Introduction about Statistics

- Biostatistics, a portmanteau word constructed from biology and statistics, is defined as per the etymology; application of statistics in biology.

- Historically the field of statistics was emerged and systematically developed to answer various problems in biology, especially on morphometry (measurement of morphological traits) and population genetics.

- It was only later that the field started having applications in various other disciplines, notably in quantitative fields of humanities (psychology and economics) such that the original meaning of statistics got steadily expanded necessitating the coinage of biostatistics to refer biological statistics.

- The term statistics now acquired a new meaning, “branch of mathematics that deals with the experimental design, the collection of numerical data, summarization of the data, analysis and interpretation of the data for drawing inferences on the basis of the probability.”

- British biologist and population geneticist, **Sir Ronald Aylmer Fisher,** is usually considered as the father of statistics because of a number of seminal contributions that he made to the discipline (for example, F-distribution and ANOVA). However, the claim is contested by many.

- Perhaps the development of the field began as early as **17th** Century when British philosopher **Francis Bacon** published *Novum Organum* in 1620 that detailed fundamentals of inductive reasoning, an attribute of statistics as we will discuss later in this module.

- Other key scientists behind the development of statistics include evolutionary biologist **Karl Pearson**, geneticist **Sewall G**. Wright, population geneticist **JBS Haldane**, geneticist **Charles Davenport**, geneticist **William Bateson**, Botanist **Wilhelm Johannsen** and morphometricians **Raphael Weldon and D'Arcy Thompson**.

- Morphometrician and anthropologist Prasanta **Chandra Mahalanobis** is considered as the father of statistics in India. He was one of the founding members of the erstwhile Planning Commission of India and founded Indian Statistical Institute, Kolkata. A long-term friend of British population geneticist JBS Haldane, Mahalanobis invited Haldane to work with him at ISI. Haldane worked from 1957 till 1961 at ISI and made several contributions to the development of statistics in India.

- On the other hand, the term 'mathematical biology' is defined as an interdisciplinary field encompassing all applications of mathematics to the biology. Development of this field is concurrent with that of biostatistics.

Scope:

- The scope of biostatistics is extensive and cover almost the whole of biology that deals with generation and analysis of numerical data. Biostatistics is used right from designing scientific experiments through the data analysis.

- The scope includes principles of scientific methodology, defining various types of data and studies, levels of measurements, descriptive statistics, inferential statistics and hypothesis testing, and correlation.

- The field also includes various predictive methods and curve/model-fitting including regression analysis, maximum likelihood, Bayesian Inference and Principal Component Analysis.

- The scope of mathematical biology is equally vast to include various applications of mathematics in biology.

- Beside biostatistics, this field also encompasses applications of other mathematical disciplines including probability, number theory, game theory, set theory, neural networks, mathematical modelling, use of calculus in biology, fractals and Fibonacci series, and so on.

- Mathematics is used very often in population genetics, environmental biology, ecology, psychology, evolutionary analysis, enzyme kinetics and so on.

Definition:

Biostatistics are the development and application of statistical methods to a wide range of topics in biology. It encompasses the design of biological experiments, the collection and analysis of data from those experiments and the interpretation of the results

Function of statistics:

- To Present **Facts** in Definite Form: We can represent the things in their true form with the help of figures.
- Precision to the **Facts**
- Comparisons:
- Formulation and Testing of Hypothesis:
- Forecasting:
- Policy Making:
- It Enlarges Knowledge:
- To Measure Uncertainty:

Limitations of statistics:

Statistics with all its wide application in every sphere of human activity has its own limitations. Some of them are given below.

1. Statistics is not suitable to the study of qualitative phenomenon: Since statistics is basically a science and deals with a set of numerical data. It is applicable to the study of quantitative measurements. As a matter of fact, qualitative aspects like empowerment, leadership, honesty, poverty, intelligence etc., cannot be expressed numerically and statistical analysis cannot be directly applied on these qualitative phenomena.

2. Statistical laws are not exact: It is well known that mathematical and physical sciences are exact. But statistical laws are not exact and statistical laws are only approximations. Statistical conclusions are not universally true. They are true only on an average.

3. Statistics table may be misused: Statistics must be used only by experts; otherwise, statistical methods are the most dangerous tools on the hands of the inexpert. The use of statistical tools by the inexperienced and untrained persons might lead to wrong conclusions.

4. Statistics is only one of the methods of studying a problem: Statistical method does not provide complete solution of the problems because problems are to be studied taking the background of the countries culture, philosophy, religion etc., into consideration. Thus the statistical study should be supplemented by other evidences.

CLASSIFICATION AND TABULATION

The collected data after their scrutiny need to be classified in order to make the data fit for analysis and interpretation the first step in the analysis and interpretation of data is classification and tabulation.

Classification is the first step in tabulation, eventhough the phrase "classification and tabulation" is used. Proper classification helps proper tabulation.

Classification:

It is the process of arranging the data on the bases of some common characteristics possessed by them.

(E.g) If sex is the basis of classification, then all the male population will be grouped together on one side and the female population on the other side.

Likewise if age is the bases of classification persons of the same age will be grouped together and so on.

objects of Classification:

The chief objects of classification are,

1. To condense the mass of data
2. To present the facts in a simple form
3. To bring out clearly the points of similarity and dissimilarity
4. To facilitate comparison
5. To bring out the relationship
6. To prepare data for tabulation
7. To facilitate the statistical treatment of the data.

Types of classification :

There are numerous ways of classifying the data The important types are

1. Geographical
2. Chronological
3 Qualitatie
4. Quantitative

1. Geographical classification (Regionwise classification).

This type of classification is based on geographical region like contries, ~~a~~ states, districts, taluks. etc.

(E.g) The yield of agricultural output per hectare for different countries in a particular year is given below.

Table : The yield of agricultured output

Country	Average output (in kg per he)
U.S.A	600
China	300
Pakistan	250
India	150.

2. Chronological Classification :-

This type of classification is based on time of its occurrence such as years, months, weeks, days, hours, etc.

(E.g) This fish (Catla) production in a particular farm over 5 years is given below.

year	Fish production (in kg per ht)
1987	1400
1988	1500
1989	1450
1990	1550
1991	1600.

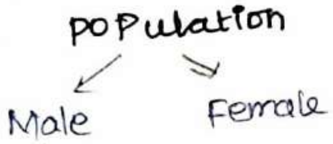
3. Qualitative Classification : (Descriptive Classification

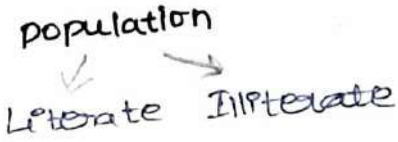
This type of classification is based on the quality or attributes such as sex, literacy, marital status, etc. so it is also called descriptive classifications.

It is further divided into two types

1. simple classification
2. Manifold classification .

i) simple classification : The data are classified into only two classes. It is dichotomy or two fold. E.g .

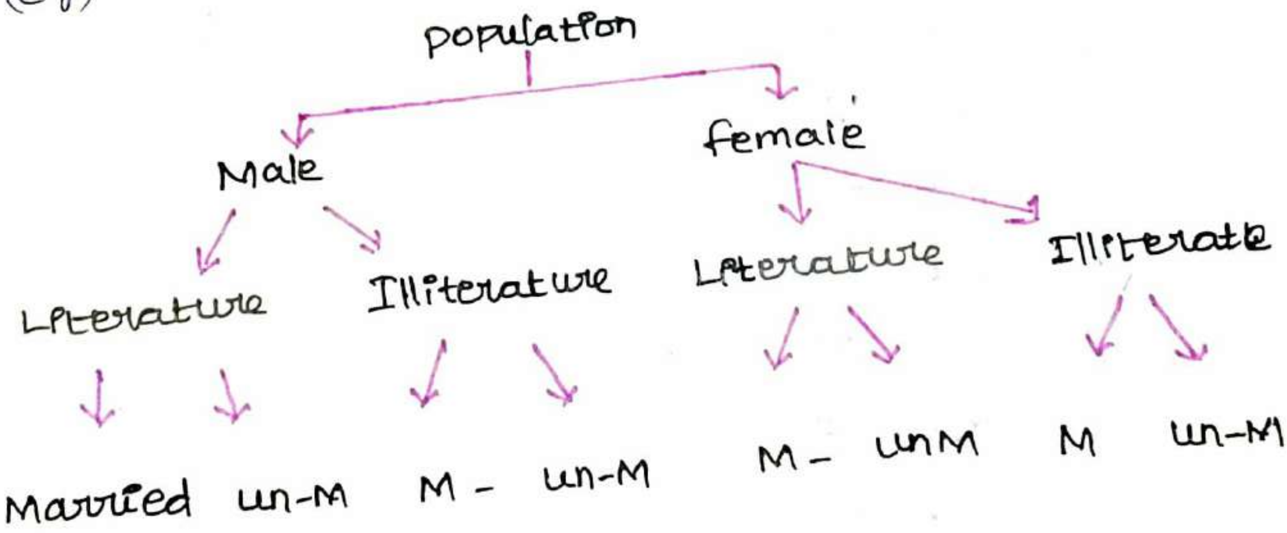




ii) Manifold classification:

Here, the data are classified into many classes.

(eg)



4. Quantitative Classification :

This type classification is based on some quantitative phenomenon such as age, height, weight, etc Here this is the quantitative phenomenon under study.

Hence this Classification is also called Classification by variables.

For e.g the weight of 100 fishes reared in a pond are given below.

Wt (in gms)	No. of fishes
0 - 100	8
100 - 200	16
200 - 300	20
300 - 400	12
400 - 500	15
500 - 600	13
600 - 700	6
700 - 800	6
800 - 900	2
900 - 1000	2
Total	100

TABULATION OF DATA

Defenition :-

Tabulation may be defened as the logical and systematic arrangement of statistical data in rows and columns.

It is designed to simplify presentation and facilitate comparison and analysis. Columns are vertical arrangements and rows are horizontal arrangements.

Objectives :

To clarify the object of investigator

To simplify the complex data

To present the facts in the minimum space

To facilitate comparison.

To detect errors and omissions in the data

To facilitate statistical processing

To help reference.

Parts of Table :

Arranging values in columns is called tabulation. A column of values is called a table. Tabulation is a presentation of data.

A table contains boxes called cells. The cells are arranged in horizontal rows and vertical columns.

A typical table has the following parts :-

1. Table number
2. Title
3. Head note
4. Caption
5. Stub
6. Body
7. Foot note
8. Source.

Table 3.4 : Food habits of III B.Sc Zoology students

Eggtarians : Egg-eaters

Table number → Head note → Title

Food habits	Number of students	
	Boys	Girls
Vegetarians	2	7
Eggtarians (Egg eaters)	1	2
Non-vegetarians	7	11

→ Caption

Stub ←

→ Body

Foot Note → Census was made during 2000-2001

Source → Data collected by the class teacher.

1. The table has a number and it is given at the top..
2. The name of the table is called title. It is given at the top.
3. Head note refers to the units of value given below the title.
4. The heading of vertical columns are called caption.
5. The headings of horizontal columns are called stub.
6. The value given in the horizontal and vertical columns are ~~be~~ called body
7. Foot note is given below the table. It gives explanation on the values.
8. Source refers to source of information. It is given at the bottom of the table.

1. Table number :

A table should always be numbered for easy identification and reference in future. The table number may be placed at the top of the table either in the centre above the tittle or in the left side of the title

2. Title :

Every table must be given a title, which usually appears at the top of the table. It should be clear, brief and self-explantory.

3. Head-note :

It is actually a part of the title. It explains certain points relating to the whole table that have not been included in the title captions or stubs.

4. Caption :

It referes to the headings of vertical columns. It has usually main heading and sub-heading. It should be clear brief and self-explaratory.

5. Stub

It refers to the headings of horizontal rows.

6. Body :

It contains numerical information arranged in accordance with caption and stub. the arrangment is generally from left to right in the horizontal rows and from top to pottom in the vertical columns.

7. Foot-note :

Anything in a table which the reader may find diffcult to understand can be explained in foot-notes.

8. Source :

It refers to the source from information has been taken. It should preferably include the name of the author, title, volume, number, page, publisher's name and the year of publication.

After collecting the data, the investigator has to undertake the task of its organization. By organization the classification and presentation of data in such a way that the data becomes easy and convenient to use and handle.

Thable are broadly classified into two types

1. Simple tables
2. complex tables.

1. simple Tables :

In a simple table, only one characteristic is shown. Hence this type of table is also known as one - way table. It has two factors placed in relation to each other. The following table showns the marks secured by students in a class test.

Marks	No. of students
0-5	2
5-10	5
10-15	10
15-20	11
20-25	9
25-30	20

2. complex tables :

In a complex table, more than two characteristics are shown. If there are two co-ordinate factors, the table is ~~three~~, called a double tables. If the number of co-ordinate group is three it is called as treble table, If it contains more than three co ordinate factors, then it is called as multiple table.

The marks of students can be classified according to the sex to get a double table

Marks	No. of students	
	Male	Female

The males and females are further classified according to their residence into hostellers or day scholars. It is a case of Treble table.

Marks	No. of Students			
	Male		female	
	Hostellers	Day Scholars	Hostellers	Day Scholars

If they are again classified as belonging to different religion, nationalities, states, etc, It will be an example for multiple table.

Marks.	Number of students											
	Male						female					
	Hostellers			Dayscholars			Hostellers			dayscholors		
	Hindu	Christian	Muslim	Hindu	Christian	Muslim	Hindu	Christian	Muslim	Hindu	Christian	Muslim.

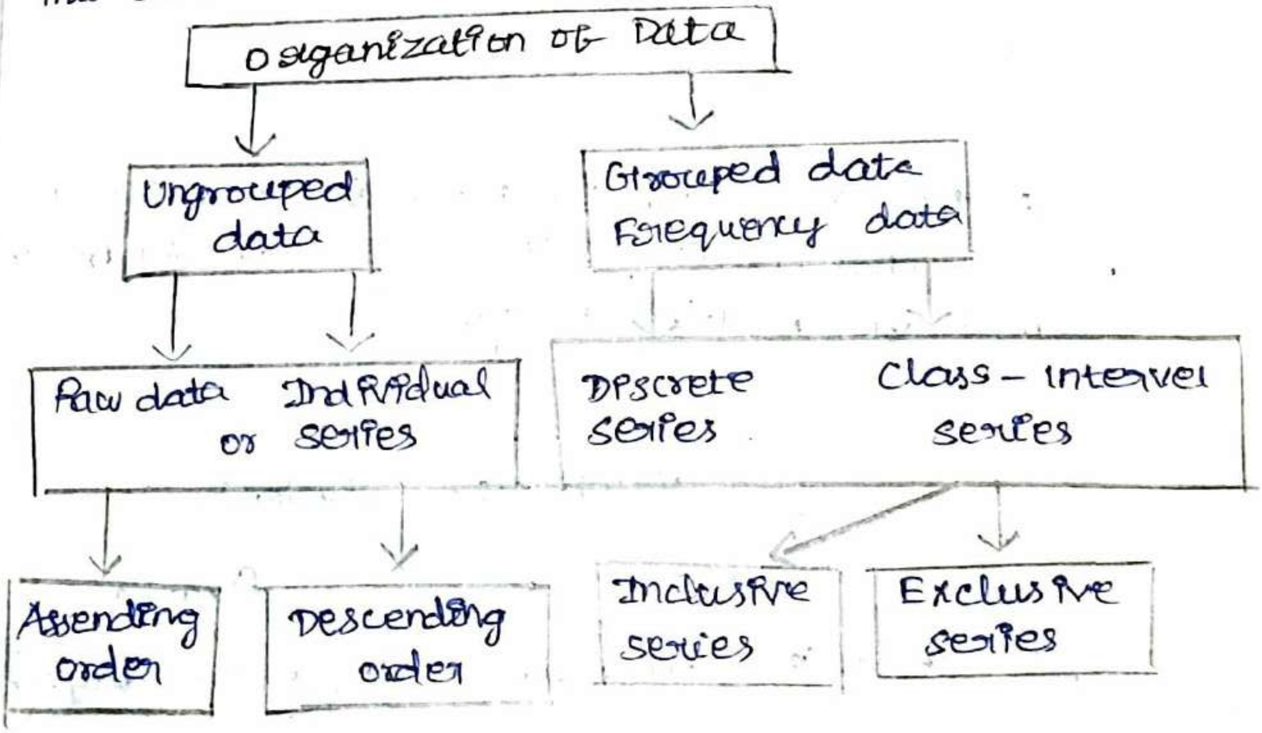
RAW DATA :

The statistical information collected from the investigation is known as raw data. suppose we are interested in the weight messurements of student of final B.Sc class. There are 40 students in this class.

Their weight measurements of 40 students table.

161	156	153	146	163	152	147	164
145	145	135	168	169	144	140	150
135	128	147	126	120	157	158	125
142	135	142	138	149	148	146	135
150	140	173	176	165	138	138	150

In the above table we have 40 observations relating to the weight measurement of 40 students this data is known as raw data.



Array data :

The data arranged in an order is called Array data

Manifold classification of the students of a class based on sex, stay and food habits.

Food habit	Males		Females	
	Day scholars	Hostellers	Day scholars	Hostellers
vegtarians	4	2	6	4
Non-vegtarians	3	1	7	3

Methods of classification:

The ungprocessed freshly collected data is called raw-data. The arranged in an order is Called array data.

The raw data is classified in three methods

i) individual series

ii) Discrete series

iii) Continuous series

i) Individual series:

In individual series, the values are written individually.

For example the weight of 10 fishes is written according to their serial number or in an ascending order in descending order.

Table

Serial No.	1	2	3	4	5	6	7	8	9	10
Weight gms	4	3	6	7	9	5	6	4	8	4

classification of individual series, Weight of 10 fishes in gms arranged in ascending order.

Weight of fishes in gms	3	4	4	4	5	6	6	7	8	9

Similarly, the students in a class can be arranged in different way e.g.

i) in the alphabetical order by their names

ii) in the serial order

iii) in the order in which they sit in the class

b. According to the size of the magnitude of the weights. The weight measurements can be recorded either in ascending or in decending order.

i) In Ascending order :

We start from the lowest value and go to the highest. The lowest weight is 119 Ib and the highest weight is 176 Ib. The arrangement is shown in table.

Arrangements of weights in asending order	120, 125, 126, 128, 135, 135, 135, 135, 136 ; 138.
Arrangements of weights in descending order	176, 173, 168, 165, 154, 147, 146.

ii) In Descending order :

We start from the highest Weight and go to the lowest. the arrangement is show in table.

The arrangement of the above said data in an asecending or desecending order of magnitude is known as an 'array'

Individual series, undoutedly, make the data relatively easier. But the story does not end here.

In our illustration we have only 40 observation and we are already finding that they are quite difficult to handle. Suppose, We have 100 or 500 or 1,000 observations it would be diffcult to handle them. We can solve the problem be organizing the data in discrete series.

2. Discrete series :

In discrete series, the data are given in group. It is also called discontinuous series

In discrete series the items having the same Values are grouped together.

For example in the above table, three fishes are having the same weight 4 gms. Similarly, 2 fishes are having the same weight 6 gms.

So the data are arranged as follows.

Classification of discrete series:

Weight in gms	3	4	5	6	7	8	9
No. of fishes	1	3	1	2	1	1	1

3. Continuous series:

In continuous series the data are presented. with class intervals.

In this method the data is devided into classes. The first class and the last class are fixed by seeing the lowest and highest values. The lowest and numbers of each class are called class limits. The lowest number is called lower limit and the highest number is called upper limit.

The class limits are made in three methods, namely

1. Exclusive method
2. Inclusive method
3. Open and class method.

In exclusive method. the class intervals are formed is following method.

0-10		0-5
10-20	or	5-10
20-30		10-15

In this method the value 10 is not included in the class 0-10, but included in the class 10-20. As the upper limit number is excluded from the class, It is called exclusive method.

Class frequencies :

The number of values of the series fall in a class are known as class frequencies.

Construction of class - interval series,

1. Determining the Range :

It is obtained by subtracting the size of the smallest item from the size of the largest item of the observed values is 119 and the highest value is 176. Thus the range will be (176 - 119) = 57.

2. Determining the Number of classes :

There is no hard and fast rule governing the number of classes in a series. It can be obtained by deviding the range by the size of the class interval.

Two things need be kept in consideration determining the number of classes.

(a) Number of classes should neither be very large nor very small.

(b) The size of the class - interval should be an easy . like 5, 10, 20 ... etc.

3. Mid values ;

The middle value of the class limits of class in interval is called mid - value of class interval.

$$\frac{100+150}{2} = \frac{250}{2} = 125$$

Weight Measurements of 40 students (Exclusive series).		
Weight (in Ib)	Tally bars	No. of stude-nts (Frequencies)
120-130	||||	4
130-140	卌 ||	7
140-150	卌 卌 |||	13
150-160	卌 卌	9
160-170	卌	5
170-180	||	2
	Σf =	40.

1. Frequency distribution with class intervals

Frequency distribution class interval is a table of values with classes. It is constructed with class intervals. It is frequency distribution of continuous series.

It invoves the following steps,

1. The data collected by the investigator is called raw data.
2. The data are arranged in an ascending order. The arranged data is called array data
3. The data is divided into 5 to 10 groups called classes.
4. The first class and the last class are fixed by seeing the lowest and highest values.
5. The lowest and highest numbers of each class are called class limits.

The lowest number of class is called lower limit. the highest number of a class is called upper limit.

6. The class limit may be made in two methods. Namely, Exclusive method and inclusive method. In the exclusive method the class limits are formed in the following way.

0-10   0-5
10-20   or   5-10
20-30   10-15

In the inclusive method, values class limpts of formed in following way.

0-9   0-4
10-19   or   5-9
20-30   10-14.

2. Cumulative Frequeny Distribution:

The cumulative frequency distribution is a statistical table the frequencies of proceeding classes are added.

A cumulative frequeny distribution for the weight of 30 fishes is constructed as follows.

1. A table of 3 columns is prepared
2. Classes are marked in the first column.
3. Frequeny is noted in the 2nd column
4 In the 3rd column the frequency for the first class. is entered as a such.

5. For the second class, the sum of the frequencies of first and second class are added and entered.

Weight	Frequency	Cumulative frequency
0-9	3	3
10-19	9	12
20-29	11	23
30-39	7	30

6. For other classes the frequencies are entered after adding the preceeding class frequency.

7. The cumulative frequency distribution containing measuring cumulative frequency is called less than cumulative distribution

8. the cumulative frequence distribution helps to find out the number of items below and above a particular weight.

PRESENTATION OF DATA.

Graphic presentation of Data:

Presenting data in the form of graphs is called graphic presentation of data.

Graph:

A graph is the geometrical image of a data.

A graph is a diagram consisting of lines of statistical data.

The graph is drawn on a graph paper

The graph has two intersecting lines called axes.

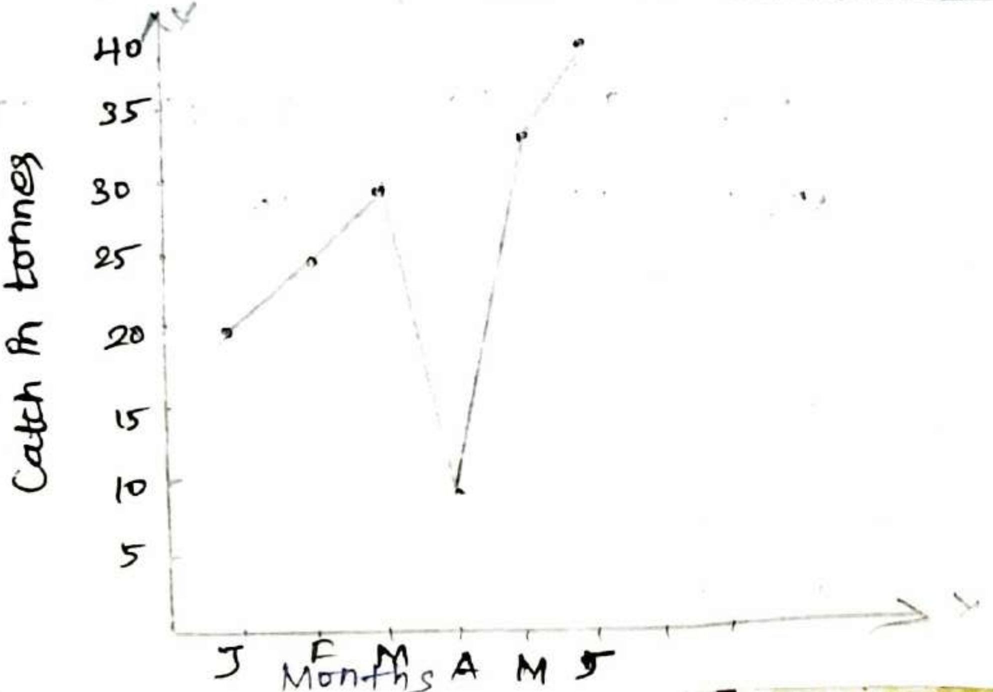
The horizontal line is called x axis. The vertical line is called y-axis

A title is given to a graph

The values corresponding to x and y axis are plotted on the paper.

Ex. Monthly fish landing in a pond

Month	Jan	Feb	Mar	Apr	May	June
Catch-ing	20	25	30	10	35	40



The graphs are classified into two types namely.

1. time series graphs
2. Frequency distributions graphs

A graph is the geometrical image of a data It is a mathematical image.

Graphs of Time Series (Line graphs).

In a line graph the data is represented in straight lines.

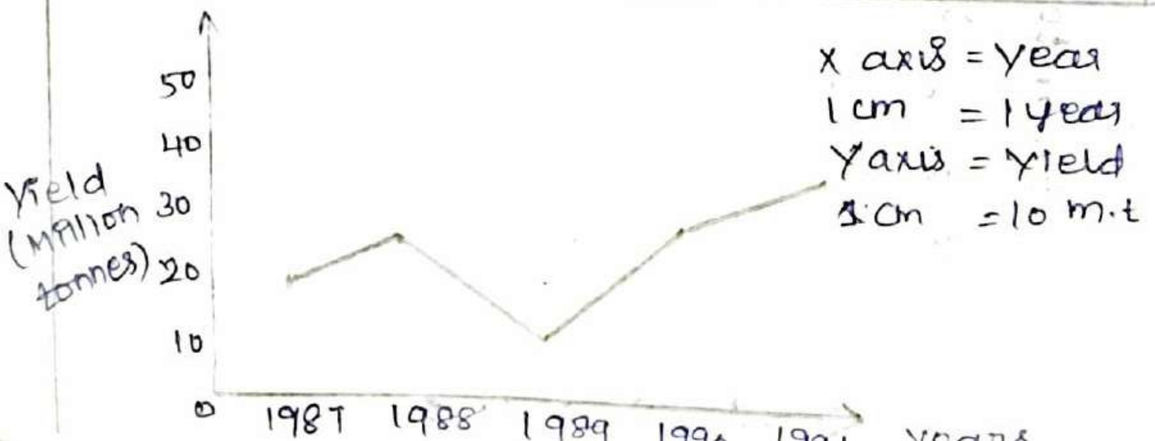
The line graph is divided into four types

They are. 1. Graph of one variable
2. Graph of two or more variables
3. Range Chart
4. Band Chart

1. Graph of one variable:

Only one variable is to be represented the desired, graph is obtained by plotting the time variable along the x is and the value of variables on y - axis on suitable scale. one Points are joined by straight lines.

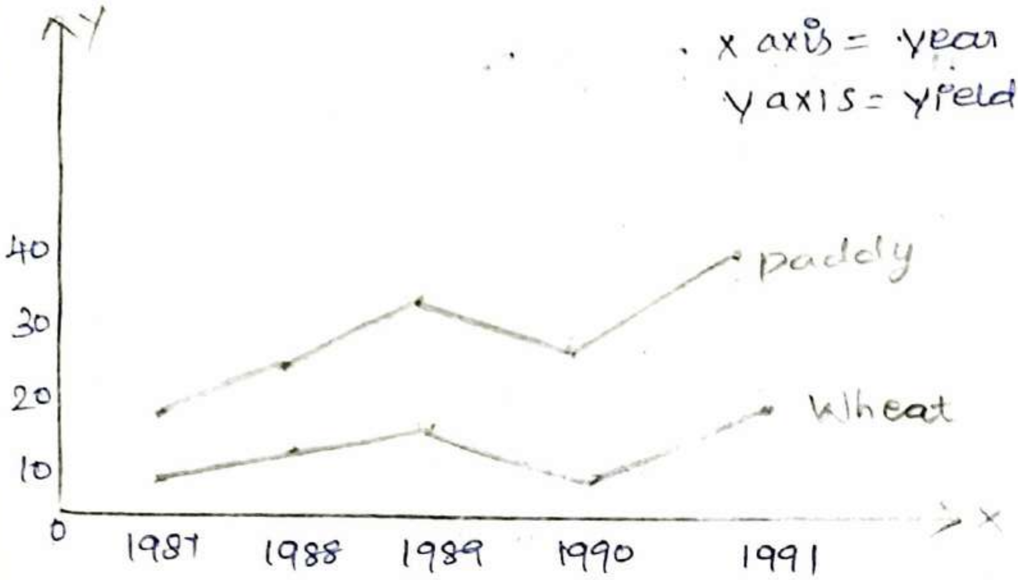
Year	1987	1988	1989	1990	1991
yield	20	25	10	30	40



2. Graph of two or more variables.

In this graph instead of one variable, two or more variable are taken time comparison is very easy in the graph.

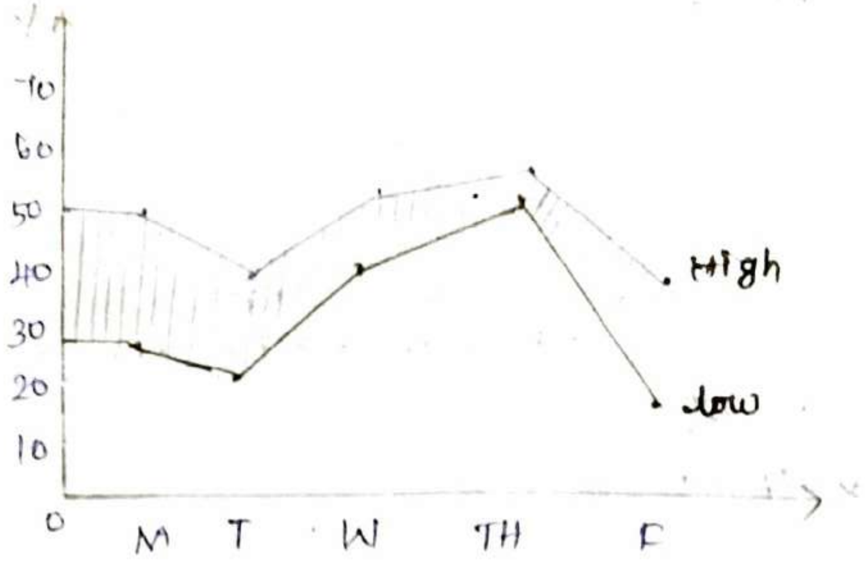
Items	1987	1988	1989	1990	1991
	Yield in Million (tonnes)				
paddy	20	25	35	30	40
Wheat	10	12	15	10	20



3. Range chart

It is used to exhibit the minimum and maximum values of a variable. For example to highlight the range of variation of the temperature on different days, the blood pressure readings of an individual in different days etc. the range chart is most appropriate method.

DAYS	Temperature °C	
	High	low
Monday	50	30
Tuesday	40	25
Wednesday	55	45
Thursday	60	55
Friday	40	20



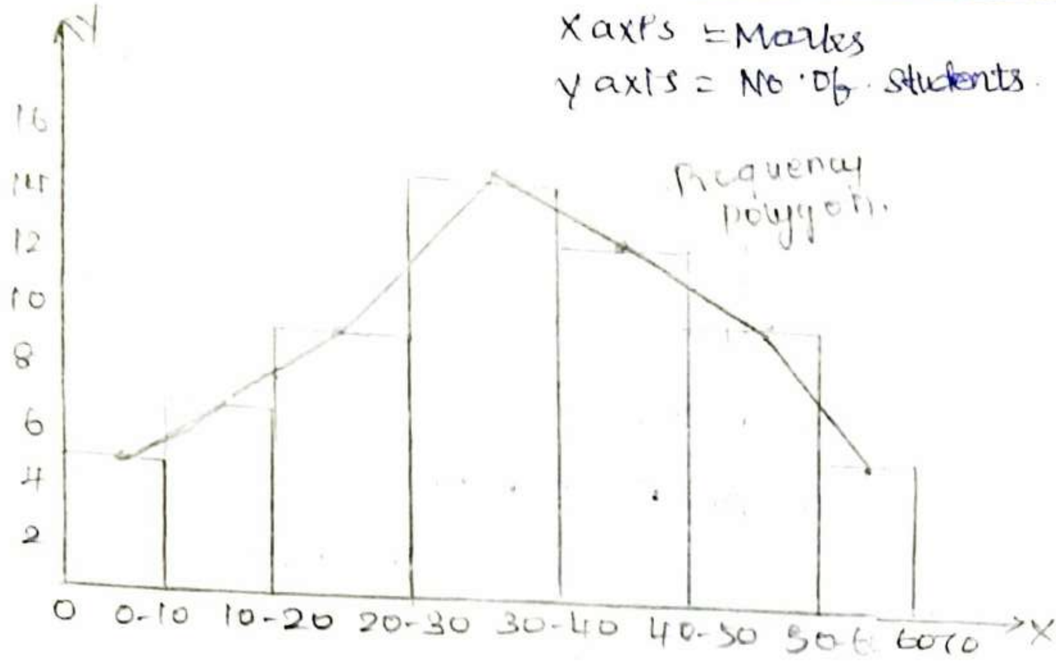
Histogram / Bar graph:

The histogram is a graph. It consists of vertical adjacent rectangles.

It is basically an area diagram. The class intervals are marked on the ox axis and the frequencies on the oy axis. The upper ends of the vertical lines are joined together. This gives rectangles.

The area of each rectangle is equal to the frequency of the class multiplied by their class intervals. In this sense, histogram is an area diagram. It is also known as block diagram or staircase chart or bar graph.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of Students	5	7	10	15	13	10	6



Histogram is a graph containing frequencie in the form of vertical rectangles.

It is an area diagram

It is a graphical presentation of frequency distribution.

The x-axis is marked with class intervals

The y-axis is marked with frequencies.

Vertical rectangles are drawn as per the height of the frequency of each class. The rectangles are drawn without any grap in between.

The width of the rectangles is equal to the range of the class.

The hight of each retangle is equal to the frequency of eachless.

The histogram is a two dimensional diagram because the hight and width of each retangle are as per data.

The histogram is different from a bar diagram. The bar diagram is one dimensional because in a bar diagram. The hight alone is as per data and the width is not as per data

Uses of Histogram :

It gives a clear proture of entir data

It simplifies a complex data

It is attractive anad impressive

It is easily memorised

Median and mode can be located

It facilitates comparison on one or more, frequency distributions on the same graph

It gives an idea of the pattern of distribution of variable in the population.

Polygon :

polygon is a histogram with straight lines joining the midpoints of the top of the rectangles.

polygon means a figure with many angles

It is an area diagram polygon is a graph. It is graphical representation of frequency distribution.

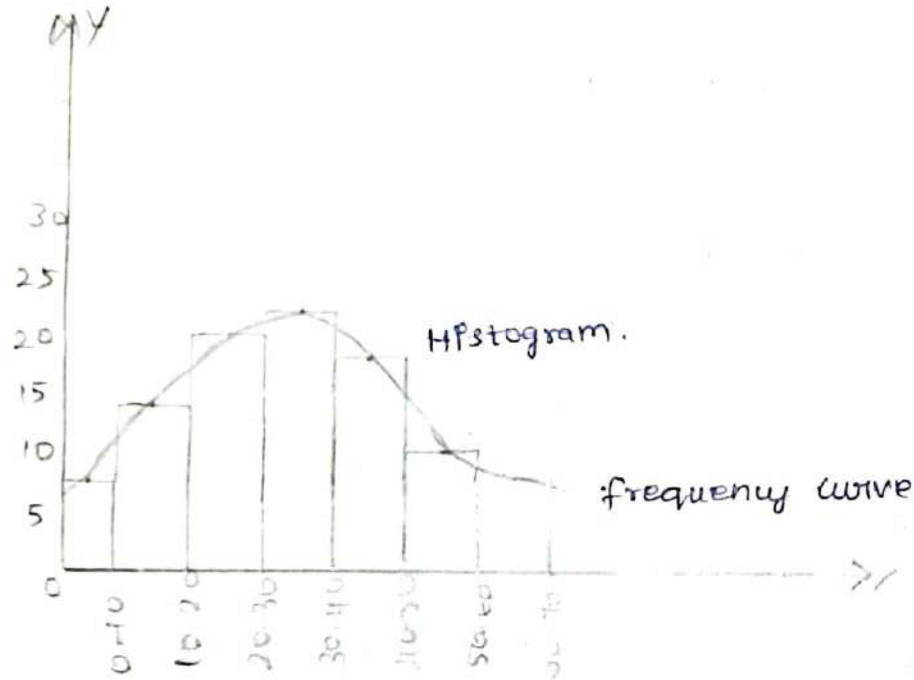
The x-axis is marked with class intervals.

The y-axis is marked ~~to~~ with frequencies.

vertical rectangles are drawn as per height of the frequency of each class. The rectangles are drawn without any gap in between.

The width of each rectangles is equal to the frequency of each class.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of Students	7	14	20	22	18	12	9



The midpoints of the top of the rectangles are joined by straight times.

The area of the polygon is equal to the area of histogram.

Uses of Frequency polygon

It gives a clear picture of the entire data.

It simplifies a complex data

It is attractive and impressive

It is easily memorised.

Median and mode can be located

It facilitates comparison of two or more frequency distributions on the same graph.

It gives an idea of the pattern of distribution of variables in the population.

Frequency curve:

Frequency curve is a graph of frequency distribution the line is smooth.

It is just like a frequency polygon. In the polygon the line is straight but in the curve the line is smooth.

It is an area diagram.

It is the graphical representation of frequency distribution.

The x-axis is marked with class intervals

The y-axis is marked with frequencies

A Histogram is drawn. the midpoints of the top of the rectangles are joined by a smooth line.

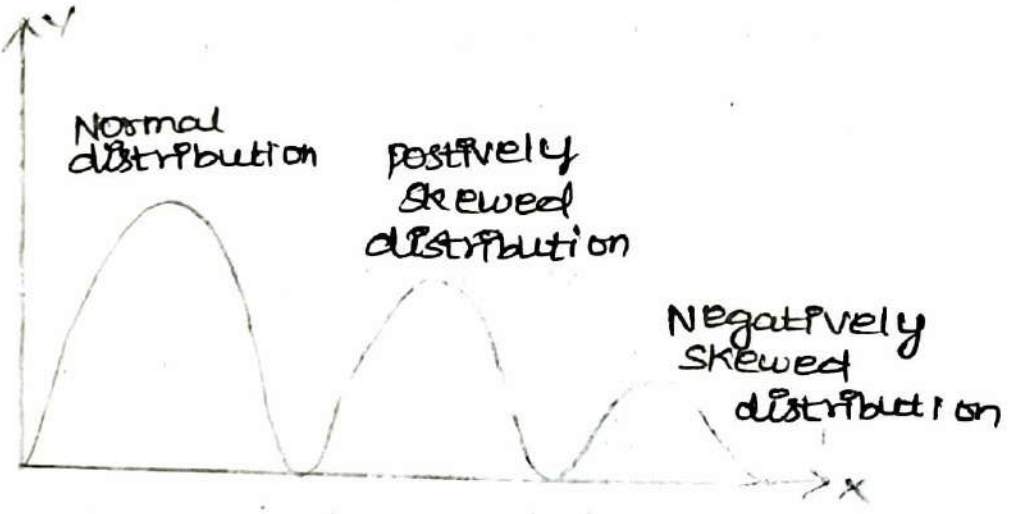
The beginning and end of the curve should touch the x-axis at the mid points of first and last class interval:

The area of the curve is equal to that of a histogram.

The frequency curve is divided into 3 types based on the shape of the curve. They are:,

1. Normal distribution curve
2. positively skewed distribution curve
3. Negatively skewed distribution curve.

The normal distribution curve is symmetrical and has an inverted bell - shape.


Y
Normal distribution
positively skewed distribution
Negatively skewed distribution
X

Frequency curves.

The positively skewed distribution curve is asymmetrical. The low values of the variables have the highest frequencies.

The negatively skewed distribution curve is also asymmetrical. High values of the variables have the highest frequencies.

Uses of Frequency curve:

It gives clear picture of the entire data.

It is easily memorised

It facilitates comparison of two or more frequency distributions on the same graph.

It gives an idea of the pattern of distribution of variables in the population.

Ogive curve:

cumulative frequencies are plotted on a graph then the frequency curve obtained is called ogive or cumulative frequence curve.

The class limits are shown along the x axis and cumulative frequencies along the y axis.

In drawing an ogive, the cumulative frequency is plotted at the upper limit of the class interval. the successive points are later joined togather to get an ogive curve. There are two types

1. Less than ogive
2. More than ogive.

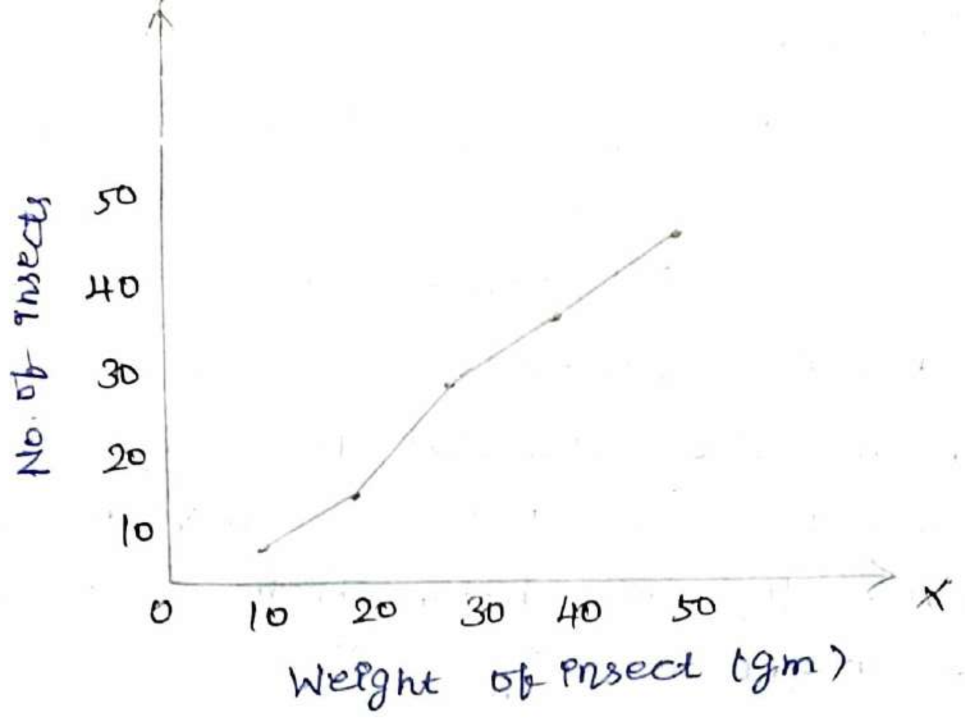
a. Less than ogive or less then cumulative Frequency Curve.

In this type of curves We cumulate the frequencies from above to below and the cumulated frequencies are poltted. The cumulated frequency is therefore minimum in the first class and gradually increases, with the result that the curve upwards from left side bottom to the right side top. so We get a rising curve.

Thus the less than cumulative frequencies are plotted against the upper class boundaries of the repective classes. The points so obtained are joined by smooth free hand curve to give "less than ogive"

Wt. of insect	0-10	10-20	20-30	30-40	40-50
Frequency	5	10	15	8	6

Weight lessthan	cumulative frequency
10	5
20	~~10~~15
30	30
40	38
50	44



b Mar than ogive or More than cumulative Frequency curve.

In this type of curves, we cumulate the frequencies from below to above and the cumulated frequencies are plotted. The cumulated frequency is therefore. maximum in the first class and gradually decreases, with the result that the curve drops from left side top right side bottom. So we get a decleining curve.

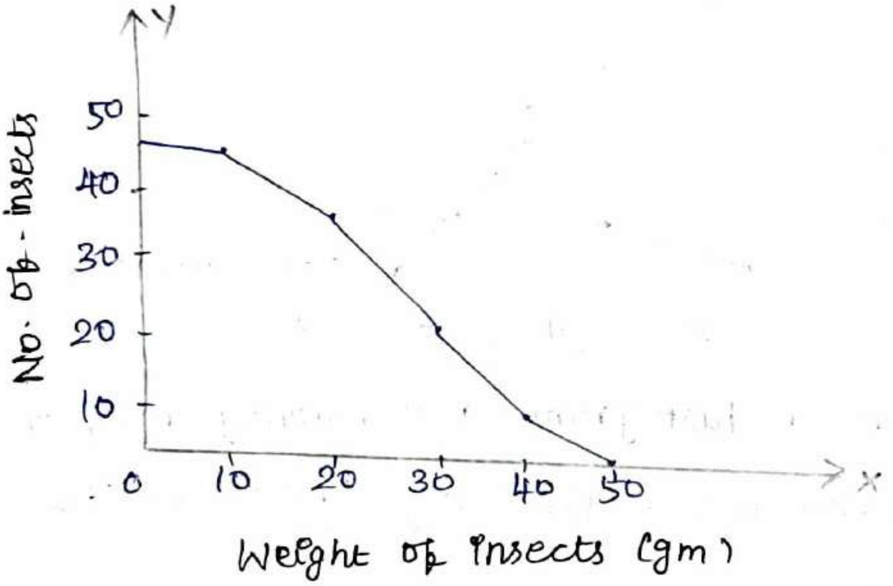
Thus the more than cumulative. frequencies are plotted against the lower class boundaries of the respective Classes.

The points so obtained are joined by a smooth free hand curve to give "more than ogive"

Wt. of insect (gm)	0-10	10-20	20-30	30-40	40-50
Frequency	5	10	15	8	6

Solution :

Weights more than	f
0	44
10	39
20	29
30	14
40	6
50	0

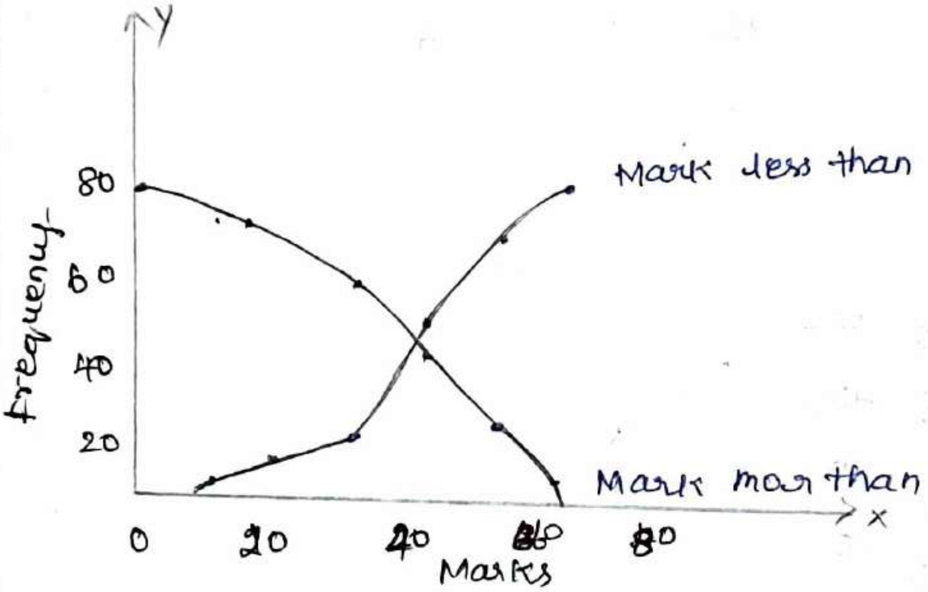


Draw less than and more than cumulative frequency curve for following data

Marks	0-10	10-20	20-30	30-40	40-50	50-60
frequency						

Solution:

Marks less than	f	Marks morethan	f
10	3	0	80
20	12	10	77
30	27	20	68
40	57	30	53
50	75	40	23
60	80	50	5

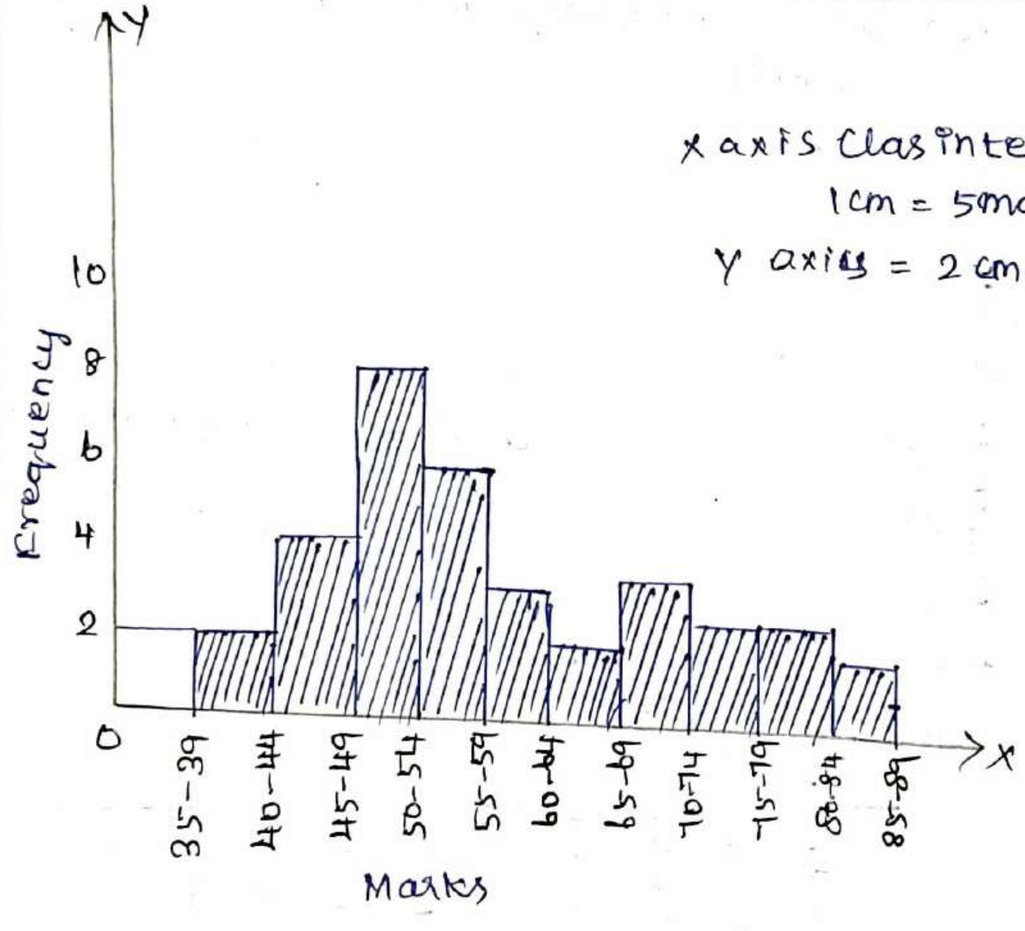


Draw a histogram a frequency polygon and a cumulative frequency curve for the following data:

Marks	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84
Frequency	2	4	6	8	6	3	2	4	3	3

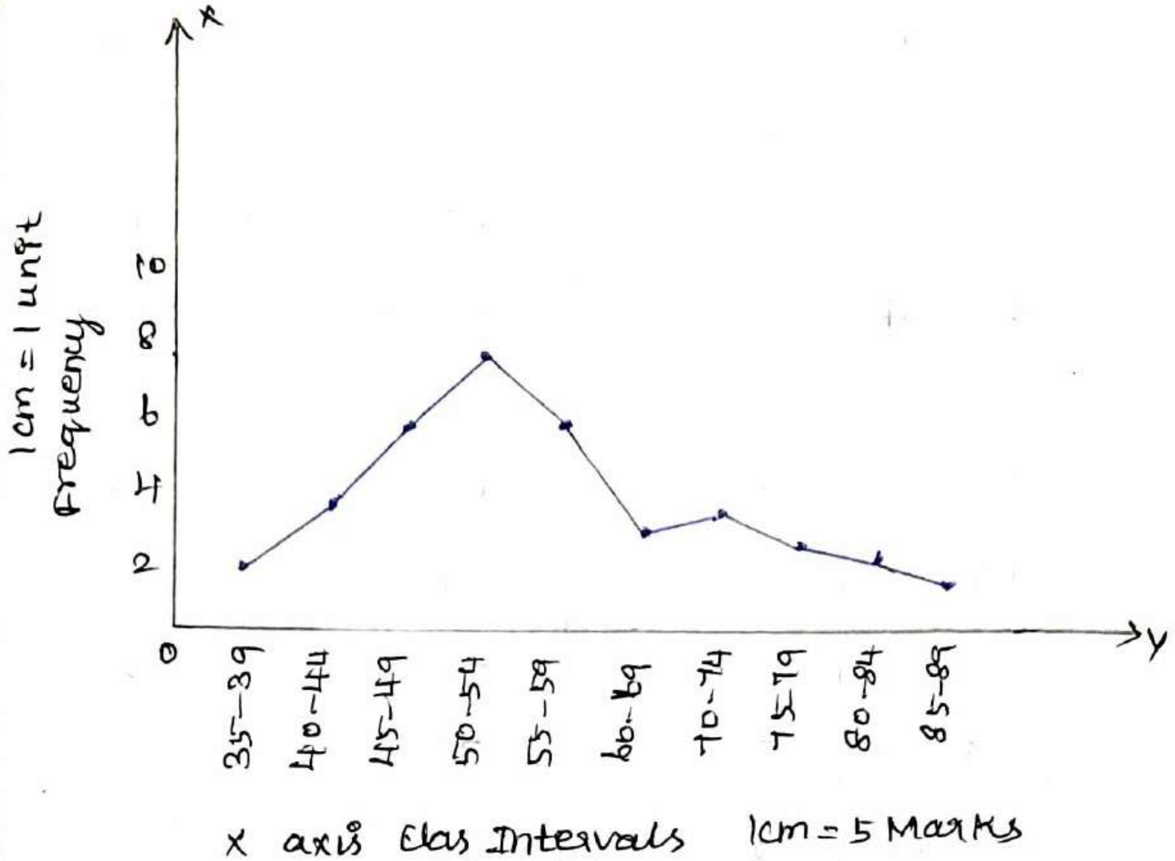
Class Interval	Frequency	Mid point (Mid pt)	Cumulative frequency (C.F)
35-39	2	37	2
40-44	4	42	6
45-49	6	47	12
50-54	8	52	20
55-59	6	57	26
60-64	3	62	29
65-69	2	67	31
70-74	4	72	35
75-79	3	77	38
80-84	3	82	41
85-89	2	87	43

1)

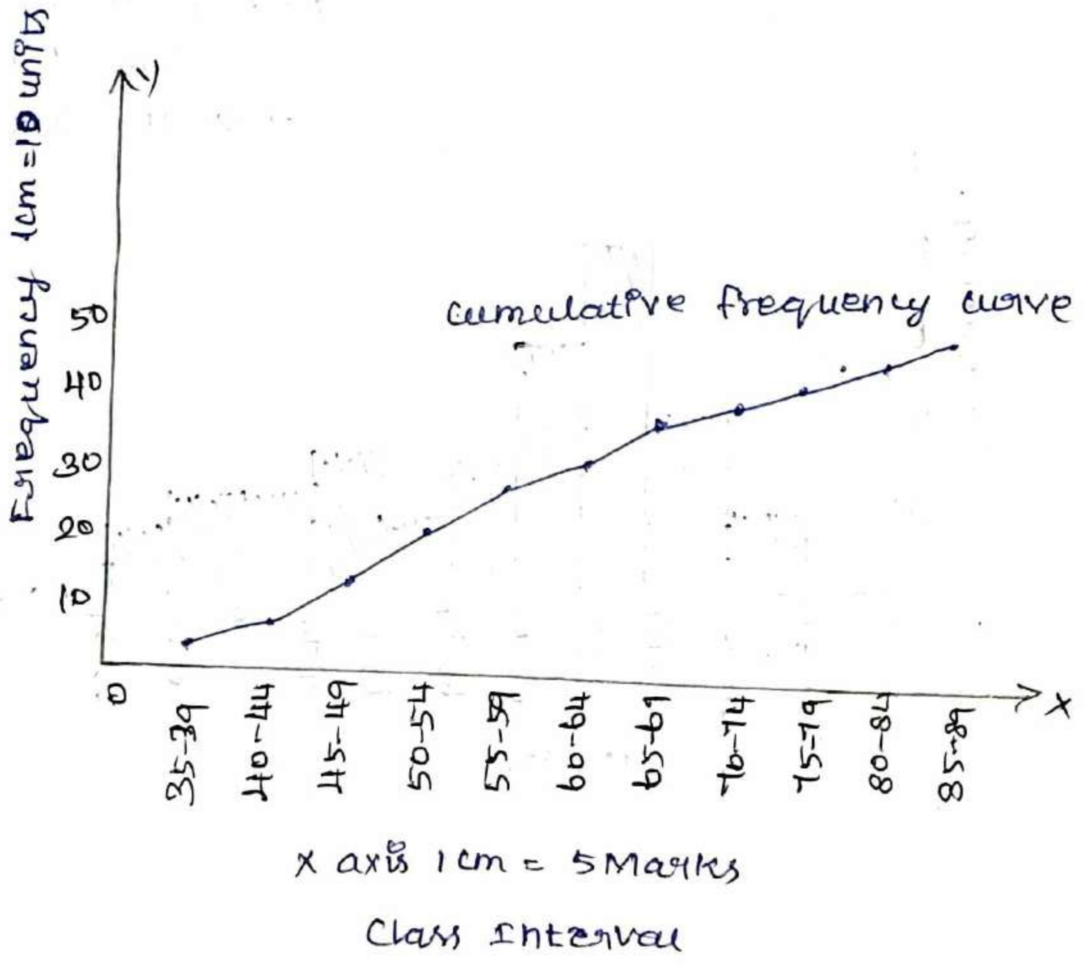


x axis Clas intervals
1cm = 5marks
Y axis = 2 cm.

2) Histogram showing marks of students



3) cumulative frequency curve showing marks of students.



Significance of graphic representation:

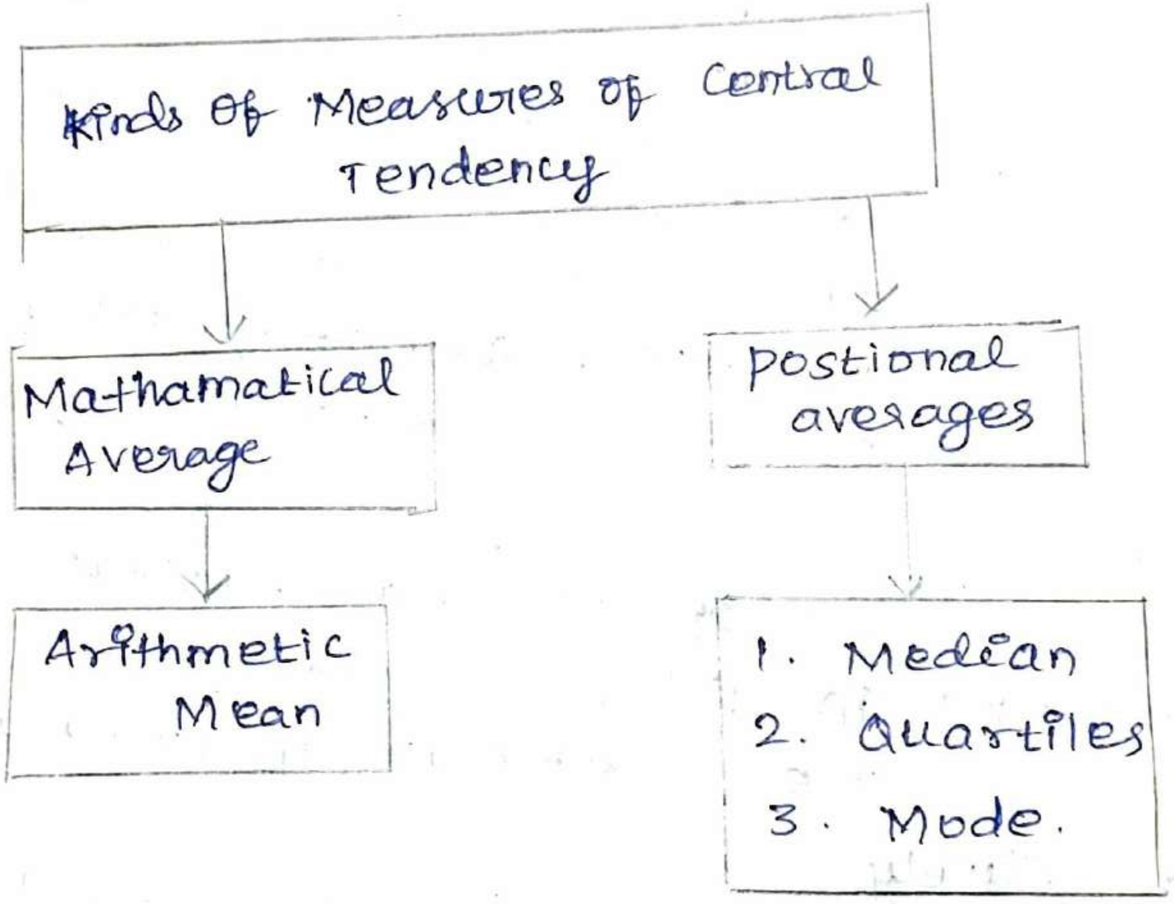
1. It is the simplest method of presentation of data.
2. They give clear cut and attractive view.
3. They make comparison of variables easy.
4. They are helpful in ascertaining certain stattical measures.
5. They save time and energy.

Limitations of graphics representation:

1. A graph simply shows tendency and fluctuations, and not the actual values.
2. Complete accuracy is not possible on a graph.
3. Graphs cannot be quoted in support of some statement.
4. only a few characteristics can be depicted on a graph. However in the case of many figures, it is difficult to follow the graph.

Meaning of central Tendency:

The measure of central tendency is defined as the statistical measure that identifies a single value as the representative of an entire distribution It aims to provide an accurate description of the entire data.



Arithmetic mean :

Arithmetic mean is the most commonly used measure of central tendency.

Arithmetic mean is computed by adding all the values in the set divided by the number of observations in it.

Individual series : (DIRECT METHOD)

If there are N observations as $x_1, x_2, x_3 \cdot x_n$ then the Arithmetic Mean (usually denoted by $\bar{x}$, which is read as X bar) in case of individual series using direct method is given by

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{N}$$

$$\bar{x} = \frac{\Sigma x}{N}$$

Individual series :- (DIRECT METHOD).

Question :

A student's marks in II Term exam were 94, 87, 98, 82. Find out average marks of the students.

Marks	Solution :-
94	$\bar{x} = \frac{\Sigma x}{N}$
87	
98	$= \frac{450}{5}$
82	
89	$\bar{x} = 90$ Marks
$\Sigma X = 450$	

Individual series (short cut Method)

This method is used when the size of items is very large.

Any of the middle values is taken as assumed average (A).

Deviation of values ~~is taken~~ of item is Calculated from A. $(d = X - A)$.

The deviations are then added (Σd) and then divided by the total number of observation. (N)

Add this value to Assumed average to get Mean.

$$\text{Mean} = A + \frac{\Sigma d}{N}$$

Individual Series (Assumed Mean Method)

Following is the pocket allowances of 5 students Find arithmetic mean using short cut Method.

Pocket Allowances (RS.)	15	20	40	65	30.

No. of. Students	Pocket Allowances (X)	$d = (X - a)$ $a = 40$
1.	15	$15 - 40 = -25$
2	20	$20 - 40 = -20$
3	40	$40 - 40 = 0$
4	65	$65 - 40 = 25$
5.	30	$30 - 40 = -10$.
N = 5.		$\Sigma d = -45 + 15$ $= 80$.

$\Sigma d / N = {-30}/5$

$= -6$

$\text{Mean } \bar{x} = A + \frac{\Sigma d}{N}$

$= 40 + (-6)$

$= 40 - 6$

$\bar{x} = 34$

∴ Arithmetic mean = RS = 34.

Discrete series (DIRECT METHOD).

Following steps are invoved in this method:

The values of all the items of a series are multiplied by their respective frequencies (fx).

These multiples are added up to get (ΣfX).

Find out total no. of items in the series (Σf)

Divide the total of value of all items (ΣfX) with no. of items (N). Thus:-

$$\bar{X} = \frac{\Sigma f(x)}{N \text{ (or) } \Sigma(f)}$$

Questions:

The following data gives the weekly wages in (RS) of 20 workers

Workers	5	2	6	4	3
Wages (Rs)	100	140	170	200	250

Wages (x)	Workers (F)	fx
100	5	500
140	2	280
170	6	1020
200	4	800
250	3	750
	$\Sigma f = 20$	$\Sigma(fx) = 3350$

Solution :

$$\bar{x} = \frac{\Sigma fx}{N}$$

$$= \frac{3350}{20}$$

$$= 167.50$$

$$\bar{X} = 167.50$$

Discrete series (Assumed Mean Method)

Any of the middle values is taken as assumed average (A).

Deviation of values of each item is calculated from A. $(d = x - A)$.

The deviations are multiplied by the respective frequencies (fd)

Add all the values to get (Σfd) and divide it by Σf.

Add this value to Assumed average to get Mean.

$$\text{Mean} = A + \frac{\Sigma fd}{N}$$

From the following data of the marks obtained by 60 students of a class find arithmetic mean using short cut method.

Marks	20	30	40	50	60	70
Students	8	12	20	10	6	4

Marks	Students	$d = (X - A)$ $a = 40$	fd
20	8	$(20-40) = -20$	-160
30	12	$(30-40) = -10$	-120
40	20	$(40-40) = 0$	0
50	10	$(50-40) = 10$	100
60	6	$(60-40) = 20$	120
70	4	$(70-40) = 30$	120
	$\Sigma f) = 60$		$\Sigma fd = 60$

$$\text{Mean} = A + \frac{\Sigma fd}{N}$$

$$= 40 + \frac{60}{60}$$

$$= 40 + 1$$

$= 41$ marks.

Continuous series : (direct Method)

The mid values of the class intervals are Calculated $m = \frac{L_1+L_2}{2}$

Mid values are multiplied by their corresponding frequencies (fm)

Σfm is divided by Σf. $\bar{x} = \frac{\Sigma fm}{\Sigma f}$.

Question :

Calculate mean marks for the students from the following information direct method.

Marks	0-10	10-20	20-30	30-40	40-50
Students	20	24	40	36	20

Marks	Mid. value (M)	Students (f)	fm
0-10	5	20	100
10-20	15	24	360
20-30	25	40	1000
30-40	35	36	1260
40-50	45	20	900
		Σ(f)=140	Σfm = 3620

$$\text{Mean} = \frac{\Sigma fm}{\Sigma f}$$

$$= \frac{3620}{140}$$

$$\bar{x} = 25.86 \text{ marks}$$

Continous Series : (Assumed Mean Method)

Mid values of Classes are determined

Deviations of the mid values from the assumed average (A) are determined by 'and indicated by 'd'.

Deviations (d) are multiplied by the frequency (f) to get 'fd'. They are added up to get Σfd.

Σfd is divided by Σf

Add this value to assumed average to get Mean $A + \frac{\Sigma fd}{\Sigma f}$

Solution :-

Marks	Mid value (m)	d = x - a
		a = 25
			Students	fd

0-10	5	-20	20	-400
10-20	15	-10	24	-240
20-30	25	0	40	·0
30-40	35	10	36	360
40-50	45	20	20	400
		$\Sigma(f)$	= 140	Σfd = 120

$$\text{Mean} = A + \frac{\Sigma fd}{\Sigma f}$$

$$= 25 + \frac{120}{140}$$

$$= 25 + 0.86$$

$$\text{Mean } \overline{x} = 25.86 \text{ marks}$$

Properties of Arthmetic mean :

The sum of the deviations, of all the values of x, from their arithmetic mean, is zero.

The product of the arithmetic mean and the number of items gives the total of all items

The sum of the squares of the deviations of the items taken from arithmetic mean is minimum

If a constant is added or subtracted to all the variables, mean increases or decreases by that constant.

If all the variables are multiplied or devided by a constant mean also gets multiplied or devided by the constant.

MEDIAN :

Median is an average.

It is a measure of central value

Median is the middle value of a data when the values are arranged in the ascending or descending order.

Medians divides a distribution into two equal halves. There will be equal number of items above and below the items.

Median is represented by the symbol md.

Median can be calculated for ungrouped data and grouped data.

Calculation of median in individual series involves the following steps:

Arrange all the values of different items of a series in the ascending or discending order.

Add up the no. of items indicated by N.

Mediam $= \left(\frac{N+1}{2}\right)$th item.

Calculation :- Even number 4, 10, 12, 18

~~N 4, 10, 12, 18~~ $x = 4, 10, 12, 18$.

$$\text{Median} = \frac{n+1}{2}$$

$$= \frac{4+1}{2} = \frac{5}{2} = 2.5^{th} \text{ item.}$$

$$\text{Median} = \frac{\text{Size of 2nd item + Size of 3 item}}{2}$$

$$= \frac{10+12}{2} = \frac{22}{2} = 11.$$

INDIVIDUAL SERIES

1. Marks : 10, 11, 15, 17, 20, 21, 32, 32, 33, 35, 41.

Ascending order

S.No	Marks
1	10
2	11
3	15
4	17
5	20
6	21
7	32
8	32
9	33
10	35
11	41
N = 11	

$M = \text{size of } \left(\frac{N+1}{2}\right)\text{th item}$

$= \text{size of } \left(\frac{11+1}{2}\right)^{\text{th}} \text{item}$

$= \frac{12}{2}$

$= 6^{\text{th}}$ item.

Medium = 21.

2. Marks :- 10, 11, 15, 17, 21, 32, 32, 33, 35, 41

S.No	Marks
1	10
2	11
3	15
4	17
5	21
6	32
7	32
8	33
9	35
10	41
N = 10	

$\text{Median} = \text{size of } \left(\frac{N+1}{2}\right)$

$= \left(\frac{10+1}{2}\right)$

$= \frac{11}{2}$

$M = 5.5^{\text{th}}$ item

$$M = \frac{5^{\text{th}} \text{ value} + 6^{\text{th}} \text{ value}}{2}$$

$= \frac{21+32}{2}$

$= 26.5$ marks.

DISCRETE SERIES

Question: Calculate Median from the following data

Marks	45	55	25	35	5	15	
No. of Students	40	30	30	50	10	20	

Solu:

Marks in an ascending order	No. of students (f)	Cumulative frequency (C.f)
5	10	10
15	20	30
25	30	60
35	50	110
45	40	150
55	30	180

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

$= \frac{180+1}{2}$

$= \frac{181}{2} = 90.5^{th}$ item

Cumulative frequency includes 90.5th item = 110

Median = size of item corresponding to 110 = 35ma

Median = 35 Marks.

Continuous series :-

The steps involved in the calculation of median are as follows :-

Step 1 : Calculate cumlative Frequencies.

Step 2 : Ascertain $N/2$th item.

Step 3 : Find out the cumlative frequency which includes $N/2$th item and corresponding class frequency. The corresponding class of this cumulative frequency is called the median class.

Formula :-

$$M = l_1 + \frac{N/2 - c.f}{f} \times i.$$

l_1 = lower limit of the median class

c.f = cumulative frequency of the class preceeding the median class, f = frequency of the median.

Weekly Expenditure	(f) No. of families	Cumlative f (c.f)
0-10	14	14
10-20	23	37
20-30	27	64
30-40	21	85
40-50	15	100

Ascertain $N/2$ th item = $100/2$ th item = 50th item lies in class interval as 20-30

Thus the median class is 20-30 Now applying formula of median

$$(M)\ \text{Median} = l + \frac{N/2 - c.f}{f} \times i$$

$$= 20 + \frac{50 - \cancel{37}64}{27} \times 10$$

$$= 20 + \frac{6 - 14}{27} \times 10$$

$$= 20 + 0.518 \times 10$$

~~= B/N'~~

$$= 20 + 5.18$$

$$\boxed{\text{Median.} = 25.18}$$

Mode :-

Mode is the value of the variable occurs most frequently in a distribution.

The value which occurs many times in the table is the mode.

It is represented by the letter Mo.

Mode is an average. It is a positional average a measure of central value.

When a data has one Concentration of frequency, it is called unimodal.

It has two concentration, it is called bimodal.

It has 3 concentration of frequency it is called trimodal.

Mode can be calculated for ungrouped data and grouped data.

To find out mode of an ungrouped data the values are arranged in an ascending order. The value which occurs maximum number of items is the mode.

Ex. 18, 21, 23, 23, 25, 25, 25, 27, 29, 29.

In the above data, 25 occurs maximum number of times. So 25 is the mode.

Calculation :-

From the following data of the hight of 100 plants in a garden determine the modal hight

Height (X)	58	60	61	62	63	64	65	66	68	70
plants (No.) (f)	4	6	5	10	20	22	24	6	2	1

soln:-

By inspection we can clearly say that the modal height is 65 cm, because the value 65 is repeated 24 times.

But in this problem, the difference between the maximum frequency and the next frequenly is very small is 24 - 22 = 2.

So prepare grouping table and analysis table. (In the case of biomodal series or trimodal series, we must prepare first grouping table and analysis table).

Steps for preparing the grouping and analysis table :-

1. Prepare a grouping table with 6 columns
2. Write the size of item in the items margin.

Grouping table

Height in Cm.	Frequencies					
	Column 1	Colu 2	Colu 3	Clau 4	Colu 5	Colu 6
58	4	10		15		
60	6		11		21	
61	5	8				35
62	10	15	30	52		
63	20	42			66	
64	22		46			52
65	24	30		32		
66	6		8		9	
68	2	3				
70	1					

In the column 4, the frequencies are grouped in threes (1, 2 and 3 / 4, 5, 6 / 7, 8, 9 and so on.

In the column 5, the frequencies are grouped in threes leaving the first frequencies are grouped in (2, 3 and 4 / 4, 5; 6, 7, / 8, 9, 10 so on).

In the Column 6, the frequencies are grouped in threes leaving the first two frequencies (3, 4 and 5/ 6,7,8 / 9,10,11 and son on).

In all the processes mark down, the maximum frequencies by a circle.

Then an analysis is table is prepared to show the exact size, which has the highest frequency.

Analysis table.

Col. No	Height in cm									
	58	60	61	62	63	64	65	66	68	70
1							1			
2					1	1				
3						1	1			
4				1	1	1				
5					1	1	1			
6						1	1	1		
Total				1	8	5	4	1		

Since the value 64. has occured the maximum number of times. i.e, 5 times.

∴ The modal heights is 64 cm.

Calculation of Mode : Continuous series

The highest frequency can be find out by inspection.

In the case of bimodal series or trimodal series we prepare, Grouping and analysis table and then find out highest frequency.

Apply the formula.

$$\text{Mode} = L + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2}\right] \times C.$$

Δ_2 = the difference between the frequency of the modal class and the succeeding modal class $(f_1 - f_2)$.

C = Class interval of the modal class

f_1 = frequency of the modal class

f_0 = frequency of the preceding modal class

f_2 = Frequency of the succeeding modal class.

Marks	No. of Students
0-10	4
40-20	9
20-30	13 f_0
30-40	15 f_1
40-50	12 f_2
50-60	8
60-70	3

find out the highest frequency in (f_1).

Highest frequency (f_1) is 15

The corresponding class is 30-40. It is the modal class ∴ 30 is the lower limit of the modal class.

f_0 is the frequency preceding the modal class =13.

f_2 is frequency the succeeding modal class =12.

C is the class interval =10.

formula

$$\text{Mode} = L + \left[\frac{\Delta_1}{\Delta_1+\Delta_2}\right] \times C$$

$\Delta_1 = f_1 - f_0 \Rightarrow 15-13 = 2$

$\Delta_2 = f_1 - f_2 \Rightarrow 15-12 = 3$

$$\text{Mode} = 30 + \left[\frac{2}{2+3}\right] \times 10$$

$$= 30 + \frac{2}{5} \times 10$$

$$= 30 + (0.4 \times 10)$$

$$= 30 + 4$$

Mode. = 34

1 i) Merits of Mean :-

It can be easily calculated

Its calculation is based on all the observations.

It is easy to understand

It is rigidly defened by the mathematical formula.

It is least affected by sampling fluctuations.

It is the best measure to compare two or more series of data.

It does not depend upon any position.

ii) Demerits of Mean :-

It may not be represented in actual data so it is theoretical.

It is affected by extreme values.

It can not be calculated if all the observations are not known.

It can not be used for qualitative data (i.e.) love, beauty, honesty, etc.

It may lead to fallacious conditions in the absence of original observations.

1iii) Uses of mean :

It is extremely used in medical statistics

Estimates are always obtained by mean.

2)i) Merits of Median :-

Simple to calculate

It can be calculated without knowing the values of all the items.

It is un affected by extreme value

It can be calculated graphically.

Not affected by extreme observations

Both for quantitative and qualitative data

ii) Demerits of Median:-

Affected more by sampling fluctuations

Not rigidly defind

Can be used for further mathematical calculations.

It is not based on all the items

It is not used as a common average

It is not used for further statistical calculation.

3) i) Merits of Mode :

Mode is readily comprehensible and easy to calculate.

Mode is not at all affacted by extreme values.

Mode can be conveniently located even if the frequency distribution has class - intervals of unequal magnitude provided the modal class and the classes preceding and succeeding it are of the same magnitude.

Demerits of mode :-

Mode is ill -deffined.

Nort always possible to find a clearly defind mode.

It is based upon all the observations,

It is not amenable to further mathematical treatment

As compared with mean, mode is affected to a great extent by fluctuations of sampling.

When data sets contain two, three or many modes, they are diffcult to interpret and compare.

Uses of Mode :-

Mode is useful for qualitative data.

Dispersion :-

Meaning of dispersion :-

In Statistics, dispersion is used commonly to mean scatter, deviation fluctuation spread or variability of data.

Dispersion is used to denote a lack of uniformity in item of a given variable.

Definition :

The degree to which the individual values of the variate scatter away from the average is called dispersion.

1. Dispersion or spread is the degree of the scatter or variation of the variable about a central value.

   -Brooks and Dick.

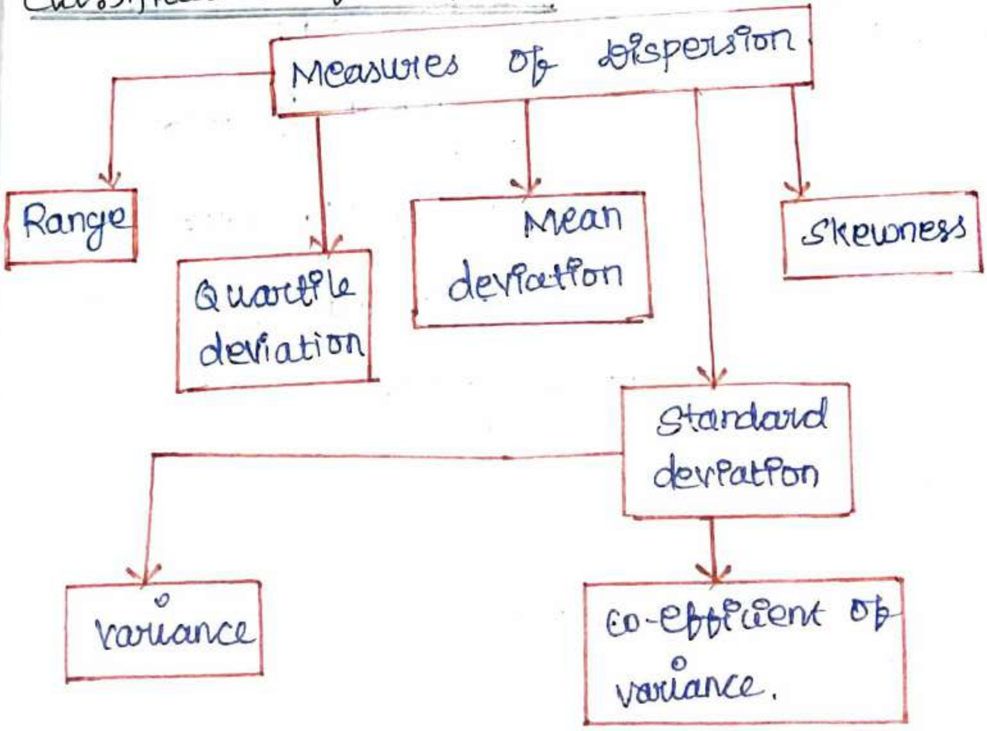
2. "Dispersion" is the measure of the variation of the items.

   - A.L. Bowley.

3. The degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data.

   - Spiegal.

Classification of dispersion



Objects of Measuring Dispersion :-

The major objects of measuring dispersion are as follows :-

1. To ~~dem~~ determine the reliability of an average.

2. To serve as a basis for the control of variability

3. To compare two or more series with regard to their variability.

4. ~~It~~ ~~show~~ To facility the use of other statistical.

Properties of a good Measure of Dispersion.

To properties of a good Measure of dispersion are as follows.

1. It should be rigidly defined

2. It should be based on all the observations of the series.

3. It should be capable of futher. algebraic treatment

4. It should be easy to calculate and simple to follows.

5. It should not be affacted by fluctuations of by sampling.

Range :

Range is the simplest possible measure of dispersion. It is a rough measure of dispersion because. it is affacted by extreme values.

definition

It is defined as the difference between the highest (largest) and the lowest (smallest) value of variable in series.

Rang can be calculated by the formulla

Rang (R) = Largest value (L) - Smallest value (S).

Coefficient of

= Ratio of the range = Relative measure of range

$$\text{Co-efficient of range} = \frac{\text{Largest value} - \text{Smallest value}}{\text{Largest value} + \text{Smallest value}}.$$

$$\text{Co-efficient of range} = \frac{L-S}{L+S}$$

Range is a crude measure of variability, as it is affacted by extreme values. But co-efficient of range is a better measure and it is used to compare the range of two series.

Location of Range in individual series :-

Illustration : Find out the range and coefficient of range of the following data.

60, 65, 50, 90, 70, 40, 110, 130, 120, 100.

Solu:-

The largest (highest) value = 130

The smallest (lowest) value = 40

$$\text{Rang } (R) = L-S$$
$$= 130-40$$
$$R. = 90$$

$$\text{Co-efficient of Range} = \frac{L-S}{L+S}$$
$$= \frac{130-40}{130+40}$$
$$= \frac{90}{170}$$

Co-efficient of Range = 0.5.

Location of Range in Discrete series:

Find out the rang and co-efficient of range of the following distribution

x	5	6	7	9	12	13	14	15
f	10	13	22	30	20	18	15	11

Solution:

Maximum value (L) = 15

Minimum value (S) = 5

$R = L - S$

$= 15 - 5$

Range = 10

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

$= \frac{15-5}{15+5} = \frac{10}{20}$

Co-efficient of Range = 0.5

Location of Range in class interval series

Size	0-10	10-20	20-30	30-40	40-50
Frequency	3	7	10	6	4

Size	Midvalue	Frequency
0-10	5	3
10-20	10	7
20-30	15	10
30-40	35	6
40-50	45	4

Range = L-S

$= 45 - 5$

R = 40

Co-e R $= \frac{L-S}{L+S}$

$= \frac{45-5}{45+5}$

$= \frac{40}{50}$

Coe. Range = 0.8.

Merits of Range

1. The range is the easiest to calculate and simplest to understand.

2. It gives a quick answer.

Demerits of range :

It gives a rough answer.

It is not ~~on~~ based on all observations

It changes from one sample to the next in a population.

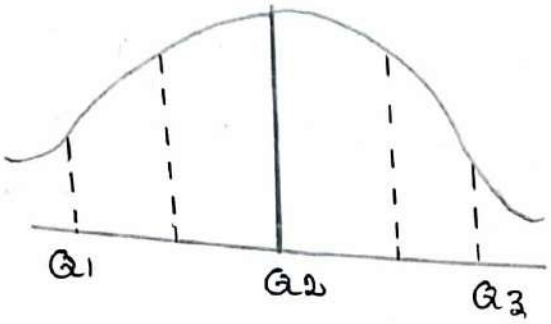
It cannot be calculated in open-end distributions.

It is affected by sampling fluctuations.

Quartile Deviation:

The half distance between 75 th percentile i.e., 3 rd quartile (Q3) and 25 th percentile i.e., 1st quantile (Q1) is called quartile deviation or Semi - interquartile range.

In normal distribution Semi - interquartile range is called probable error (PE).



The Figure shows that Q1, Q2 and Q3 divide the range of a variable into 4 parts.

Q1 divides the range of a variable into 25 % and 75 % observation. Q2

Q2 divides the range of variable into 50 % and 50 % observation and Q3 divides the range of variable into 75 % and 25 % observations.

$$\text{Formula: } \frac{(Q_3 - Q_2) + (Q_2 - Q_3)}{2} = \frac{Q_3 - Q_1}{2}$$

The Semi - interquartile range takes into account only the middle half of the data between Q3 and Q1. If Q is larger there is greater scatter in the inter - quartile range.

If Q is smaller there is greater concentration in the middle.

If the distribution is symmetrical, $Q_3 - Q_2 = Q_2 - Q_1$.

25%. of the observations above and below lie equidistant from the median.

Semi - inter quartile deviation or quartile deviation for ungrouped data:

example: The range of a variable such as height between first quartile - Q_1 and third Quartile - Q_3 is 166.84 cm and 174.93 cm respectively. Second quartile - Q_2 viz, median is 171.46 cm. Find out the Semi - interquartile deviation.

solution : $Q = \frac{Q_3 - Q_1}{2}$

$\therefore Q = \frac{170.90 - 160.80}{2}$

$= \frac{10.1}{2} = \underline{\underline{5.05}}$ cm.

Quartile deviation for grouped data:

To obtain quartile deviation from grouped data (discrete series) one has to obtain Q_1 and Q_2 first. Formula to calculate Q_1 and Q_3 is given below

$$Q_1 = L + \frac{\left(\frac{N}{4} - F\right)}{fq} \times i \; ; \; Q_3 = L + \frac{\left(\frac{3N}{4} - F\right)}{fq} \times i$$

Here. L = Lower limit of that class interval, where $Q_1 \left(\frac{N}{4}\right)$ (or) $Q_3 \left(\frac{3N}{4}\right)$ falls.

F = Cumulative frequency just above that class interval where

$Q_1\left(\frac{}{4}\right)$ or $Q_3\left(\frac{}{4}\right)$

fq = frequency of that class interval where Q_1 and Q_2 falls

i = length of class interval.

example: Length of earthworms in cm and their frequency was studied and given in the form of following table. Find, the quartile deviation.

L. of earthworms in class interval	16-20	21-25	26-30	31-35	36-40	41-45	46-50	51-55	56-60	61-65
Frequency	4	3	8	9	14	3	3	2	2	2

solution:

class interval	16-20	21-25	26-30	31-35	36-40	41-45	46-50	51-55	56-60	61-65
Frequency	4	3	8	9	14	3	3	2	2	2
Cum. frequency	4	7	15	24	38	41	44	46	48	50
			$\boxed{Q_1}$		$\boxed{Q_3}$					

To find out Q we have to calculate Q_1 & Q_3.

Now calculate $\frac{N}{4}$ and $\frac{3N}{4}$

Here $N = 50$ and therefore, $\frac{N}{4} = \frac{50}{4} = 12.5$

$$\frac{3N}{4} = \frac{3 \times 50}{4} = 37.5$$

Hence Q_1 will fall in 3rd class interval i.e., in 26-30.

Hence Q_3 will fall in 5th class interval i.e., in 36-40.

Now $\therefore Q_1 = L + \frac{\left(\frac{N}{4} - F\right)}{fq} \times i$

$$\therefore Q = 25.5 + \frac{12.5 - 7}{8} \times 5$$

$$= 25.5 + \frac{5.5}{8} \times 5$$

$$= 25.5 + (0.687) \times 5$$

$$= 25.5 + 3.437 = \underline{\underline{28.93}}$$

$$\therefore Q_3 = L \frac{\left(\frac{3N}{4} - F\right)}{fq} \times i$$

$$Q_3 = 35.5 + \frac{37.5 - 24}{14} \times 5$$

$$= 35.5 + \frac{13.5}{14} \times 5$$

$$= 35.5 + (0.964) \times 5$$

$$= 35.5 + 4.821 = \underline{\underline{40.32}}$$

Since $Q = \frac{Q_3 - Q_1}{2}$

$$Q = \frac{40.32 - 2893}{2}$$

$$= \frac{11.39}{2} = \underline{\underline{5.69}}$$

Coefficient of quartile deviation:

It is calculated by this formula $\Big\}$: Coefficient of $Q = \frac{Q_3 - Q_1}{Q_3 + Q_1}$

Merits Of quartile deviation:

i) The quartile deviation is better measure of dispersion as it is not based on two extreme values like range but rather on middle 50 % observations.

ii) It is the only measure of dispersion which can be used for open end distribution.

Mean Deviation

Definition :-

The mean deviation is the average of the absolute values of the deviation from the mean (or median or mode).

Let us understand the fact that each variate of the variable deviates from the mean. The scatter above or below the mean (unless it happens to concide with the mean where deviation is zero) is regarded as deviation. Deviations above the mean are negative deviation and below the mean are positive deviation.

Computation of mean deviation (denoted as δ) for ungrouped data (Individual series).

$$\text{Mean Deviation or M.D or } \delta = \frac{\Sigma|x|}{N}$$

(for ungrouped data).

Here,

M.D or δ = Mean deviation

x = deviation from actual mean

Σx = Sum of all deviations of distribution.

$||$ = Not considering sign (+ve or -ve) while summation is done.

Following formula is applied to obtain deviation i.e x

$$\text{Deviation} = \text{Score} - \text{Mean} \ (\text{or}) \ x = X - \overline{X}$$

[X is $\overset{s}{\text{ued}}$ for score or variable where as x is used for deviation].

Porocedure:

Calculate mean

Calculate deviation from mean ignoring the sign +ve or -ve.

N is the total no. of items. Divide Σx by the N.

Calculation:-

1. Egg laid by a species of brids were counted as 5, 7, 8, 10, 14, 12, 13, 5, 8, 8. Compute the Mean Deviation of the distribution. [Hypothetical data].

$$\text{Mean} \cdot \bar{x} = \frac{\Sigma x}{N}$$

$$\bar{x} = \frac{5+7+8+10+14+12+13+5+8+8}{10}$$

$$= \frac{90}{10} = 9$$

$$\boxed{\text{Mean } \bar{x} = 9}$$

Now following table 1 of 3 columns is prepared to get values of deviations. 1st coloumn for scrose, 2nd Coloumn for score - mean $(x - \bar{x})$ and 3rd for deviationx.

Table :-

Score	5	7	8	10	14	12	13	5	8	8
Score-Mean	5-9	7-9	8-9	10-9	14-9	12-9	13-9	5-9	8-9	8-9
Deviation.	-4	-2	-1	+1	+5	+3	+4	-1	-1	

Now sum up all deviations ignoring their +ve or -ve signs.

$$\Sigma x = 4+2+1+1+5+3+4+1+1 = 26$$

$$\because \text{M.D (or) } \delta = \frac{\Sigma |x|}{N} = \frac{26}{10} = 2.6.$$

Computation of Mean deviation for grouped [illegible]
(Discrete series)

Following formula is applied.

$$\delta = \frac{\Sigma|fx|}{\Sigma f}$$

Here, Σfx = is the sum of multiplication of frequency and deviation of each variable.

Σf = sum of all frequencies

$||$ = not considering + ve or -ve signs.

Following steps are taken to find out Mean deviation.

1. Calculate the mean.
2. Calculate the deviation from the mean ignoring the signs and denote them by small x.
3. Multiply these deviation by respective frequencies and obtain the data $\Sigma f.x$.
4. Divide the total Σfx by Σf.

Length of 50 butterflies of a species and their frequency was observed in a survey. Find out Mean Deviation of this distribution.

Length in cm	2	2.5	2.7	2.9	3	3.1	3.3	3.7	3.9	4	4.6	4.8	4.9	5	5.5	5.9	6	6.1	6.7	6.9
Frequency	2	1	1	2	3	1	3	2	4	3	2	3	3	3	2	3	3	3	3	3.

Solution

First of all make a table of 5 columns. First column for variable, 2nd column for frequency 3rd column for frequency x variable (fx). 4th for deviation (x) and fifth for frequency x Deviation (fx)

VARIABLE (X)	Frequency (f)	Variable × Freq. (f.X)	Deviation x	Frq × deviation $f.x$.
2	2	4	-2.62	5.24
2.5	1	2.5	-2.12	2.12
2.7	1	2.7	-1.92	1.92
2.9	2	5.8	-1.72	3.44
3	3	9	-1.62	4.48
3.1	1	3.1	-1.52	1.52
3.3	3	9.9	-1.32	3.96
3.7	2	7.4	-0.92	1.84
3.9	4	15.6	-0.72	2.88
4.0	3	12.6	-0.62	1.86
4.6	2	9.2.0	-0.02	0.04
4.8	3	14.4	0.18	0.54
4.9	3	14.7	0.28	0.84
5.0	3	15.0	0.38	1.14
5.5	2	11.0	0.88	1.78
5.9	3	17.7	0.28	3.84
6.0	3	18.0	1.32	4.12
6.1	3	18.3	1.48	4.26
6.7	3	20.1	2.08	6.32
6.9	3	20.7	2.28	6.43
	50	231.1		59.46

$$\text{Mean } \bar{x} = \frac{\Sigma f x}{\Sigma f} = \frac{231.1}{50} = 4.62$$

$$\delta = \frac{\Sigma |f x|}{\Sigma f} = \frac{59.46}{50} = 1.18.$$

Computation of Mean deviation for grouped data ? (continuous series).

In this case same above formula(as in the case of discrete series) is applied but mid points of each class interval is obtained.

Find out the mean deviation from the given data in continuous series :-

Class Interval	2 - 2.9	3-3.9	4-4.9	5-5.9	6-6.9
Frequency	6	13	11	8	12.

Solution:-

First of all prepare a table of six columns. 1st column for class interval. 2nd for mid points of each class interval. 3rd for frequency of each class interval. 4th for multiplication of mid points with their respective frequency. 5th column is for deviation of each variable. 6th column multiplication of frequency with their deviation

Class Interval	Mid points (X)	frequency (f)	frequ x Mid.point fx	$x = X - \overline{X}$ Deviation x	Fre x deviation fx
2-2.9	2.45	6	14.7	-2.14	-12.84
3-3.9	3.45	13	44.85	-1.14	-14.82
4-4.9	4.45	11	48.95	0.86	6.88
5-5.9	5.45	8	43.6	0.86	6.88
6-6.9	6.45	12	77.4	1.86	22.32.
	$\Sigma X = 22.25$	$\Sigma f = 50$	$\Sigma fX = 229.5$		58.4

$\overline{x} = \Sigma f$

$$\text{Mean } \bar{x} = \frac{\Sigma fX}{\Sigma f}$$

$$= \frac{229.5}{50}$$

$$= 4.59.$$

$$\text{M.D } \delta = \frac{\Sigma |fx|}{\Sigma f} = \frac{58.4}{50} = 1.168.$$

Mean deviation = 1.168

Coefficient of Mean deviation

Coefficient of mean deviation denoted by c of $\delta \bar{x}$ is obtained by formula.

$$c \text{ of } \delta \bar{x} = \frac{\delta \bar{x}}{\bar{x}}$$

$$= \frac{1.168 \times 4.59}{4.59.}$$

$$= \frac{34.088}{4.59} \doteq 7.42$$

$$c \text{ of } \delta \bar{x} = 7.42.$$

Merits of Mean deviation :- Mean deviation is easy to Calculate but since mean deviation has less mathematical value.

It is simple to understand and easy to compute.

It is based on each and every item of the data

M.D is less affacted by the values of extreme items than the standard deviation.

Demerits of Mean Deviation :-

The greatest drawback of this method is that algebraic signs are ignared while taking the deviations of the item.

It is not capable of further algebraic treatment

It is much less popular as compared to standard deviation.

STANDARD DEVIATION (S.D).

Introduction :-

* Karl pearson Introduced the concept of standard deviation in 1893. It is the most important measure of dispersion and is widely used in many statistical formulae. It is also called rootmean sequare deviation or mean error or mean square error. It provides accurate results.

Definition :-

It is defined as the positive square-root of the arithmetic mean of the squares of the deviations of the given observation from their arithmetic mean. It is represented by the Greek letter σ (small sigma).

A Calculation of S.D Individual observation :-

There are two methods.

1. Direct Method : Deviation taken from actual mean.
2. Short-Cut Method : Deviation taken from assumed mean.

1. Direct Method

STEPS 1 : Find out the actual mean of the series ($\bar{x}$).

2 : Find out the deviation of each value from the mean ($x-\bar{x}$).

3 : Square the deviation of each value and take the total of squared deviations $\Sigma(x-\bar{x})^2$

4 : Divide the total by the number of observations $\frac{\Sigma(X-\bar{X})^2}{N}$.

5. Find out the square root of the product.

$$\text{Formula } S.D = \sqrt{\frac{\Sigma(x-\bar{x})^2}{N}} \text{ (or) } \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

1. Illustration :

The following data are the height of 10 students Calculate standard deviation

60, 60, 61, 62, 63, 63, 63, 64, 70.

Solution :-

STEP 1. Find out mean $(\bar{x})$

2. Find out deviation of each value from the mean $(x-\bar{x})$.

3. Square the deviation of each value and take the total of squared deviations $(x-\bar{x})^2$ and then $\Sigma(x-\bar{x})^2$

Height (x)	$\bar{x}=63$ $(x-\bar{x})$	$(x-\bar{x})^2$
60	$(60-63) = -3$	9
60	$(60-63) = -3$	9
61	$(61-63) = -2$	4
62	$(60-62) = -1$	1
63	$(63-63) = 0$	0
63	$(63-63) = 0$	0
63	$(63-63) = 0$	0
64	$(64-63) = 1$	1
64	$(64-63) = 1$	1
70.	$(70-63) = 7$	49
$\Sigma x = 630$		$\Sigma(x-\bar{x})^2 = 74$

$$\bar{x} = \frac{\Sigma x}{N}$$

$$= \frac{630}{10}$$

$$\bar{x} = 63$$

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

Because the observation is less then 30

$$= \sqrt{\frac{74}{9}}$$

$$= \sqrt{8.22}$$

$$= 2.86$$

$\bar{x} =$

$S.D = 2.86$

2. Short - cut Method :

STEP 1 : Instead of finding actual mean, we take assumed mean. So the calculation is simplified The processes are the same.

Illustration:-

The following data are the height of 10 students Calculate S.D. 60, 60, 61, 62, 63, 63, 63, 64, 64, 70.

Solution:-

STEP 1 : Assume any one of the values in the data as assumed mean (A).

Find out deviation of each value from the assumed mean (x - A = d).

Square the deviation of each value and take the total of squared deviations (d^2).

Apply the formula.

$$S.D = \sqrt{\frac{\Sigma d^2}{N} - \left[\frac{\Sigma d}{N}\right]^2}$$

Height (x)	A = 62 x - A = d	d^2
60	(60 - 62) = -2	4
60	(60 - 62) = -2	4
61	(61 - 62) = -1	1
62	(62 - 62) = 0	0
63	(63 - 62) = 1	1
63	(63 - 62) = 1	1
63	(63 - 62) = 1	1
64	(64 - 62) = 2	4
64	(64 - 62) = 2	4
70	(70 - 62) = 8	64
	Σd = 10	$\Sigma d^2 = 84$

$$S.D = \sqrt{\frac{\Sigma d2}{N} - \left[\frac{\Sigma d}{N}\right]^2}$$

$$= \sqrt{\frac{84}{10} - \left(\frac{10}{10}\right)^2}$$

$$= \sqrt{8.4 - (1)^2}$$

$$= \sqrt{8.4 - 1}$$

$$= \sqrt{7.4}$$

S.D = 2.72

B. Calculation of S.D Discrete series :-

1. Direct Method - Actual Mean Method

2. Short - cut Method - Assumed Mean Method.

1. Direct Method : Actual Mean Method

STEPS 1 : Calculate the mean

2 : Find out deviation of the various values from the mean value $(x-\bar{x})$.

3 : Square the deviations $(x-\bar{x})^2$

4 : Multiply $(x-\bar{x})^2$ with the respective frequencies (f) against various values and add all such values $\Sigma f(x-\bar{x})^2$

5 : Divide $\Sigma f(x-\bar{x})^2$ by the number of items N or Σf Apply the formula. $S.D = \sqrt{\dfrac{\Sigma f(x-\bar{x})^2}{N}}$

x	6	7	8	9	10	11	12
y//	3	6	9	13	8	5	4

Solution :

x	f	fx	$\bar{x}=9$ $(x-\bar{x})$	$(x-\bar{x})^2$	$f(x-\bar{x})^2$
6	3	18	-3	9	27
7	6	42	-2	4	24
8	9	72	-1	1	9
9	13	117	0	0	0
10	8	80	1	1	8
11	5	55	2	4	20
12	4	48	3	9	36
	$\Sigma f=48$	$\Sigma fx=$ 432			$\Sigma f(x-\bar{x})^2=$ 124

$$\bar{x} = \frac{\Sigma fx}{N}$$

$$= \frac{432}{48}$$

$$\bar{x} = 9.$$

$$S.D = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{N}} = \sqrt{\frac{124}{48}}$$

$$= \sqrt{2.58}$$

$$\boxed{S.D = 1.6}$$

2. Short-cut Method - Assumed Mean Method:

STEPS 1: Assume any one of the values in the data an assumed mean(A)

2: Find out deviations of each value from the assumed mean $(x-A)=d$.

3: Square the deviation d^2

4: Multiply d^2 with the respective frequencies (f) against various values and add all such values Σfd^2

5: Apply the formula.

$$S.D = \sqrt{\frac{\Sigma fd^2}{N} - \left[\frac{\Sigma fd}{N}\right]^2}$$

1. Find S.D for the following data?

x	18	19	20	21	22	23	24	25	26	27
f	3	7	11	14	18	17	13	8	5	4

x	f	A = 22 x − A = d	fd	d^2	fd2
18	3	−4	−12	16	48
19	7	−3	−21	9	63
20	11	−2	−22	4	44
21	14	−1	−14	1	14
22	18	0	0	0	0
23	17	1	17	1	17
24	13	2	26	4	52
25	8	3	24	9	72
26	5	4	20	16	80
27	4	5	20	25	100
	$N = \Sigma f$ $= 100$		$\Sigma fd =$ 38		$\Sigma fd^2 =$ 490

$$S.D = \sqrt{\frac{\Sigma fd}{N} - \left[\frac{\Sigma fd}{N}\right]^2}$$

$$= \sqrt{\frac{490}{100} - \left[\frac{38}{100}\right]^2}$$

$$= \sqrt{4.9 - 0.144}$$

$$= \sqrt{4.756}$$

$S.D. = 2.2$

C. Calculation of S.D - Continous series.

There are 3 methods

1. Direct Method - Actual mean Method

2. short cut Method - Assumed mean Method

3 step deviation Method

Formula for the Direct Method :

$$S.D = \sqrt{\frac{\Sigma f(\text{mid } x - \bar{x})^2}{N}}$$

Formula for the short-cut Method:

$$S.D = \sqrt{\frac{\Sigma f(\text{mid}x^2)}{N} - \left[\frac{\Sigma f \text{ mid}x}{N}\right]^2}$$

Formula is $S.D = \sqrt{\frac{\Sigma f d'^2}{N} - \left[\frac{\Sigma f d'}{N}\right]^2} \times c.$

$$d = \frac{\text{Mid } x - A}{c}, \quad c = \text{common factors.}$$

STEPS 1 : Find out the mid - value of each class.

Assume one of the mid - values as an assumed mean (A)

Find out deviation of each mid-value from the assumed mean ($\text{mid } x - A = d$)

Divide each deviation by a common factor $\frac{d}{c} = d'$

Square the deviations (d'^2)

Multiply these deviation d'^2 by the respective frequencies and add all such values Σfd^2

Apply the formula $SD = \sqrt{\frac{\Sigma fd'^2}{N} - \left[\frac{\Sigma fd'}{N}\right]^2} \times C$

Illustration :

Compute the SD for the following data?

X	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Y	5	10	20	40	30	20	10	4

Solution :-

class(x)	mid x	frequen (y)	A = 35 mid x−A = d	d/c = d' C = 10	d'^2	fd'	fd'^2
0-10	5	5	-30	-3	9	-15	45
10-20	15	10	-20	-2	4	-30	40
20-30	25	20	-10	-1	1	-20	20
30-40	35	40	0	0	0	0	0
40-50	45	30	10	1	1	30	30
50-60	55	20	20	2	4	40	80
60-70	65	10	30	3	9	30	90
70-80	75	4	40	4	16	16	64
		N=Σf= 139				Σf'd= 61	$\Sigma fd'^2$= 369

$$S.D = \sqrt{\frac{\Sigma fd'^2}{N} - \left[\frac{\Sigma fd'}{N}\right]^2} \times C$$

$$= \sqrt{\frac{369}{139} - \left[\frac{61}{139}\right]^2} \times 10$$

$$= \sqrt{2.65 - (0.4388)^2 \times 10}$$

$$= \sqrt{2.65 - 0.1925 \times 10}$$

$$= \sqrt{2.4575} \times 10$$

$$= 1.57 \times 10$$

$$S.D = 15.7$$

Merits of Standard Deviation:-

It is rigidly and it is based on all the observations.

It is possible for further algebraic treatment

It is most important and widty used measures of dispersion.

It is less affected by sampling fluctuations

squaring the deviations make all of them positive

It provides the unit of measurement for the normal distribution.

Demerits of S.D :

It is not easy to understand and it is difficult to calculate.

It is affected by the value of every items in the series.

Uses :

It is the best measure of dispersion.

It is widely used in statistics, sampling theory and biology.

Standard deviation formula :-

Standard Error

The Standard deviation of the Sampling distribution is called Standard error. The word "error" is used to emphasize the variation among Sample mean is due to Sampling error.

Characteristic of Standard Error.

1. It is an instrument for Studying the characteristics of population for testing hypothesis.
2. It helps to Study or obtain the probable limits between true value and observed value.
3. It is used to obtain point estimates of the population parameter.
4. It gives an idea about the unreliability of a sample.
5. It is used to test the Significance of the difference between two Indepent Sample estimates of the same population parameter,

Formula:

$$SE = \frac{\sigma}{\sqrt{N}}$$

Steps: for Individual Series:

1. Calculate the mean $(\bar{x})$
2. Find out the deviations of each value from the mean i.e., $d = (x - \bar{x})$
3. Square the deviations and take the sum of squared deviations.
4. Divide the sum of squared deviations by no. of observation (N) which will give Variance.
5. Take the square root for the variance to get Standard Deviation.
6. Divide the SD by $\sqrt{N}$ (of the sample size) to get the SE.

Example: 1

Sl. No: 1, 2, 3, 4, 5, 6, 7, 8, 9 & 10

Blood Sugar}: 50, 55 60 62 58 54 56 58 54 & 43.

mg | 100ml

Sl. no	X	d $(x-\bar{x})$	d^2
1	43	-12	144
2	50	-5	25
3	54	-1	1
4	54	-1	1
5	55	0	0
6	56	1	1
7	58	3	9
8	58	3	9
9	60	5	25
10	62	7	49
	550		264

$$\bar{x} = \frac{550}{10} = 55$$

$$\text{Variance} = \frac{\Sigma d^2}{N} = \frac{264}{10} = 26.4$$

$$SD = \sqrt{\text{Variance}} = \sqrt{26.4} = 5.14$$

$$SE = \frac{\sigma}{\sqrt{N}} = \frac{5.14}{\sqrt{10}} = \frac{5.14}{3.16} = 1.63$$

Steps for Continuous Series:

1. Find out the mean by fixing the midpoint
2. Find out the deviation of each midpoint from the mean.
3. Square the deviations (d^2)
4. Multiply the squared deviations by respective frequencies (fd^2)
5. Divide the sum of fd^2 by sum of 'f' to get the variance
6. Square root of the. Variance will give SD.
7. Divide the standard deviation by square roots of sample size to get the standard error.

Example : 2

wt. in g

x	10-14	14-18	18-22	22-26	26·30
f	4	13	20	10	3

x	m	f	fm	d (m - $\bar{x}$)	d^2	fd^2
10-14	12	4	48	-7·6	57·76	231·04
14-18	16	13	208	-3·6	12·96	168·48
18-22	20	20	400	0·4	0·16	3·2
22-26	24	10	240	4·4	19·36	193·6
26-30	28	3	84	8·4	70·56	211·68
		50	980			808·00

$$\bar{x} = \frac{\Sigma fm}{\Sigma f} \quad \frac{980}{50} = 19·6 g.$$

$$\text{Variance} = \frac{\Sigma fd^2}{\Sigma f} \quad \frac{808}{50} = 16·16$$

$$SD = \sqrt{\text{Variance}} \quad \sqrt{16·16} = \underline{\underline{4·02}}$$

$$SE = \frac{SD}{\sqrt{N}} \quad \frac{4·02}{\sqrt{10}} = \frac{4·02}{3·16} = \underline{\underline{1·27}}$$