CHAPTER-1

Classification and Presentation of Business Data

1. **Variable (or Variate)** — A quantity which can vary from one individual to another is called a variable or variate. For example, age, sex, height, weight, income, no. of persons, the number of accidents per week, beauty, habits, intelligence etc.

2. **Types of Variables** — The variables may be classified into the two categories as follows —
   (i) **Qualitative Variables** — A qualitative variable is that which can't be measure in terms of magnitude and can be expressed in terms of quality or kind. These are called attributes. For example, sex, nationality, occupation, religion, marital status, literacy, beauty, honesty etc.
   (ii) **Quantitative Variables** — A quantitative variable is that which can be measure numerically on a scale. For example, age, height, income, speed, weight, no. of persons, no. of accidents per week, marks of students in a class etc.

3. **Types of Quantitative Variables** — The quantitative variables may be further classified into two categories —

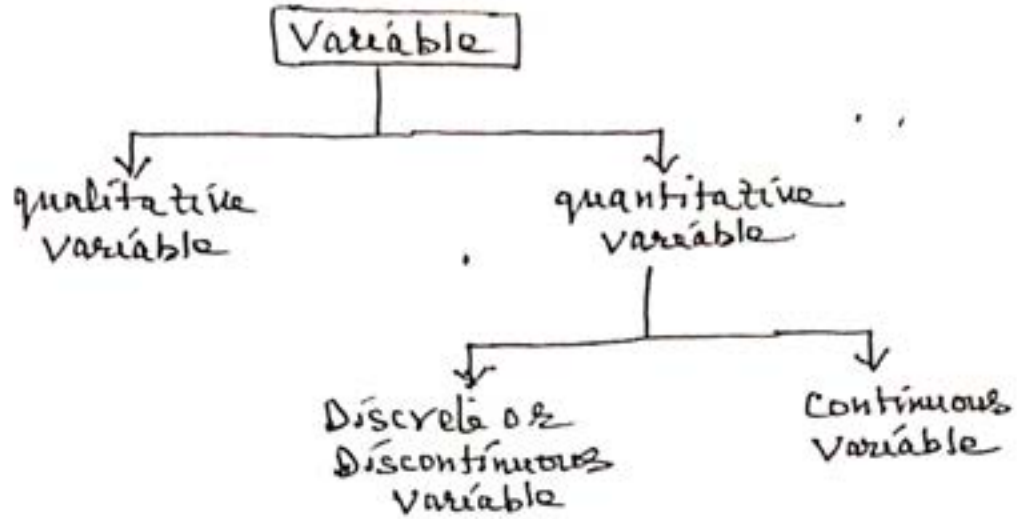
(i) Discrete or Discontinuous Variables —

A variable which can take only the finite number of values within a certain range is called discrete or discontinuous variable. In other words a discrete variable is one where the values of the variable may differ from one another by definite amount. For example, the number of children in families, the number of accidents per week in a particular city, the marks obtained by students in a particular subject etc. In all these examples each value of variable differs from the nearest by a finite value 1. Further, if we consider the daily expenditure of a particular person then expenditure of each day may differ from its nearest value by Rs. 0.01 or 1 paisa (a finite amount).

(ii) Continuous Variable —

A variable which can take any numerical value within a certain range is called continuous variable. In other words, a continuous variable can take the infinite values within a certain range. For example, age, distance, height, weight etc.

The various categories of a variable may also be shown in the tree diagram as shown below —



4. Primary and Secondary Data — Primary data is original and first hand information. In other words, the data which is collected by the investigator himself is known as primary data. For example the data in a population census obtained by the Census commissioner is known as primary data.

The data is termed as secondary when it is collected from records or it is already available. In other words, the secondary data is one which has been already collected by a source other than the present investigator. For example, population census data is primary for the office of the Census Commissioner whereas for other organisation who use such data, it is secondary.

5. Classification — We know that statistics is the science of collection, organisation, presentation, analysis and interpretation of numerical data. The collection of data is the first step of statistical investigation. The collected data

(i) Chronological or Temporal classification—

In this case the collected data is arranged according to the order of time expressed in year, months, weeks etc. The data is generally classified in ascending order of time. For ex., the data related with population, sales of a firm, import and export of a country etc during various years/months/weeks is always chronological classification. As an illustration, the following classification showing the birth rates in India during (1970-1976) is a chronological classification

Year	1970	1971	1972	1973	1974	1975	1976
Birth rate	36.8	36.9	36.6	34.6	35.0	40.0	40.5

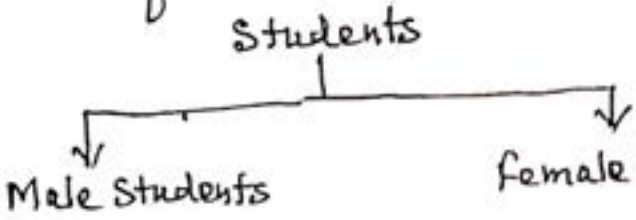
(ii) Geographical or Spatial Classification—

In this type of classification, the data is classified according to geographical region or place. For example, the production of wheat in different countries, the number of students in different cities etc. As an illustration, the following classification is a geographical or spatial classification—

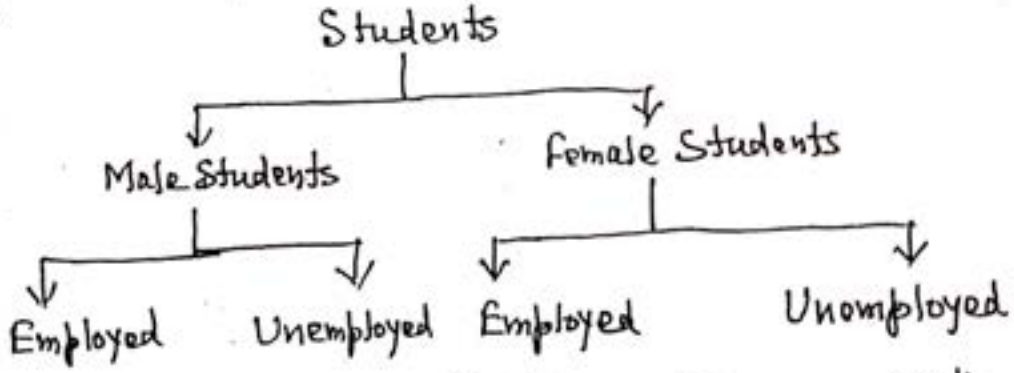
Country	America	China	Denmark	France	India
Yield of Wheat in Kg/acre	1925	893	225	439	862

(III) Qualitative Classification —

In this case the data is classified on the basis of some attributes or quality like sex, literacy, religion, cast etc. For example, if the students in a class are to be classified in respect of sex then we can classify the students into two classes namely - males and females as follows —



Each of the above two classes may further be classified in respect of the another attributes (quality) namely 'employment' as follows —



(IV) Quantitative Classification or Frequency Dis^{n} —

In this classification, the data is based on some quantitative characteristics like height, weight, age, income, sales etc. Let us consider the marks obtained by 40 students of a class in Mathematics out of 50 marks as follows —

5, 6, 7, 8, 11, 15, 20, 8, 11, 25
30, 15, 17, 11, 6, 22, 25, 20, 22, 15
30, 32, 32, 8, 20, 25, 22, 22, 35, 37
40, 20, 11, 25, 20, 10, 10, 15, 35, 42

On a perusal of the above data we observe that the data does not give any useful information. It is rather confusing to mind. The above is called raw data. From this raw data we can't draw any conclusion immediately. In order to draw some important conclusions this data can be arranged in the following manner —

Marks	Tally bar	No. of Students
5 - 10	𝍸 II	7
10 - 15	𝍸 I	6
15 - 20	𝍸	5
20 - 25	𝍸 IIII	9
25 - 30	IIII	4
30 - 35	IIII	4
35 - 40	III	3
40 - 45	II	2

In this table, the students are divided into 8 classes or groups according to their marks. This grouping gives a clear picture of the data and is known as grouped frequency distribution. Here the marks have been divided into classes as shown in first column and are known as class-intervals. The difference between the upper and lower values of a class interval is called width of the class interval which may be same or unequal for

different class intervals. Here, the width of each class interval is same i.e. 5. The middle value of a class interval is called mid-point of the class. The number of students corresponding to each class-interval, shown in column-3 is known as the class frequency (f). The tally bars corresponding to each class interval are used in column-2 to prepare the column-3.

The following criteria should be taken into consideration in order to obtain a grouped frequency distribution.

(i) As far as possible the width of each class interval should be equal.

(ii) There is no hard and fast rule to decide the no. of classes or groups. However, the number of classes should be between 5 and 30. With less no. of classes, accuracy is lost and with more no. of classes the computations become tedious.

(iii) The observations corresponding to the common point of two classes should always be included in the higher class eg. if 20 is an element of the data and 10-20 and 20-30 are two classes, then 20 is to be put in the class 20-30. This is to say, every class should be regarded as open to the right.

(iv) The total no. of all frequencies must be equal to the no. of observations in the raw data.

5.2 Cumulative Frequency :—

Some times one may be interested, at a glance to know the number of observations less than or greater than a particular value. This is done by finding cumulative frequency. The cumulative frequency (some times called less than c.f.) corresponding to a class is the sum of the frequencies of that class all of all classes prior to that class. A table displaying the class-intervals, frequencies and cumulative frequencies is known as cumulative frequency table. For the example considered in section 5.1, the cumulative frequency table is as follows —

Class-intervals	Tally bar	Frequency function	Cumulative freq. (less than
5-10	𝍸 ||	7	7
10-15	𝍸 |	6	13
15-20	𝍸	5	18
20-25	𝍸 ||||	9	27
25-30	||||	4	31
30-35	||||	4	35
35-40	|||	3	38
40-45	||	2	40

From this table, it is obvious that the no. of students who got the marks

(i) less than 10 is 7, greater than 10 is 40-7 = 33.

(ii) less than 15 is 13, greater than 15 is 40-13 = 27.

(iii) less than 20 is 18, greater than 20 is 40-18 = 22

and so on.

However, the number of observations greater than a particular value may also be obtained directly from a more than cumulative frequency table. The more than cumulative frequency corresponding to the initial class is the sum of the frequencies of all classes i.e. the total no. of observations. Now to obtain the more than cumulative frequency for second class, we subtract the previous (initial) class frequency from its more than cumulative frequency. Thus in general, the more than cumulative frequency for a class is equal to the difference of frequency and more than cumulative frequency of previous class. In the above example, the more than cumulative frequency table is given below—

class-interval (Marks)	Frequency function	cumulative freq. (More than
5 - 10	7	40
10 — 15	6	40-7=33
15 - 20	5	33-6=27
20 — 25	9	27-5=22
25 - 30	4	22-9=13
30 - 35	4	13-4=9
35 -40	3	9-4=5
40 -45	2	5-3=2

This table clearly indicates that the no. of students who got marks more than 5 is 40, more than 10 is 33 and so on.

5.3 Exclusive and Inclusive Class-Intervals

The class-intervals open to the right i.e. of the type $[a, b)$ are called exclusive since they exclude the upper limit of the class. The following is an example of exclusive class-intervals.

Income (in Rs.)	No. of Persons
250-500	75
500-750	70
750-1000	52
1000-1250	33
1250-1500	20

In exclusive class-intervals, the upper limit of each class is the lower limit of the next class. If the income of a person is exactly 750 then the person is considered in the class 750-1000.

The class-intervals closed to the right i.e. of the type $[a, b]$ are called inclusive since they include the upper limit of the class. The following frequency distribution is based on inclusive class intervals.

Income (in Rs.)	No. of Persons
250-499	60
500-749	43
750-999	25
1000-1249	16
1250-1499	06

However, for further statistical analysis it is desirable that the class-intervals be exclusive. To convert inclusive class-intervals into exclusive one is to make an adjustment as follows—

Adjustment – Find the difference between the lower limit of second class and the upper limit of the first class. Subtract the half of this difference from all the lower limits and add to all the upper limits.

In the above example, the adjustment factor is $\frac{500-499}{2} = 0.5$. So the frequency distribution with adjusted class intervals is as follows –

Income (in Rs)	No. of Persons
249.5 – 499.5	60
499.5 – 749.5	43
749.5 – 999.5	25
999.5 – 1249.5	16
1249.5 – 1499.5	06

The width of each class-interval before adjustment was 249 but after adjustment it is 250.

5.4 Various Types of Series – For the statistical analysis of data we come across the following three types of series –

(i) Individual Observations – In this case the frequencies are not given. The series of observation is of the form –

$x : x_1, x_2, x_3, \ldots, x_n$

(ii) Discrete Series (Un-grouped frequency dis^n)

It is a series of observations of the form as follows –

Variate (x) : $x_1, x_2, x_3 \cdots, x_n$

Frequency (f) : $f_1, f_2, f_3, \cdots, f_n$

(iii) Continuous Series (Grouped frequency dis^n)

It is a series of observations of the form —

Class-interval : $a_1 - a_2, a_2 - a_3, a_3 - a_4, \cdots, a_n - a_{n+1}$

Frequency : $f_1, f_2, f_3 \cdots f_n$

For the further statistical analysis, the mid-point of each class is taken to represent the class. Thus, if x_i is the mid-point of ith class, then $x_i = \frac{a_i + a_{i+1}}{2}$ and the above series takes the form as follows —

Mid-value (x) : $x_1, x_2, x_3, \cdots, x_n$

Frequency (f) : $f_1, f_2, f_3, \cdots, f_n$

6. Graphical Representation of data —

We have seen that the grouped frequency distribution provides a better idea of the data as compared to the ungrouped frequency distribution. Further, if the distribution is represented by graphs a more clear visual impression about the data is obtained as Graphs are the good visual aid. The various important types of graphs are as follows —

(i) Histogram
(ii) Frequency Polygon
(iii) Frequency Curve
(iv) Cumulative Frequency Curve or Ogive.

6.1 Histogram — To draw the histogram of a given grouped frequency distribution, the following points are to take into consideration.

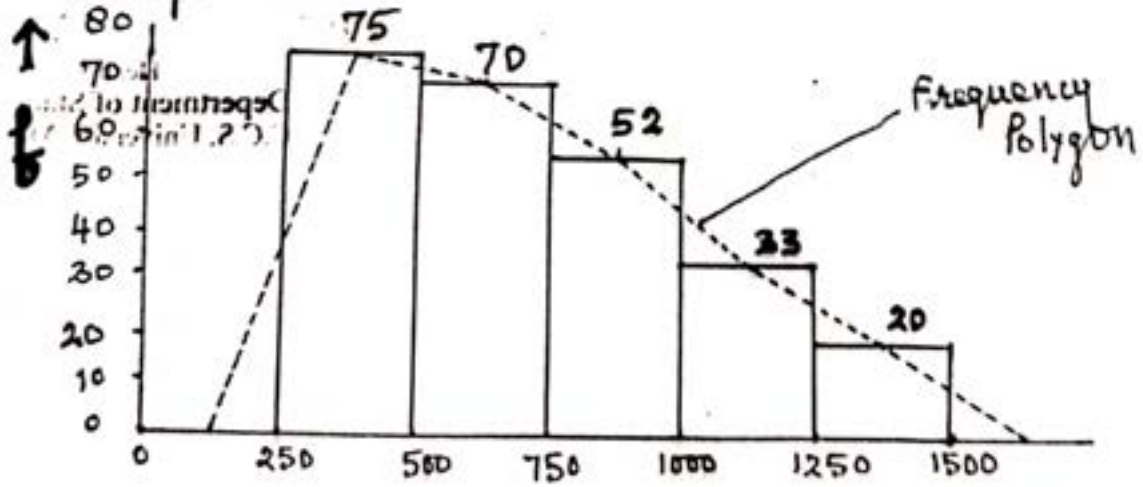
(a) Mark off along the x-axis all the class intervals on a suitable scale.

(b) Mark frequencies along y-axis on a suitable scale.

(c) It is not necessary to assume the same scale for both the axis. Different scales for different axis may be consider.

(d) Construct rectangles on each class interval as bases having the right equal to the corresponding frequency.

A diagram with all these rectangles is called a Histogram. Following is the Histogram of example considered in Section 5.3. (Smooth Line)



6.2 Frequency Polygon — For a grouped frequency distribution with equal class-intervals, a frequency polygon is obtained by joining the middle points of the upper side (top) of the adjacent rectangles of the histogram by means of straight lines. To complete the polygon, the mid point at each end are joined to the immediately lower and higher mid-points at zero frequency i.e, on the X-axis.

For the example considered in section 5.3 the frequency polygon is shown by dotted lines in section 6.1.

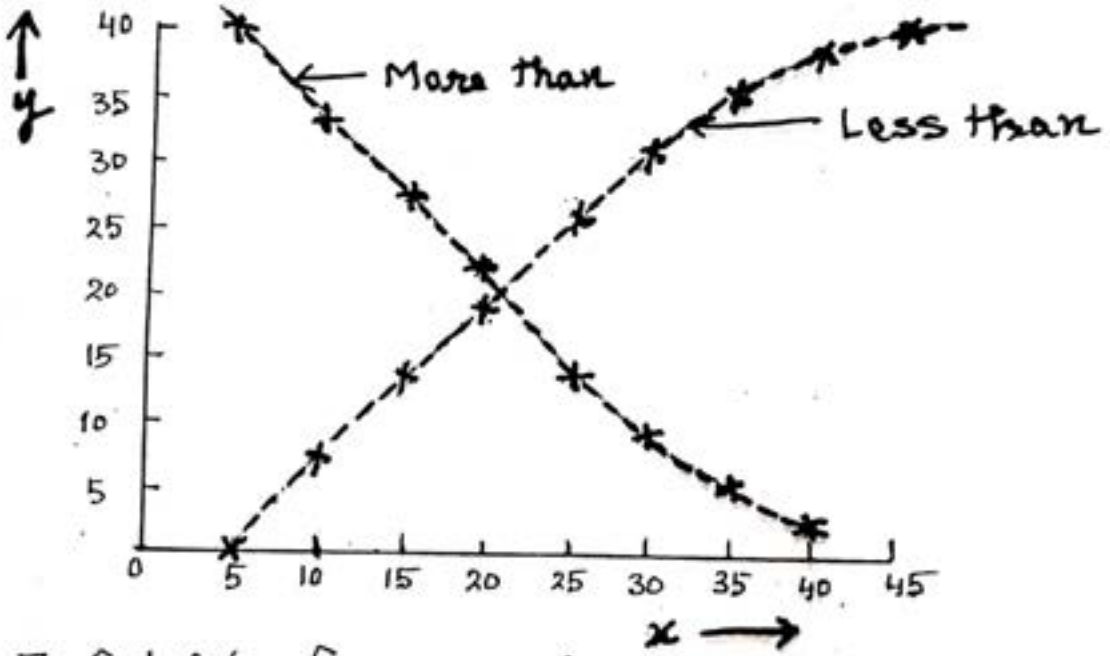
6.3 Frequency Curve — If in a grouped frequency distribution the width of the class-intervals become very small so that the number of observations increases and the histogram or the frequency polygon shall tend to a smooth free-hand continuous curve, known as frequency curve.

6.4 Cumulative Frequency Curve or Ogive —

The curve obtained by plotting cumulative frequencies is called a cumulative frequency curve or an ogive. There are two types of ogive (i) Less than ogive (ii) More than ogive. To obtain less than ogive we plot the points corresponding the upper limits on X-axis and

less than cumulative frequencies on y-axis. A free-hand smooth curve drawn by joining these points is called less than ogive. To obtain more than ogive we plot the points considering the lower limits on x-axis and more than cumulative frequencies on y-axis. A free-hand smooth curve drawn by joining these points is called more than ogive.

The less than and more than ogive for the example taken in the section 5.2 are shown below—



7. Relative Frequency Distribution—

Some times it is desirable to express the frequencies of each class in the form of proportion or percentage of the total frequency. This may be useful for comparing the distributions of

frequencies for two sets of data. The relative frequency of a class is obtained on dividing the class frequency by the total frequency. Further, if we multiply each relative frequency by 100, we get the percentage distribution of the frequencies. For the data considered in section 5.1, the relative frequencies and percentage frequencies are as follows:

Class Interval (Marks)	Frequencies	Relative Frequencies	Percentage Frequencies
5 – 10	7	0.175	17.5
10 – 15	6	0.150	15.0
15 – 20	5	0.125	12.5
20 – 25	9	0.225	22.5
25 – 30	4	0.100	10.0
30 – 35	4	0.100	10.0
35 – 40	3	0.075	7.5
40 – 45	2	0.050	5.0
Total	40	1.000	100

8. Diagrammatic Presentation of data

The diagrammatic presentation of data provides a good visual impression of the important features of the whole data. The various types of diagrams used for presenting

data are as follows—

(i) One-dimensional diagrams

(ii) Two-dimensional diagrams

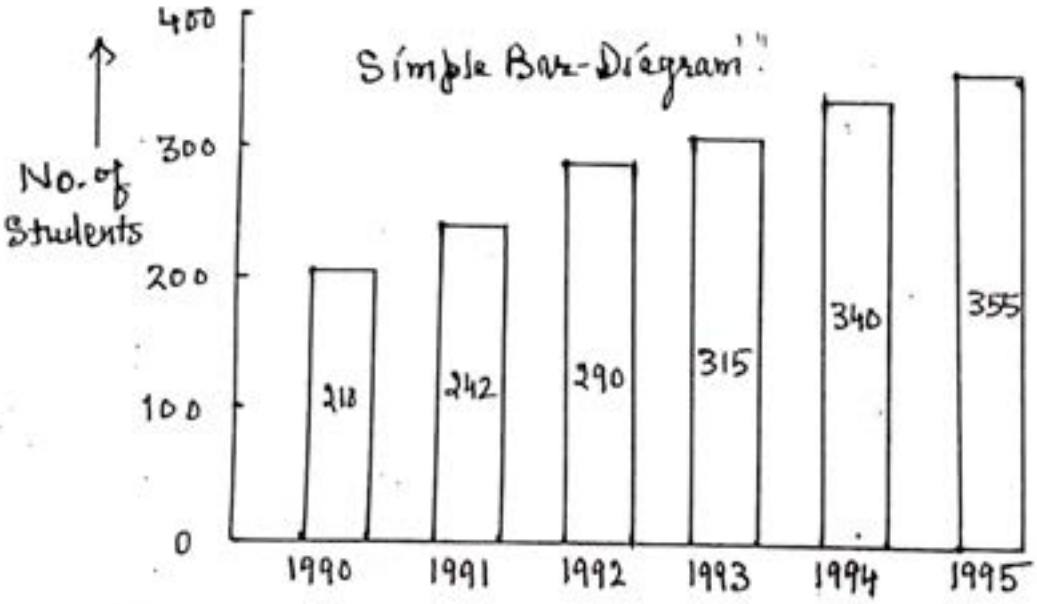
(iii) Pictorial diagrams

(iv) Cartograms or Maps.

8·1. One dimensional diagrams— These are also called Bar-diagrams. Various types of bar-diagrams are as follows—

(i) Simple Bar-diagram — Consider the example

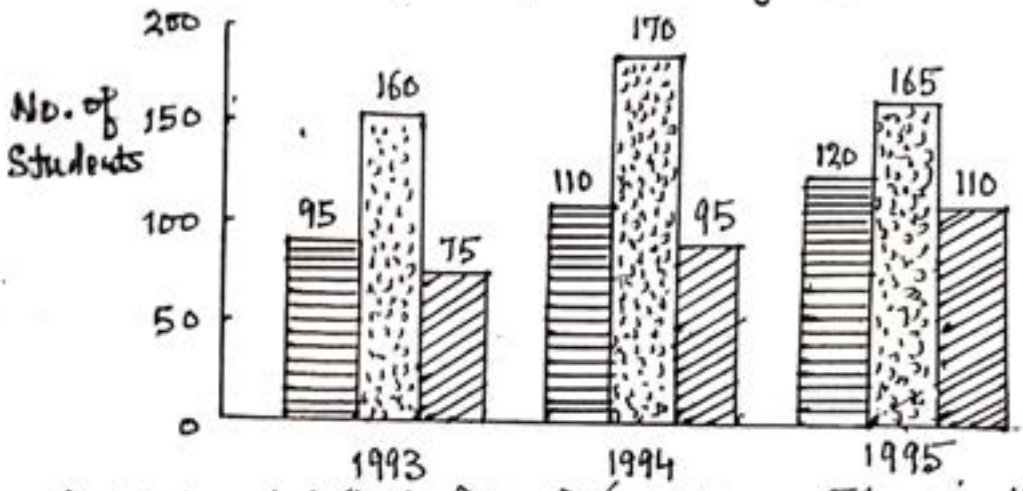
Year	1990	1991	1992	1993	1994	1995
No. of Students	210	242	290	315	340	355



(ii) Multiple Bar Diagram — It is used when a comparison is to be made. Consider the following example—

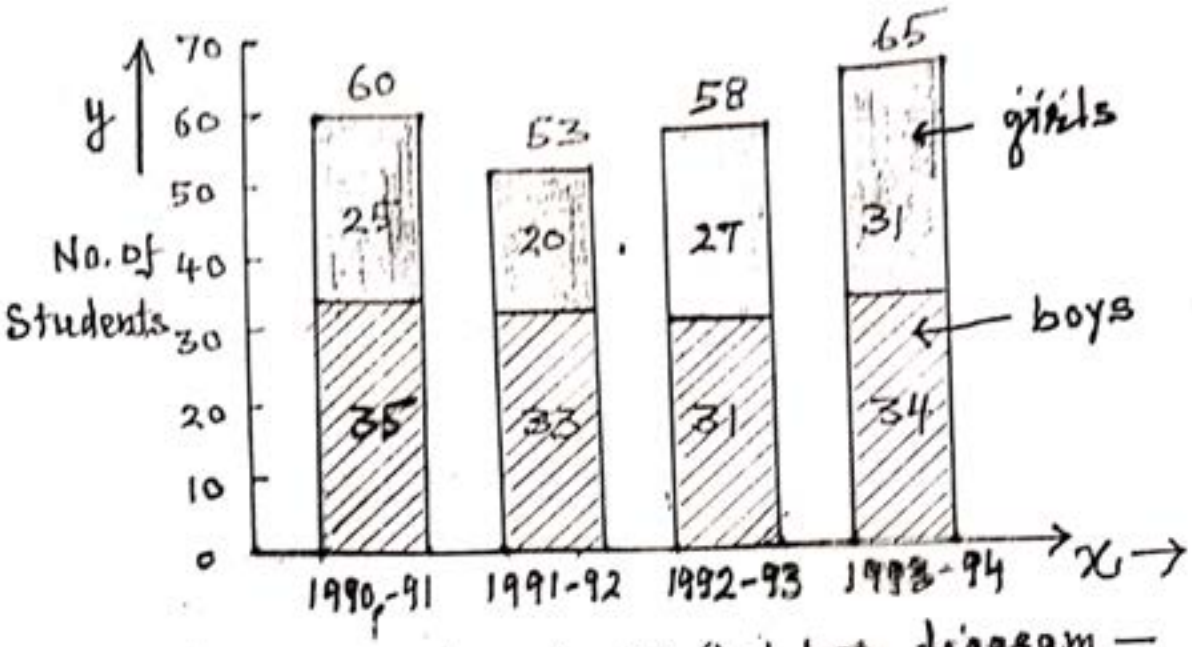
Year	1993	1994	1995
No. of Arts Students	95	110	120
No. of Science Students	160	170	165
No. of Commerce Students	75	95	110

(Multiple-bar diagram)



(iii) Sub-divided Bar Diagram — It is used when it is necessary to show the break-up of one variable —

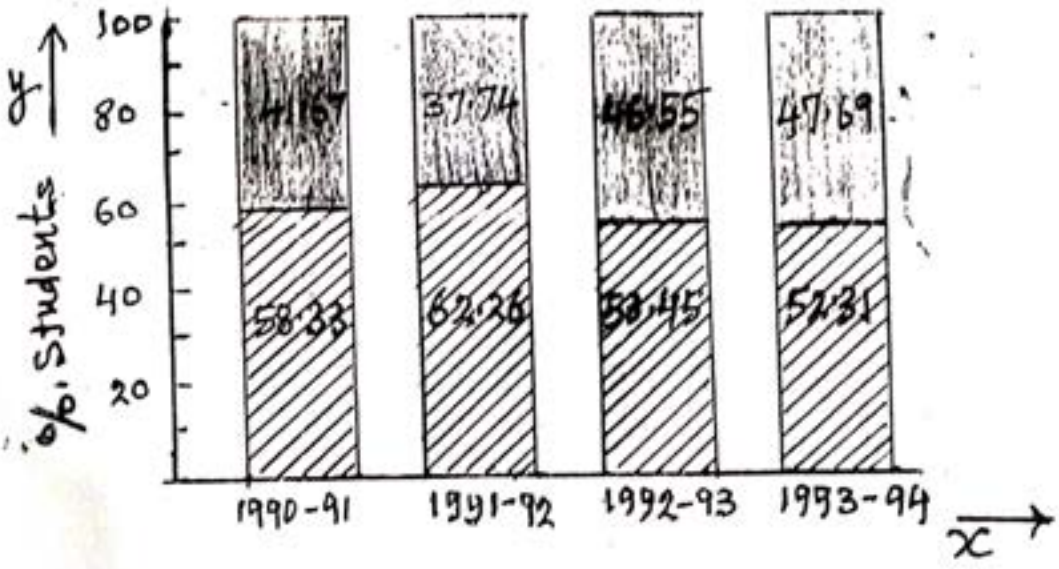
Year	1990-91	1991-92	1992-93	1993-94
No. of girls	25	20	27	31
No. of boys	35	33	31	34
Total	60	53	58	65



(IV) Percentage sub-divided bar diagram —

It is similar to the sub-divided bar, diagram. The only difference between the two is that, in the percentage sub-divided bar diagram the component parts are transformed into percentage of the total. Let us consider the example of sub-divided bar diagram as follow —

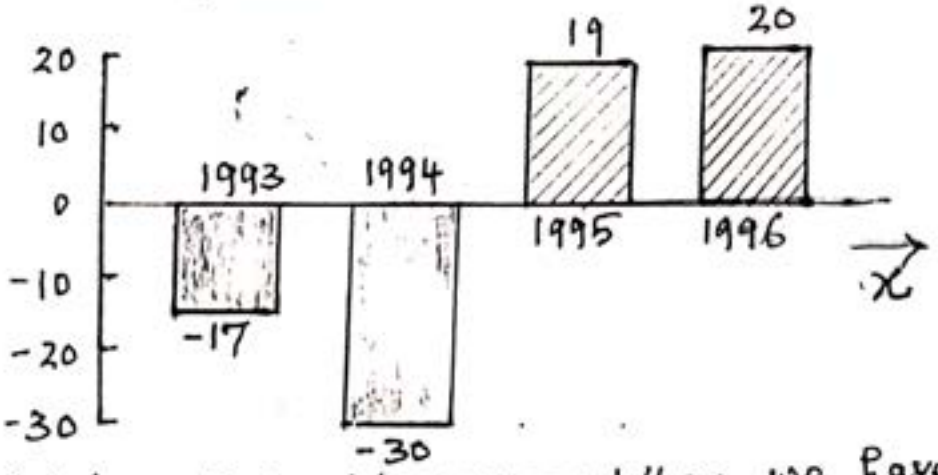
Year	1990-91	1991-92	1992-93	1993-94
% of girls	41.67	37.74	46.55	47.69
% of boys	58.33	62.26	53.45	52.31
Total	100	100	100	100



(V) Deviation Bar-diagram—

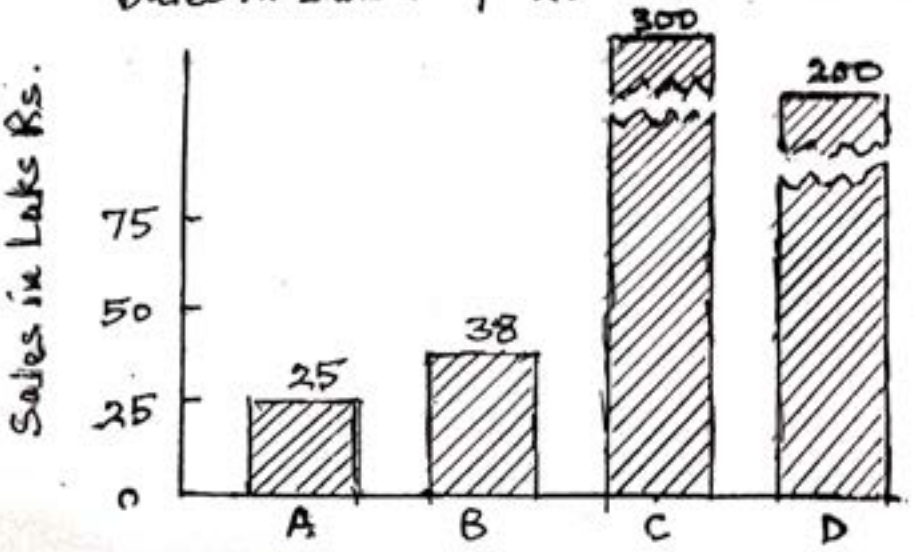
It is used to show the net magnitudes of a phenomenon i.e. net profit/loss, net export/import etc. Consider the following example—

Year	Export	Import	Balance of trade = Export - Import
1993	98	115	-17
1994	110	140	-30
1995	115	96	+19
1996	120	100	+20



(VI) Broken Bar-diagram — When we have wide variation in the values i.e. some values are very small as compared to others, we use broken bar diagram. Consider the following data—

Firms	A	B	C	D
Sales in Laks Rs	25	38	300	200

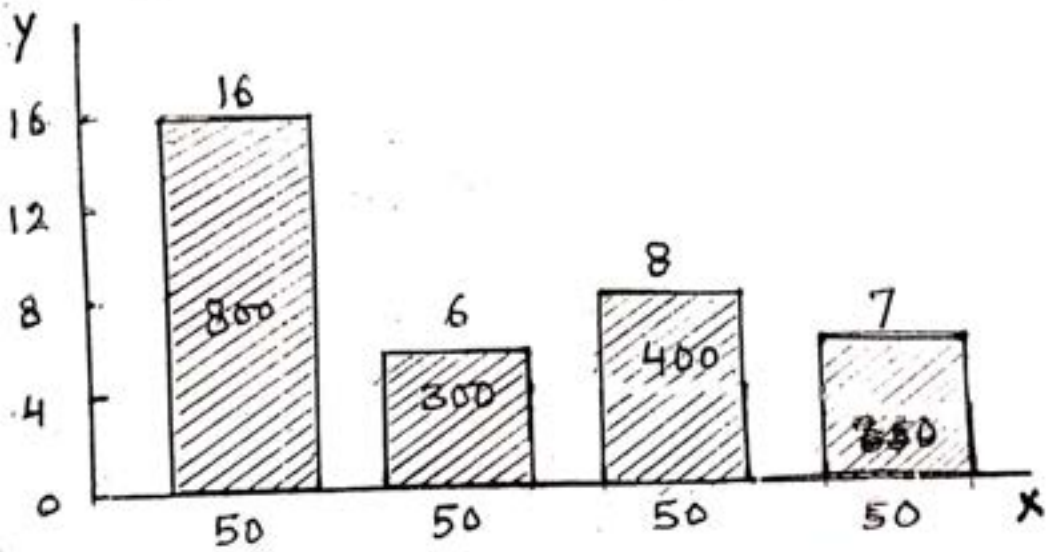


8.2. Two-Dimensional or Area Diagram —

In one-dimensional diagrams we consider only the length of bar to represent data where as in two-dimensional or area diagrams we consider the area of the geometric figure in proportion of the data. Some of the important two-dimensional diagrams are as follows —

(i) **Rectangle Diagram** — In rectangle diagram the length as well as width of bars are considered to represent the data. Consider the following example —

Items	: Food	Clothing	Rent	Miscellaneous
Expenditure (Rs)	: 800	300	400	350



(ii) **Square and Circle Diagrams** —

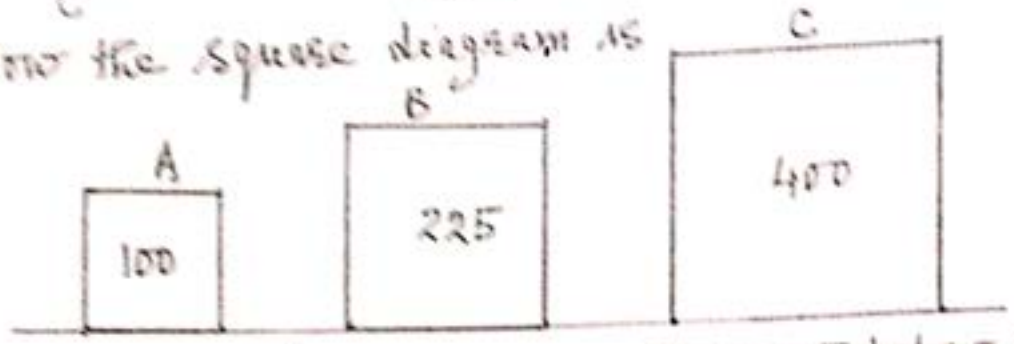
consider the example —

City	: A	B	C
Population in Lakh	: 100	225	400

For square diagram, first find the sq. root of the data as follows —

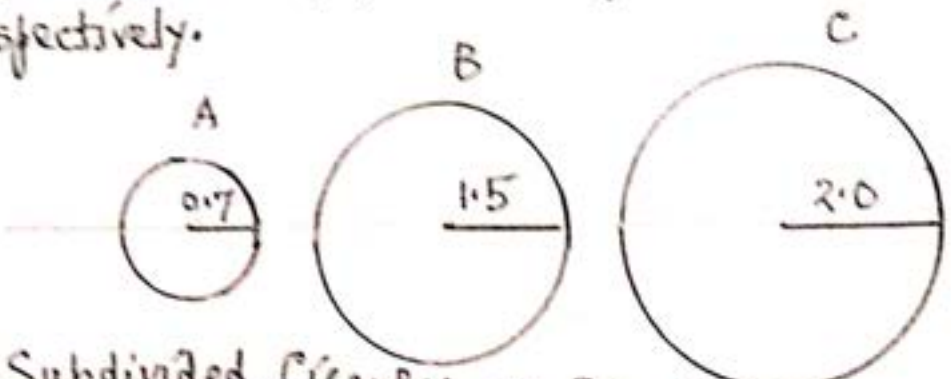
City	Population in Lakhs	Sq. Roots
A	100	10
B	225	15
C	400	20

Now the square diagram is



Scale : 5 Lakhs = 1 cm

For circle diagram of the above example, we remember the formula Area $= \pi r^2$. Therefore, the radious (r) of cities A, B and C are $\frac{7}{\sqrt{\pi}}$, $\frac{15}{\sqrt{\pi}}$ and $\frac{20}{\sqrt{\pi}}$ respectively. Now taking the scale $\frac{10}{\sqrt{\pi}}$ Lakhs = 1 c.m., we draw the circles of radious 0.70, 1.50 and 2.0 c.m. to represent the populations of cities A, B and C respectively.



(iii) Subdivided Circular or Pie-Diagram —

Here we divide a circle into different sectors in order to represent the component

part such that the areas of these sectors are proportional to the data of different component parts. We find

The angle proportional to the value of a component

$$= \frac{\text{Value of component}}{\text{Total value of all component}} \times 360^\circ$$

In above example, the angles corresponding to the values of cities A, B and C are computed as follows –

(i) Angle corresponding to value of city A

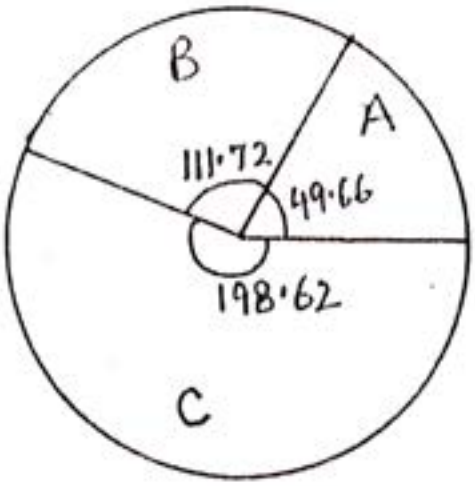
$$= \frac{100}{725} \times 360 = 49.66$$

(ii) Angle corresponding to value of city B

$$= \frac{225}{725} \times 360 = 111.72$$

(iii) Angle corresponding to value of city C

$$= \frac{400}{725} \times 360 = 198.62$$

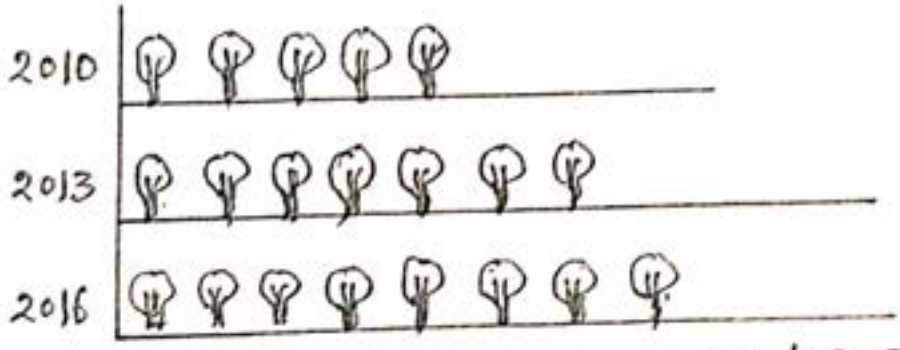


8.3 Pictogram — The pictogram is a device to represent statistical data in pictures. It is being widely used by government as well as by private organisations. The chief advantage of this representation is its attractiveness, even a common man can understand it easily. Consider the example—

Year	Production of Bulbs (Millions)
2010	25
2013	35
2016	40

The data can be represented with the help of pictogram as follows—

1 bulb = 5 millions bulbs.



8.4 Cartogram — A cartogram presents the numerical facts in the form of a map. For example, the clouds in a particular month say July may be represented by blue colour in different states in the map of India.

which is also known as raw data or ungrouped data is always in an unorganised form and need to be organised in a systematic and meaningful form so that its further statistical analysis can be made. So, it is essential for an investigator to condense a mass of data into more and more systematic and organised form. Thus, classification is a grouping of data according to their identity, similarity or resemblances. For example, students in a class may be grouped in respect of sex, age, marital status etc. Similarly, letters in the post office may be classified according to their destinations viz. Delhi, Jaipur, Agra, Kanpur etc. Obviously, the classification or grouping of the collected data is the first step of an organisation. The following are the main objectives of classification—

(i) It provides the compactness and closeness in data.

(ii) It eliminates the unnecessary details.

(iii) It shows a comparison between the various categories and highlights the significant aspect of data.

5.1 Types of Classifications— According to the nature of data, there are four basic types of classifications—

UNIT-2

MEASURES OF CENTRAL TENDENCY

INTRODUCTION

A measure of central tendency is a single value that attempt to describe a set of data by identifying the central position within that set of data. As such measures of central tendency are sometimes called measure of central location.

OR

The measuring of central tendency is a way of summarizing the data in the form of a typical or representative value. There are several statistical measures of central tendency or "averages". The commonly used averages are:

- Arithmetic Mean
- Median
- Mode
- Geometric Mean
- Harmonic Mean

Definitions:

1. According to Clark **"Average is an attempt to find one single figure to describe whole of figure."**
2. In the words of A.E. Waugh **"An average is a single value selected from a group of values to represent them in a same way—a value which is supposed to stand for whole group of which it is a part, as typical of all the values in the group."**
3. J.P. Guilford has pointed out that **"an average is a central value of a group of observations or individuals."**

Uses of Measure of Central Tendency:

The central tendency is needed for the following reasons:

1. Average provides the overall picture of the series. We cannot remember each and every facts relating to a field of enquiry.

2. Average value provides a clear picture about the field under study for guidance and necessary conclusion.

3. It gives a concise description of the performance of the group as a whole and it enables us to compare two or more groups in terms of typical performance.

ARITHMETIC MEAN

Arithmetic mean is the most commonly used measure of central tendency. It is defined as the sum of the values of all observations divided by the number of observations and is usually denoted by $\bar{X}$. In general, if there are N observations as $X_1, X_2, ..., X_N$ then the Arithmetic Mean is given

$$\bar{X} = \frac{X_1 + X_2 + ... + X_N}{N}$$

The right hand side can be written

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

Here i is an index which takes successive values 1, 2, 3,...N.

For convenience, this will be written in simpler form without the index i. Thus

$$\bar{X} = \frac{\sum X_i}{N}$$

Where $\sum X_i =$ sum of all the observations

And N = total number of observations.

Example 1: Suppose the monthly income (in Rs) of six families is given as: 1600, 1500, 1400, 1525, 1625, 1630.

Solution : The mean family income is obtained by adding up the incomes and dividing by the number of families.

$$\bar{X} = \frac{1600 + 1500 + 1400 + 1525 + 1625 + 1630}{6}$$

= Rs 1, 546.66 = Rs 1, 547 (approximate)

It implies that on an average, a family earns Rs 1,547.

Example 2: Calculate Arithmetic Mean from the data showing marks of students in a class in an economics test: 40, 50, 55, 78, 57.

Solution: The average marks of students is

$$\bar{X} = \frac{40 + 50 + 55 + 78 + 57}{5} = 56.2$$

Example 3. Find the mean of the first five prime numbers.

Solution: The first five prime numbers are 2, 3, 5, 7 and 11.

$\bar{X}$ = Sum of the first five prime numbers/number of prime numbers

= (2 + 3 + 5 + 7 + 11)/5

= 28/5

= 5.6

Hence, their mean is 5.6

Find the missing Observation:

Example 4: If the mean of 9, 8, 10, x, 12 is 15, find the value of x.

Solution: Mean of the given numbers = (9 + 8 + 10 + x + 12)/5 = (39 + x)/5

According to the problem, mean = 15 (given).

Therefore, (39 + x)/5 = 15

⇒ 39 + x = 15 × 5

⇒ 39 + x = 75

⇒ 39 - 39 + x = 75 - 39

⇒ x = 36

Hence, x = 36.

Example 5: If the mean of five observations x, x + 4, x + 6, x + 8 and x + 12 is 16, find the value of x.

Solution: Mean of the given observations

= x + (x + 4) + (x + 6) + (x + 8) + (x + 12)/5

= (5x + 30)/5

According to the problem, mean = 16 (given).

Therefore, (5x + 30)/5 = 16

⇒ 5x + 30 = 16 × 5

⇒ 5x + 30 = 80

⇒ 5x + 30 - 30 = 80 - 30

⇒ 5x = 50

⇒ x = 50/5

⇒ x = 10

Hence, x = 10.

Example 6: The mean of 40 numbers was found to be 38. Later on, it was detected that a number 56 was misread as 36. Find the correct mean of given numbers.

Solution: Calculated mean of 40 numbers = 38.

Therefore, calculated sum of these numbers = (38 × 40) = 1520.

Correct sum of these numbers

= [1520 - (wrong item) + (correct item)]

= (1520 - 36 + 56)

= 1540.

Therefore, the correct mean = 1540/40 = 38.5.

Example 7: The mean of the heights of 6 boys is 152 cm. If the individual heights of five of them are 151 cm, 153 cm, 155 cm, 149 cm and 154 cm, find the height of the sixth boy.

Solution: Mean height of 6 boys = 152 cm.

Sum of the heights of 6 boys = (152 × 6) = 912 cm

Sum of the heights of 5 boys = (151 + 153 + 155 + 149 + 154) cm = 762 cm.

Height of the sixth boy

= (sum of the heights of 6 boys) - (sum of the heights of 5 boys)

= (912 - 762) cm = 150 cm.

Hence, the height of the sixth girl is 150 cm.

Example 8 : The mean of 16 items was found to be 30. On rechecking, it was found that two items were wrongly taken as 22 and 18 instead of 32 and 28 respectively. Find the correct mean.

Solution: Calculated mean of 16 items = 30.

Incorrect sum of these 16 items = (30 × 16) = 480.

Correct sum of these 16 items

= (incorrect sum) - (sum of incorrect items) + (sum of actual items)

= [480 - (22 + 18) + (32 + 28)]

= 500.

Therefore, correct mean = 500/16 = 31.25.

Hence, the correct mean is 31.25.

Assumed Mean Method

If the number of observations in the data is more and/or figures are large, it is difficult to compute arithmetic mean by direct method. The computation can be made easier by using assumed mean method.

In order to save time in calculating mean from a data set containing a large number of observations as well as large numerical figures, you can use assumed mean method. Here you assume a particular figure in the data as the arithmetic mean on the basis of logic/experience. Then you may tahe deviations of the said assumed mean from each of the observation. You can, then, take the summation of these deviations and divide it by the number of observations in the data. The actual arithmetic mean is estimated by taking the sum of the assumed mean and the ratio of sum of deviations to number of observations. Symbolically,

Let, A = assumed mean

X = individual observations

N = total numbers of observations

d= deviation of assumed mean from individual observation, i.e. d=X-A

Then sum of all deviations is taken as $\sum d = \sum (X-A)/c$

Therefore, $\bar{X} = A + \frac{\sum d}{N} \times c$

It should remember that any value, whether existing in the data or not, can be taken as assumed mean. However, in order to simplify the calculation, centrally located value in the data can be selected as assumed mean.

Example 9: The following data shows the weekly income of 10 families.

Family	A	B	C	D	E	F	G	H	I	J
Weektq income	850	700	100	750	5000	80	420	2500	400	360

Compute mean family income.

Solution:

Computation of arithmetic mean by assumed arithmetic mean

Family	Income	d'=X-850	d=d'/10
A	850	0	0
B	700	-150	-15
C	100	-750	-75
D	750	-100	-10
E	5000	4150	415
F	80	-770	-77
G	420	-430	-43
H	2500	1650	165
I	400	-450	-45
J	360	-490	-49
Total	11160	2660	266

The arithmetic mean by assumed mean method

$\bar{X} = 850 + 266 = 1116$

Calculation of arithmetic mean for Grouped data

Discrete Series: In case of discrete series, frequency against each observation is multiplied by the value of the observation. The values, so obtained, are summed up and divided by the total number of frequencies. Symbolically,

$$\bar{X} = \frac{\sum f X}{\sum f}$$

Where $\sum f X$ = sum of product of variables and frequencies.

$\sum f$ = sum of frequencies.

Example 10: Plots in a housing colony come in only three sizes: 100 sq. metre, 200 sq. meters and 300 sq. metre and the number of plots are respectively 200, 50 and 10.

Solution: Direct Method

Plot size	No. of plots F	fX
100	200	20000
200	50	10000
300	10	3000
Total	260	33000

So the arithmetic mean direct method $\bar{X} = \frac{\sum fX}{\sum f}$

$$\bar{X} = \frac{33000}{260} = 143.47$$

Therefore, the mean plot size in the housing colony is 143.47 Sq. metre.

Continuous *Series:* Here, class intervals are given. The process of calculating arithmetic mean in case of continuous series is same as that of a discrete series. The only difference is that the mid-points of various class intervals are taken. We have already known that class intervals may be exclusive or inclusive or of unequal size. Example of exclusive class interval is, say, 0—10, 10—20 and so on. Example of inclusive class interval is, say, 0—9, 10—19 and so on. Example of unequal class interval is, say, 0-20, 20-50 and so on. In all these cases, calculation of arithmetic mean is done in a similar way.

Example11: Calculate average marks of the following students.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of students	5	12	15	25	8	3	2

Solution:

Marks (x)	No. of students (f)	Mid point (m)	Fm
0-10	5	5	25
10-20	12	15	180
20-30	15	25	375
30-40	25	35	875
40-50	8	45	360
50-60	3	55	165
60-70	2	65	130

Total	70		2110

The average marks =2110/70 = 30.14 marks

Properties of Arithmetic Mean

(1) The sum of deviations of items about arithmetic mean is always equal to zero. Symbolically, $\sum(X-\bar{X})=0$.

(2) Arithmetic mean is affected by extreme values. Any large value, on either end, can push it up or down.

(3) The sum of the square of the deviations of a set of values is minimum when taken about mean.

(4) The mean of n observations $x_1, x_2, \ldots, x_n$ is $\bar{X}$. If each observation is increased by p, the mean of the new observations is $(\bar{X} + p)$.

(5) The mean of n observations $x_1, x_2, \ldots, x_n$ is $\bar{X}$. If each observation is decreased by p, the mean of the new observations is $(\bar{X} - p)$.

(6) The mean of n observations $x_1, x_2, \ldots, x_n$ is $\bar{X}$. If each observation is multiplied by a nonzero number p, the mean of the new observations is $p\bar{X}$.

(7) The mean of n observations $x_1, x_2, \ldots, x_n$ is $\bar{X}$. If each observation is divided by a nonzero number p, the mean of the new observations is $(\bar{X}/p)$.

Merits and Demerits of arithmetic mean

Merits:

1. It is rigidly defined.
2. It is easy to understand and easy to calculate.
3. It is based on all observations.
4. Of all the averages, arithmetic mean is affected least by fluctuation of sampling.

Demerits :

1. It cannot be determined by inspection nor it can be located graphically.
2. Arithmetic mean cannot be used if we are dealing with qualitative characteristics which can not be measured quantitively; such as intelligence, honesty, beauty etc.

WEIGHTED ARITHMETIC MEAN

Weighted Mean is an average computed by giving different weights to some of the individual values. If all the weights are equal, then the weighted mean is the same as the arithmetic mean. It represents the average of a given data. The Weighted mean is similar to arithmetic mean or sample mean. The Weighted mean is calculated when data is given in a different way compared to an arithmetic mean or sample mean.

The weights cannot be negative. Some may be zero, but not all of them; since division by zero is not allowed. Weighted means play an important role in the systems of data analysis, weighted differential and integral calculus.

Formula of weighted Mean

The Weighted mean for given set of non-negative data $x_1, x_2, ..., x_N$ with non-negative weights $w_1, w_2, ..., w_N$ can be derived from the formula.

$$\bar{X} = \frac{x_1 w_1 + x_2 w_2 + ... + x_N \ w_N}{w_1 + w_2 + ... + w_N}$$

Where,
x is the repeating value
w is the number of occurrences of weight
$\bar{X}$ is the weighted mean

Example: Here's the sample data for the survey:

Number of TVs per Household	Number of Households
1	73
2	378
3	459
4	90

As many of the values in this data set are repeated multiple times, you can easily compute the sample mean as a weighted mean. Follow these steps to calculate the weighted arithmetic mean:

Solution: Step 1: Assign a weight to each value in the dataset:

$x_1 = 1,\ w_1 = 73$

$x_2 = 2,\ w_2 = 378$

$x_3 = 3,\ w_3 = 459$

$x_4 = 4,\ w_4 = 90$

Step 2: Compute the numerator of the weighted mean formula.

Multiply each sample by its weight and then add the products together:

$$\sum_{i=1}^{4} x_i w_i = x_1 w_1 + x_2 w_2 + x_3 w_3 + x_4 w_4$$

$$= (1)(73)+(2)(378)+(3)(459)+(4)(90)$$

$$=2566$$

Step 3: Now, compute the denominator of the weighted mean formula by adding the weights together.

$$\sum_{i=1}^{4} w_i = w_1 + w_2 + w_3 + w_4$$

$$= 73+378+459+90$$

$$=1000$$

Step 4: Divide the numerator by the denominator

$$\bar{X} = \frac{\sum_{i=1}^{4} x_i\, w_i}{\sum_{i=1}^{4} w_i}$$

$$= \frac{2566}{1000}$$

$$= 2.566$$

The mean number of TVs per household in this sample is 2.566.

Example 2: A student obtained the marks 40, 50, 60, 80, and 45 in math, statistics, physics, chemistry and biology respectively. Assuming weights 5, 2, 4, 3, and 1 respectively for the above mentioned subjects, find the weighted arithmetic mean per subject.

Solution:

Subject	**MarkObtained** (x)	**Weight** (w)	xw
Math	40	5	200
Statistics	50	2	100
Physics	60	4	240
Chemistry	80	3	240
Biology	45	1	45
Total		$\sum_{i=1}^{5} w_i = 15$	$\sum_{i=1}^{5} x_i w_i = 825$

Now we will find the weighted arithmetic mean as:

$$\bar{X} = \frac{\sum_{i=1}^{5} x_i w_i}{\sum_{i=1}^{5} w_i}$$

$$= \frac{825}{15}$$

=55 marks/ subject.

MEDIAN (OR POSITIONAL AVERAGE)

Median is that positional value of the variable which divides the distribution into two equal parts, one part comprises all values greater than or equal to the median value and the other comprises all values less than or equal to it. *the* Median is *the* ~~–m~~iddle" *element* when *the* data *set is* arranged in *order* of *the* magnitude. Since *the* median is determined by *the position of different values, it remains* unaffected if, say, *the* size *of the largest val*ue *increases*.

Example 1: Suppose we have the following observation in a data set: 5, 7, 6, 1, 8, 10, 12, 4, and 3.

Solution: Arranging the data, in ascending order you have:

1, 3, 4, 5, 6, 7, 8, 10, 12.

The ~~"m~~iddle *score"* is 6, so the median is 6. Half of the scores are larger than 6 and half of the scores are smaller. If there are even numbers in the data, there will be two observations which fall in the middle. The median in this case is computed as the arithmetic mean of the two middle values.

When the number of observations (N) is odd.

Then, median is $(N + 1)/2^{th}$ observation.

Example 2: Find the median of the data 25, 37, 47, 18, 19, 26, 36.

Solution: Arranging the data in ascending order, we get 18, 19, 25, 26, 36, 37, 47

Here, the number of observations is odd, i.e., 7.

Therefore, median $=(n + 1/2)^{th}$ observation.

$= (7 + 1/2)^{th}$ observation.

$= (8/2)^{th}$ observation

$= 4^{th}$ observation.

4^{th} observation is 26.

Therefore, median of the data is 26.

When the number of observations (N) is even.

Then median is the mean of $(N/2)^{th}$ and $((N + 1)/2)^{th}$ observation.

i.e., Median $= \dfrac{(N/2)^{th} \text{ observation} + ((N+ 1)/2)^{th} \text{ observation}}{2}$

Example 3: Find the median of the data 24, 33, 30, 22, 21, 25, 34, 27.

Solution: Here, the number of observations is even, i.e., 8.

Arranging the data in ascending order, we get 21, 22, 24, 25, 27, 30, 33, 34

Therefore, median $= \{(n/2)^{th}$ observation $+ (n + 1/2)^{th}$ observation$\}/2$

$= (8/2)^{th}$ observation $+ (8/2 + 1)^{th}$ observation

$= 4^{th}$ observation $+ (4 + 1)^{th}$ observation

$= \{25 + 27\}/2$

= 52/2

= 26

Therefore, the median of the given data is 26.

Discrete Series

In case of discrete series the position of median i.e. $((N+1)/2)^{th}$ item can be located through cumulative frequency. The corresponding value at this position is the value of median.

Example 4: The frequency distributsion of the number of persons and their respective incomes (in Rs) are given below. Calculate the median income.

Income (in *Rs):*	10	20	30	40
Number ofpersoris:	2	4	10	4

Solution: In order to calculate the median income, you may prepare the frequency distribution as given below.

Income	No. of persons	Cumulative frequency
10	2	2
20	4	6
30	10	16
40	4	20

The median is located in the $(N+1)/2 = (20+1)/2 = 10.5^{th}$ observation. This can be easily located through cumulative frequency. The 10.5^{th} observation lies in the c.f. of 16. The income corresponding to this is Rs 30, so the median income is Rs 30.

Continuous Series

In case of continuous series you have to locate the median class where $N/2^{th}$ item [not $(N+1)/2^{th}$ item] lies. The median can then be obtained as follows:

$$\text{Median} = L + \frac{(N/2 - \text{c.f.})}{f} h$$

Where, L = lower limit of the median class,

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

f= frequency of the median class,

h = magnitude of the median class interval.

No adjustment is required if frequency is of unequal size or magnitude.

Example 5: Following data relates to daily wages of persons working in a factory. Following data relates to daily wages of persons working in a factory.

Daily Wages(in Rs)	55-60	60-65	65-70	70-75	75-0	80-85	85-90	90-95
No. of Workers	7	13	15	20	30	33	28	14

Solution: The data is arranged in descending order here. In the above illustration median class is the value of $(N/2)^{t"}$ item (i.e. 160/2) = 80^{th} item of the series, which lies in 35—40 class interval. Applying the formula of the median as:

Daily wages(in Rs)	No. of workers	Cumulative frequences
55-60	7	14
60-65	13	42
65-70	15	75
70-75	20	105
75-80	30	125
80-85	33	140
85-90	28	153
90-95	14	160

$$\text{Median} = L + \frac{(N/2 - \text{c.f.})}{f} h$$

$$= 35 + \frac{(80 - 75)}{30} 5$$

$= 35.83$ Rs.

Properties of the Median:

1. The median is used to find the center or middle value of a data set.

2. The median is used when it is necessary to find out whether the data values fall into the upper half or lower half of the distribution.

3. The median is used for an open-ended distribution.

4. The median is affected less than the mean by extremely high or extremely low values.

Merits and Demerits of arithmetic mean

Merits:

1. It is rigidly defined.

2. It is easy to understand and easy to calculate.

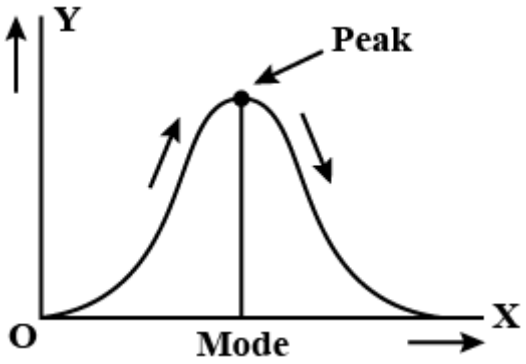
3. It is not affected by extreme values.
4. Of all the averages, arithmetic mean is affected least by fluctuation of sampling.

Demerits :

1. It case of even number of observations median cannot be determinate exactly.
2. It is not amenable to algebraic treatments.

MODE (MOST FASHIONABLE)

The mode is the value that occurs the most frequently in your data set. The **mode** of a set of data values is the value that appears most often.



Example 1: Find the mode of the given set of number

2, 2, 3, 5, 4, 3, 2, 3, 3, 5

Solution: Arranging the number with same values together, we get

2, 2, 2, 3, 3, 3, 3, 4, 5, 5

We observe that 3 occurs maximum number of times, i.e., four times.

Therefore, mode of this data is 3.

Example 2: The data 2, 5, 1, 3, 5, 7, 6, 3, 8 have two modes 3 and 5.

Therefore, each is repeated two times which is maximum.

Example 3: Let us find the Mode of the following data

4, 89, 65, 11, 54, 11, 90, 56

Solution: Here in these varied observations the most occurring number is 11, hence the Mode = 11

Example 4 : The height of 50 plants in a garden are given below.

Height (cm)	10	25	30	40	45
Number Plants	13	15	12	8	2

Find the mode of the data.

Solution: The frequency of 25 is maximum.

So, the mode of this data is 15.

Properties of the mode:

1. The mode is used when the most typical case is desired.

2. The mode is the easiest average to compute.

3. The mode can be used when the data are nominal or categorical, such as religious preference, gender, or political affiliation.

4. The mode is not always unique. A data set can have more than one mode, or the mode may not exist for a data set.

5. It is not affected by a few very high or low values

Merits and Demerits of the Mode

Merits:

1. The mode is easy to understand and calculate.

2. The mode is not affected by extreme values.

3. The mode is easy to identify in un-grouped data and discrete frequency distribution.

4. The mode is useful for qualitative data.

5. The mode can be computed in an open-ended frequency table.

6. The mode can be located graphically.

Demerits:

1. The mode is not well defined.

2. The mode is not based on all values.

3. The mode is stable for large values and will not be well defined if the data consist of a small number of values.

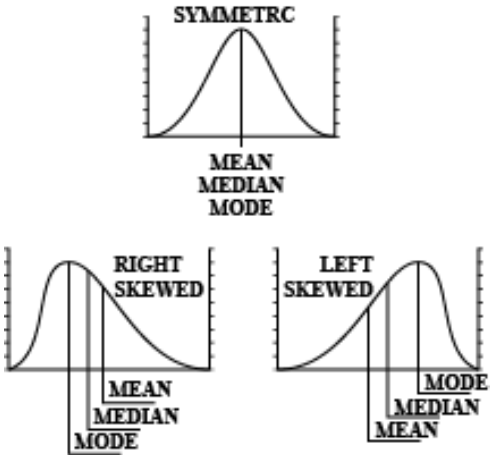
4. The mode is not capable of further mathematical treatment.

5. Sometimes data have one mode, more than one mode, or no mode at all.

Relation between Mean, Median, and Mode

There is an inter-relation between the measures of central tendency. Professor Karl Pearson has suggested an empirical relationship between Mean, Median, and Mode. Via this equation, if the values of two measures are known we can find the third measure. The equation is as follows

Mean – Mode = 3 [Mean – Median]



GEOMETRIC MEAN

The Geometric Mean (G.M) of a series containing n observations is the nth root of the product of the values. Consider, if x_1 , x_2 …. X_n are the observation, then the G.M is defined as:

$$G.M=\sqrt[n]{x_1.x_2.....x_n}$$

Or

$$G.M=\left(x_1.x_2.....x_n\right)^{1/n}$$

This can also be written as;

$$\text{Log GM} = \frac{1}{n}\log(x_1 x_2 .....x_n)$$

$$= \frac{1}{n}(\log(x_1)+\log(x_2)+.....+\log(x_n))$$

$$= \frac{\sum_{i=1}^{n}\log(x_i)}{n}$$

Therefore, Geometric Mean,

$$\text{GM= Antilog } \frac{\sum_{i=1}^{n}\log(x_i)}{n}$$

It is also represented as:

$$\text{G.M}=\sqrt[n]{\prod_{i=1}^{n} x_i}$$

For any Grouped Data, G.M can be written as;

$$\text{GM= Antilog } \frac{\sum_{i=1}^{n} f_i \log(x_i)}{n}$$

Where $n = f_1 + f_2 + ... + f_n$.

Difference between Geometric Mean And Arithmetic Mean

There is a difference between both the means like G.M and arithmetic mean for the given data set how they are calculations are done using G.M and arithmetic mean formula.

Arithmetic Mean	**Geometric Mean**
The arithmetic mean or mean can be found by adding all the numbers for the given data set divided by the number of data points in a set.	It can be found by multiplying all the numbers in the given data set and take the nth root for the obtained result.
For example, the given data sets are: 5, 10, 15 and 20 Here, the number of data points = 4	For example, consider the given data set, 4, 10, 16, 24 Here n= 4

Arithmetic mean or mean = (5+10+15+20)/4 Mean = 50/4 =12.5	Therefore, the G.M = 4th root of (4 ×10 ×16 × 24) = 4th root of 15360 G.M = 11.13

Geometric Mean Properties

Some of the important properties of the G.M are:

- The G.M for the given data set is always less than the arithmetic mean for the data set
- If each object in the data set is substituted by the G.M, then the product of the objects remains unchanged.
- The ratio of the corresponding observations of the G.M in two series is equal to the ratio of their geometric means
- The products of the corresponding items of the G.M in two series are equal to the product of their geometric mean.

Application of Geometric Mean

The greatest assumption of the G.M is that data can be really interpreted as a scaling factor. Before that, we have to know when to use the G.M. The answer to this is, it should be only applied to positive values and often used for the set of numbers whose values are exponential in nature and whose values are meant to be multiplied together. This means that there will be no zero value and negative value which we cannot really apply. Geometric mean has a lot of advantages and it is used in many fields. Some of the applications are as follows

- It is used in stock indexes. Because many of the value line indexes which is used by financial departments use G.M.
- It is used to calculate the annual return on the portfolio.
- It is used in finance to find the average growth rates which are also referred to the compounded annual growth rate.
- It is also used in studies like cell division and bacterial growth etc.

Example 1 : Find the G.M of the values 10, 25, 5, and 30

Solution : Given 10, 25, 5, 30

We know that,

$$GM = \sqrt[4]{10\times25\times5\times30}$$
$$= \sqrt[4]{37500}$$
$$= 13.915$$

Therefore, the geometric mean = 13.915

Example 2 : Find the geometric mean of the following data.

Weight of ear head x (g)	Log x
45	1.653
60	1.778
48	1.681
100	2.000
65	1.813
Total	8.925

Solution: Here n=5

$$GM = \text{Antilog}\ \frac{\sum_{i=1}^{n}\log(x_i)}{n}$$
$$= \text{Antilog}\ 8.925/5$$
$$= \text{Antilog}\ 1.785$$
$$= 60.95$$

Therefore the G.M of the given data is 60.95

Example 3: Find the geometric mean of the following grouped data for the frequency distribution of weights.

Weights of ear heads (g)	**No of ear heads (f)**
60-80	22
80-100	38
100-120	45

120-140	35
140-160	20
Total	**160**

Solution:

Weights of ear heads (g)	**No of ear heads (f)**	**Mid x**	**Log x**	**f log x**
60-80	22	70	1.845	40.59
80-100	38	90	1.954	74.25
100-120	45	110	2.041	91.85
120-140	35	130	2.114	73.99
140-160	20	150	2.716	43.52
Total	**160**			**324.2**

From the given data, n = 160

We know that the G.M for the grouped data is

$$GM = \text{Antilog} \frac{\sum_{i=1}^{n} \log(x_i)}{n}$$

$$GM = \text{Antilog} (324.2 / 160)$$

$$GM = \text{Antilog} (2.02625)$$

$$GM = 106.23$$

Therefore, the GM = 106.23

Example 4: What is the geometric mean of 4,8.3,9 and 17?

Solution: First, multiply the numbers together and then take the 5th root (because there are 5 numbers) = $(4 \times 8 \times 3 \times 9 \times 17)^{(1/5)}$

$= 6.81$

Example 5: What is the geometric mean of 1/2, 1/4, 1/5, 9/72 and 7/4?

Solution: First, multiply the numbers together and then take the 5th root:

GM= $(1/2 \times ¼ \times 1/5 \times 9/72 \times 7/4)^{(1/5)}$

$= 0.35.$

HARMONIC MEAN

The Harmonic Mean (HM) is defined as the reciprocal of the arithmetic mean of the given data values. It is based on all the observations, and it is rigidly defined. Harmonic mean gives less weightage to the large values and large weightage to the small values to balance the values properly. In general, the harmonic mean is used when there is a necessity to give greater weight to the smaller items. It is applied in the case of times and average rates.

Harmonic mean formula

Since the harmonic mean is the reciprocal of the arithmetic mean, the formula to define the harmonic mean "HM" is given as follows:

If $x_1, x_2, x_3, \ldots, x_n$ are the individual items up to n terms, then,

$$\text{Harmonic Mean} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots\ldots + \frac{1}{x_n}}$$

Weighted Harmonic Mean

Calculating weighted harmonic mean is similar to the simple harmonic mean. It is a special case of harmonic mean where all the weights are equal to 1. If the set of weights such as $w_1, w_2, w_3, \ldots, w_n$ connected with the sample space $x_1, x_2, x_3, \ldots, x_n$, then the weighted harmonic mean is defined by

$$\text{Weighted Harmonic Mean} = \frac{w_1 + w_2 + \ldots\ldots + w_n}{\frac{w_1}{x_1} + \frac{w_2}{x_2} + \frac{w_3}{x_3} + \ldots\ldots + \frac{w_n}{x_n}} = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{x_i}}$$

If the frequencies "f" is supposed to be the weights "w", then the harmonic mean is calculated as follows:

If $x_1, x_2, x_3, \ldots, x_n$ are n items with corresponding frequencies $f_1, f_2, f_3, \ldots, f_n$, then the weighted harmonic mean is

$$\text{Harmonic Mean} = \frac{f_1 + f_2 + \ldots\ldots + f_n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \frac{f_3}{x_3} + \ldots\ldots + \frac{f_n}{x_n}} = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} \frac{f_i}{x_i}}$$

Note:

1. f values are considered as weights
2. For continuous series, mid-value = (Lower limit + Upper limit)/2 is taken as x

Example 1:

Calculate the harmonic mean for the following data:

X	1	3	5	7	9	11
F	2	4	6	8	10	12

Solution: The calculation for the harmonic mean is shown in the below table:

X	**F**	**1/x**	**f/x**
1	2	1	2
3	4	0.333	1.333
5	6	0.2	1.2
7	8	0.143	1.143
9	10	0.111	1.111
11	12	0.091	1.091
Total	**Σ f =42**		**Σ f/x = 7.878**

The formula for weighted harmonic mean is

$$HM_w = N / [(f_1/x_1) + (f_2/x_2) + (f_3/x_3) + \ldots.(f_n/x_n)]$$

$$HM_w = 42 / 7.878$$

$$HM_w = 5.331$$

Therefore, the harmonic mean, HM_w is 5.331

Harmonic Mean Uses

The main uses of harmonic means are as follows:

- The harmonic mean is applied in the finance to the average multiples like the price-earnings ratio
- It is also used by the market technicians in order to determine the patterns like the Fibonacci Sequences.

Merits and Demerits of Harmonic Mean

Merits:

The following are the merits of the harmonic mean:

- It is rigidly confined.
- It is based on all the views of a series. It means that it cannot be computed by ignoring any item of a series.
- It is able to advance the algebraic method.
- It provides a more reliable result when the results to be achieved are the same for the various means adopted.
- It provides the highest weight to the smallest item of a series.
- It can be measured also when a series holds any negative value.
- It produces a skewed distribution a normal one.
- It produces a curve straighter than that of the A.M and G.M.

Demerits

The demerits of the harmonic series are as follows:

- The harmonic mean is greatly affected by the values of the extreme items
- It cannot be able to calculate if any of the items is zero
- The calculation of the harmonic mean is cumbersome, as it involves the calculation using the reciprocals of the number.

Relationship Between Arithmetic Mean, Geometric Mean and Harmonic Mean

The three means such as arithmetic mean, geometric mean, harmonic means are known as Pythagorean means. The formulas for three different types of means are:

Arithmetic Mean = $(x_1 + x_2 + x_3 + ... + x_n) / n$

Harmonic Mean = $\frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + ... + \frac{1}{x_n}}$

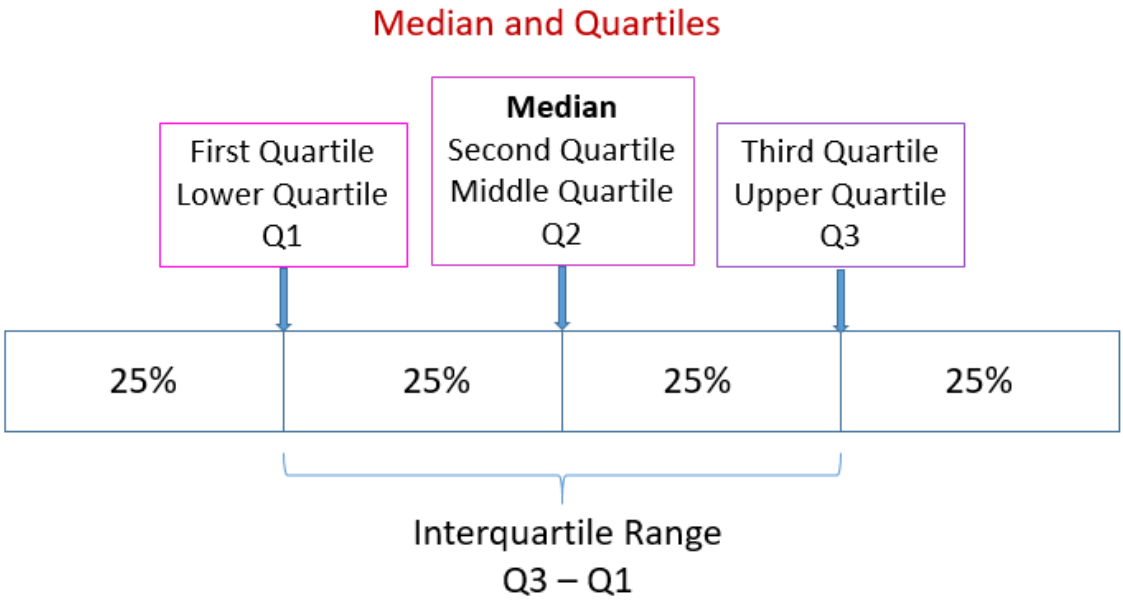
Geometric Mean = $\sqrt[n]{(x_1 + x_2 + x_3 + ... + x_n)}$

If GM is the geometric mean, HM is the harmonic mean, and AM is the arithmetic mean, then the relationship joining them is given by

$$GM^2 = AM \times HM$$

QUARTILES

Quartiles in statistics are values that divide your data into quarters. **However, quartiles aren't shaped like pizza slices;** Instead they divide your data into four segments according to where the numbers fall on the number line. The four quarters that divide a data set into quartiles are:



A quartile divides data into three points – a lower quartile, median, and upper quartile – to form four groups of the data set. The lower quartile or first quartile is denoted as Q1 and is the middle number that falls between the smallest value of the data set and the median. The second quartile, Q2, is also the median. The upper or third quartile, denoted as Q3, is the central point that lies between the median and the highest number of the distribution.

Now, we can map out the four groups formed from the quartiles. The first group of values contains the smallest number up to Q1; the second group includes Q1 to the median; the third set is the median to Q3; the fourth category comprises Q3 to the highest data point of the entire set.

Each quartile contains 25% of the total observations. Generally, the data is arranged from smallest to largest:

1. First quartile: the lowest 25% of numbers
2. Second quartile: between 25.1% and 50% (up to the median)
3. Third quartile: 51% to 75% (above the median)
4. Fourth quartile: the highest 25% of numbers

Quartile Formula

The Quartile Formula For Q1 $= \frac{1}{4}(n+1)^{th}$ term

The Quartile Formula For Q3 $= \frac{3}{4}(n+1)^{th}$ term

The Quartile Formula For Q2 $=$ Q3 – Q1 (Equivalent to Median)

Example 1: Find the median, lower quartile, upper quartile, interquartile range and range of the following numbers.

12, 5, 22, 30, 7, 36, 14, 42, 15, 53, 25, 65

Solution:

First, arrange the data in ascending order:

5, 7, 12, 14, 15, 22, 25, 30, 36, 42, 53, 65

lower quartile or first quartile — median or second quartile — upper quartile or third quartile

Lower quartile or first quartile $= \left(\frac{12+1}{4}\right)^{th}$ term $= \frac{12+14}{2} = 13$

Median or second quartile $= \left(\frac{2\times(12+1)}{4}\right)^{th}$ terms $= \frac{22+25}{2} = 23.5$

Upper quartile or third quartile = $\left(\frac{3\times(12+1)}{4}\right)^{th}$ term = $\frac{36+42}{2}=39$

Interquartile range = Upper quartile – lower quartile

= 39 – 13 = 26

Range = largest value – smallest value

= 65 – 5 = 60

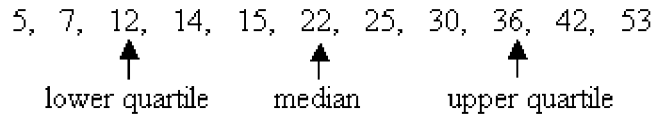
When evaluating the quartiles, always remember to first arrange the data in increasing order.

Example 2:

Find the median, lower quartile and upper quartile of the following numbers.

12, 5, 22, 30, 7, 36, 14, 42, 15, 53, 25

Solution: First, arrange the data in ascending order:



Median (middle value) = 22

Lower quartile (middle value of the lower half) = 12

Upper quartile (middle value of the upper half) = 36

PERCENTILE (OR CENTILE)

A **percentile** (or a **centile**) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations falls. For example, the 20th percentile is the value (or score) below which 20% of the observations may be found. Similarly, 80% of the observations are found above the 20th percentile.

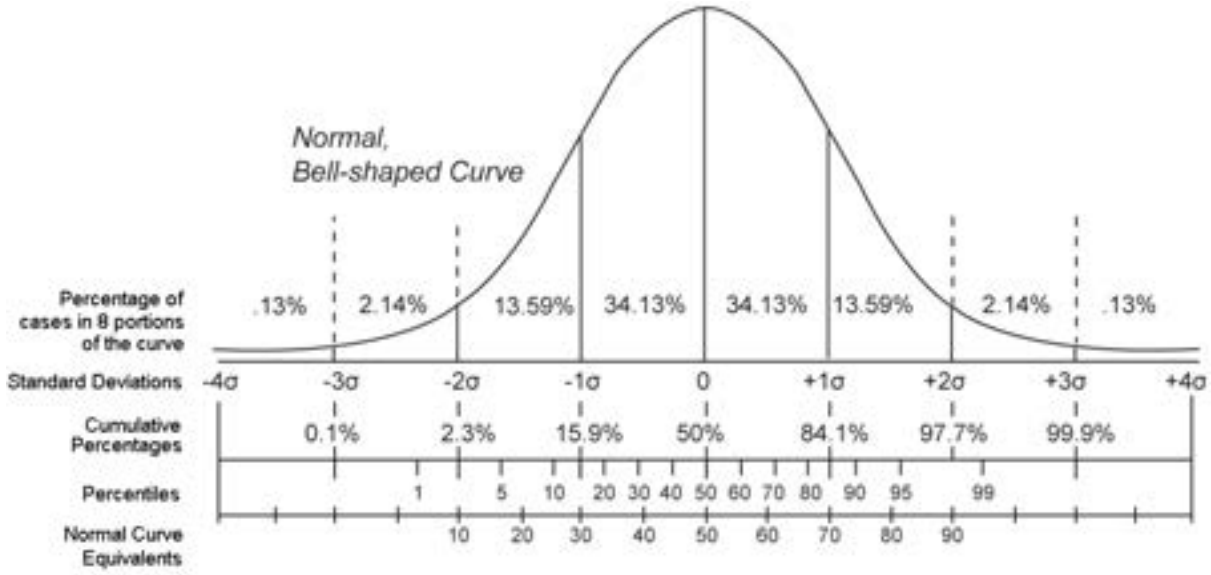
We give names to special percentiles. The 50th percentile is the median. This value splits a dataset in half.

Half the values are below the 50th percentile, and half are above it. The median is a measure of central tendency in statistics.

Quartiles are values that divide your data into quarters, and they are based on percentiles.

- The first quartile, also known as Q1 or the lower quartile, is the value of the 25th percentile. The bottom quarter of the scores fall below this value, while three-quarters fall above it.
- The second quartile, also known as Q2 or the median, is the value of the 50th percentile. Half the scores are above and half below.
- The third quartile, also known as Q3 or the upper quartile, is the value of the 75% percentile. The top quarter of the scores fall above this value, while three-quarters fall below it.

Percentiles can be calculated using the formula n = (P/100) x N, where P = percentile, N = number of values in a data set (sorted from smallest to largest), and n = ordinal rank of a given value.



UNIT -3
MEASURE OF DISPERSION

INTRODUCTION

In statistics, the measure of central tendency gives a single value that represents the whole value; however, the central tendency cannot describe the observations fully. The **measure of dispersion** helps us to study the variability of the items. In a statistical sense, dispersion has two meanings:

Ist: it measures the variation of the items among themselves, and

2nd: second, it measures the variation around the average.

OR

As the name suggests, the measure of dispersion shows the scatterings of the data. It tells the variation of the data from one another and gives a clear idea about the distribution of the data. The measure of dispersion shows the homogeneity or the heterogeneity of the distribution of the observations.

SOME STANDARD DEFINITIONS:

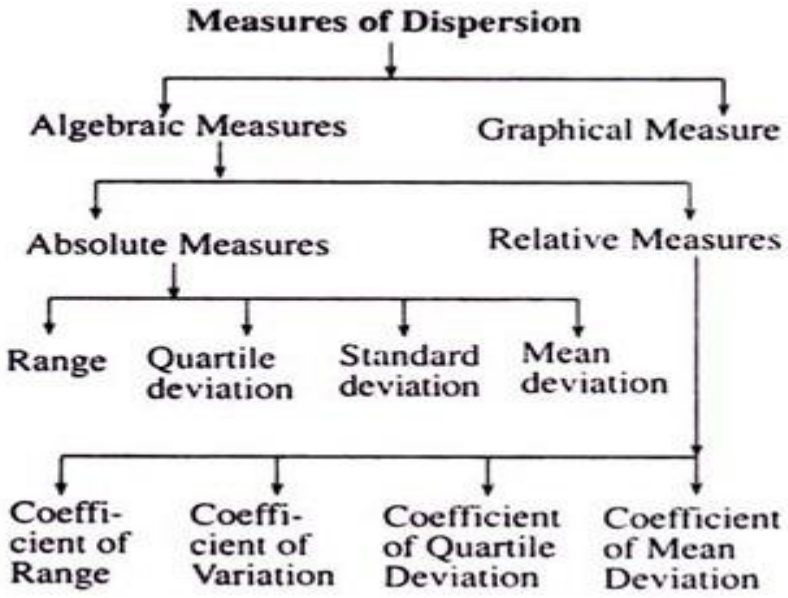
> *"Dispersion is the measure of the variation of the items."*
> *—A.L. Bowley*
>
> *"Dispersion is a measure of the extent to which the individual items vary."* *—L.R. Connor*
>
> *"Dispersion or spread is the degree of the scatter or variation of the variables about a central value."*
> *—B.C. Brooks and W.F.L. Dicks*
>
> *"The degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data."* *—Spiegal*

TYPES OF MEASURES OF DISPERSION

There are two main types of dispersion methods in statistics which are:

- Absolute Measure of Dispersion
- Relative Measure of Dispersion

As shown in figure



ABSOLUTE MEASURE OF DISPERSION

An absolute measure of dispersion contains the same unit as the original data set. Absolute dispersion method expresses the variations in terms of the average of deviations of observations like standard or means deviations. It includes range, standard deviation, quartile deviation, etc.

The types of absolute measures of dispersion are:

1. Range
2. Variance
3. Standard Deviation
4. Quartiles and Quartile Deviation
5. Mean Deviation

RELATIVE MEASURE OF DISPERSION

The relative measures of depression are used to compare the distribution of two or more data sets. This measure compares values without units. Common relative dispersion methods include:

1. Coefficient of Range
2. Coefficient of Variation
3. Coefficient of Standard Deviation
4. Coefficient of Quartile Deviation
5. Coefficient of Mean Deviation

If the difference between the value and average is high, then dispersion will be high. Otherwise it will be low.

CHARACTERISTICS OF A GOOD MEASURE OF DISPERSION

An ideal measure of dispersion is expected to possess the following properties

1. It should be rigidly defined

2. It should be based on all the items.

3. It should not be unduly affected by extreme items.

4. It should lend itself for algebraic manipulation.

5. It should be simple to understand and easy to calculate

RANGE

This is the simplest possible measure of dispersion and is defined as the difference between the largest and smallest values of the variable.

In symbols, Range = L – S.

Where L = Largest value.

S = Smallest value.

In individual observations and discrete series, L and S are easily identified.

In continuous series, the following two methods are followed.

Method 1

L = X_{max} = Upper boundary of the highest class

S = X_{min} = Lower boundary of the lowest class.

Method 2

L = Mid value of the highest class.

S = Mid value of the lowest class.

Example1 : The yields (kg per plot) of a cotton variety from five plots are 8, 9, 8, 10 and 11. Find the range?

Solution: L=11, S = 8.

Range = L – S

= 11- 8

= 3

Example 2: Calculate range from the following distribution.

Size	60-63	63-66	66-69	69-72	72-75
Number	**5**	1	42	27	8

Solution:

L = Upper boundary of the highest class = 75

S = Lower boundary of the lowest class = 60

Range = L – S

= 75 – 60

= 15

COEFFICIENT OF RANGE

The coefficient of range is defined as

$$\frac{L-S}{L+S} \text{x} 100$$

Example 3: Let us consider an example, Find out the range and the coefficient of range in the following data,

Data = 8, 5, 6, 7, 3, 2, 4

Solution:

Step 1: Find Range

Range = Maximum Value - Minimum Value

Range = 8 - 2

Range = **6**

Step 2: Find Range Coefficient

Coefficient of Range=

$$\frac{(\text{Maximum Value - Minimum Value})}{(\text{Maximum Value + Minimum Value})} \text{x}100$$

= ((8 - 2) / (8 + 2)) x100

= 6 x10

Coefficient of Range = **60**

MERITS AND DEMERITS OF RANGE

Merits

1. It is simple to understand.
2. It is easy to calculate.
3. In certain types of problems like quality control, weather forecasts, share price analysis, etc., range is most widely used.

Demerits

1. It is very much affected by the extreme items.
2. It is based on only two extreme observations.
3. It cannot be calculated from open-end class intervals.

4. It is not suitable for mathematical treatment.

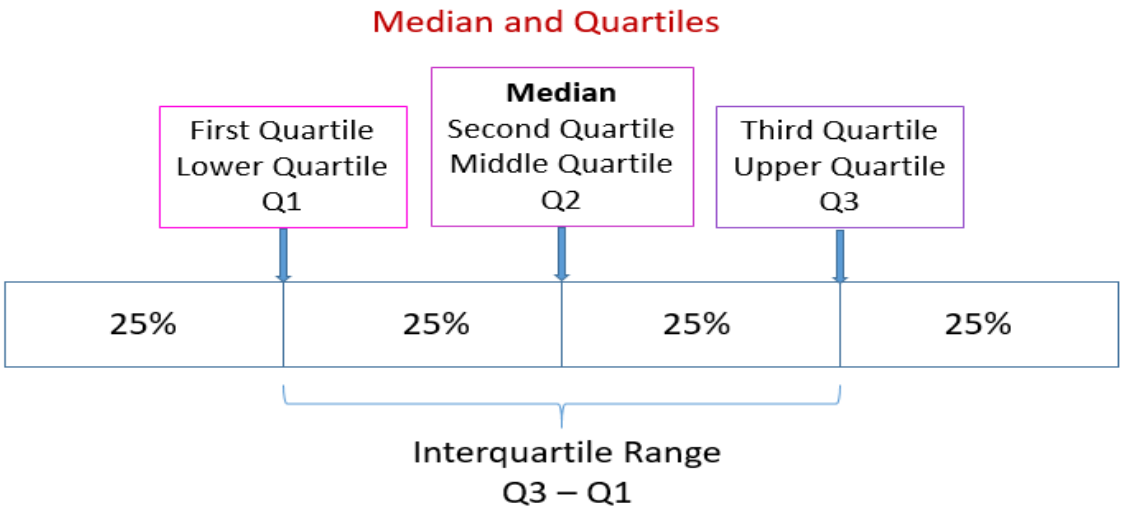
5. It is a very rarely used measure.

INTER-QUARTILE RANG

Inter-quartile range is defined as the difference between the upper and lower quartile values in a set of the values. The **inter-quartile range** (IQR) is a measure of variability, based on dividing a data set into quartiles.

The inter-quartile range is equal to Q3 minus Q1.

$$IQR = Q_3 - Q_1$$



Example 1: Consider the following numbers: 1, 3, 4, 5, 5, 6, 7, 11 and find the inter-quartile range.

Solution: Q1 = (3 + 4)/2 or Q1 = 3.5.

And , Q3 = (6 + 7)/2 or Q3 = 6.5.

Thus the inter-quartile range is Q3 minus Q1,

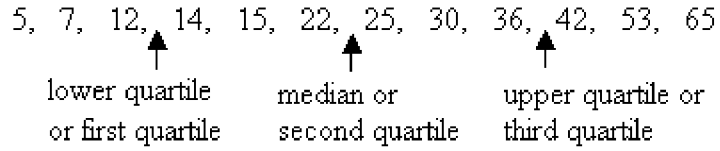
so IQR = 6.5 - 3.5 = 3.

Example 2 : Find the median, lower quartile, upper quartile, inter-quartile range and range of the following numbers.

12, 5, 22, 30, 7, 36, 14, 42, 15, 53, 25, 65

Solution:

First, arrange the data in ascending order:



Lower quartile or first quartile $= \left(\frac{12+1}{4}\right)^{th}$ term $= \frac{12+14}{2} = 13$

Median or second quartile $= \left(\frac{2\times(12+1)}{4}\right)^{th}$ terms $= \frac{22+25}{2} = 23.5$

Upper quartile or third quartile $= \left(\frac{3\times(12+1)}{4}\right)^{th}$ term $= \frac{36+42}{2} = 39$

Inter-quartile range = Upper quartile – lower quartile

= 39 – 13 = 26

Range = Largest value – small value

= 65-5 =60

THE QUARTILE DEVIATION

Formally, the Quartile Deviation is equal to the half of the Inter-Quartile Range and thus we can write it as –

$$QD = \frac{Q_3 - Q_1}{2}$$

Therefore, we also call it the ***Semi Inter-Quartile Range***.

- If the scale of the data is changed, the QD also changes in the same ratio.
- It is the best measure of dispersion for open-ended systems (which have open-ended extreme ranges).
- Also, it is less affected by sampling fluctuations in the dataset as compared to the range (another measure of dispersion).
- Since it is solely dependent on the central values in the distribution, if in any experiment, these values are abnormal or inaccurate, the result would be affected drastically.

COEFFICIENT OF QUARTILE DEVIATION

Based on the quartiles, a relative measure of dispersion, known as the Coefficient of Quartile Deviation, can be defined for any distribution. It is formally defined as :

$$\text{Coefficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100$$

Since it involves a ratio of two quantities of the same dimensions, it is unit-less.

Example 1: The number of vehicles sold by a major Toyota Showroom in a day was recorded for 10 working days. The data is given as –

Day	Frequency
1	20
2	15
3	18
4	5
5	10
6	17
7	21
8	19
9	25
10	28

Find the Quartile Deviation and its coefficient for the given discrete distribution case.

Solution: We first need to sort (arrange) the frequency data given to us before proceeding with the quartiles calculation :

Sorted Data – 5, 10, 15, 17, 18, 19, 20, 21, 25, 28

n(number of data points) = 10

Now, to find the quartiles, we use the logic that the first quartile lies halfway between the lowest value and the median; and the third quartile lies halfway between the median and the largest value.

First Quartile Q_1 =(n+1/4)th term.

= (10+1)/4th term = 2.75th term

= 2nd term + 0.75 × (3rd term – 2nd term)

= 10 + 0.75 × (15 – 10)

= 10 + 3.75

= 13.75

Third Quartile Q_3 = 3(n+1)/4th term.

= 3(10+1)4th term = 8.25th term

= 8th term + 0.25 × (9th term – 8th term)

= 21 + 0.25 × (25 – 21)

= 21 + 1

= 22

Using the values for Q1 and Q3, now we can calculate the Quartile Deviation and its coefficient as follows –

Quartile Deviation = Semi-Inter Quartile Range

= (Q3–Q1)/2

= (22–13.75)/2

=8.25/2

= 4.125

Coefficient of Quartile Deviation $= \frac{Q_3 - Q_1}{Q_3 + Q_1} 100$

$= \frac{22\text{-}13.75_1}{22+13.75_1} 100$

$= 8.2535.75 \times 100$

≈ 23.08

Example 2: Consider a data set of following numbers: 22, 12, 14, 7, 18, 16, 11, 15, 12. You are required to calculate the Quartile Deviation.

Solution: First, we need to arrange data in ascending order to find Q3 and Q1 and avoid any duplicates.

7, 11, 12, 13, 14, 15, 16, 18, 22

Calculation of Q1 can be done as follows,

Q1 = ¼ (9 + 1)

=¼ (10)

Q1=2.5 Term

Calculation of Q3 can be done as follows,

Q3=¾ (9 + 1)

=¾ (10)

Q3= 7.5 Term

Calculation of quartile deviation can be done as follows,

- Q1 is an average of 2nd which is11 and adds the product of the difference between 3rd & 4th and 0.5 which is (12-11)*0.5 = 11.50.
- Q3 is 7th term and product of 0.5 and the difference between 8th and 7th term which is (18-16)*0.5 and the result is 16 + 1 = 17.

Q.D. = (Q3 – Q1)/ 2

Using quartile deviation formula , we have (17-11.50) / 2

=5.5/2

Q.D.=2.75

Coefficient of quartile deviation is $= \frac{Q_3 - Q_1}{Q_3 + Q_1} 100$

$= \frac{5.5}{28.5} 100$ =19.29

VARIANCE

It is defined as the arithmetic mean of the Square of the deviations of the given observation from their arithmetic mean.

OR

Deduct the mean from each data in the set then squaring each of them and adding each square and finally dividing them by the total no of values in the data set is the variance. Variance

$$\sigma^2 = \frac{\sum (X_i - \bar{X})^2}{n}$$

The formula for the variance is given below.

$$\mathbf{Var(x) = (\sigma^2) = E[(X - E(X)^2)]}$$

$$\mathbf{Var(x) = (\sigma^2) = E[\ X^2 - 2X\ E(X) + (E(X))^2]}$$

$$\mathbf{Var(x) = (\sigma^2) = E(X^2) - 2\ E(X)\ E(X) + (E(X))^2}$$

OR $$\mathbf{Var(x) = (\sigma^2) = E(X^2) - (E(X))^2}$$

OR $$\mathbf{Var(x)} = \sigma^2 = \frac{\sum (X_i - \bar{X})^2}{n}$$

Variance for the Grouped Data:

There are many ways of writing the formula for the standard deviation. The one above is for a basic list of numbers. The formula for the variance when the data is grouped is as follows.

$$\mathbf{Mean} = \frac{\Sigma fx}{\Sigma f}$$

$$\mathbf{Variance,}\ \sigma^2 = \frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2$$

VARIANCE PROPERTIES

The variance, var(X) of a random variable X has the following properties.

1. Var(X + C) = Var(X), where C is a constant.
2. $Var(CX) = C^2.Var(X)$, where C is a constant.

3. $Var(aX + b) = a^2.Var(X)$, where a and b are constants.

4. If $X_1, X_2, \ldots\ldots, X_n$ are n independent random variables, then

$Var(X_1 + X_2 + \ldots\ldots + X_n) = Var(X_1) + Var(X_2) + \ldots\ldots..+Var(X_n)$.

MERITS OR DEMERITS OF VARIANCE

Merits

1. It is rigidly defined and its value is always definite and based on all the observations and the actual signs of deviations are used.

2. As it is based on arithmetic mean, it has all the merits of arithmetic mean.

3. It is the most important and widely used measure of dispersion.

4. It is possible for further algebraic treatment.

5. It is less affected by the fluctuations of sampling and hence stable.

6. It is the basis for measuring the coefficient of correlation and sampling.

Demerits

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

2. It gives more weight to extreme values because the values are squared up.

3. As it is an absolute measure of variability, it cannot be used for the purpose of comparison.

STANDARD DEVIATION

Standard Deviation It is defined as the positive square-root of the arithmetic mean of the Square of the deviations of the given observation from their arithmetic mean. The standard deviation is denoted by s in case of sample and Greek letter σ (sigma) and defined as:

$$\textbf{S.D.} = s = \sqrt{\textbf{variance}}\ .$$

STEPS INVOLVED IN PROCESS:

What the formula means:

(1) $(X_i - \bar{X})$ means take each value in turn and subtract the mean from each value.

(2) $(X_i - \bar{X})^2$ means square each of the results obtained from step (1). This is to get rid of any minus signs.

(3) $\sum(X_i - \bar{X})^2$ means add up all of the results obtained from step (2).

(4) Divide step (3) by n, which is the number of numbers

(5) For the standard deviation, square root the answer to step (4).

PROPERTIES OF STANDARD DEVIATION

1. Standard deviation is only used to measure spread or dispersion around the mean of a data set.
2. Standard deviation is never negative.
3. Standard deviation is sensitive to outliers.
4. For data with approximately the same mean, the greater the spread, the greater the standard deviation.
5. If a constant, k, is added to each number in a set of data, the mean will be increased by k and the standard deviation will be unaltered (since the spread of the data will be unchanged).
6. If the data is multiplied by the constant k, the mean and standard deviation will both be multiplied by k.

Example 1: Find the variance and standard deviation of the following numbers: 1, 3, 5, 5, 6, 7, 9, 10 .

Solution: The mean = 46/ 8 = 5.75

(Step 1): (1 - 5.75), (3 - 5.75), (5 - 5.75), (5 - 5.75), (6 - 5.75), (7 - 5.75), (9 - 5.75), (10- 5.75)

= -4.75, -2.75, -0.75, -0.75, 0.25, 1.25, 3.25, 4.25

(Step 2): 22.563, 7.563, 0.563, 0.563, 0.063, 1.563, 10.563, 18.063

(Step 3): 22.563 + 7.563 + 0.563 + 0.563 + 0.063 + 1.563 + 10.563 + 18.063

= 61.504

(Step 4): n = 8,

therefore variance = 61.504/ 8

= 7.69

(Step 5): standard deviation = 2.77

Example 2: You and your friends have just measured the heights of your dogs (in millimeters):

The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

Find out the Mean, the Variance, and the Standard Deviation.

Solution : First step is to find the Mean:

$$\begin{aligned} \text{Mean} &= \textit{600 + 470 + 170 + 430 + 300}\mathbf{5} \\ &= \textit{1970}\mathbf{5} \\ &= 394 \end{aligned}$$

so the mean (average) height is 394 mm.

To calculate the Variance, take each difference, square it, and then average the result:

Variance

$$\begin{aligned} \sigma^2 &= \textit{206}^2 + \textit{76}^2 + \textit{(−224)}^2 + \textit{36}^2 + \textit{(−94)}^2\mathbf{5} \\ &= \textit{42436 + 5776 + 50176 + 1296 + 8836}\mathbf{5} \\ &= \textit{108520}\mathbf{5} \\ &= 21704 \end{aligned}$$

So the Variance is **21,704**

And the Standard Deviation is just the square root of Variance, so:

$$\begin{aligned} \sigma &= \sqrt{21704} \\ &= 147.32... \\ &= \mathbf{147}\ \textit{(to the nearest mm)} \end{aligned}$$

Example 3: Suppose you have to calculate the standard deviation of the following values:

5, 10, 25, 30, 50

Solution:

X	$(X-\bar{X})$	$(X-\bar{X})^2$
5	-19	361
10	-14	196
25	1	1
30	6	36
50	26	676
Totoal	0	1270

Following formula is used:

therefore variance = 1270/ 5

= 254

standard deviation = 15.937

COEFFICIENT OF VARIATION

The coefficient of variation (CV) is the ratio of the standard deviation to the mean. The higher the coefficient of variation, the greater the level of dispersion around the mean. It is generally expressed as a percentage. Without units, it allows for comparison between distributions of values whose scales of measurement are not comparable.

When we are presented with estimated values, the CV relates the standard deviation of the estimate to the value of this estimate. The lower the value of the coefficient of variation, the more precise the estimate.

OR

The coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean. The coefficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from one another.

CV = (Standard Deviation (σ) / Mean (μ))

$$= \frac{\sigma}{\mu}$$

Example 1 : Calculate the relative variability (coefficient of variance) for the samples 60.25, 62.38, 65.32, 61.41, and 63.23 of a population

Solution:

Step by step calculation:

Step 1: calculate mean

Mean = (60.25 + 62.38 + 65.32 + 61.41 + 63.23)/5

= 312.59/5

= 62.51

Step 2: calculate standard deviation

$= \sqrt{((1/(5 - 1)) * (60.25 - 62.51799)^2 + (62.38 - 62.51799)^2 + (65.32 - 62.51799)^2 + (61.41 - 62.51799)^2 + (63.23 - 62.51799)^2)}$

$= \sqrt{((1/4) * (-2.26799^2 + -0.13798999^2 + 2.80201^2 + -1.10799^2 + 0.71201^2))}$

$= \sqrt{((1/4) * (5.14377 + 0.01904 + 7.85126 + 1.22764 + 0.50695))}$

$= \sqrt{3.68716}$

$\sigma = 1.92$

Step 3: calculate coefficient of variance

CV = (Standard Deviation / Mean)

= 1.92 / 62.51

= 0.03071

Example 2: Find the coefficient of the variation of the data set 10,30,20,23.

Solution :

Mean = (10 + 30 + 20 + 23)/4

= 83/4

Mean = 20.75

Standard Deviation $\sigma = \sqrt{(1/4 - 1) \times ((10 - 20.75)^2 + (30 - 20.75)^2 + (20 - 20.75)^2 + (23 - 20.75)^2)}$

$= \sqrt{(1/3) \times ((-10.75)^2 + (9.25)^2 + (-0.75)^2 + (2.25)^2)}$

= √(0.3333) x ((115.5625) + (85.5625) + (0.5625) + (5.0625))

= √(0.3333) x (206.75)

= √(68.909775)

= 8.3016

Coefficient of Variance

=8.3016/20.75

Coefficient of Variance = 0.4001

MEAN DEVIATION

To understand the dispersion of data from a measure of central tendency, we can use mean deviation. It comes as an improvement over the range. It basically measures the deviations from a value. This value is generally mean or median. Hence although mean deviation about mode can be calculated, mean deviation about mean and median are frequently used.

OR

―Average deviation is the average amount of scatter of the items in a distribution from either the mean or the median, ignoring the signs of the deviation. The average that is taken of the scatter is an arithmetic mean, which account for the fact that this measure is often called the mean-deviation". —**Clark and Schkade**

Mean deviation= Sum of absolute values of deviations from ‗average' ' The number of observations

=∑|x - Average | / n

And the coefficient of variation is defined as

COEFFICIENT OF MEAN DEVIATION

A relative measure of dispersion applying mean-deviation is given by

Coefficient of Mean Deviation =(Mean deviation/Average)x100

Mean-deviation takes its minimum value when the deviations are taken from the median.

Also mean-deviation remains unchanged due to a change of origin but changes in the same ratio due to a change in scale

i.e. if $y = a + bx$, a and b being constants,

then MD of $y = |b| \times$ MD of x

Example 1 : What is the mean-deviation about mean for the following numbers?

5, 8, 10, 10, 12, 9

Solution :

The mean is given by

$x = (5 + 8 + 10 + 10 + 12 + 9) / 6$

$x = 54 / 6$

$\mathbf{x = 9}$

Thus mean-deviation about mean is given by

$\sum|x - x| / n = 10 / 6 = 1.67$

Hence, mean-deviation for the given data is 1.67

Example 2 : Find mean-deviation about median and also the corresponding coefficient for the following observations.

82, 56, 75, 70, 52, 80, 68

Solution :

The given observations are in ascending order.

Number of observations = 7

Median = (n+1) / 2 th value

Median = (7+1) / 2 th value

Median = 8 / 2 th value

Median = 4 th value

Median = 70

Thus mean-deviation about median is given by

$\sum|x - \text{median}| / n = 61 / 7 = 8.71$

Hence, mean-deviation for the given data is 8.71

Example 3: Anubhav scored 85, 91, 88, 78, 85 for a series of exams. Calculate the mean deviation for his test scores?

Ans: Given test score; 85, 91, 88, 78, 85

Mean, $\bar{x}$ = (85+91+88+78+85)/5

= 85.4

Subtracting mean from each score;

X	$x-\bar{x}$	$\lvert x-\bar{x} \rvert$
85	-0.4	0.4
91	5.6	5.6
88	2.6	2.6
78	-7.4	7.4
85	-0.4	0.4

Mean deviation = 16.4/5 = 3.28

Example 4: Calculate the mean deviation, and the coefficient of mean deviation of the given dataset –

Marks	0 – 10	10 – 20	20 – 30	30 – 40
Number of students	2	5	1	3

Solution:

To find the coefficient of mean deviation, we first need to know the mean of the distribution. We can calculate it as –

Let us find the required values now –

Marks	x_i (mid-point of the class)	Number of students (f_i)	$f_i x_i$
0 – 10	5	2	10
10 – 20	15	5	75
20 – 30	25	1	25
30 – 40	35	3	105
Total		$\Sigma f_i = n = 11$	$\Sigma f_i x_i = 215$

Then,

$\bar{x}$ (Arithmetic Mean) = 215/11

= 19.54 marks

Now, let us analyze the data to find the mean deviation from μ.

x_i	$\lvert x-\bar{x} \rvert$	f_i	$f_i \lvert x-\bar{x} \rvert$
5	14.54	2	29.08
15	4.54	5	22.7
25	5.46	1	5.46
35	15.46	3	46.38
Total			$\Sigma f_i \lvert x-\bar{x} \rvert = 103.62$

Now, we find, the mean deviation using the mean deviation formula –

$$\text{M.D.} = \frac{\sum f_i |x-\bar{x}|}{n}$$

=103.6211/11

=9.42 marks

Also, we can find the coefficient of mean deviation (about the mean) as –

Coefficient of Mean Deviation=9.42/19.54

=0.4820

=48.2 percent

Thus, this concludes our discussion on range and mean deviation. Also, we learnt the application of mean deviation formula.

PROPERTIES OF MEAN DEVIATION:

1) Mean-deviation takes its minimum value when the deviations are taken from the median.

2) Mean-deviation remains unchanged due to a change of origin but changes in the same ratio due to a change in scale

i.e. if $y = a + bx$, a and b being constants,

then MD of $y = |b| \times$ MD of x

3) It is rigidly defined

4) It is based on all the observations and not much affected by sampling fluctuations.

5) It is difficult to comprehend and its computation

6) Furthermore, unlike SD, mean-deviation does not possess mathematical properties.

Merits or Demerits of Mean Deviation:

Merits

1. It is simple to understand and easy to compute.

2. It is based on each and every item of the data.

3. MD is less affected by the values of extreme items than the Standard deviation.

Demerits

1. The greatest drawback of this method is that algebraic signs are ignored while taking the deviations of the items.

2. It is not capable of further algebraic treatments.

3. It is much less popular as compared to standard deviation.

SKEWNESS AND KURTOSIS

A fundamental task in many statistical analyses is to characterize the location and variability of a data set. A further characterization of data includes skewness and kurtosis.

SKEWNESS

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

OR

It is the *degree of distortion* from the symmetrical bell curve or the normal distribution. It measures the lack of symmetry in data distribution. It differentiates extreme values in one versus the other tail. A symmetrical distribution will have a skewness of 0.

DEFINITIONS:

According to ***Croxton & Cowden*** –When a series is not symmetrical it is said to be asymmetrical or skewed".

According to ***Simpson & Kafka*** –Measures of skewness tell us the direction and the extent of skewness. In a symmetrical distribution the mean, median and mode are identical. The more we move away from the mode, the larger the asymmetry or skewness".

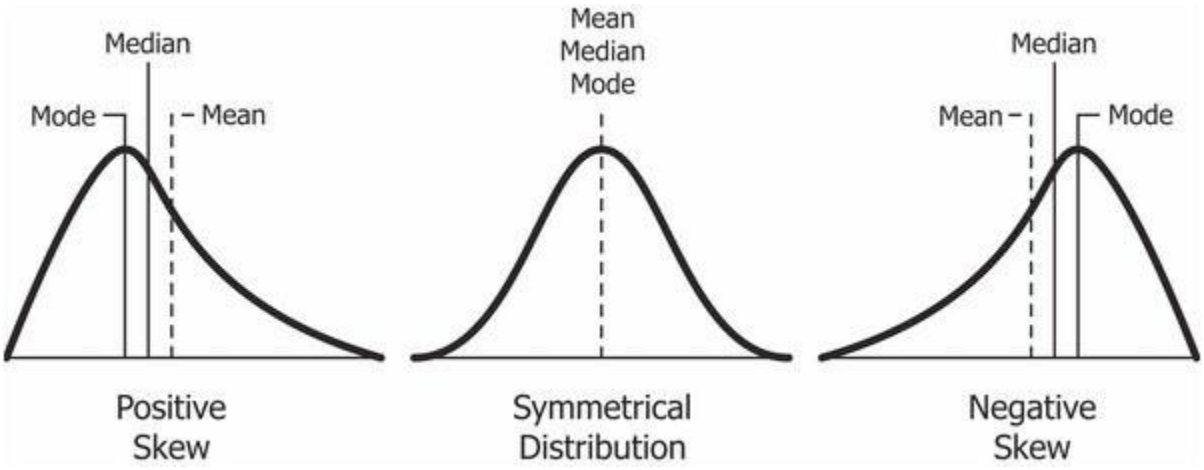
In the words of ***Riggleman and Frisbee*** –Skewness is the lack of symmetry when a frequency distribution is plotted on a chart, skewness present in the items tends to be dispersed more on one side of the mean than on the other".

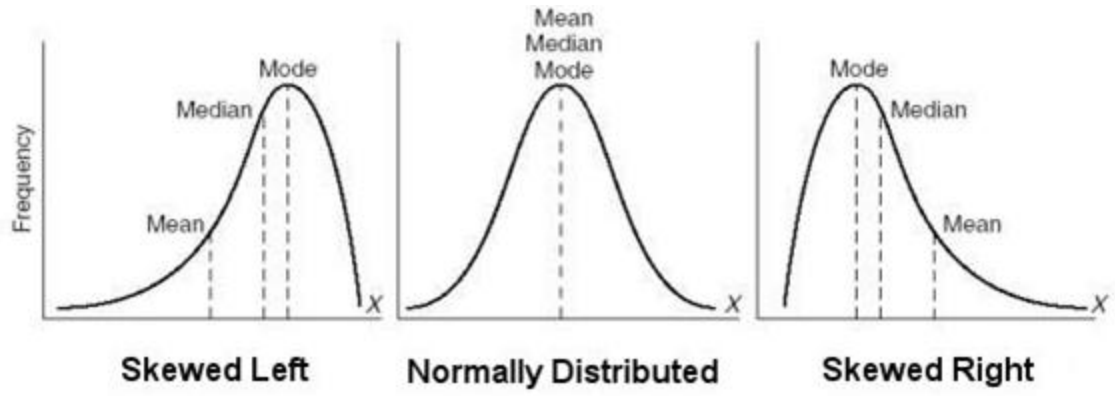
There are two types of Skewness:

1. Positive or Right Skewed
2. Negative or Left Skewed

Positive Skewness: means when the tail on the right side of the distribution is longer or fatter. The mean and median will be greater than the mode.

Negative Skewness: is when the tail of the left side of the distribution is longer or fatter than the tail on the right side. The mean and median will be less than the mode.





So, when is the skewness too much?

- If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.
- If the skewness is between -1 and -0.5(negatively skewed) or between 0.5 and 1(positively skewed), the data are moderately skewed.
- If the skewness is less than -1(negatively skewed) or greater than 1(positively skewed), the data are highly skewed.

Karl Pearson's coefficient of skewness

The first coefficient of skewness as defined by Karl Pearson is

$$\text{Coefficient of skewness} = \frac{\text{Mean}-\text{Mode}}{\text{Std.deviation}} = \frac{M-M_0}{\sigma}$$

This measure is based on the fact that the mean and the mode are drawn widely apart. Skewness will be positive if mean > mode and negative if mean < mode. There is no limit to this measure in theory and this is a slight drawback. But in practice the value given by this formula is rarely very high and its value usually lies between -1 and +1.

It may also be written as $\frac{3(\text{Mean}-\text{Median})}{\sigma}$ as Mode = 3 Median - 2 Mean

This coefficient is a pure number without units since both numerator and denominator have the same dimensions. The value of this coefficient lies between -3 and +3.

Bowley's Coefficient of Skewness

Prof. A.L. Bowley's Coefficient of Skewness is based on quartiles and is given by:

$$\text{Coefficient of Skewness} = \frac{(Q_3 - \text{Median}) - (\text{Median} - Q_1)}{Q_3 - \text{Median}) + (\text{Median} - Q_1)} = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

This is also known as Coefficient of Skewness based on quartiles and is especially useful in situations where quartiles and median are used viz.

- When the mode is ill-defined and extreme observations are present in the data.
- When the distribution has open end classes or unequal class intervals.
- This coefficient is a pure number without units since both numerator and denominator have the same dimensions. The value of this coefficient lies between -1 and +1.

Kelly's Coefficient of Skewness

The drawback of Bowley's Coefficient of Skewness is that it ignores the 50% of the data which can be partially removed by taking two deciles or percentiles equidistant from the median value. The refinement was suggested by Kelly.

$$\text{Coefficient of Skewness} = \frac{P_{90} + P_{10} - 2\,\text{Median}}{(P_{90} - P_{10})} = \frac{D_9 + D_1 - 2\text{Median}}{(D_9 - D_1)}$$

Example: Let us take a very common example of house prices. Suppose we have house values ranging from $100k to $1,000,000 with the average being $500,000.

If the peak of the distribution was left of the average value, portraying a *positive skewness* in the distribution. It would mean that many houses were being sold for less than the average value, i.e. $500k. This could be for many reasons, but we are not going to interpret those reasons here.

If the peak of the distributed data was right of the average value, that would mean a *negative skew.* This would mean that the houses were being sold for more than the average value.

KURTOSIS

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be the extreme case.

OR

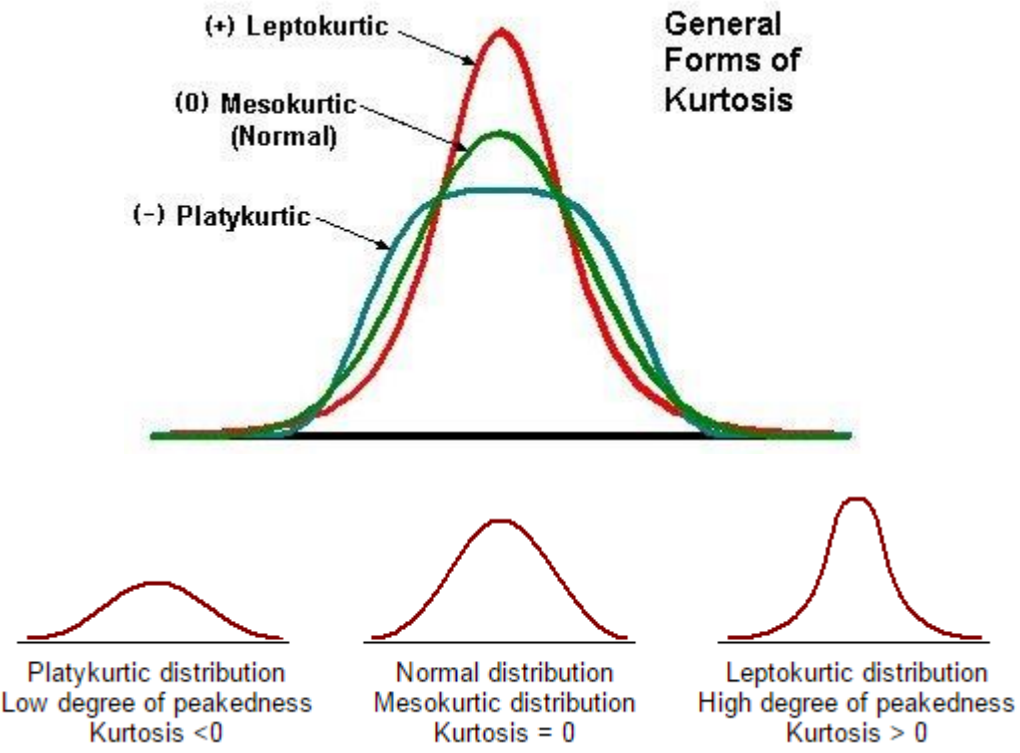
Kurtosis is all about the tails of the distribution — not the peakedness or flatness. It is used to describe the extreme values in one versus the other tail. It is actually the *measure of outliers* present in the distribution.

–Kurtosis is the degree of peakedness of a distribution" – **Wolfram MathWorld**

–We use kurtosis as a measure of peakedness (or flatness)" – **Real Statistics Using Excel**

High kurtosis in a data set is an indicator that data has heavy tails or outliers. If there is a high kurtosis, then, we need to investigate why do we have so many outliers. It indicates a lot of things, maybe wrong data entry or other things.

Low kurtosis in a data set is an indicator that data has light tails or lack of outliers. If we get low kurtosis(too good to be true), then also we need to investigate and trim the dataset of unwanted results.



Mesokurtic: This distribution has kurtosis statistic similar to that of the normal distribution. It means that the extreme values of the distribution are similar to that of a normal distribution

characteristic. This definition is used so that the standard normal distribution has a *kurtosis of three.*

Leptokurtic (*Kurtosis > 3***):** Distribution is longer, tails are fatter. Peak is higher and sharper than Mesokurtic, which means that data are heavy-tailed or profusion of outliers. Outliers stretch the horizontal axis of the histogram graph, which makes the bulk of the data appear in a narrow (~~"sk~~inny") vertical range, thereby giving the ~~"sk~~inniness" of a leptokurtic distribution.

Platykurtic: (*Kurtosis < 3***):** Distribution is shorter, tails are thinner than the normal distribution. The peak is lower and broader than Mesokurtic, which means that data are light-tailed or lack of outliers.

The reason for this is because the extreme values are less than that of the normal distribution.

COEFFICIENT OF DISPERSION

The coefficients of dispersion are calculated along with the measure of dispersion when two series are compared which differ widely in their averages. The dispersion coefficient is also used when two series with different measurement unit are compared. It is denoted as C.D.

The common coefficients of dispersion are:

C.D. In Terms of	Coefficient of dispersion
Range	C.D. = $(X_{max} - X_{min}) / (X_{max} + X_{min})$
Quartile Deviation	C.D. = $(Q3 - Q1) / (Q3 + Q1)$
Standard Deviation (S.D.)	C.D. = S.D. / Mean
Mean Deviation	C.D. = Mean deviation/Average

Chapter VI

Correlation and Regression

1. **Bivariate and Multivariate Distribution** —

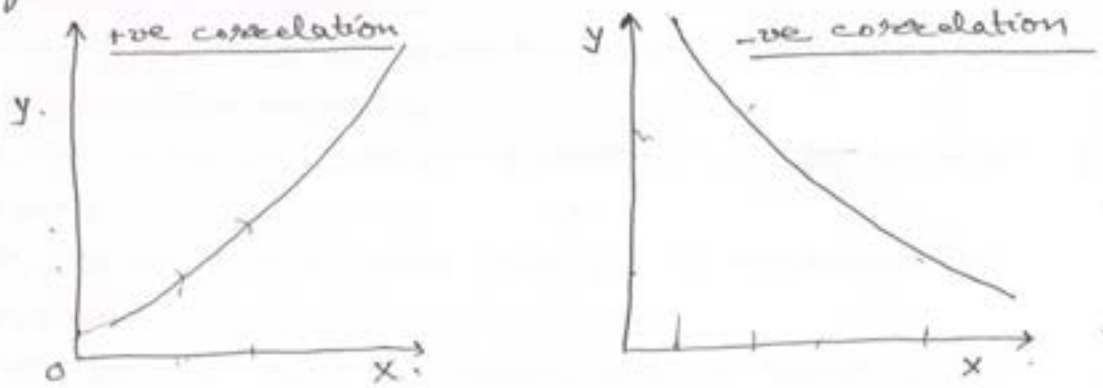
Distributions involving only one quantitive variable say X are called univariate distributions. For example, the distribution of marks obtained by a class of students in a particular subject in an examination, the distribution of incomes of a particular group of workers in a factory.

The joint distributions involving two variables say x and y are called bivariate distributions. For example, the joint distribution of the variables heights and weights of a particular group of students, joint distribution of the two variables income and expenditure of a certain group of families etc. Similarly, the joint distributions involving more than two variables are called Multivariate distributions.

Thus in case of bivariate distribution we have a pair of observations (x, y) corresponding to each unit of the group under study and in case of multivariate distribution we have more than two observations corresponding to each unit of the group.

2. **Correlation** — In a bivariate distribution one may be interested to find if there is any relationship between two variables. If the change in one variable affects a change in the other variable, the variables are called correlated.

if the increase (or decrease) in the value of one variable results the increase (or decrease) in the value of other variable i.e. the two variables deviate in the same direction, the correlation is said to be direct or positive. For example, the correlation between income and expenditure, the correlation between height and weight etc. Similarly, if two variables deviate in opposite direction i.e. the increase (or decrease) in the value of one variable results the decrease (or increase) in the value of other, the correlation is said to be inverse or negative. For example, the correlation between volume and pressure, the correlation between price and demand etc. The situations of positive and negative correlations are shown in the following figures —



The correlation between two variables is said to be linear if corresponding to a unit change in one variable, there is a constant change in the other variable over the entire range of the values.

For example, let us consider the following data :

X :	1	2	3	4	5	6	7	8
Y :	4	7	10	13	16	19	22	25

Here for a unit change in the value of X there is a constant change in the corresponding values of Y and we observe

$$y = 3x + 1$$

In general, two variables X and Y are said to be linearly correlated if there exists a relationship between X and Y of the form

$$y = ax + b.$$

In case of linear correlation, if the values of the two variables are plotted on the xy-plane, we get a straight line.

On the other hand, the correlation between two variables is said to be non-linear or curvilinear if corresponding to a unit change in one variable there is not a constant change in the other variable. For example, let us consider the following data

X :	1	2	3	4	5	6	7	8
Y :	1	4	9	16	25	36	49	64

In this case we observe $y = x^2$ which is not the equation of a straight line. Also, if the given data are plotted on the xy-plane we do not get a straight line.

3. Methods of Studing Correlation —

The various methods of measuring the correlation between two variables in case of un-grouped data are as follows —

1. Scatter or dot diagram method.
2. Karl-Pearson's Coefficient of correlation.
3. Spearman's coefficient of rank-correlation
4. Concurrent deviations method.

3.1 Scatter or dot diagram method -

It is the simplest method of the diagrammatic representation of bivariate data. Suppose we are given n pairs of values (x_i, y_i); $i = 1, 2, \ldots, n$. of two variables X and Y. Let the values of the variables X and Y be plotted along the x-axis and y-axis on a suitable scale. Then corresponding to every ordered pair there corresponds a point or dot on the xy-plane. The diagram of dots so obtained is called a dot or scatter diagram. To observe the correlation between the two variables, we note the following points -

(i) If all the points lie on a straight line and the trend is upward from left to right, we have a perfect positive correlation between two variables as shown in fig.

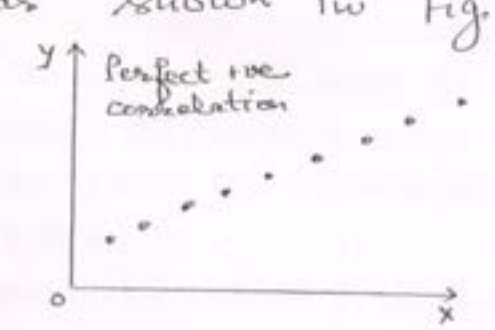
(ii) If all the points lie on a straight line and the trend is downward from left to right, we have a perfect negative correlation between the two variables as shown in fig.

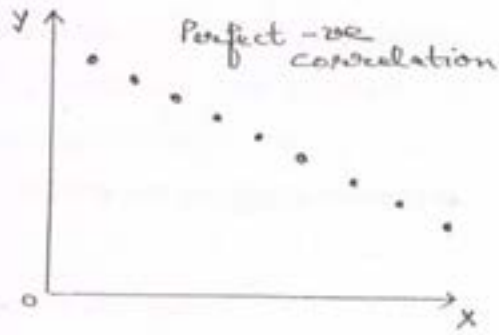
(iii) If the points are very close to each other and they reveal any trend (upward or downward) then high (positive or negative) correlation may be expected between the two

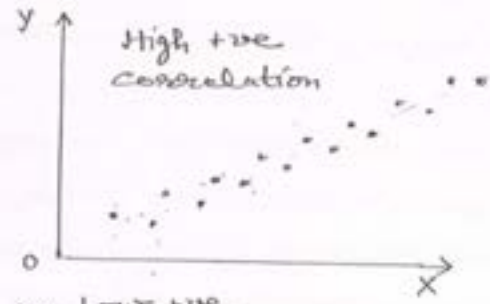
variables as shown in figs.

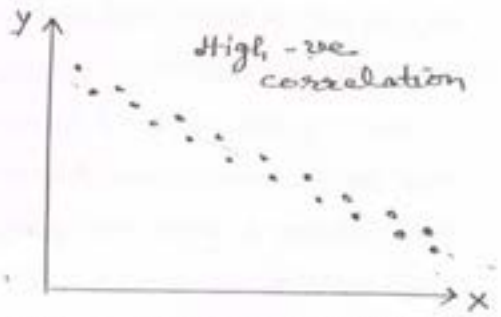
(IV) If the points are widely scattered to each other then a low (poor) positive or negative correlation may be observed according as the trend is upward or downward sides as shown in figs.

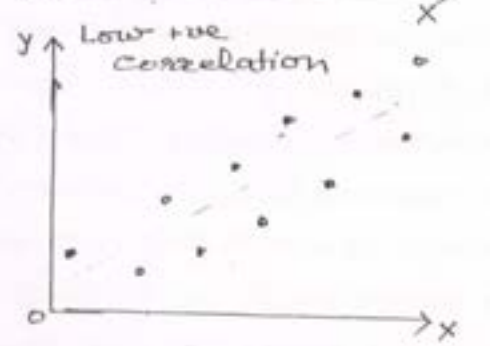
(V) If the dots do not follow any trend then the variables are said to be uncorrelated as shown in fig.

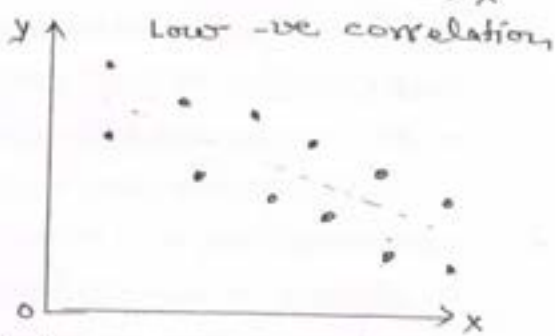


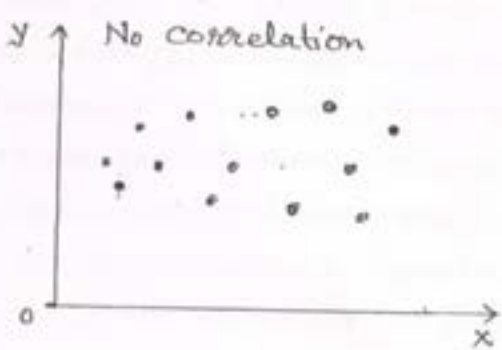


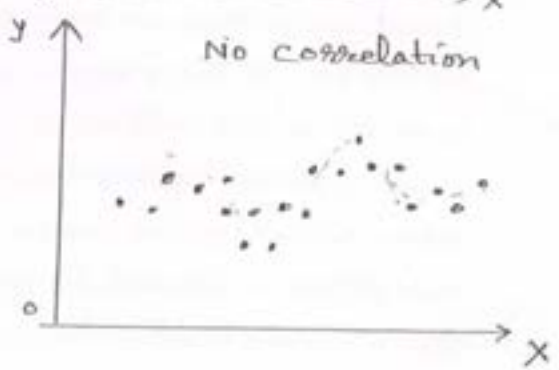












2 Karl Pearson's Coefficient of Correlation –

Correlation coefficient between two variables x and y, usually denoted by r_{xy} is a numerical unit less measure of linear relationship between them and is defined as

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

The value of correlation coefficient always lies between -1 and +1. If $r = +1$, the correlation is perfect positive. If $r = -1$, the correlation is perfect negative. $r = 0$ means no linear relationship between the two variables. If the value of r is close to 1 then the correlation is said to be high positive. Similarly, if the value of r is close to -1 then the correlation between two variables is high negative.

The coefficient of correlation is independent of the change of origin and scale.

For computation purpose, the above formula may also be written as

$$r_{xy} = \frac{\frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y}}{\sqrt{\left(\frac{1}{n}\sum_i x_i^2 - \bar{x}^2\right)\left(\frac{1}{n}\sum_i y_i^2 - \bar{y}^2\right)}}$$

Making the change of origin and scale i.e.

$$u = \frac{x - a}{h} \quad \text{and} \quad v = \frac{y - b}{k}$$

then

$$r_{xy} = r_{uv} = \frac{\frac{1}{n}\sum_i u_i v_i - \bar{u}\bar{v}}{\sqrt{\left(\frac{1}{n}\sum_i u_i^2 - \bar{u}^2\right)\left(\frac{1}{n}\sum_i v^2 - \bar{v}^2\right)}}$$

Further, if $dx = x - \bar{x}$ and $dy = y - \bar{y}$

then $r_{xy} = r_{dx\,dy} = \frac{\sum dx\,dy}{\sqrt{\sum dx^2 \sum dy^2}}$

Example -1. Calculate the coefficient of correlation for the following data:

x:	1	5	4	6	2	3	8	6	2	8
y:	2	10	8	8	5	3	13	10	6	11

Solution—

x	y	$u = x-5$	$v = y-10$	u^2	v^2	uv
1	2	-4	-8	16	64	32
5	10	0	0	0	0	0
4	8	-1	-2	1	4	2
6	8	1	-2	1	4	-2
2	5	-3	-5	9	25	15
3	3	-2	-7	4	49	14
8	13	3	3	9	9	9
6	10	1	0	1	0	0
2	6	-3	-4	9	16	12
8	11	3	1	9	1	3
		$\Sigma u_i = -5$	$\Sigma v_i = -24$	$\Sigma u_i^2 = 59$	$\Sigma v_i^2 = 172$	$\Sigma u_i v_i = 85$

$$\bar{u} = \frac{\Sigma u_i}{n} = \frac{(-5)}{10} = -0.5$$

$$\bar{v} = \frac{\Sigma v_i}{n} = \frac{(-24)}{10} = -2.4$$

Therefore

$$r_{uv} = \frac{\frac{1}{n}\Sigma u_i v_i - \bar{u}\bar{v}}{\sqrt{\left(\frac{1}{n}\Sigma u_i^2 - \bar{u}^2\right)\left(\frac{1}{n}\Sigma v_i^2 - \bar{v}^2\right)}}$$

$$= \frac{\frac{1}{10}(85) - (-0.5)(-2.4)}{\sqrt{\left[\frac{59}{10} - (-0.5)^2\right]\left[\frac{172}{10} - (-2.4)^2\right]}}$$

$$= \frac{8.5 - 1.2}{\sqrt{(5.9 - 0.25)(17.2 - 5.76)}} = \frac{7.3}{\sqrt{(5.65)(11.44)}}$$

$$= \frac{7.3}{8.04} = 0.92$$

Hence $r_{xy} = r_{uv} = 0.92$ Ans.

Example-2 Ten students got the following percentage of marks in the two subjects Mathematics and Statistics:

Marks in Maths: 78 36 98 25 75 82 90 62 65 39

Marks in Stat.: 84 51 91 60 68 62 86 58 53 47

Calculate the coefficient of correlation.

Solution – Let the r.vs x and y denote the marks in Mathematics and Statistics. Now we have the following table –

x	y	$u = x-65$	$v = y-66$	u^2	v^2	uv
78	84	13	18	169	324	234
36	51	-29	-15	841	225	435
98	91	33	25	1089	625	825
25	60	-40	-6	1600	36	240
75	68	10	2	100	4	20
82	62	17	-4	289	16	-68
90	86	25	20	625	400	500
62	58	-3	-8	9	64	24
65	53	0	-13	0	169	0
39	47	-26	-19	676	361	494

$\Sigma u_i = 0$, $\Sigma v_i = 0$, $\Sigma u_i^2 = 5398$, $\Sigma v_i^2 = 2224$, $\Sigma u_i v_i = 2704$

Therefore,

$$r_{uv} = \frac{\frac{1}{n}\Sigma u_i v_i - \bar{u}\bar{v}}{\sqrt{\left(\frac{1}{n}\Sigma u_i^2 - \bar{u}^2\right)\left(\frac{1}{n}\Sigma v_i^2 - \bar{v}^2\right)}}$$

$$= \frac{\frac{1}{10}(2704)}{\sqrt{\left(\frac{5398}{10}\right)\left(\frac{2224}{10}\right)}}$$

$$= \frac{270.4}{\sqrt{539.8 \times 222.4}} = \frac{270.4}{346.48}$$

$$= 0.78$$

Hence $r_{xy} = r_{uv} = 0.78$ Ans

Example-3 Calculate the coefficient of correlation for the following data :

X :	1	2	3	4	5	6	7	8	9	10
Y :	1	4	9	16	25	36	49	64	81	100

Solution –

x	y	$u = x-5$	$v = y-40$	u^2	v^2	uv
1	1	-4	-39	16	1521	156
2	4	-3	-36	9	1296	108
3	9	-2	-31	4	961	62
4	16	-1	-24	1	576	24
5	25	0	-15	0	225	0
6	36	1	-4	1	16	-4
7	49	2	9	4	81	18
8	64	3	24	9	576	72
9	81	4	41	16	1681	164
10	100	5	60	25	3600	300

$\Sigma u_i = 5$, $\Sigma v_i = -15$, $\Sigma u_i^2 = 85$, $\Sigma v_i^2 = 10533$, $\sum_i u_i v_i = 900$

Now

$$r_{uv} = \frac{\frac{1}{n}\Sigma u_i v_i - \bar{u}\,\bar{v}}{\sqrt{\left(\frac{1}{n}\Sigma u_i^2 - \bar{u}^2\right)\left(\frac{1}{n}\Sigma v_i^2 - \bar{v}^2\right)}}$$

$$= \frac{\frac{1}{10}\times 900 - \left(\frac{5}{10}\right)\left(\frac{-15}{10}\right)}{\sqrt{\left[\frac{85}{10} - \left(\frac{5}{10}\right)^2\right]\left[\frac{10533}{16} - \left(\frac{-15}{10}\right)^2\right]}}$$

$$= \frac{90 + 0.75}{\sqrt{(8.5 - 0.25)(1053.3 - 2.25)}} = \frac{90.75}{\sqrt{8.25 \times 1051.05}}$$

$$= \frac{90.75}{93.12} = 0.97$$

Therefore,

$r_{xy} = r_{uv} = 0.97$ Ans –

Problem-1 Calculate the coefficient of correlation for the following heights in inches of fathers (X) and their sons (Y).

X :	65	66	67	67	68	69	70	72
Y :	67	68	65	68	72	72	69	71

Ans: 0.603

Problem-2 Calculate the coefficient of correlation for the following ages of husbands (X) and their wives (Y):

X :	23	27	28	28	29	30	31	33	35	36
Y :	18	20	22	27	21	29	27	29	28	29

Ans: 0.82

Problem-3 Find the coefficient of correlation for the following table —

X :	10	14	18	22	26	30
Y :	18	12	24	6	30	36

Ans: 0.60

3.3 Calculation of Coefficient of Correlation for a bivariate frequency distribution —

If the bivariate data on two variables is presented on a two way correlation table and f is the frequency given in the table, then

$$r_{xy} = \frac{\Sigma f xy - \frac{1}{n}\Sigma f x \Sigma f y}{\sqrt{\left[\Sigma f x^2 - \frac{1}{n}(\Sigma f x)^2\right]\left[\Sigma f y^2 - \frac{1}{n}(\Sigma f y)^2\right]}}$$

Since change of origin and scale do not affect the coefficient of correlation, therefore

$$r_{xy} = r_{uv}$$

where, $u = \frac{x-a}{h}$ and $v = \frac{y-b}{k}$

a, b, h and k are arbitrary numbers.

Example-4 The following table gives according to age the frequency of marks obtained by 100 students in an intelligence test:

Age (in years) / Marks	18	19	20	21	Total
10-20	4	2	2	0	8
20-30	5	4	6	4	19
30-40	6	8	10	11	35
40-50	4	4	6	8	22
50-60	0	2	4	4	10
60-70	0	2	3	1	6
Total	19	22	31	28	100

Calculate the coefficient of correlation between age and intelligence.

Solution – Let age and intelligence be denoted by x and y respectively.

Consider $u = x - 20$ and $v = \frac{y-45}{10}$

Mid Value	y \ x	18	19	20	21	f	v	fv	fv^2	fu	fuv
15	10-20	4	2	2		8	-3	-24	72	-10	30
25	20-30	5	4	6	4	19	-2	-38	76	-10	20
35	30-40	6	8	10	11	35	-1	-35	35	-9	9
45	40-50	4	4	6	8	22	0	0	0	-4	0
55	50-60		2	4	4	10	1	10	10	2	2
65	60-70		2	3	1	6	2	12	24	-1	-2
	f	19	22	31	28	100	Total →	-75	217	-32	59
	u	-2	-1	0	1	Total ↓					
	fu	-38	-22	0	28	-32					
	fu^2	76	22	0	28	126					
	fv	-28	-16	-18	-13	-75					
	fvu	56	16	0	-13	59					

Now
$$r_{xy} = r_{uv} = \frac{\Sigma fuv - \frac{1}{n}\Sigma fu \Sigma fv}{\sqrt{\left[\Sigma fu^2 - \frac{1}{n}(\Sigma fu)^2\right]\left[\Sigma fv^2 - \frac{1}{n}(\Sigma fv)^2\right]}}$$

$$= \frac{59 - \frac{1}{100}(-32)(-75)}{\sqrt{\left[126 - \frac{1}{100}(-32)^2\right]\left[217 - \frac{1}{100}(-75)^2\right]}}$$

$$= \frac{59 - 24}{\sqrt{(126 - 10.24)(217 - 56.25)}} = \frac{35}{\sqrt{115.76 \times 160.75}}$$

$$= \frac{35}{136.41} = 0.26 \quad \text{Ans}$$

Problem-4 From the following bivariate frequency table calculate the value of correlation coefficient.

Wife's Age (Yrs) \ Husband Age (Yrs)	20 – 25	25 – 30	30 – 35	35 – 40
15 – 20	20	10	3	2
20 – 25	4	28	6	4
25 – 30	–	5	11	–
30 – 35	–	–	2	–
35 – 40	–	–	–	5

Ans: 0.613

3.4 Spearman's Coefficient of Rank Correlation

Some times we come across situations when variables under consideration can't be measure quantitatively but qualitative assesment is possible. Let a group of n individuals be assesed in respect of two characteristics say intelligence and beauty and assigned the rank to each individuals for intelligence and for beauty. It is not necessary that the most intelligent individual may be the most beautiful and vice-versa. Thus an individual who is ranked at the top for the characteristic 'intelligent' may be ranked at the

bottom for the characteristic 'beauty'. Let (x_i, y_i), $i = 1, 2, \ldots, n$ be the ranks of the n individuals in the pair for the two characteristics. Then Spearman's coefficient of rank correlation between two characteristics is given by

$$r_{xy} = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)}$$

where $d_i = x_i - y_i$

Example-5 The ranking of ten students in two subjects A and B are as follows—

A : 3 5 8 4 7 10 2 1 6 9

B : 6 4 9 8 1 2 3 10 5 7

What is the coefficient of rank correlation?

Solution — $n = 10$

Rank in A (x_i)	3	5	8	4	7	10	2	1	6	9	Total
Rank in B (y_i)	6	4	9	8	1	2	3	10	5	7	
$d_i = x - y$	-3	1	-1	-4	6	8	-1	-9	1	2	0
d^2	9	1	1	16	36	64	1	81	1	4	214

Therefore

$$r_{xy} = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 214}{10(100-1)}$$

$$= 1 - \frac{1284}{990} = 1 - 1.297$$

$$= -0.297 \quad \text{Ans}$$

Problem-5 Ten students got the following percentage of marks in two subjects Mathematics and Statistics :

Maths :	78	36	98	25	75	82	90	62	65	39
Stat. :	84	51	91	60	68	62	86	58	63	47

Calculate the rank correlation coefficient.

Ans : 0.84

Problem-6 The ranks of same 16 students in Mathematics and Physics are as follows. Two numbers within bracket denote the ranks of the students in Mathematics and Physics.

(1,1), (2,10), (3,3), (4,4), (5,5), (6,7), (7,2), (8,6)
(9,8), (10,11), (11,15), (12,9), (13,14), (14,12), (15,16), (16,13).

Calculate the rank correlation coefficient for proficiencies of this group of students in two subjects.

Ans : 0.80

3.5 Rank Correlation for repeated ranks —

If two or more individuals have same rank in the series of marks then the above formula for computing the rank correlation fails. In such cases, each individual having the same rank is given an average rank. which can be obtained by taking the average of the ranks which these individuals would have assumed if they were slightly different from each other. Now the formula for computing the rank correlation is corrected as follows —

$$r = 1 - \frac{6\left\{\sum_i d_i^2 + \sum_j \frac{m_j(m_j^2-1)}{12}\right\}}{n(n^2-1)}$$

where $m_1, m_2, \ldots$ each takes the value equal to the number of observations having equal ranks.

Example-6 The ranking of ten students marks in two subjects: Mathematics and Statistics are as follows:

Maths:	4	5	2	9	5	1	2	10	8	5
Stats:	5	7	3	10	1	6	3	9	8	2

Solution- In the series of maths two students are assigned same rank 2. These two students would have assumed the 2nd and 3rd ranks if they were got marks slightly different. Therefore, the common rank given to these two students is $\frac{2+3}{2} = 2.5$. Similarly, in this series three students are assigned same rank 5. The common rank given to these three students is $\frac{5+6+7}{3} = 6$.

In the series of stats two students are assigned same rank 3. The common rank given to these two students in stats is $\frac{3+4}{2}$ $= 3.5$. Now we have the following table -

Maths x_i	4	6	2.5	9	6	1	2.5	10	8	6	Total
Stats y_i	5	7	3.5	10	1	6	3.5	9	8	2	
$d_i = x_i - y_i$	-1	-1	-1	-1	5	-5	-1	1	0	4	0
d_i^2	1	1	1	1	25	25	1	1	0	16	72

Thus we have the three values of m say $m_1 = 2$, $m_2 = 3$ and $m_3 = 2$.

Therefore, $$\rho = 1 - \frac{6\left\{\sum_i d_i^2 + \frac{m_1(m_1^2-1)}{12} + \frac{m_2(m_2^2-1)}{12} + \frac{m_3(m_3^2-1)}{12}\right\}}{n(n^2-1)}$$

$$= 1 - \frac{6\left\{72 + \frac{2(4-1)}{12} + \frac{3(9-1)}{12} + \frac{2(4-1)}{12}\right\}}{10(100-1)}$$

$$= 1 - \frac{6\left(72 + \frac{1}{2} + 2 + \frac{1}{2}\right)}{990} = 1 - \frac{450}{990} = \frac{54}{99}$$

$$= 0.545 \quad \text{Ans}$$

Problem-7 From the following data, calculate the coefficient of rank correlation between x and y.

x :	33	56	50	65	44	38	44	50	15	26
y :	50	35	70	25	35	58	75	60	55	26

Ans: 0·076

4. Regression - Regression is the estimation or prediction of unknown values of one variable from known values of another variable.

After studying the correlation between two variables one may be interested to know the nature of relationship between the two variables. If two variables x and y are correlated, i.e. there exists an association or relationship between them, then the scatter diagram will be more or less concentrated around a curve. This curve is called the curve of regression and the relationship is said to be expressed by means of curvilinear regression. In the particular case, when the curve is a straight line, it is called a line of regression and the regression is said to be linear.

4·1 Analysis of Linear Regression - If the two variables are correlated then by the analysis of linear regression we mean to obtain a average linear relationship between two variables in terms of the original values of the data.

If the line of regression is so chosen that the sum of squares of deviations parallel to the axis of y is minimised by the method of least square (see fig. a), it is called the line of regression of y on x and it gives the best estimate of y for any given value of x.

$\Sigma (Y_i - y_i)^2$

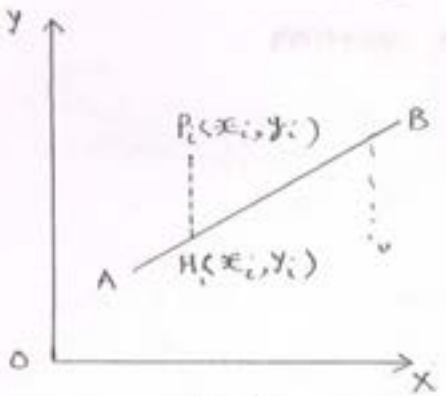


Fig. a

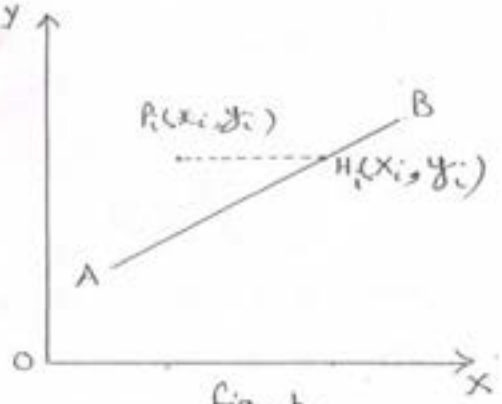


Fig. b

If the line of regression is so chosen that the sum of squares of deviations parallel to the y axis of x is minimised by the method of least square (see fig. b), it is called the line of regression of x on y and it gives the best estimate of x for any given value of y. Thus we have the two regression line namely 'y on x and 'x on y'.

Let $y = a + bx$ be the equation of the line of regression of y on x. From the fig. a we note that the observed value of variable y at the ith point is y_i and expected value of y at the ith point is Y_i. Therefore, the sum of squares of deviations of observed value of the variable y from its expected value is given by

$$S^2 = \sum_{i=1}^{n} (y_i - Y_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Now by the method of least square we minimize S^2 and find out the values of a and b as follows —

$$a = \bar{y} - \bar{x} \cdot \frac{r\sigma_y}{\sigma_x}$$

$$b = \frac{r\sigma_y}{\sigma_x}$$

So that the line of regression of y on x is

$$\boxed{y - \bar{y} = \frac{r\sigma_y}{\sigma_x}(x - \bar{x})}$$

Similarly, the equation of the line of regression of x on y is given by

$$\boxed{x - \bar{x} = \frac{r\sigma_x}{\sigma_y}(y - \bar{y})}$$

where $\bar{x}$ and σ_x are the mean and S.D. of the values given on variable x, $\bar{y}$ and σ_y are the mean and S.D. of the values given on variable y and r is the coefficient of correlation between the two variables.

Note-1 In the line of regression y on x, the term $\frac{r\sigma_y}{\sigma_x}$ is called the regression coefficient of y on x. Similarly, in the line of regression x on y the term $\frac{r\sigma_x}{\sigma_y}$ is called the regression coefficient of x on y. These are denoted by b_{yx} & b_{xy} respectively.

Note-2 If $r = 0$, the two lines of regression become $y = \bar{y}$ and $x = \bar{x}$ which are two straight lines parallel to x and y axis respectively and the point $(\bar{x}, \bar{y})$ is their intersection point.

Note-3 If $r = \pm 1$, the two lines of regression will coincide.

4.2 Properties of Regression Coefficients –

Property-1 Correlation coefficient is the geometric mean of the regression coefficients.

Property-2 If one of the regression-coefficient is greater than unity then the other must be less than unity.

Property-3 Regression coefficients are independent of the change of origin but not of change of scale.

Property-4 The regression coefficients are either both positive or both negative.

Example-7 If the two regression coefficients are 0.8 and 0.2, what would be the value of coefficient of correlation.

Solution – We know, the coefficient of correlation is the geometric mean of regression coefficients.

$$\therefore r = \pm\sqrt{0.8 \times 0.2} = \pm\sqrt{0.16}$$

$$= \pm 0.4$$

Example-8 Is the following statement correct? Give reasons.

The regression coefficient of y on x is 0.8 and that of x on y is 3.2.

Solution – We know that if one regression regression coefficient is greater than 1 than other should be less than 1. Now, we have

$$r^2 = b_{xy} \times b_{yx} = 3.2 \times 0.8$$

$$= 2.56 > 1$$

Since $r^2 > 1$, the given statement is false.

Example-9 If the two lines of regression are

$$4x - 5y + 30 = 0 \quad \text{and} \quad 20x - 9y - 107 = 0$$

which of these is the line of regression of x on y. Find b_{xy}, b_{yx} and r.

Solution – We are given the regression lines as

$$4x - 5y + 30 = 0 \qquad (i)$$

$$\text{and } 20x - 9y - 107 = 0 \qquad (ii)$$

Let (i) be the eqⁿ of the line of regression of x on y and (ii) be the eqⁿ of the line of regression of y on x.

From (i) we get

$$x = \frac{5}{4}y - \frac{30}{4} \Rightarrow b_{xy} = \frac{5}{4}$$

From (ii) we get

$$y = \frac{20}{9}x - \frac{107}{9} \Rightarrow b_{yx} = \frac{20}{9}$$

Now

$$r^2 = b_{xy} \times b_{yx}$$

$$= \frac{5}{4} \times \frac{20}{9} = \frac{25}{9} = 2.77$$

Since $r^2 > 1$, our supposition is wrong.

Hence, (i) is the line of regression of y on x and (ii) is the line of regression of x on y.

Now, writing (i) as

$$y = \frac{4}{5}x + 6 \Rightarrow b_{yx} = \frac{4}{5}$$

Similarly, writing (ii) as

$$x = \frac{9}{20}y + \frac{107}{20} \Rightarrow b_{xy} = \frac{9}{20}$$

$$\therefore r^2 = \frac{9}{20} \times \frac{4}{5} = \frac{9}{25}$$

$$\Rightarrow r = \pm \frac{3}{5} = \pm 0.6$$

Since both regression coefficients are positive, r must be positive.

Hence $r = 0.6$.

Example-10 Find the coefficient of correlation and the equations of two regression lines for the following data —

x	1	2	3	4	5
y	2	5	3	8	7

Solution — We have

$$r^2 = b_{xy} \times b_{yx}$$

where

$$b_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

and

$$b_{yx} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(y_i - \bar{y})^2}$$

CBCS Course Notes

Unit-5: PROBABILITY THEORY

Statistical experiment: Statistical experiment has following 3 common features:

1. Experiment has more than one possible outcome
2. Each outcome can be specified in advance
3. Outcome of the experiment depends on chance. Sum of probabilities for all outcome is 1.

Example: If, an experiment can have three possible outcomes (A, B, and C), Then,

$$P(A) + P(B) + P(C) = 1.$$

Example of Statistical Experiment is Tossing a coin, in which we have

1. More than one possible outcome
2. Each outcome is known in advance
3. The outcome depends on chance

Sample Space: The set of all possible outcomes for an experiment is called a sample space, and is denoted by S.

Event: An event is a collection of one or more of the outcomes of an experiment.

Types of Events:

1. Simple Event: An event that includes one and only one of the (final) outcomes for an experiment is called a simple event.

2. Compound Event: A compound event is a collection of more than one outcome for an experiment. For Ex- if a coin is tossed two times then we have

S= {HH, HT, TH, TT}. All these four events are compound events.

3. Equally Likely Events: The events are said to be equally likely if all events have equal chance of their occurrences. That is no preference is given to any event. For Ex- In the experiment of throwing a fair dice, where

A is the event of getting 1.

B is the event of getting 2.

C is the event of getting 3.

D is the event of getting 4.

E is the event of getting 5.

F is the event of getting 6.

The events A, B, C, D, E and F are equally likely. All these events have the same chance of occurrence.

4. Exhaustive Events: The events are said to be exhaustive when they are such that at least one of the events compulsorily occurs.

5. Mutually exclusive events: The events are mutually exclusive if they cannot occur at same time. For Ex.- In a coin tossing experiment, the events head and tail are mutually exclusive.

6. Disjoint Events: Mutually exclusive events are also known as disjoint events.

7. Independent Events: Two events A and B are independent if happening of the event A does not depend on the happening of the event B and vice-versa. Otherwise the events are dependent.

Set: Set is a well-defined collection of distinct objects For ex- The collection of the number 2,4,6 i.e., {2,4,6} is a set.

The basic operations that we can perform on sets are as follows:

1. Union of sets
2. Intersection of sets
3. Difference of sets

1. **Union of Sets:** The union of two sets A and B is equal to the set of elements which are present in set A, in set B, or in both the sets A and B. For ex- Suppose we have two sets A=(2,4,5,7) and B=(4,8,5,9). Then their union is given by

$A \cup B$=(2,4,5,7,8,9).

2. **Intersection of Sets:** The intersection of two sets A and B is the set which consists of all those elements which are common to both A and B. The intersection of the above two sets A and B is $A \cap B$=(4, 5).

3. **Difference of sets:** Difference of two sets A and B is the set of elements which are present in A but not in B. It is denoted as A-B. The difference of the above two sets A and b is A-B=(2, 7). Similarly B-A=(8, 9).

4. **Complement of Set:** If S is a universal set and A be any subset of S then the complement of A is the set of all members of the universal set S which are not the elements of A. For ex- If S=(1,2,3,4,5,6) and A=(1,3,5) is a subset of S. The the complement of A is $\overline{A}$ =(2,4,6).

Remark: (i) The complement of the union of the sets A and B is equal to the intersection of the complements of the sets A and B i.e, $\overline{A \cup B} = \overline{A} \cap \overline{B}$.

(ii) The complement of the intersection of two sets is equal to the union of their complements i.e, $\overline{A \cap B} = \overline{A} \cup \overline{B}$.

What is Probability: A probability is simply a number between 0 **and** 1 that measures the uncertainty of a particular event. In other words, the probability of an event refers to the likelihood that the event will occur. It is expressed as a number between 0 and 1. Probability of event A is represented by P(A).

Interpret of Probability:

Probability	Interpretation
P(A) = 0.0	Impossible event i.e., Event will almost definitely not occur.
P(A) = 0.25	25 % chance that event A occurs.
P(A) = 0.5	50 % chance that event A occurs.
P(A) = 0.75	75 % chance that event A occurs.
P(A) = 1.0	Possible Events i.e., Event A will almost definitely occur.

Problem: A coin is tossed three times. What is the probability that it lands on heads exactly one time?

Solution: Here, Sample Space S contains 2^3 =8 outcomes.

S={HHH, TTT, HTT, HTH, HHT, TTH, THT, THH}

Let A be the event that the coin lands on heads exactly one time. Then the favourable cases to the event A are 3 i.e., {HTT, TTH, THT} as these cases have exactly one head. So

P(A)=Number of favourable cases/Total number of cases

=3/8=0.375

Classical Definition of Probability: If a statistical experiment has finite number of outcomes say, n, and each outcome is equally likely. Then the probability of an event A is given by

P(A)= k/n

Where k is the number of favourable cases to an event A.

Problem: Suppose an urn contains 12 balls in which 3 are red, 4 are blue and 5 are black. One ball is randomly drawn from the urn. What is the probability that this ball is of black colour.

Solution: Let A be the event that the drawn ball is of black colour. Then

P(A)=5/12

Relative Frequency as an Approximation of Probability: If an experiment is repeated *n* times and an event *A* is observed *f* times, then, according to the relative frequency concept of probability, The probability of an even A is

P(A)=f/n

Problem: Out of the 3000 families who live in a given apartment complex in Meerut City, 600 paid no income tax last year. What is the probability that a randomly selected family from these 3000 families did pay income tax last year?

Solution: Since out of 3000 families, 600 did not pay income tax last year. So the probability that a randomly chosen family paid no income tax last year is given by 600/3000=0.2.

Subjective probability is the probability assigned to an event based on subjective judgment, experience, information and belief. It is used when the classical probability rule and the relative frequency concept of probability cannot be applied.

Conditional Probability: The probability of event A, given event B is called Conditional Probability, and is denoted by P(A|B). Similarly, P(B|A) is the conditional probability of an even B given event A.

Mathematically, P(A|B)= $\frac{P(A \cap B)}{P(B)}$

And P(B|A)= $\frac{P(A \cap B)}{P(A)}$

Example- Suppose Sample Space S=(1,2,3,5,6,7,8,10) and the events A=(3,6,5,7) and B=(1,2,5,7), then $A \cap B$=(5,7). Here, n(S)=8, n(A)=4, n(B)=4, n($A \cap B$)=2.

So P(B|A)= $\frac{P(A \cap B)}{P(A)} = \frac{2/8}{4/8} = 1/2 = 0.5$

Probability of Intersection: The probability that event A and B both occur is the probability of the intersection of A and B. It is denoted by $P(A \cap B)$.

For ex- We have two sets A=(1,3,5,6,7) and B=(3,5,9,8,12). Then $A \cap B$=(3,5). So the probability of intersection of the events A and B is

P($A \cap B$)=2/10= 0.20

Note that (i) For mutually exclusive events A and B, the probability of the interaction A and B is zero. For ex- Let us consider two events A=(1,3,5) and B=(2,4,6), then $A \cap B$=null set i.e., this set contains no event, and hence P($A \cap B$)=0.

(ii) For independent events A and B, P($A \cap B$)=P(A) P(B).

Probability of Union: The probability of the union of two events is the probability that either or both events occur. It is denoted by P($A \cup B$).

For ex- For two sets A=(1,3,5,6,7) and B=(3,5,9,8,12), we have $A \cup B$=(1,3,5,6,7,8,9,12). So the probability of union of the events A and B is

P($A \cup B$)=8/10=0.8

Note that for mutually exclusive events A and B, P($A \cup B$)=P(A)+P(B).

For ex- Let S=(1,2,3,4,5,6,7,8), A=(1,3,5) and B=(2,4,6). Then $A \cup B$=(1,2,3,4,5,6).

So P($A \cup B$)=6/8=0.75
P(A)=3/8 and P(B)=3/8
Hence P(A)+P(B)=3/8+3/8=6/8=0.75

Rules of Probability:

1. Rule of Subtraction: The probability that an event A will occur is equal to one minus the probability that event A will not occur, i.e., $P(A) = 1 - P(\bar{A})$, Here $\bar{A}$ is the complementary event of A.
For ex- Let S=(1,2,3,4,5,6), A=(1,3,5), then $\bar{A}$ =(2,4,8), so P(A)=1-P($\bar{A}$)=1-1/2=1/2.

2. Rule of Multiplication: The probability that the events A and B both occur is equal to the probability of A multiplied by the conditional probability of B given A or is equal to the probability of b multiplied by the conditional probability of A given B.
That is P($A \cap B$)=P(A) P(B|A) or
P($A \cap B$)=P(B) P(A|B).

Problem: An urn contains 4 red balls and 5 black balls. Two balls are drawn without replacement from the urn. What is the probability that both of the balls are red?

Solution: Let A = the event that the first red is red; and let B = the event that the second ball is red. Here we know that in the beginning, there are 9 balls in the urn, 4 of which are red. Therefore, P(A)=4/9

After the first draw, there are 8 ballss in the urn, 3 of which are red. Therefore,

P(B|A)=3/8

So by the rule of multiplication, P($A \cap B$)=P(A) P(B|A)=(4/9).(3/8)=1/6.

Remarks: 1. The rule of multiplication applies to the situation when we want to know the probability of the intersection of two events; that is, we want to know the probability that two events A and B both occur.

2. For two independent events A and B, the probability of the intersection A and B is equal to the probability of A multiplied by the probability of B i.e., P($A \cap B$)=P(A) P(B).

3. Rule of Addition: The probability that either event A or event B or both occur is equal to the probability of A plus the probability of B, minus the probability of the intersection of A and B i.e.,

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

Problem: A card is drawn randomly from a deck of playing cards. We win Rs. 200 if the card is a spade or a king. What is the probability that we will win the game?

Solution: There are 52 cards out of which 13 are spades and 4 are kings. So probability of getting spade P(S)=13/52, and the probability of getting king P(K)=4/52. Also the probability of getting king of spade P($S \cap K$)=1/52 as there are only one king of spade.

So the probability that we will win the game is equal to the probability that we get the card of either spade or king i.e., P($S \cup K$)=P(S)+P(K)-P($S \cap K$)

=13/52+4/52-1/52=16/52=4/13

Remarks: 1. The rule of addition applies when we want to know the probability that either event occurs.

2. For two mutually exclusive (disjoint) events A and B, P($A \cup B$)=P(A)+P(B).

Random Variable: A random variable is a rule that assign a numerical value to each outcome of a random experiment. In other words, a random variable is a measurable function defined on a probability space that maps from the sample space to the real numbers. Random variables can be discrete or continuous. Ex- the age of a randomly selected person is a random variable.

Discrete Random Variable: A discrete random variable can take countable number of possible values. For ex- the number of getting success in a certain competition in say 20 attempts; the number of T-shirts sold in a day; out of 50 students, the number of students come to the class today, etc.

Continuous Random Variable: A continuous random variable can take any value in some interval. For ex- height of a student, lifetime of a device, the time a person spends waiting in a line at the grocery shop, etc.

Probability Distributions of Random Variables:

Discrete: For a discrete random variable X, there is a probability mass function (PMF) attached to each possible values of X. Suppose X can take values from 0 to n, then we have

$P[X=k]=p_k$; k=0, 1, 2,……, n.

with $\sum_{k=0}^{n} p_k = 1.$ (sum of the total probabilities should be equal to1.)

Example: Suppose a fair coin is tossed three times, and let the random variable X denote the number of times getting head. Then we have the following events S=[HHH, TTT, HTT, HHT, HTH, THT, TTH, THH]. Thus, the probability mass function of X is

Event	[TTT]	[HTT, THT, TTH]	[HHT, HTH, THH]	[HHH]
x	0	1	2	3
P(x)=P[X=x]	1/8	3/8	3/8	1/8

Note that: The random variable is denoted by capital letter X whereas its particular value is denoted by small letter x. We can denote the possible *n* values of X by $x_1, x_2, x_3, ..., x_n.$

Continuous: Let X be a continuous random variable. Then the probability density function (PDF) of X is a function f(x) such that for any two numbers a and b, a<b, we have

$$P[a \le X \le b] = \int_a^b f(x)dx$$

= probability that the random variable X lies in the interval (a, b).

Cumulative Distribution Function (CDF): The cumulative distribution function, denoted by F(x), is defined as the cumulative probability up to the point x i.e., F(x)= $P[X \le x]$.

1. For a discrete random variable X, we have F(k)= $P[X \le k] = \sum_{x=0}^{k} p_x$.
2. For a continuous random variable X, we have F(k)= $P[X \le k] = \int_0^k f(x)\,dx$. Here we suppose that X can take values from zero to infinity.

Expectations of Random Variables: The expectation of a random variable X is defined as the expected value (mean value) of X, and is denoted by E(X).

1. For a discrete random variable X ($0 < X < \infty$), the expected value of X, if it exists, is given by E(X)= $\sum_{x=0}^{\infty} x\,P[X = x]$. Similarly, we have

   $$E(X^r) = \sum_{x=0}^{\infty} x^r\,P[X = x].$$

   If the random variable X has values $x_1, x_2, x_3, ..., x_n$ with corresponding probabilities $p_1, p_2, p_3, ..., p_n$, then we have

   E(X)= $x_1p_1, +x_2p_2 + x_3p_3 + ... + x_np_n$.
2. For a continuous random variable X ($-\infty < X < \infty$), the expectation of X is given by E(X)= $\int_{-\infty}^{\infty} x\,f(x)dx$. Similarly, we have $E(X^r) = \int_{-\infty}^{\infty} x^r\,f(x)dx$.

Variance of Random Variable: The variance of a random variable is given by

$$\begin{aligned} Var(X) &= E[X - E(X)]^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

Note: The expectation and variance of a random variable are generally denoted by μ and σ^2.

Example: Suppose a random variable X takes values 1, 2, 3, 4, 5, 6 with probabilities 0.2, 0.1, 0.3, 0.15, 0.1, 0.15 respectively. Find E(X) and Var(X).

Solution: E(X)=1x0.2+2x0.1+3x0.3+4x0.15+5x0.1+6x0.15

= 3.3

$E(X^2)$ =1x1x0.2+2x2x0.1+3x3x0.3+4x4x0.15+5x5x0.1+6x6x0.15

=13.6

Var(X)=13.6-3.3x3.3=2.71

Some Important Distributions: The following are some important discrete and continuous distributions:

Discrete Distributions:

1. Binomial Distribution
2. Poisson Distribution

Continuous Distribution:

3. Normal Distribution (Gaussian Distribution)

1. Binomial Distribution: The binomial distribution is used to model the outcomes of the random experiment with two possible outcomes (success and failure, head and tail, etc.). The experiment is repeated a fixed number of times (say, n times), each trial results in any one of the two possible outcomes (either success or failure). Each trial is assumed to be independent and the probability of getting success or failure in each trail is same.

Definition: Let a random experiment with two possible outcomes either success or failure is repeated n times. Let X denote the number of success in n trails with probability of being success p. Then the probability mass function of the random variable X is given by

$$P[X=k]=C_k^n p^k (1-p)^{n-k} \; ; k=0,1,.....,n;\; 0<p<1$$

Here, $C_k^n = \frac{n!}{k!(n-k)!}$ and $n! = n(n-1)(n-2).....2.1$

The mean and variance of X are np and np(1-p). We can write X~B(n, p), n and p are the parameters of binomial distribution.

Application Areas of Binomial Distribution:

- Taking a survey of opinions (like and dislike) from the public about the product or show/play.
- The number of male and female members working in teaching profession.
- The number of matches win by a particular team in a game.

Properties of Binomial Distribution

- There are two possible outcomes: true or false, success or failure, yes or no.
- There are 'n' number of independent trials.
- The probability of success or failure is same for each trial
- Every trial is an independent trial, which means the outcome of one trial does not affect the outcome of another trial.

Problem: Suppose a coin with probability of head ¼ and probability of tail ¾ is tossed 3 times. Then obtain (i) the probability of getting exactly 2 heads. (ii) the probability of getting at least two tails. (iii) the probability of getting at most two heads.

Solution: Let X and Y denote the number of heads and tails respectively. Given that P(H)=1/4 and P(T)=3/4.

Here the random variable X follows binomial distribution with parameters n=3 and p=1/4. The random variable Y follows binomial distribution with parameters n=3 and p=3/4. Thus we have

(i) P[getting exactly two heads]=P[X=2]

$$= C_2^3 (1/4)^2 (1-1/4)^{3-2}$$

$$= \frac{3!}{2!(3-2)!}(1/4)^2(3/4)$$

=9/64

(ii) P[getting at least two tails]= $P[Y \geq 2]$

=P[Y=2]+P[Y=3]

$$= C_2^3 (3/4)^2 (1-3/4)^{3-2} + C_3^3 (3/4)^3 (1-3/4)^{3-3}$$

= 27/64+27/6=27/32

(iii) P[getting at most two heads]= $P[X \leq 2]$ =P[X=0]+P[X=1]+P[X=2]

=1-P[X=3] (Since sum of the total probability is equal to 1.)

$$=1-C_3^3 (1/4)^3 (1-1/4)^{3-3}$$ =1-1/64=63/64

2. Poisson Distribution: This distribution is useful in studying rare events. It can also be applied in a situation where the events occur at certain points in time. The binomial distribution can be approximated by Poisson distribution when n is large and p is small with np=b. The probability mass function of Poisson distribution is

$$P[X=k]=\frac{e^{-b}\,b^{k}}{k!};\ k=0,1,2,.....,\infty.$$

P[X=k] is the probability of observing k number of events in a fixed time. The mean and variance of Poisson distribution are equal i.e., E(X)=b and Var(X)=b. Here, b is the expected number of events per unit time t.

Properties of Poisson Distribution: It has the following characteristics:

- Each occurrence are independent to each other.
- It describes discrete occurrences over an interval.
- The occurrences in each interval can be from zero to infinity.
- The mean number of occurrences must be constant throughout the experiment.

Application Areas: It is used to study:

- The number of road accidents in a day.
- The number of blowballs in a square meter plot of land.
- The number of patients arriving in an emergency room in the morning.
- The number of typing mistakes in a page.
- The number of telephone calls arrive in a system.

Problem: Ravi makes on average 2 mistake per 3 pages when he types. What is the probability to type 6 pages without making any mistake?

Solution: Here, b=$2.(6/3)$=4. Let X denote the number of mistakes per page. Then the probability of typing 6 pages without any mistake is

P[X=0]= $\frac{e^{-4}4^{0}}{0!}=e^{-4}$

Problem: The average number of goals in a Football match is approximately 1.5. What is the probability to score at most one goal in a match?

Solution: Here, b=1.5. Let X denote the number of goals in a match. Then the probability to score at most one goal is

$$P[X\le 1]=P[X=0]+P[X=1]$$

$$=\frac{e^{-1.5}1.5^{0}}{0!}+\frac{e^{-1.5}1.5^{1}}{1!}=e^{-1.5}+1.5e^{-1.5}=2.5e^{-1.5}$$

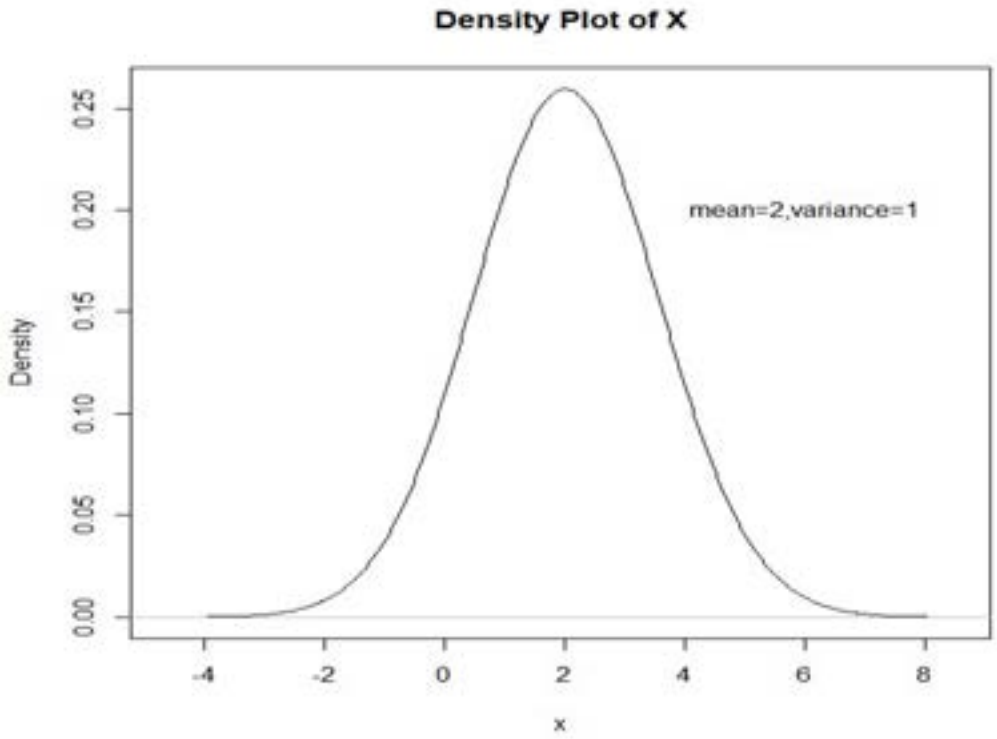
3. Normal Distribution: It is one of the important continuous distributions used in many fields including business, quality control, economics, marketing, etc. In real world, many of the variables studied is normally distributed. Most of the statistical tests

used for decision-making are based on the assumption that the data is normally distributed. The density function of normal distribution has bell-shaped curve with a single peaked. The distribution is symmetric about its mean. The normal distribution has two parameters μ (mean) and σ (standard deviation).

Definition: A random variable X is said to follow normal distribution if its probability density function is

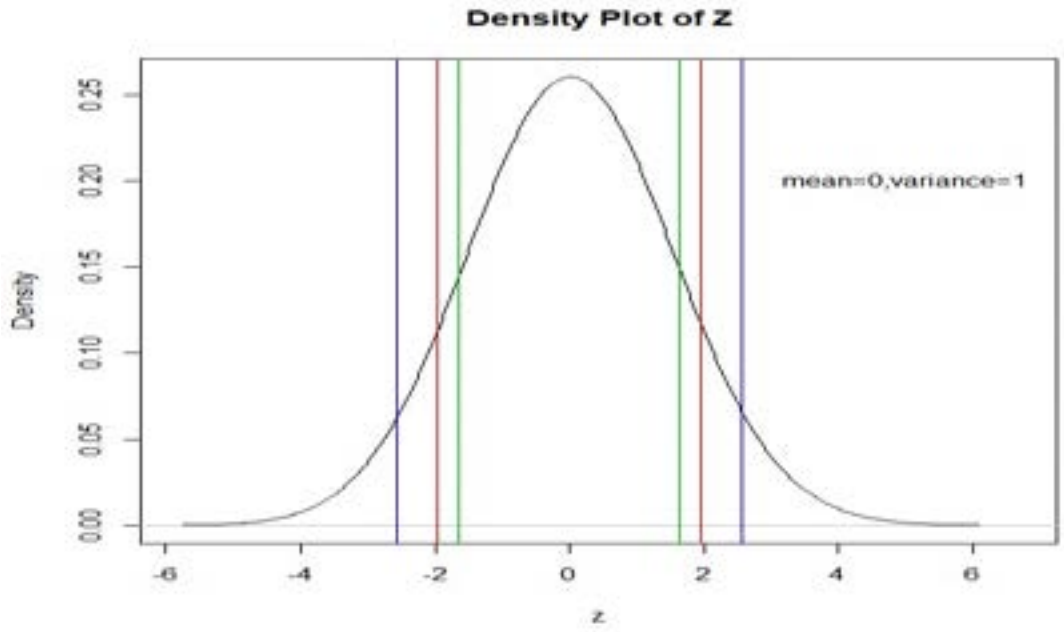
$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad ;-\infty<x<\infty,-\infty<\mu<\infty,\sigma>0.$$

Here, Mean=E(X)= μ and Var(x)= σ^2 .



Properties of Normal Distribution:

1. The normal curve has one maximum point which occur at mean.
2. Here, Mean=Median=Mode.
3. The normal variate X with mean μ and variance σ^2 can be standardized to a standard normal variate Z by making the transformation $Z=\frac{X-\mu}{\sigma}$. The random variable Z will follow normal distribution with mean zero and variance one.



Here, (i) P[-1.64<Z<1.64]=.90 (Area under green vertical lines)

(ii) P[-1.96<Z<1.96]=.95 (Area under red vertical lines)

(iii) P[-2.57<Z<2.57]=.99 (Area under blue vertical lines)

4. Let $X_1, X_2, ..., X_n$ be a random sample from normal distribution with mean μ and variance σ^2. Then for sufficiently large n, the sample mean $\bar{X}$ will follow normal distribution with mean μ and variance σ^2 / n.
5. Mean deviation = 4/5 standard deviation.
6. Quartile deviation = 5/6 mean deviation.
7. The normal distribution is also called a distribution of error.