

Program written in high level programming languages like Java, C++, etc. is called **source code**. Set of instructions in machine readable form is called **object code** or **machine code**. **System software** that converts source code to object code is called **language processor**. There are three types of language interpreters−

- **Assembler** − Converts assembly level program into machine level program.
- **Interpreter** − Converts high level programs into machine level program line by line.
- **Compiler** − Converts high level programs into machine level programs at one go rather than line by line.

Device Drivers

System software that controls and monitors functioning of a specific device on computer is called **device driver**. Each device like printer, scanner, microphone, speaker, etc. that needs to be attached externally to the system has a specific driver associated with it. When you attach a new device, you need to install its driver so that the OS knows how it needs to be managed.

Application Software

Software that performs a single task and nothing else is called **application software**. Application software are very specialized in their function and approach to solving a problem. So spreadsheet software can only do operations with numbers and nothing else.

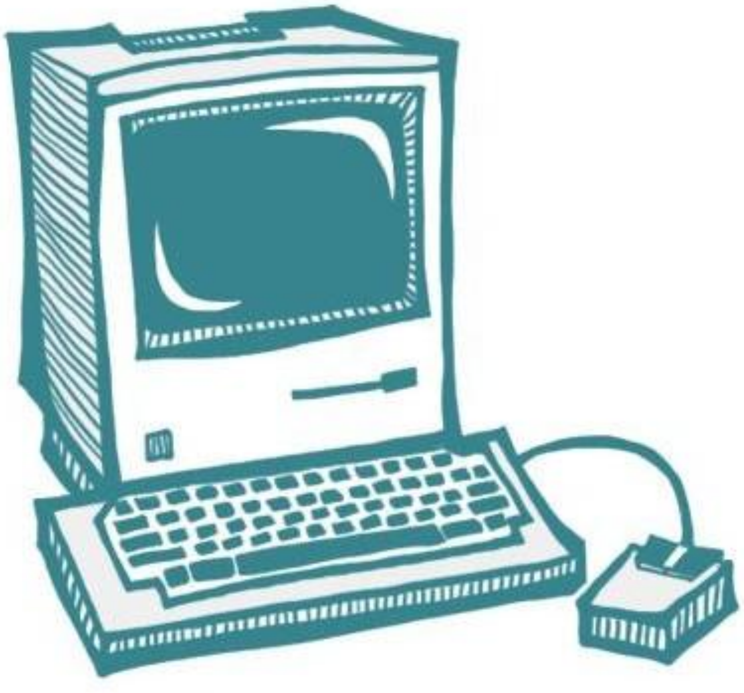
Here are some commonly used application software −

- Word processing
- Spreadsheet
- Presentation
- Database management
- Multimedia tools

Utility Software

Application software that assists system software in doing their work is called **utility software**. Thus utility software is actually a cross between system software and application software. Examples of utility software include −

- Antivirus software
- Disk management tools
- File management tools
- Compression tools
- Backup tools



Question 1: Define computer.

Question 2: Write down the name of different input and output devices.

Question 3: What is software? Give the name of different types of software

Question 4: What do you mean by computer memory? Discuss about the different types of primary and secondary memories.

EXERCISE No. 12

COMPUTER OPERATIONS

Internet

The **Internet** is the global system of interconnected computer networks that use the Internet protocol suite (TCP/IP) to link devices worldwide. It is a *network of networks* that consists of private, public, academic, business, and government networks of local to global scope, linked by a broad array of electronic, wireless, and optical networking technologies. The Internet carries a vast range of information resources and services, such as the inter-linked hypertext documents and applications of the World Wide Web (WWW), electronic mail, telephony, and file sharing.

One of the fundamental Internet technologies started in the early 1960s in the work of **Paul Baran.**

Electronic mail (email or e-mail):

Email is a method of exchanging messages between people using electronics. It is a service which allows us to send the message in electronic mode over the internet. It offers an efficient, inexpensive and real time mean of distributing information among people. Email operates across computer networks, which today is primarily the Internet.

Today's email systems are based on a store-and-forward model. Email servers accept, forward, deliver, and store messages. Neither the users nor their computers are required to be online simultaneously; they need to connect only briefly, typically to a mail server or a webmail interface, for as long as it takes to send or receive messages.

Microsoft Excel:

Microsoft Excel has the basic features of all spreadsheets, using a grid of cells arranged in numbered rows and letter-named columns to organize data manipulations like arithmetic operations. In addition, it can display data as line graphs, histograms and charts, and with a very limited three-dimensional graphical display. It allows sectioning of data to view its dependencies on various factors from different perspectives (using pivot tables and the scenario manager).

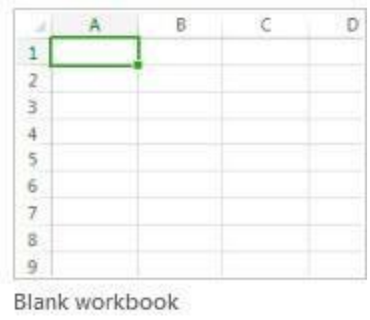
Basic Tasks in Excel

Create a new workbook

Excel documents are called workbooks. Each workbook has sheets, typically called spreadsheets. You can add as many sheets as you want to a workbook, or you can create new workbooks to keep your data separate.

1. Click **File**, and then click **New**.

2. Under **New**, click the **Blank workbook**.



Enter your data

Click an empty cell.

For example, cell A1 on a new sheet. Cells are referenced by their location in the row and column on the sheet, so cell A1 is in the first row of column A.

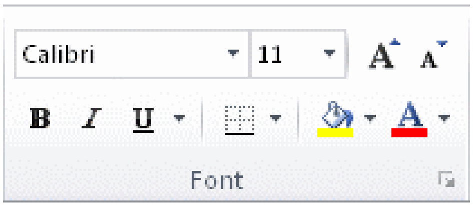
1. Type text or a number in the cell.

2. Press Enter or Tab to move to the next cell.

Apply cell borders

Select the cell or range of cells that you want to add a border to.

1. On the **Home** tab, in the Font group, click the arrow next to Borders, and then click the border style that you want.



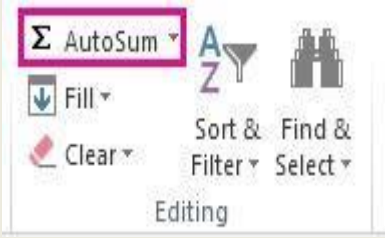
Apply cell shading

1. Select the cell or range of cells that you want to apply cell shading to.
2. On the **Home** tab, in the **Font** group, choose the arrow next to **Fill Color** , and then under **Theme Colors** or **Standard Colors**, select the color that you want.

Use AutoSum to add your data

When you‘ve entered numbers in your sheet, you might want to add them up. A fast way to do that is by using AutoSum.

1. Select the cell to the right or below the numbers you want to add.
2. Click the **Home** tab, and then click **AutoSum** in the **Editing** group.



AutoSum adds up the numbers and shows the result in the cell you selected.

Create a simple formula

Adding numbers is just one of the things you can do, but Excel can do other math as well. Try some simple formulas to add, subtract, multiply, or divide your numbers.

1. Pick a cell, and then type an equal sign (=).

   That tells Excel that this cell will contain a formula.

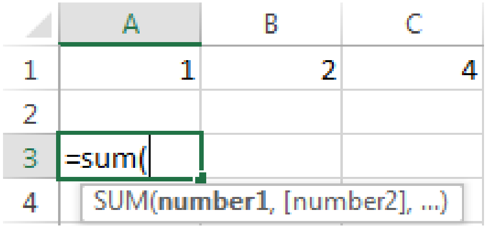
2. Type a combination of numbers and calculation operators, like the plus sign (+) for addition, the minus sign (-) for subtraction, the asterisk (*) for multiplication, or the forward slash (/) for division.

   For example, enter **=2+4**, **=4-2**, **=2*4**, or **=4/2**.

3. Press Enter.

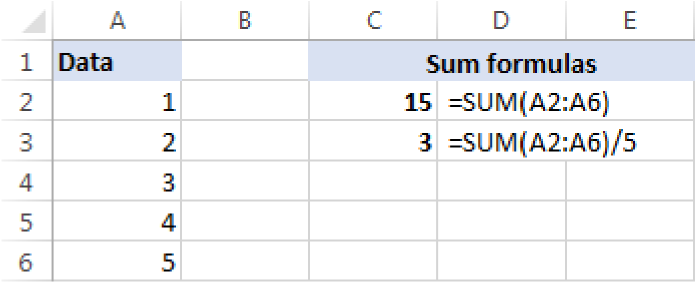
   This runs the calculation.

   You can also press Ctrl+Enter if you want the cursor to stay on the active cell.

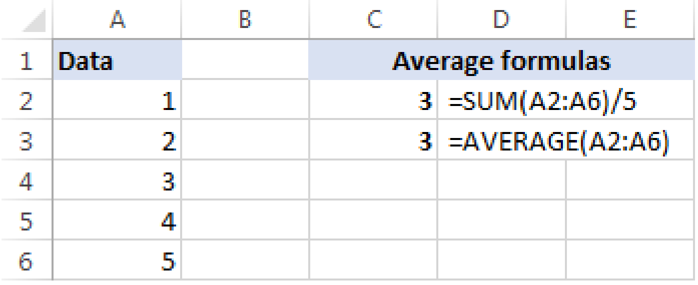


	A	B	C
1	1	2	4
2			
3	=sum(		
4			

SUM(**number1**, [number2], ...)



	A	B	C	D	E
1	**Data**		**Sum formulas**		
2	1		**15**	=SUM(A2:A6)	
3	2		**3**	=SUM(A2:A6)/5	
4	3				
5	4				
6	5				



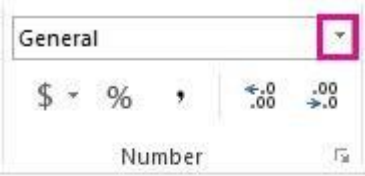
	A	B	C	D	E
1	**Data**		**Average formulas**		
2	1		**3**	=SUM(A2:A6)/5	
3	2		**3**	=AVERAGE(A2:A6)	
4	3				
5	4				
6	5				

Apply a number format

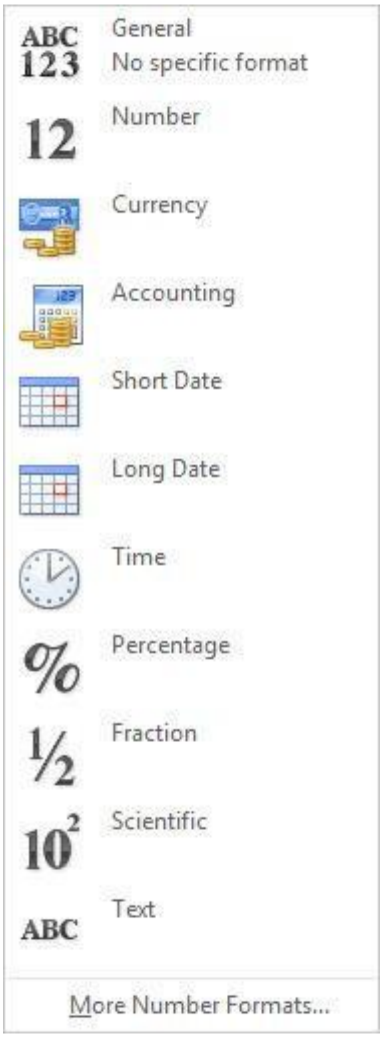
To distinguish between different types of numbers, add a format, like currency, percentages, or dates.

1. Select the cells that have numbers you want to format.

2. Click the **Home** tab, and then click the arrow in the **General** box.



3. Pick a number format.



If you don‘t see the number format you‘re looking for, click **More Number Formats**.

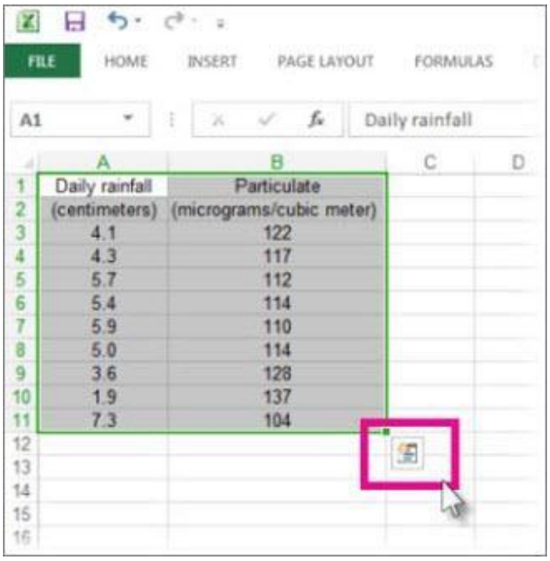
Put your data in a table

A simple way to access Excel‘s power is to put your data in a table. That lets you quickly filter or sort your data.

1. Select your data by clicking the first cell and dragging to the last cell in your data.

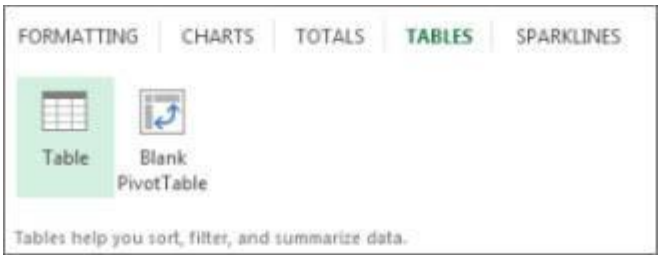
   To use the keyboard, hold down Shift while you press the arrow keys to select your data.

2. Click the **Quick Analysis** button in the bottom-right corner of the selection.



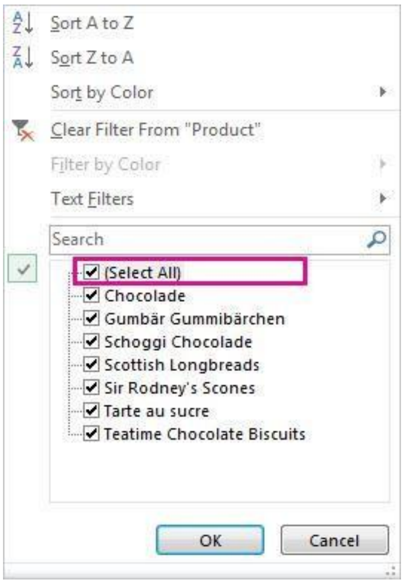
	A	B
1	Daily rainfall	Particulate
2	(centimeters)	(micrograms/cubic meter)
3	4.1	122
4	4.3	117
5	5.7	112
6	5.4	114
7	5.9	110
8	5.0	114
9	3.6	128
10	1.9	137
11	7.3	104

3. Click **Tables**, move your cursor to the **Table** button to preview your data, and then click the **Table** button.

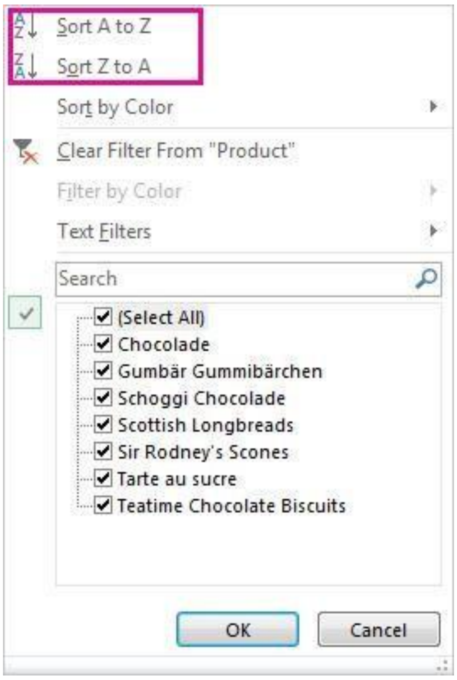


4. Click the arrow ▾ in the table header of a column.

5. To filter the data, clear the **Select All** check box, and then select the data you want to show in your table.



6. To sort the data, click **Sort A to Z** or **Sort Z to A**.



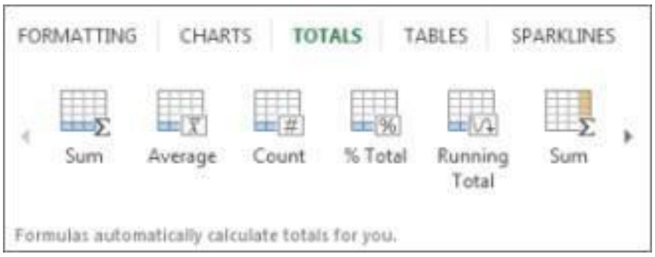
☑ (Select All)
☑ Chocolade
☑ Gumbär Gummibärchen
☑ Schoggi Chocolade
☑ Scottish Longbreads
☑ Sir Rodney's Scones
☑ Tarte au sucre
☑ Teatime Chocolate Biscuits

7. Click **OK**.

Show totals for your numbers using Quick Analysis

The Quick Analysis tool (available in Excel 2016 and Excel 2013 only) let you total your numbers quickly. Whether it‘s a sum, average, or count you want, Excel shows the calculation results right below or next to your numbers.

1. Select the cells that contain numbers you want to add or count.
2. Click the **Quick Analysis** button in the bottom-right corner of the selection.
3. Click **Totals**, move your cursor across the buttons to see the calculation results for your data, and then click the button to apply the totals.

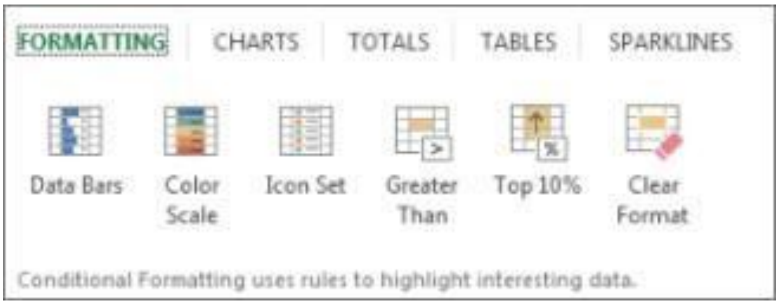


Formulas automatically calculate totals for you.

Add meaning to your data using Quick Analysis

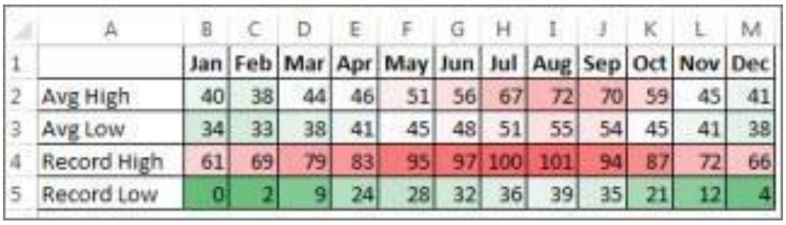
Conditional formatting or sparklines can highlight your most important data or show data trends. Use the Quick Analysis tool (available in Excel 2016 and Excel 2013 only) for a Live Preview to try it out.

1. Select the data you want to examine more closely.
2. Click the **Quick Analysis** button in the bottom-right corner of the selection.
3. Explore the options on the **Formatting** and **Sparklines** tabs to see how they affect your data.



Conditional Formatting uses rules to highlight interesting data.

For example, pick a color scale in the **Formatting** gallery to differentiate high, medium, and low temperatures.



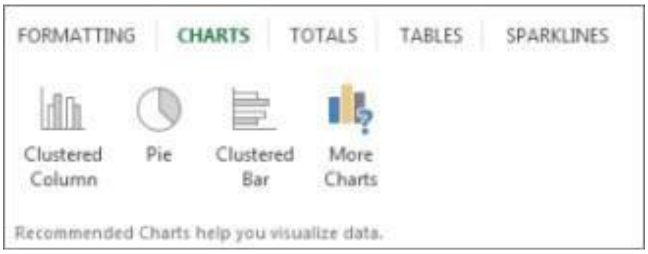
	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2	Avg High	40	38	44	46	51	56	67	72	70	59	45	41
3	Avg Low	34	33	38	41	45	48	51	55	54	45	41	38
4	Record High	61	69	79	83	95	97	100	101	94	87	72	66
5	Record Low	0	2	9	24	28	32	36	39	35	21	12	4

4. When you like what you see, click that option.

Show your data in a chart using Quick Analysis

The Quick Analysis tool (available in Excel 2016 and Excel 2013 only) recommends the right chart for your data and gives you a visual presentation in just a few clicks.

1. Select the cells that contain the data you want to show in a chart.
2. Click the **Quick Analysis** button in the bottom-right corner of the selection.
3. Click the **Charts** tab, move across the recommended charts to see which one looks best for your data, and then click the one that you want.



Note: Excel shows different charts in this gallery, depending on what‘s recommended for your data.

Sort your data: To quickly sort your data

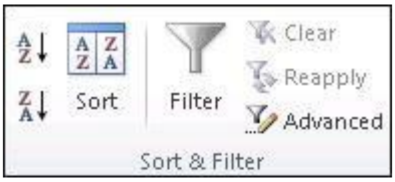
1. Select a range of data, such as A1:L5 (multiple rows and columns) or C1:C80 (a single column). The range can include titles that you created to identify columns or rows.
2. Select a single cell in the column on which you want to sort.
3. Click AZ↓ to perform an ascending sort (A to Z or smallest number to largest).
4. Click ZA↓ to perform a descending sort (Z to A or largest number to smallest).

To sort by specific criteria

1. Select a single cell anywhere in the range that you want to sort.
2. On the **Data** tab, in the **Sort & Filter** group, choose **Sort**.
3. The **Sort** dialog box appears.
4. In the **Sort by** list, select the first column on which you want to sort.
5. In the **Sort On** list, select either **Values**, **Cell Color**, **Font Color**, or **Cell Icon**.
6. In the **Order** list, select the order that you want to apply to the sort operation — alphabetically or numerically ascending or descending (that is, A to Z or Z to A for text or lower to higher or higher to lower for numbers).

Filter your data

1. Select the data that you want to filter.
2. On the **Data** tab, in the **Sort & Filter** group, click **Filter**.

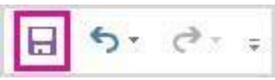


3. Click the arrow ▾ in the column header to display a list in which you can make filter choices.

4. To select by values, in the list, clear the **(Select All)** check box. This removes the check marks from all the check boxes. Then, select only the values you want to see, and click **OK** to see the results.

Save your work

1. Click the **Save** button on the **Quick Access Toolbar**, or press Ctrl+S.



If you‘ve saved your work before, you‘re done.

2. If this is the first time you've save this file:

a. Under **Save As**, pick where to save your workbook, and then browse to a folder.

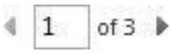
b. In the **File name** box, enter a name for your workbook.

c. Click **Save**.

Print your work

1. Click **File**, and then click **Print**, or press Ctrl+P.

2. Preview the pages by clicking the **Next Page** and **Previous Page** arrows.



The preview window displays the pages in black and white or in color, depending on your printer settings.

If you don‘t like how your pages will be printed, you can change page margins or add page breaks.

3. Click **Print**.

Activate and use an add-in

1. On the **File** tab, choose **Options**, and then choose the **Add-Ins** category.

2. Near the bottom of the **Excel Options** dialog box, make sure that **Excel Add-ins** is selected in the **Manage** box, and then click **Go**.

3. In the **Add-Ins** dialog box, select the check boxes the add-ins that you want to use, and then click **OK**.

If Excel displays a message that states it can't run this add-in and prompts you to install it, click **Yes** to install the add-ins.

Find and apply a template

Excel allows you to apply built-in templates, to apply your own custom templates, and to search from a variety of templates on Office.com. Office.com provides a wide selection of popular Excel templates, including budgets.

Arithmetic Precedence

Microsoft Excel follows the rules of arithmetic precedence when evaluating formulas.

( )	operations enclosed in parentheses are evaluated first; nested parentheses are evaluated from the inside out
^	exponentiation
* and /	multiplication and division, evaluated from left to right
+ and -	addition and subtraction, evaluated from left to right

Example

- the formula **=5*3-4^2** evaluates as **-1**
- the formula **=(5*(3-4)^2)** evaluates as **5**

Deleting a Column or a Row

- Click on the column or row header to highlight the entire column or row to be deleted. Right-click on any cell in the highlighted column or row. Click on **Delete** from the menu.

Inserting a Column

- Click on the column header *directly to the right* of where you want to insert a new column. Right-click on any cell in the highlighted column. Click on **Insert** from the menu.

Inserting a Row

- Click on the row header *directly below* where you want to insert a new row. Right-click on any cell in the highlighted row. Click on **Insert** from the menu.

Question 1: Discuss the utility or importance of Microsoft excel and Microsoft word.

Question 2: Write down the formula used in Microsoft excel

a. **Average**

b. **Sum**

c. **Correlation**

d. **Variance**

e. **Standard deviation**

Question 3: Applications of E-mail