INTRODUCTION OF STATISTICS UNIT-2

Statistics:—

→ Statistic is the science of collection, organization, presentation, analysis and interpretation of numerical data.

→ Father of Modern Statistic is "Prof. Ronald Alfred fisher."

→ Father of Indian statistic is "Prof. Prasanta Chandra Mohalanobis."

Data:—

It means numerical information about a particular inquiry.

Measures of central tendency / Average:—

It is a single value which represents the entire distribution. In other words it gives the idea about the central part of the distribution.

Characteristics of Average:—

→ It should be rigidly defined.

→ It should be based on all the observation.

→ It should be capable of further mathematical treatment. → It is easy to calculate and understand.

→ It should not be affected by extreme items.

Various Measures of Central Tendency:—

There are different types of average. The most commonly used average are

1. Arithmatic Mean (A.M)
2. Median
3. Mode
4. Geometric Mean (G.M)
5. Harmonic Mean (H.M)

Variable :—

The quantitative phenomenon under study like marks in a test, height, weight, number of student in a class etc. is known as variable.

→ There are two types of variables.

1. Discrete variable
2. Continuous variable

Discrete variable :—

Those variables which can not take all possible values within a given specified range is known as Discrete variable.

EX:— Marks in a test; number of student in a class

Continuous variable :—

Those variables which can take all possible values with in a given specified range is known as continuous variable.

EX:— Age, height, weight etc.

Frequency :—

The number of times an item occur in a series.

EX:— 18, 19, 20, 18, 19, 30

Age	Frequency
18	2
19	2
20	1
30	1

Frequency Distribution:—

→ The primary data are generally in raw form and are produced in an unorganised manner and it leads to be organized and processed.

→ Frequency distribution is one such technique to process the raw data.

→ Frequency distribution is a least table or graph that displays the frequency of various outcomes in a sample.

→ Each entry in the table contains the frequency of the occurances of values within a particular group or interval.

→ There are different types of frequency distribution.

1. Discrete / ungrouped frequency distribution
2. Grouped frequency distribution
3. Continuous frequency distribution

1. Discrete / ungrouped frequency Distribution:—

In discrete frequency distribution the values of the variable are determined individually.

EX:—

Age	frequency
18	2
19	2
20	2
30	1

2. Grouped Frequency Distribution:—

In grouped frequency distribution the data are classified in to different class intervals with gaps and their respective frequency are assigned as the class intervals.

Ex:—

Age	No. of Student
1-9	6
10-19	5
20-29	3
30-39	4

3. Continuous Frequency Distribution:—

In Continuous frequency Distribution, the data are classified in to different class intervals without gaps and their respective frequency are assigned as the class intervals.

Ex:—

Age	No. of student
10-20	10
20-30	5
30-40	3
40-50	4

Class interval:—

→ Class interval is a numerical width of a class in a frequency Distribution each class is specified by two extreme values is called as class limits.

→ The smaller one is known as Lower limit & larger one is known as upper limit.

→ There are two types of class interval.

1. Inclusive
2. Exclusive

1. Inclusive:—

The classes of the type like 1-9, 10-19, 20-29 etc. in which both Lower and upper limits are included

in the class are known as Inclusive class interval.

2. Exclusive:—

The classes of the type like 10-20, 20-30, 30-40 etc. in which the Lower limit is included and the Upper limit is excluded from the respective classes and immediately included in the next class is known as Exclusive.

Mid value:—

The mid value of any class is obtained on dividing the sum of the upper and lower limits.

$$\text{Mid value} = \frac{\text{Lower Limit} + \text{Upper Limit}}{2}$$

Ex:—

Age	Mid value
10-20	15
20-30	25
30-40	35

Class Boundaries:—

→ In Inclusive type classification or grouped frequency distribution there are gap between the upper limit of any class and lower limit of the succeeding class. So, there is need to convert the data in to a continuous frequency Distribution.

→ The upper and Lower limit of the new exclusive type classes are called as class boundaries.

→ If d is the gap between the upper limit of the any class & lower limit of the succeeding class, the class Boundaries of any class is given by

Upper class Boundary = Upper class Limit + $\frac{d}{2}$

Lower class Boundary = Lower class limit − $\frac{d}{2}$

($\frac{d}{2}$ is the correction factor)

EX:—

Age	New class Interval
1-9	1-0.5 − 9+0.5 = 0.5-9.5
10-19	10-0.5 − 19+0.5 = 9.5-19.5
20-29	20-0.5 − 29+0.5 = 19.5-29.5
30-39	30-0.5 − 39+0.5 = 29.5-39.5

Cumulative frequency Distribution:—

Cumulative frequency Distribution is obtained on successively adding the frequencies the values of the variables.

EX:—

Age	Frequency	cumulative frequency (C.F)
0-10	5	5
10-20	3	8
20-30	2	10
30-40	1	11
40-50	2	13

Arithmetic Mean (A.M):—

→ It is defined by sum of the observations divided by the total no. of observation.

→ Generally A.M is denoted as ($\bar{X}$)

So, Mathematically we can write if a series consist of n observation, say $X_1, X_2, \ldots\ldots\ldots X_n$,

→ The Arithmetic Mean ($\bar{x}$) is given by

$$\bar{x} = \frac{\text{Sum of the observation}}{\text{Total no. of observation}} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\boxed{\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}}$$

→ In case of Discrete Frequency Distribution: If a series consist of n observation say $x_1, x_2, \ldots x_n$ & the corresponding frequency is $f_1, f_2, \ldots f_n$.

The Arithmetic Mean is given by

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \cdots + f_nx_n}{f_1 + f_2 + \cdots + f_n}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

$$\boxed{\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N}}, \left(\text{where } N = \sum_{i=1}^{n} f_i\right)$$

→ In case of continuous frequency Distribution x is taken as the mid value of corresponding classes

Example-1 :— (Discrete series)

Calculate the Arithmetic Mean the wages of 100 workers of a firm form the following frequency distribution.

Wages (in Rs.) (X)	No. of workers Frequency (f)	fX
80	5	400
82	8	656
85	17	1445
90	30	2700
95	18	1710
98	11	1078
100	8	800
110	3	330
	$\Sigma f = 100$	$\Sigma fX = 9119$

$$\bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{9119}{100} = \text{Rs. } 91.19$$

Example-2 :— (Continuous series)

The marks of 200 school student is given in the following frequency distribution calculate the average Marck.

Marcks	No. of Students (f)	New Marcks	Mid value (X)	fX
11-20	8	10.5-20.5	15.5	124
21-30	12	20.5-30.5	25.5	306
31-40	35	30.5-40.5	35.5	1242.5
41-50	63	40.5-50.5	45.5	2866.5
51-60	45	50.5-60.5	55.5	2497.5
61-70	27	60.5-70.5	65.5	1768.5
71-80	10	70.5-80.5	75.5	755
	$\Sigma f = 200$			$\Sigma fX = 9560$

$$\bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{9560}{200} = 47.8 \text{ Marks}$$

Example-3:—

Calculate the Arithmetic Mean of Marks of 10 students are 50, 60, 52, 70, 57, 52, 53, 63, 62, 70.

Ans:— $\bar{X} = \frac{\text{sum of observation}}{\text{Total no. of observation}}$

$$= \frac{50+60+52+70+57+52+53+63+62+70}{10}$$

$$= \frac{589}{10} = 58.9$$

Change of origin and scale:—

→ When the values of the variable x in the data are large or the value involves decimal figures, the computation of arithmetic mean becomes time consuming and tedious (difficult)

→ So, the computation can be made simple by setting the origin to a convenient point and reducing the magnitude of the figures by changing the Shortcut method or assumed Mean Method or step deviation method.

→ It is given by (Individual Series)

$$\bar{X} = A + h\bar{d}$$

where, $\bar{d} = \frac{\Sigma fd}{\Sigma f}$

where, $d = \frac{X-A}{h}$

A is any Arbitary value

h is the magnitude of the Class Interval.

Note:-1 (Continuous series)

If we take $d = x - A$

So, A.M is given by $\bar{x} = A + \bar{d}$

then, $\bar{d} = \frac{\Sigma d}{n}$ (n = no. of observation).

Example :—

Find the average of the salary 210 persons of a firm given in the following frequency distribution.

Monthly salary in Rs.	No. of person (f)	Mid value (x)	$d = \frac{x-A}{h}$	fd
3000 - 3999	15	3499.5	-3	-45
4000 - 4999	13	4499.5	-2	-26
5000 - 5999	46	5499.5	-1	-46
6000 - 6999	67	6499.5	0	0
7000 - 7999	32	7499.5	1	32
8000 - 8999	23	8499.5	2	46
9000 - 9999	14	9499.5	3	42
	$n = 210$			$\Sigma fd = 3$

$A = 6499.5$

$h = 1000$

$\bar{x} = A + h \cdot \frac{\Sigma fd}{\Sigma f}$

$= 6499.5 + 1000 \cdot \frac{3}{210}$

$= 6513.78$

Discrete series:—

$d = x - A$

$\bar{x} = A + \frac{\Sigma fd}{\Sigma f}$

$N = \Sigma f$

Properties of arithmetic mean:—

① The sum of the deviations of set of values from their arithmetic mean is set to zero.

So, mathematically we can write

$$\Sigma f_i (x_i - \bar{x}) = 0$$

OR $$\Sigma (x_i - \bar{x}) = 0$$

Proof:—

$$L.H.S = \Sigma f_i (x_i - \bar{x})$$

$$= \Sigma f_i x_i - \bar{x} \Sigma f_i$$

$$= \bar{x} \Sigma f_i - \bar{x} \Sigma f_i$$

$$= N\bar{x} - N\bar{x}$$

$$= 0 = R.H.S$$

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}$$

$$\Rightarrow \Sigma f_i \bar{x} = \Sigma f_i x_i$$

$$(N\bar{x})$$

② Let $\bar{x}_1$ be the arithmetic mean of a set of data having small n_1 observations & $\bar{x}_2$ be the arithmetic Mean of another set of data having n_2 observations the arithmetic mean of the combine data of $(n_1 + n_2)$ observations is given by

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

Example:—

① In a class of 2 sections A & B the average of the marks in statistics of the students of section A is 62 & of section B is 60 if the no. of student of section A & B is 80 & 70 find the average marks of all 150 students taken together.

Ans:—

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

$$= \frac{62 \times 80 + 60 \times 70}{150} = \frac{4960 + 4200}{150} = \frac{9160}{150} = 61.066$$

(2) The average weekly wages of all the workers in a factory in Rs-600 & the average weekly wages of male & female workers are Rs-630 & Rs-530. find the percentages of male & female workers in a factory.

Ans:— Given $\overline{X} = 600$

$\overline{X_1} = 630$

$\overline{X_2} = 530$

$$\overline{X} = \frac{n_1\overline{X_1} + n_2\overline{X_2}}{n_1 + n_2}$$

$$\Rightarrow 600 = \frac{n_1 \times 630 + n_2 \times 530}{n_1 + n_2}$$

$$\Rightarrow 600n_1 + 600n_2 = 630n_1 + 530n_2$$

$$\Rightarrow 600n_1 - 630n_1 = 530n_2 - 600n_2$$

$$\Rightarrow -30n_1 = -70n_2$$

$$\Rightarrow \frac{n_1}{n_2} = \frac{70}{30} = \frac{7}{3}$$

% of male workers $= \frac{7}{7+3} \times 100 = 70\%$

% of female workers $= \frac{3}{7+3} \times 100 = 30\%$

Proof :—

We know that if $\overline{X}$ is the mean of n observation,

$$\overline{X} = \frac{\Sigma X}{n}$$

$\Rightarrow \Sigma X = n\overline{X}$ (i.e the sum of n observations $= n \times A.M$)

If $\bar{x}_1$ is the mean of n_1 observations of the 1st group.

The sum of n_1 observations in the 1st group $= n_1\bar{x}_1$

similarly, $\bar{x}_2$ is the mean of n_2 observations of the 2nd group.

The sum of n_1+n_2 observations of the combine group $= n_1\bar{x}_1 + n_2\bar{x}_2$

Hence, the mean $\bar{x}$ of the combine group n_1+n_2 observation is given by

$$\bar{x} = \frac{\text{sum of } (n_1+n_2) \text{ observation}}{n_1+n_2}$$

$$\boxed{\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2}}$$ (proved)

(3) The sum of the squares of deviation of the given set of observations is minimum, when taken from Arithmetic mean.

Mathematically, we can write for a giving frequency distribution the sum,

$$\boxed{S = \Sigma F (X-A)^2}$$

which represents the sum of the squares of deviation of given observation from any arbitary value 'A' is minimum when we $\boxed{A=\bar{x}}$

Proof (By using the Principle of maxima and minima in differential calculus). :—

For S to be minimum,

$$\frac{dS}{dA} = 0 \quad \& \quad \frac{d^2S}{dA^2} > 0$$

Differentiating 'S' w.r.t A, we get

$$\frac{ds}{dA} = \frac{d}{dA}[\Sigma f(X-A)^2]$$

$$= 2\Sigma f(X-A)\cdot\frac{d}{dA}(X-A)$$

$$= 2\Sigma f(X-A)(-1)$$

$$= -2\Sigma f(X-A) \quad \text{——— (1)}$$

$$\frac{ds}{dA} = 0 \Rightarrow \Sigma f(X-A) = 0$$

$$\Rightarrow \Sigma fX - \Sigma fA = 0$$

$$\Rightarrow \Sigma fX - A\Sigma f = 0$$

$$\Rightarrow \Sigma fX - AN = 0$$

$$\Rightarrow \Sigma fX = NA$$

$$\Rightarrow \frac{\Sigma fX}{N} = A$$

$$\Rightarrow \boxed{\bar{X} = A}$$

Differentiating eq^n (1) w.r.t, A, we get

$$\frac{d^2S}{dA^2} = 2\Sigma f$$

$$= 2N > 0 \quad \text{(It is always positive)}$$

(Proved)

Weighted Arithmetic Mean:—

→ For computing Arithmetic Mean are based on the assumption their all the items in the distribution are of equal importance. But in real life it can not happen.

→ In such cases Proper weightage is should be given to various items. The weights attach to each items bring proportional to the important of the item in the distribution.

→ Let $w_1, w_2, w_3 \ldots\ldots w_n$ be the weights attached to variables say $x_1, x_2, x_3 \ldots\ldots x_n$.

The weighted Arithmetic mean usually denoted by $\bar{x}_w$ is given by

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + w_nx_n}{w_1 + w_2 + \ldots + w_n}$$

$$= \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Example:—

① A candidate obtained the following Percentages of marks in an examination.

English: 63, Hindi: 75, Math: 63, Phy: 59, Chem: 55

Find the weighted Arithmetic Mean if weights 1, 2, 1, 3 & 3 repectively are allocated in the subject.

Sub	x (% of marks)	weights (w)	w·x
Eng	60	1	60
Hin	75	2	150
Math	63	1	63
Phy	59	3	177
Chem	55	3	165
		$\Sigma w = 10$	$\Sigma wx = 615$

$$\bar{x}_w = \frac{\Sigma wx}{\Sigma w} = \frac{615}{10} = 61.5$$

② The no.s 3.2, 5.8, 7.9, 4.5 & have frequency $x, x+2, x-3, x+6$ if the arithmetic mean is 4.876 · 50, find the value of x.

N10.(x)	f	fx
3·2	x	3·2x
5·8	x+2	5·8x + 11·6
7·9	x-3	7·9x - 23·7
4·5	x+6	4·5x + 27
	$\Sigma f = 4x+5$	

$$\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{3\cdot2x + 5\cdot8x + 11\cdot6 + 7\cdot9x - 23\cdot7 + 4\cdot5x + 27}{4x+5}$$

$$\Rightarrow 4\cdot876 = \frac{21\cdot4x + 14\cdot9}{4x+5}$$

$$\Rightarrow 19\cdot504x + 24\cdot38 = 21\cdot4x + 14\cdot9$$

$$\Rightarrow 19\cdot504x - 21\cdot4x = 14\cdot9 - 24\cdot38$$

$$\Rightarrow -1\cdot896x = -9\cdot48$$

$$\Rightarrow x = \frac{9\cdot48}{1\cdot896} = 5$$

Merits of Arithmetic Mean :—

(i) It should be rigidly define

(ii) It should be easy to understand and simple to calculate.

(iii) It should be based on all the observation.

(iv) It should be suitable for further mathematical treatment.

(v) It should not be affected much by extreme observation.

Demerits of Arithmetic Mean:—

(i) The strongest drawback of arithmetic mean is that it is very much affected by extreme observations. Two or three very large values of the variable may unduly affect the value of the arithmetic mean.

(ii) It can not be determined by inspection nor can it be located graphically.

(iii) Arithmetic mean can not be used if we are dealing with qualitative characteristics which can not be measured quantitatively such as intelligence, honesty etc.

(iv) Arithmetic mean cannot be obtained if a single observation is missing or lost or is illegible unless we drop it out and compute the arithmetic means of the remaining values.

(v) In extremely asymmetrical distribution, usually arithmetic mean is not representative of the distribution and hence is not a suitable measure of location.

(vi) Arithmetic mean may lead to wrong conclusion if the details of the data from which it is obtained are not available. In this connection it is worthwhile to quote the words of H. secrist:

"If an average is taken as a substitute for the details, then the arithmetic mean, in spite of the simplicity and ease of calculation, has little to recommend when series are non-homogeneous."

Median:—

Median is the middle most observation which divide the series into two equal parts. It is the value such that the no. of observation above it and no. of observations below it. The median is only positional average that is its value depends on the position occupied by a value in the frequency distribution.

Calculation of Median:— (Individual series)

→ If the no. of observation is odd then the median is $(n+1/2)^{th}$ item.

→ If the no. of observation is even then the median is obtain by the A.M of two middle observations after they are arranged in either ascending or descending order of magnitude.

Ex:—

Find the median of 5 observations 35, 12, 40, 8, 60.

$n = 5$ Descending order of 5 observations

8, 12, 35, 40, 60.

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

$$= \left(\frac{5+1}{2}\right)^{th}$$

$$= \left(\frac{6}{2}\right)^{th}$$

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

Find the median of 6 observations 60, 40, 12, 8, 35, 50

$n = 6$ Descending order of 6 observations

8, 12, 35, 40, 50, 60

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

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

$$= \frac{3^{rd} \text{ item} + 4^{th} \text{ item}}{2}$$

$$= \frac{35+40}{2}$$

$$= \frac{75}{2} = 37.5$$

∴ Median is 37.5.

* If the no. of observation is odd then median is $\left(\frac{n+1}{2}\right)^{th}$ item. After the observations having arranged in ascending or descending order of magnitude.

* In case of even no. of observations median is the arithmetic mean of $\left(\frac{n}{2}\right)^{th}$ observation & $\left(\frac{n}{2}+1\right)^{th}$ observation.

Discrete series:—

→ If the series consist of 'n' observation $x_1, x_2, \ldots, x_n$ the corresponding frequency are $f_1, f_2, \ldots, f_n$ $N = \Sigma f_i$ the median is the size of the capital $\left(\frac{N+1}{2}\right)^{th}$ item.

→ In case if you use cumulative frequency distribution the steps for calculating median are

(1) Prepare less than cumulative frequency distribution.

(2) find $\frac{N}{2}$

(3) See the cumulative frequency just greater than $\frac{N}{2}$.

(iv) The corresponding value of the variable gives median.

Ex:— Calculate median.

8 coins were tossed together and the no. of heads(x) resulting was noted the operation was repeated 256 times and the frequency distribution of no. of head is given below.

No. of heads(x)	Frequency(F)	Less than C.F
0	1	1
1	9	10
2	26	36
3	59	95
4	72	167
5	52	219
6	29	248
7	7	255
8	1	256

$N = \Sigma F = 256$

$\frac{N}{2} = \frac{256}{2} = 128$

The cumulative frequency is just greater than 128 is 167. The value of X corresponding to 167 is 4. Hence, the median no. of X is 4.

Continuous series :—

(i) Prepare 'less than' cumulative frequency distribution.

(ii) Find $\frac{N}{2}$.

(iii) See the cumulative frequency just greater than $\frac{N}{2}$.

(iv) The corresponding class contains the median value it is called the median class. The median is given by

$$\text{Median} = l + \frac{h}{f}\left(\frac{N}{2} - C\right)$$

where,

l is the lower limit of the median class,

f is the frequency of the median class,

h is the magnitude or width of the median class,

$N = \Sigma f$, is the total frequency,

C is the cumulative frequency of the class preceeding the median class

EX:—

Find the missing frequency from the following distribution of daily sales of shops, given that the median sale of shops is Rs-2,400.

Sales in hundred Rs.	No. of shops (f)	cumulative frequency
0-10	5	5
10-20	25	30
20-30	a	30+a
30-40	18	48+a
40-50	7	55+a

Let the missing frequency be 'a'.

Given, the median sale is 2400.

20-30 is the median class.

$$\text{Median} = l + \frac{h}{f}(N/2 - C)$$

where, $l = 20$, $h = 10$, $f = a$, $\frac{N}{2} = \frac{55+a}{2}$, $C = 30$.

$$\text{Median} = 20 + \frac{10}{a}\left(\frac{55+a}{2} - 30\right)$$

$$\Rightarrow 24 = 20 + \frac{10}{a}\left(\frac{55+a-60}{2}\right)$$

$$\Rightarrow 24 = 20 + \frac{10}{a}\left(\frac{a-5}{2}\right)$$

$$\Rightarrow 24 = 20 + \frac{10a-50}{2a}$$

$$\Rightarrow 24 = \frac{40a+10a-50}{2a}$$

$$\Rightarrow 24 \times 2a = 50a - 50$$

$$\Rightarrow 48a = 50a - 50$$

$$\Rightarrow 48a - 50a = -50$$

$$\Rightarrow -2a = -50$$

$$\Rightarrow a = \frac{-50}{-2} = 25$$

∴ Hence, the missing frequency is 25.

EX:—

In the frequency distribution of 100 families given below the number of families corresponding to expenditure groups 20-40 & 60-80 are missing from the table. However, the median is known to be 50. find the missing frequencies.

Expenditure (in Rupees)	No. of families (f)	Cumulative frequency (C.F) less than
0-20	14	14
20-40	f_1	$14+f_1$
40-60	27	$41+f_1$
60-80	f_2	$41+f_1+f_2$
80-100	15	$56+f_1+f_2$
	$N = 100 = 56+f_1+f_2$	

Given, $N = 100$

$N = 56 + f_1 + f_2$

$\Rightarrow 100 = 56 + f_1 + f_2$

$\Rightarrow f_1 + f_2 = 100 - 56$

$\Rightarrow f_1 + f_2 = 44$ ——— (1)

Since median is given to be 50, which lies in the class 40-60, therefore, 40-60 is the median class.

where, $l = 40$, $h = 20$, $f = 27$, $c = 14 + f_1$

$$\text{Median} = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$$

$$\Rightarrow 50 = 40 + \frac{20}{27}[50 - (14 + f_1)]$$

$$\Rightarrow 50 - 40 = \frac{20}{27}(36 - f_1)$$

$$\Rightarrow 10 = \frac{20}{27}(36 - f_1)$$

$$\Rightarrow 270 = 720 - 20f_1$$

$$\Rightarrow 270 - 720 = -20f_1$$

$$\Rightarrow -450 = -20f_1$$

$$\Rightarrow f_1 = \frac{-450}{-20} = \frac{45}{2} = 22.5 = 23$$

$$f_2 = 44 - f_1 = 44 - 23 = 21.$$

Merits:——

(i) It is rigidly defined.

(ii) Median is easy to understand and easy to calculate for a non-mathematical person.

(iii) Since median is a positional average, it is not affected at all by extreme observations and as such is very useful in the case of skewed distributions, J-shaped or inverted J-shaped distribution such as the distribution of wages, incomes & wealth. So in case

of extreme observations, median is a better average to use than the arithmetic mean since the latter gives a distorted picture of the distribution.

(iv) Median can be computed while dealing with a distribution with open end classes.

(v) Median can sometimes be located by simple inspection and can also be computed graphically.

Demerits:—

(i) In case of even number of observations for an ungrouped data, median cannot be determined exactly. We merely estimate it as the arithmetic mean of the two middle terms. In fact any value lying between the two middle observations can serve the purpose of median.

(ii) Median, being a positional average, is not based on each and every item of the distribution depends on all the observations only to the extent whether they are smaller than or greater than it; the exact magnitude of the observations being immaterial.

(iii) Median is not suitable for further mathematical treatment i.e. given the sizes and the median values of different groups, we cannot compute the median of the combined group.

(iv) Median is relatively less stable than mean, particularly for small samples since it is affected more by fluctuations of sampling as compared with arithmetic mean.

Mode :—

Mode is the value of a series which is predominant in it. In other words mode is the value which occur most frequently in a set of observations. In case of discrete series mode is the value of variable corresponding to the maximum frequency.

Age	Frequency	
15	20	
16	30	
17	35	→ Maximum frequency
18	15	

Mode is 17.

→ In case of continuous series mode is defined by

$$\text{Mode} = l + \frac{h(f_1 - f_0)}{2f_1 - f_0 - f_2}$$

The class corresponding to the maximum frequency is called as modal class.

where 'l' is the lower limit of the class
'f_1' is the frequency of the modal class
'f_0' is the frequency of class preceeding the modal class
'f_2' is the frequency of the class suceeding the modal class
'h' is the magnitude of modal class.

calculate mode.

(1)

X	f
1	3
2	1
3	18
4	25
5	40
6	30

Here maximum frequency is 40
the value corresponding to maximum frequency is 5.
Hence, mode is 5.

7	22
8	10
9	6

(2)

Class	f
0 - 10	5
10 - 20	12
20 - 30	25
30 - 40	10
40 - 50	9
50 - 60	8

Here the maximum frequency is 25, the class corresponding to maximum frequency is 20-30. The modal class is 20-30.

where, $l = 20$

$h = 10$

$f_1 = 25$

$f_0 = 12$

$f_2 = 10$

$$\text{Mode} = l + \frac{h(f_1 - f_0)}{2f_1 - f_0 - f_2}$$

$$= 20 + \frac{10(25 - 12)}{2 \times 25 - 12 - 10}$$

$$= 20 + \frac{130}{28}$$

$$= \frac{560 + 130}{28}$$

$$= \frac{690}{28}$$

$$= 24$$

$\therefore$ Mode is 24.

Merits :—

(i) Mode is easy to calculate & understand. In some cases it can be located merely by inspection. It can also be estimated graphically from a histogram.

(ii) Mode is not all affected by extreme observations and as such is preferred to arithmetic mean while dealing with extreme observations.

(iii) It can be conveniently obtained in the case of open end classes which do not pose any problems here.

Demerits :—

(i) Mode is not rigidly defined.

(ii) Since mode is the value of x corresponding to the maximum frequency, it is not based on all the obs^{n} of the series. Even in the case of the continuous frequency distribution mode depends on the frequencies of modal class and the classes preceeding and succeeding it.

(iii) Mode is not suitable for further mathematical treatment.

(iv) As compared with mean, mode is affected to a greater extent by the fluctuations of sampling.

Geometric Mean :—

If the series consists of 'n' observations say $x_1, x_2, \ldots\ldots x_n$ then geometric mean is defined as n^{th} root of their product.

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

Taking logarithms both sides, we get

$$\log G.M = \log [(x_1, x_2, \ldots\ldots x_n)^{1/n}]$$

$\Rightarrow \log G.M = \frac{1}{n}[\log x_1 + \log x_2 + \cdots + \log x_n]$

$\Rightarrow \log G.M = \frac{1}{n}\sum_{i=1}^{n} \log x_i$

$\Rightarrow G.M = \text{Antilog}\left[\frac{1}{n}\sum_{i=1}^{n} \log x_i\right]$

*In case of frequency distribution geometric mean is defined by

$$G.M = \text{Antilog}\left[\frac{1}{N}\sum_{i=1}^{n} f_i \log x_i\right]$$

Ex:— Find the Geometric mean of 24, 8, 12, 16 & 24.

x	$\log x_i$
2	$\log 2 = 0.301$
4	$\log 4 = 0.6020$
8	$\log 8 = 0.9030$
12	$\log 12 = 1.079$
16	$\log 16 = 1.204$
24	$\log 24 = 1.380$
	$\Sigma \log x_i = 5.469$

$G.M = \text{Antilog}\left[\frac{1}{n}\Sigma \log x\right]$

$= \text{Antilog}\left[\frac{1}{6} \times 5.469\right]$

$= \text{Antilog}\ (0.9115)$

$= 8.11$

② Find the Geometric mean for the following frequency distribution

Marks	No. of students(f)	Mid value (x)	$\log x$	$f.\log x$
0–10	5	5	$\log 5 = 0.699$	3.495
10–20	7	15	$\log 15 = 1.176$	8.232
20–30	15	25	$\log 25 = 1.398$	20.97
30–40	25	35	$\log 35 = 1.544$	38.6
40–50	8	45	$\log 45 = 1.653$	13.224

$N = \Sigma f = 60$ $\quad \Sigma f \log x = 84.521$

$G.M = \text{Antilog}\left[\frac{1}{N}\Sigma f_i \log x_i\right]$

$= \text{Antilog}\left[\frac{1}{60} \times 84.521\right]$

$= \text{Antilog}(1.4087)$

$= 25.628$

Merits :—

(i) Geometric mean is rigidly defined.

(ii) It is based on all the observations.

(iii) It is suitable for further mathematical treatment. If G_1 & G_2 are the geometric means of two groups of sizes n_1 & n_2 respectively, then the geometric mean G of the combined group of size n_1+n_2 is given by

$$\log G = \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}$$

(iv) As compared with mean, G.M is affected to a lessere extent by extreme observations.

(v) It is not affected much by fluctuations of sampling.

Demerits :—

(i) Because of its abstract mathematical character, geometric mean is not easy to understand to calculate for a non-mathematical person.

(ii) If any one of the observation is zero, geometric mean becomes zero and if any one of the observation is negative, geometric mean becomes imaginary regardless of the magnitude of the others items.

Harmonic mean:—

Harmonic mean is the reciprocal of arithmatic mean and reciprocal of given value.

→ We can write, if $x_1, x_2, \ldots x_n$ is a set of n observation the Harmonic mean is given by

$$H.M = \frac{1}{\frac{1}{n}\left[\sum_{i=1}^{n} \frac{1}{x_i}\right]} = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)}$$

→ In case of frequency distribution, the Harmonic mean is given by

$$H.M = \frac{1}{\frac{1}{N}\left[\sum_{i=1}^{n} \frac{f_i}{x_i}\right]} = \frac{N}{\sum_{i=1}^{n}\left(\frac{f_i}{x_i}\right)} \quad (\text{where } N = \Sigma f_i)$$

Ex:—

The following table if the weights of 31 persons in a sample inquiry. Calculate the mean weight using Harmonic mean.

Weights(x)	No. of Persons(f)	(f/x)
130	3	0.0230
135	4	0.0296
140	6	0.0428
145	6	0.0414
146	3	0.0205
148	5	0.0338
149	2	0.0134
150	1	0.0066
157	1	0.0063
	$\Sigma f = N = 31$	$\Sigma(f/x) = 0.2177$

$$H.M = \frac{N}{\Sigma(f/x)} = \frac{31}{0.2177} = 142.36$$

(2) A cyclist pedals from his house to his college at a speed of 10 km. P. hrc. and back from the college to his house at 15 km. P hrc. Find the average speed.

⇒ Let the distance from the house to college be x kms

In going from house to college, the distance is covered in $x/10$ hours.

While in coming from college to house, the distance is covered in $x/15$ hours.

Thus a total distance of $2x$ kms is covered in $\left(\frac{x}{10} + \frac{x}{15}\right)$ hours.

Hence, average speed $= \frac{\text{Total distance travelled}}{\text{Total time taken}}$

$$= \frac{2x}{\left(\frac{x}{10} + \frac{x}{15}\right)} = \frac{2}{\left(\frac{1}{10} + \frac{1}{15}\right)}$$

$= 12$ km P.hrc.

EX:—

Given below is the frequency distribution of marks obtained by 90 students. Compute the arithmetic mean, median and mode.

Marcks	Marcks	x	f	fx	C.f
15-19	14.5-19.5	17	6	102	6
20-24	19.5-24.5	22	14	308	20
25-29	24.5-29.5	27	12	324	32
30-34	29.5-34.5	32	10	320	42
35-39	34.5-39.5	37	10	370	52
40-44	39.5-44.5	42	9	378	61
45-49	44.5-49.5	47	9	423	70
50-54	49.5-54.5	52	10	520	80
55-59	54.5-59.5	57	5	285	85
60-64	59.5-64.5	62	4	248	89
65-69	64.5-69.5	67	1	67	90
			$\Sigma f=90$	$\Sigma fx=3345$	

Mean $\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{3345}{90} = 37.166$

Median $= l + \frac{h}{f}(N/2 - C)$

where, $\frac{N}{2} = \frac{90}{2} = 45$, $f = 10$, $C = 42$

$h = 5$ $l = 34.5$

C.f is just greater than 52. and the corresponding median class is 34.5-39.5.

Median $= 34.5 + \frac{5}{10}(45-42)$

$= 34.5 + \frac{1}{2}(3)$

$= 34.5 + 1.5$

$= 36$

Mode = 3median - 2mean

= 3(36) - 2(37.16)

= 108 - 74.32

= 33.68

Relationship between mean, Median & mode:—

→ In case of symmetrical distribution

Mean = Median = Mode ;

$M = M_d = M_o$

$\text{Mean} - \text{Median} = \frac{1}{3}(\text{Mean} - \text{Mode})$

$\Rightarrow \boxed{\text{Mode} = 3\,\text{Median} - 2\,\text{Mean}}$

Relationship between AM, GM & HM.

Prove that $AM \geq GM \geq HM$.

Proof:—

Let a & b be two real positive numbers.

i.e $a > 0, b > 0$.

$A.M = \frac{a+b}{2}$

$G.M = (a.b)^{1/2} = \sqrt{ab}$

$HM = \frac{1}{\frac{1}{2}\left(\frac{1}{a}+\frac{1}{b}\right)} = \frac{2ab}{a+b}$

$AM - GM = \frac{a+b}{2} - \sqrt{ab}$

$= \frac{a+b-2\sqrt{ab}}{2}$

$= \frac{1}{2}\left(\sqrt{a}-\sqrt{b}\right)^2 \geq 0$

i.e the square of real quantity is always non-negative.

$$AM - GM \geq 0$$

$$AM \geq GM \quad \text{---} \quad (1)$$

$$GM - HM = \sqrt{ab} - \frac{2ab}{a+b}$$

$$= \sqrt{ab}\left(1 - \frac{2\sqrt{ab}}{a+b}\right)$$

$$= \sqrt{ab}\left(\frac{a+b-2\sqrt{ab}}{a+b}\right)$$

$$= \sqrt{ab}\left(\frac{(\sqrt{a}-\sqrt{b})^2}{a+b}\right) \geq 0$$

i.e the square of a real quantity is always non-negative.

$$GM - HM \geq 0$$

$$GM \geq HM \quad \text{---} \quad (II)$$

Combining eq^n (1) & (II)

$$AM \geq GM \geq HM \quad \text{(Proved)}$$

Partition value:—

The value which devide the series in to number of equal parts are known as Partition value.

→ There are 3 types of Partition value.

1. Quartiles
2. Deciles
3. Percentiles

Quartiles:—

→ The three points which divide the series in to four equal parts are known as quartiles.

→ There are 3 points Q_1, Q_2, Q_3 such that $Q_1 \leq Q_2 \leq Q_3$. known as the quartiles.

→ Q_1 is the lower or 1st quartile is a value which has 25% of the items of the distribution below it & consequently 75% of the item are greater than it.

→ Q_2 is the 2nd quartile & coincides with the median. It has an equal no. of observation above it & below it.

→ Q_3 is the 3rd quartile it has 75% of the observation below it and 25% of the observation above it.

Calculation of quartiles:—

1. Individual series:—

In case of ungrouped (Individual) data having n observation arranged in the ascending or descending order ith quartile is given by

$$Q_i = \text{value of } \frac{i(n+1)}{4} \text{th item (where } i=1,2,3)$$

2. Discrete series:—

(i) Find $\frac{iN}{4}$, where $N = \Sigma f_i$.

(ii) C is the cumulative frequency just greater than $\frac{iN}{4}$.

(iii) The corresponding value of x keeps the value of Q_i.

3. Continuous series:—

In case of continuous or grouped frequency distribution the i^{th} Quartile is given by

$$Q_i = l + \frac{h}{f}\left(\frac{iN}{4} - c\right) \quad (\text{where } i=1,2,3)$$

Ex:—

Marks secured by 11 students in statistics are as follows 34, 27, 64, 89, 75, 82, 54, 49, 68, 84, 45 then calculate Q_1, Q_2 & Q_3.

Ans:— Arranging in Descending order =

27, 34, 45, 49, 54, 64, 68, 75, 82, 84, 89

Q_i = value of $\frac{i(n+1)}{4}$ th item

Q_1 = value of $\frac{1(11+1)}{4}$ th item

= value of $\frac{1 \times 12^3}{4}$ th item

= value of 3rd item = 45

Q_2 = value of $\frac{2(11+1)}{4}$ th item

= value of $\frac{2 \times 12^3}{4}$ th item

= value of 6th item = 64

Q_3 = value of $\frac{3(11+1)}{4}$ th item

= value of $\frac{3 \times 12^3}{4}$ th item

= value of 9th item = 82

Deciles:—

→ The 9 values which divides the series in to 10 equal parts. So, there are 9 deciles such as $D_1, D_2, \ldots D_9$, such that $D_1 \leq D_2 \leq \ldots \leq D_9$.

→ D_5 coincide with median

Calculation of Deciles:—

1. Individual Series:—

In case of ungrouped data the deciles is

given by

$$D_i = i^{th} \text{ decile} = \text{value of } \frac{i(n+1)}{10}\text{-th item}$$

(where $i = 1, 2, 3, 4, \ldots 9$)

Discrete series:—

(i) Find $\frac{iN}{10}$, where $N = \Sigma f$

(ii) C is the cumulative frequency just greater than $\frac{iN}{10}$.

(iii) The corresponding value of x keeps the value of D_i.

Continuous series:—

In case of continuous series ith deciles is given by

$$D_i = l + \frac{h}{f}\left(\frac{iN}{10} - C\right) \quad (i = 1, 2, 3, \ldots, 9)$$

Percentiles:—

→ The 99 values which divides the series in to 100 equal parts. So, there are 99 percentiles are $P_1, P_2, \ldots P_{99}$. such that $P_1 \leq P_2 \leq \ldots \leq P_{99}$

→ P_{50} coincide with median.

Calculation of Percentiles:—

1. Individual series:—

In case of ungrouped data the percentiles is given by

$$P_i = i^{th} \text{ percentile} = \text{value of } \frac{i(n+1)}{100}\text{th item}$$

(where $i = 1, 2, 3, \ldots, 99$)

Discrete series:—

(i) Find $\frac{iN}{100}$, where $N = \Sigma f$

(ii) c is the cumulative frequency just greater than $\frac{iN}{100}$.

(iii) The corresponding value of x keeps the value of P_i.

Continuous series:—

In case of continuous series the i^{th} percentile is given by :

$$P_i = l + \frac{h}{F}\left(\frac{iN}{100} - c\right)$$

Ex:—

Annual sales	No. of Stores
0-100	4
100-200	28
200-300	35
300-400	25
400-500	7
500-600	1

Compute the Quartile 6th Deciles and 20th Percentiles.

Ans:—

Annual sales	No. of Stores	C.F
0-100	4	4
100-200	28	32
200-300	35	67
300-400	25	92
400-500	7	99
500-600	1	100
	N = 100	

(i) $Q_1 = \frac{N}{4} = \frac{100}{4} = 25$

C.f just greater than $\frac{N}{4}$ is 32.

The 1st Quartile lies in the class interval 100-200.

$$Q_1 = l + \frac{h}{f}\left(\frac{iN}{4} - c\right)$$

$$= 100 + \frac{100}{28}(25 - 4)$$

$$= 100 + \frac{100}{28} \times 21$$

$$= \frac{2800 + 2100}{28} = \frac{4900}{28} = 175$$

(ii) $Q_2 = \frac{100}{2} = 50$

C.f just greater than $\frac{N}{2}$ is 67.

The 2nd Quartile lies in the class interval 200-300.

$$Q_2 = l + \frac{h}{f}\left(\frac{iN}{2} - c\right)$$

$$= 200 + \frac{100}{35}(50 - 32)$$

$$= 200 + \frac{100}{35}(18)$$

$$= \frac{7000 + 1800}{35} = 251.43$$

(iii) $Q_3 = \frac{3(N+1)}{4} = \frac{3(100+1)}{4} = \frac{303}{4} = 75.75$

C.f is just greater than $\frac{3(N+1)}{4}$ is 92.

The 3rd Quartile lies in the class interval 300-400.

$$Q_3 = l + \frac{h}{f}\left(\frac{iN}{4} - c\right)$$

$$= 300 + \frac{100}{25}(775 - 67)$$

$$= 300 + \frac{100}{25} \times 8.85$$

$$= \frac{7500 + 800}{25} = 332$$

(iv) $D_6 = ?$

$$\frac{6N}{10} = \frac{6 \times 100}{10} = 60$$

The cumulative frequency is just greater than 67. The corresponding interval class is 200-300

$$D_6 = l + \frac{h}{f}\left(\frac{6N}{10} - c\right)$$

$l = 200$, $h = 100$, $f = 35$, $\frac{6N}{10} = 60$, $c = 32$

$$D_6 = 200 + \frac{100}{35}(60 - 32)$$

$$= 200 + \frac{100}{35}(28)$$

$$= \frac{7000 + 2800}{35}$$

$$= 280$$

(v) $P_{20} = ?$

$$\frac{20N}{100} = \frac{20 \times 100}{100} = 20$$

The cumulative frequency just greater than 20 is 32. The corresponding interval class is 100-200.

$$P_{20} = l + \frac{h}{f}\left(\frac{20N}{100} - c\right)$$

$l = 100$, $h = 100$, $f = 28$, $\frac{20N}{100} = 20$, $c = 4$

$$P_{20} = 100 + \frac{100}{28}(20 - 4)$$

$$= 100 + \frac{100}{28} \times 16$$

$$= \frac{2800 + 1600}{28}$$

$$= 157.14$$

MEASURES OF DISPERSION

Measures of Dispersion:—

→ The Measures of Dispersion is a measure of extent to which individual items vary from central value.

→ Literally, dispersion means scatteredness.

→ If the dispersion is small it means that the given data values are closer to the central value.

Objectives or significance of the Measures of Dispersion:—

* The main objectives of studying dispersion may be summarised as :—

(1) To find out the reliability of an average.

(2) To control the variation of the data from the central value.

(3) To compare two or more sets of data regarding their variability.

(4) To obtain other statistical measures for further analysis of data.

Characteristics for an ideal Measure of Dispersion:—

(i) It should be rigidly defined.

(ii) It should be easy to calculate & easy to understand.

(iii) It should be based on all the observations.

(iv) It should be amenable to further mathematical treatment.

(v) It should be affected as little as possible by fluctuations of sampling.

(vi) It should not be affected much by extreme observations.

Types of Dispersion :—

(1) Absolute measures of dispersion

(2) Relative measures of dispersion

(1) Absolute measures of dispersion :—

→ The measures of dispersion which are expressed in terms of the original units of a series are known as Absolute measures of dispersion.

→ Such measures are not suitable for comparing the variability of the two distributions which are expressed in different units of measurement.

(2) Relative measures of dispersion :—

→ The Relative measures of dispersion are obtained as ratios or percentage and it is independent of units of measurement for comparing of the variability of the two distributions, we can use Relative measures of dispersion instead of Absolute measures of dispersion.

Various Measures of dispersion :—

→ The various measures of dispersion are :—

(i) Range

(ii) Quartile deviation or semi-Interquartile range

(iii) Mean deviation

(iv) Standard deviation

(v) Variance

Range :—

→ Range is defined as the difference between the maximum value and minimum value of the distribution.

Range = Maximum value − Minimum value

Relative Measures of Range:—

→ If we want to compare the variability of the distribution given in different units of measurements we can use Relative measures. So, this relative measures is called as Co-efficient of range.

Let Maximum value = A

Minimum value = B

$$\text{Co-efficient of range} = \frac{\text{Maximum value} - \text{Minimum value}}{\text{Maximum value} + \text{Minimum value}}$$

$$\Rightarrow \text{Co-efficient of range} = \frac{A-B}{A+B}$$

Example:—

Calculate the Range and the co-efficient of range of A.S monthly earning for a year.

Ans:

Month	Monthly earning (in hundred)
1	13900
2	15000
3	15100
4	15100
5	15700
6	15800
7	16000
8	16100
9	16200
10	16200
11	17300
12	17500

Maximum earning → 17500

Minimum earning → 13900

* Range = Maximum earning − Minimum earning

= 17500 − 13900 = 3600.

* Co-efficient of Range

$$= \frac{\text{Maximum earning} - \text{Minimum earning}}{\text{Maximum earning} + \text{Minimum earning}}$$

$$= \frac{3600}{17500+13900} = \frac{3600}{31400} = 0.115$$

Quartile Deviation or semi Inter-quartile Range:—

→ It is given by

$$Q.D = \frac{Q_3 - Q_1}{2}$$

where, Q_3 = upper quartile

Q_1 = lower quartile

→ For comparative studies of variability of two distributions we need a relative measure which is known as co-efficient of Quartile Deviation.

$$\text{Co-efficient of } Q.D = \frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_3 + Q_1}{2}} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Merits of Quartile Deviation:—

→ Quartile deviation is quite easy to understand and calculate.

→ It has a number of obvious advantages over range as a measure of dispersion.

Ex:—

(a) As against range which was based on two observations only, Q.D. makes use of 50% of the data as such is obviously a better measure than Range.

(b) Since Q.D ignores 25% of the data from the beginning of the distribution and another 25% of the data from the top end, it is not affected at all by extreme observations.

(c) Q.D can be computed from the frequency distribution with open end classes. In fact, Q.D is the only measure of dispersion, which can be obtained while dealing with a distribution open and classes.

Demerits of Quartile deviation:—

→ Q.D is not based on all the observations since it ignore 25% of the data at the lower end and 25% of the data at the upper end of the distribution. Hence, it can not be regarded as a reliable measure of variability.

→ Q.D is affected considerably by fluations of sampling.

→ Q.D is not suitable for further mathematical treatment.

Mean Deviation:— (Average deviation)

Individual series:—

Mean deviation is given by

$$M.D = \frac{1}{n} \sum_{i=1}^{n} |X - A|$$

where, n = no. of Observation

A = Mean / Mode / Median

Absolute Measures of Mean deviation:—

→ In case of Frequency distribution

$$M.D = \frac{1}{N} \sum_{i=1}^{n} f_i |X_i - A| \qquad N = \sum f$$

Mean deviation about mean:—

$$M.D = \frac{1}{N} \sum f |X - \bar{X}|$$

Mean deviation about median:—

$$M.D = \frac{1}{N} \sum f |X - M_d|$$

Mean deviation about mode:—

$$M.D = \frac{1}{N} \sum f |X - M_o|$$

Relative Measures of Mean deviation:—

→ Relative measures of dispersion of mean deviation is called the co-efficient of mean deviation.

→ It is given by

$$\text{Co-efficient of M.D} = \frac{\text{Mean Deviation}}{\text{Average about which it is calculated}}$$

$$\text{Co-efficient of M.D about mean} = \frac{\text{M.D}}{\text{Mean}}$$

$$\text{Co-efficient of MD about Median} = \frac{\text{M.D}}{\text{Median}}$$

NOTE:—

Mean deviation about mean is always greater than mean deviation about median.

OR

Mean deviation is least when taken about median.

Merits of Mean deviation:—

→ Mean deviation is rigidly defined.

→ It is easy to understand and calculate.

→ Mean deviation is based on all observations.

→ It is thus definitely a better measure of dispersion than the range and quartile deviation.

→ Mean deviation provides an accurate and true measure of dispersion.

→ As compared with standard deviation, it is less affected by extreme observation.

Demerits of Mean deviation:—

→ The strongest objection against mean deviation is that while computing its value, we take the absolute value of the deviations about an average and ignore the signs of the deviations.

→ It is rarely used in sociological studies.

→ It can not be computed for distributions with open end classes.

→ Mean deviation tends to increase with the size of the sample though not proportionately and not so rapidly as range.

EX:—

Calculate the M.D from the median for the following data.

Marks	Frequency	C.f	Mid value(x)	$\|x-Md\|$	$f\|x-Md\|$
0-10	5	5	5	41.43	207.15
10-20	8	13	15	31.43	251.44
20-30	7	20	25	21.43	150
30-40	12	32	35	11.43	137.16
40-50	28	60	45	1.43	40.04
50-60	20	80	55	8.57	171.14
60-70	10	90	65	18.57	185.70
70-80	10	100	55	28.57	285.70
	$N=\Sigma f=100$				$\Sigma f\|x-Md\|=1428.6$

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

The cumulative frequency is greater than equal to 50 is 60. The corresponding class 40-50 is the median class.

$l = 40,\ h = 10,\ f = 28,\ c = 32,\ N/2 = 50$

$$\text{Median} = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$$

$$= 40 + \frac{10}{28}(50 - 32)$$

$$= 40 + \frac{10}{28} \times 18$$

$$= 40 + 6.43$$

$$= 46.43$$

$$\text{M.D about Median} = \frac{1}{N}\Sigma f\,|X - Md|$$

$$= \frac{1}{100} \times 1428.6$$

$$= \frac{1428.6}{100} = 14.286 = 14.29$$

Standard deviation:—

→ Standard deviation is defined as the positive square root of the arithmetic mean of the square of the deviation of the given deviations from their arithmetic mean.

→ It is usually denoted by 'σ'.

Individual series:—

In case of Individual series the Standard deviation is given by

$$\boxed{\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X - \bar{X})^2}}$$

$\bar{X} = \frac{1}{n}\Sigma X$, n = no. of observation

↓

(Arithmetic mean of the given values)

Frequency distribution:—

In case of frequency distribution the standard deviation is given by

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{2} f(x-\bar{x})^2}$$

where, $N = \Sigma f$, $\bar{x} = \frac{\Sigma fx}{\Sigma f}$

$$\text{Co-efficient of standard Deviation} = \frac{S.D}{\text{mean}} = \frac{\sigma}{\bar{x}}$$

Variance:—

→ The square of the standard deviation is known as variance.

→ It is denoted by 'σ^2'.

Individual series:—

In case of Individual series variance is given by

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{2}(x_i-\bar{x})^2$$

Frequency distribution:—

In case of frequency distribution variance is given by

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{2} f_i(x_i-\bar{x})^2$$

Co-efficient of variation:—

→ Hundred times the co-efficient of dispersion based on standard deviation is called the co-efficient of variance.

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

→ For comparing the variability of two series we

calculate the co-efficient of variation for each series. The series having greater C.V is said to be more variability than the other series. And the series having lesser C.V is said to be more consistent or homogenous than the other.

Combined standard deviation:—

Let $\sigma_1, \sigma_2, \ldots \sigma_k$ are the S.D $\bar{x}_1, \bar{x}_2, \ldots \bar{x}_k$ are the A.M and $n_1, n_2, \ldots n_k$ are the sizes of k group then the S.D σ of the combined group up size $N = n_1 + n_2 + n_3 + \ldots + n_k$ is given by

$$\sigma = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2) + \ldots + n_k(\sigma_k^2 + d_k^2)}{n_1 + n_2 + \ldots + n_k}}$$

where $d_1 = \bar{x}_1 - \bar{x}$

$d_2 = \bar{x}_2 - \bar{x}$

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

Moments:—

→ Moment gives the idea about the central part of the distribution and their variability. It also gives the idea about the skewness and kurtosis of the distribution.

→ There are 4 types of moments.

(i) Raw moment
(ii) Central moment
(iii) Factorial moment
(iv) Absolute moment

Raw moment:—

The r^{th} raw moment of a variable x about any point $x = A$ is usually denoted by μ'_r is given by

$$\mu_r' = \frac{1}{N}\sum f_i(x_i - A)^r \quad \text{---- (1)}$$

→ Raw moment is denoted by 'μ_r''.

putting $r = 1$

$$\mu_1' = \frac{1}{N}\sum f_i(x_i - A)^1$$

$$= \frac{1}{N}\sum f_i x_i - \frac{1}{N}\sum f_i A$$

$$= \bar{x} - A$$

putting $A = 0$, $\boxed{\mu_1' = \bar{x}}$

i.e the 1st order of raw moment is equivalent to mean.

Taking $r = 2, 3, 4$ in eq (1)

we get,

μ_2' (2nd order raw moment about A) $= \frac{1}{N}\sum f(x-A)^2$

μ_3' (3rd order raw moment about A) $= \frac{1}{N}\sum f(x-A)^3$

μ_4' (4th order raw moment about A) $= \frac{1}{N}\sum f(x-A)^4$

Central moment:—

The rth moment of a variable x about the mean $\bar{x}$ usually denoted as μ_r.

It is given by

$$\mu_r = \frac{1}{N}\sum f_i(x_i - \bar{x})^2 \quad \text{---- (2)}$$

putting $r = 0$, $\mu_0 = \frac{1}{N}\sum f_i(x_i - \bar{x})^0 = 1$

$r = 1$, $\mu_1 = \frac{1}{N}\sum f_i(x_i - \bar{x})^1 = 0$

$r = 2$, $\mu_2 = \frac{1}{N}\sum f_i(x_i - \bar{x})^2 = \sigma^2$ (variance)

The 2nd order of central moment is equivalent to variance.

Relation between raw moment & central moment:-

$$\mu_2' = \mu_2 + (\mu_1')^2$$

$$\mu_3' = \mu_3 + 3\mu_2\mu_1' + (\mu_1')^3$$

$$\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2(\mu_1')^2 + (\mu_1')^4$$

Relation between central moment & raw moment:-

$$\mu_2 = \mu_2' - (\mu_1')^2$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$

$$\mu_4 = \mu_4' + 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$$

Factorial moment:-

Factorial moment of order r about the origin of the distribution x_i / f_i $(i=1,2,\ldots n)$ is obtained as

$$\boxed{\mu_{(r)}' = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^{(r)}} \quad \text{---(i)}$$

Where, $x^{(r)} = x(x-1)(x-2)\ldots\ldots(x-r+1)$

and $N = \sum_{i=1}^{n} f_i$

Thus, the factorial moment of order r about any point $x=a$ is given by,

$$\mu_{(r)}' = \frac{1}{N}\Sigma f_i (x_i - a)^{(r)}$$

Where $(x-a)^{(r)} = (x-a)(x-a-1)(x-a-2)\ldots\ldots(x-a-r+1)$ ---(ii)

Put $r = 1, 2, 3, 4$ in eq(i), we get

$$\mu_{(1)}' = \frac{1}{N}\Sigma f_i x_i = \mu_1' \text{ (about origin)} = \text{Mean}(\bar{X})$$

$$\mu_2' = \frac{1}{N}\Sigma f_i x_i^{(2)}$$

$$= \frac{1}{N}\Sigma f_i x_i(x_i - 1) = \frac{1}{N}\Sigma f_i x_i^2 - \frac{1}{N}\Sigma f_i x_i$$

$$\boxed{\mu_{(2)}' = \mu_2' - \mu_1'}$$

$$\mu_3' = \frac{1}{N}\sum f_i x_i^{(3)}$$

$$= \frac{1}{N}\sum f_i x_i (x_i - 1)(x_i - 2)$$

$$= \frac{1}{N}\sum f_i \left[(x_i^2 - x_i)(x_i - 2)\right]$$

$$= \frac{1}{N}\sum f_i \left[x_i^3 - 2x_i^2 - x_i^2 + 2x_i\right]$$

$$= \frac{1}{N}\sum f_i x_i^3 - 3\frac{1}{N}\sum f_i x_i^2 + 2\frac{1}{N}\sum f_i x_i$$

$$\boxed{\mu_{(3)}' = \mu_3' - 3\mu_2' + 2\mu_1'^2}$$

$$\mu_4' = \frac{1}{N}\sum f_i x_i^{(4)}$$

$$= \frac{1}{N}\sum f_i \left[x_i(x_i - 1)(x_i - 2)(x_i - 3)\right]$$

$$= \frac{1}{N}\sum f_i x_i \left[x_i^3 - 6x_i^2 + 11x_i - 6\right)$$

$$= \frac{1}{N}\sum f_i x_i^4 - 6\frac{1}{N}\sum f_i x_i^3 + 11\frac{1}{N}\sum f_i x_i^2 - 6\frac{1}{N}\sum f_i x_i$$

$$\boxed{\mu_{(4)}' = \mu_4' - 6\mu_3' + 11\mu_2' - 6\mu_1'}$$

Conversly, we will get,

$$\mu_1' = \mu_{(1)}'$$

$$\mu_2' = \mu_{(2)}' + \mu_{(1)}'$$

$$\mu_3' = \mu_{(3)}' + 3\mu_{(2)}' - 2\mu_{(1)}'$$

$$\mu_4' = \mu_{(3)}' + 3(\mu_{(2)}' + \mu_{(1)}') - 2\mu_{(1)}'$$

$$\boxed{\mu_{(3)}' = \mu_{(3)}' + 3\mu_{(3)}' + \mu_{(1)}'}$$

$$\mu_4' = \mu_{(4)}' + 6\mu_{(3)}' - 11\mu_{(2)}' + 6\mu_{(1)}'$$

$$= \mu_{(4)}' + 6(\mu_{(3)}' + 3\mu_{(2)}' + \mu_{(1)}') - 11(\mu_{(2)}' + \mu_{(1)}') - 2\mu_{(1)}'$$

$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu'^2_1 - 3\mu'^4_1$$

Absolute moments :—

For the frequency distribution x_i/f_i $(i=1,2,3,\ldots n)$ the r^{th} absolute moment of the variable x about the origin is given by,

$$\frac{1}{N}\sum_{i=1}^{n} f_i|x_i^r|, \quad N=\sum f_i$$

where, $|x_i^r|$ represents the absolute / modulus value of x_i^r.

The r^{th} absolute moment of the variable about the mean $\bar{x}$ is given by,

$$\frac{1}{N}\sum_{i=1}^{n} f_i|x_i - \bar{x}|^r$$

NOTE :—

The moments are used to describe the various characteristics of a frequency distribution like central tendency, variation, skewness and kurtosis.

Sheppard's correction :—

→ While calculating moments it is assumed that all the values of a variable in a class interval are concentrated at the center of that interval ~~that is~~ mid value.

→ So, this assumption is an approximation to facilitate calculation and it introduced some error which is known as grouping error.

Conditions for applying Sheppard's correction:—

(i) Total frequency should be large.

(ii) The frequency distribution should be continuous.

(iii) The distribution must have finite range.

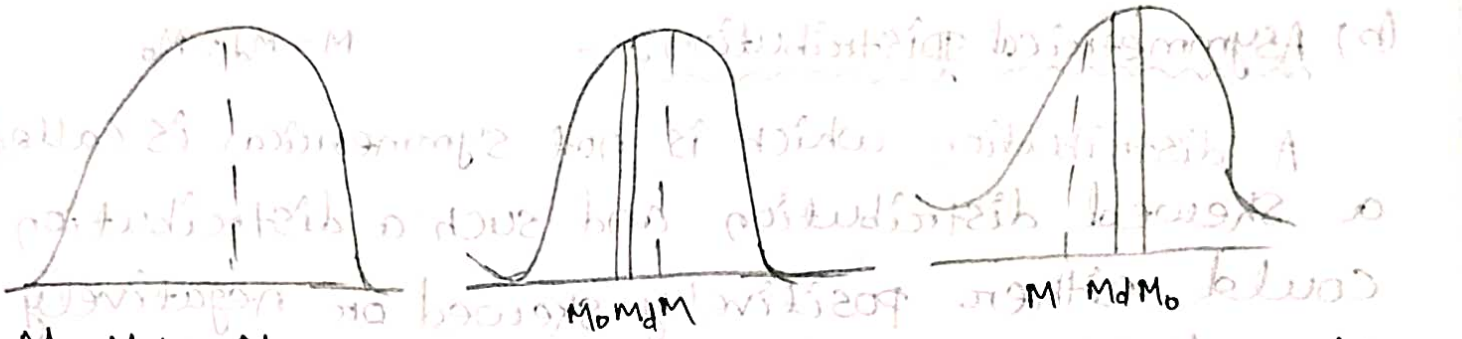
→ The effect due to grouping at the mid point of the interval can be corrected by the following formulas. So, this formula is known as Sheppard's correction.

$$\mu_2 \text{ (Corrected)} = \mu_2 - \frac{h^2}{12}$$

$$\mu_3 \text{ (Corrected)} = \mu_3$$

$$\mu_4 \text{ (Corrected)} = \mu_4 - \frac{h^2}{2} \times \mu_2 + \frac{7}{240} \times h^4$$

where, h is the width of the class interval.



Skewness:—

→ When a series is not symmetrical it said to be asymmetrical or skewed.

→ Literally, skewness means lack of symmetry, we study skewness to have an idea about the shape of the curve which can draw with the help of the given data.

→ A distribution is said to be skewed if

(i) Mean, Median and Mode fall at different points. i.e. Mean ≠ Median ≠ Mode

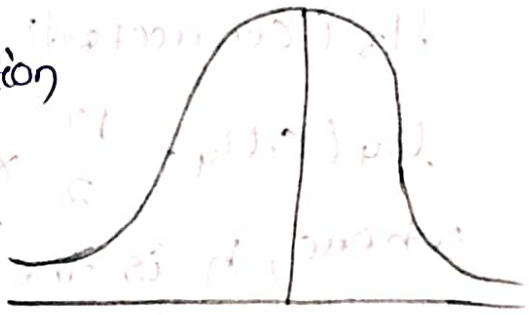
(ii) Quartiles are not equidistant from median

i.e $Q_3 - M_d \neq M_d - Q_1$

(iii) The curve drawn with the help of the given data is not symmetrical but stretched more to one side them to the other.

(a) Symmetrical Distribution:—

In symmetrical distribution the values of mean, median and mode coincide.



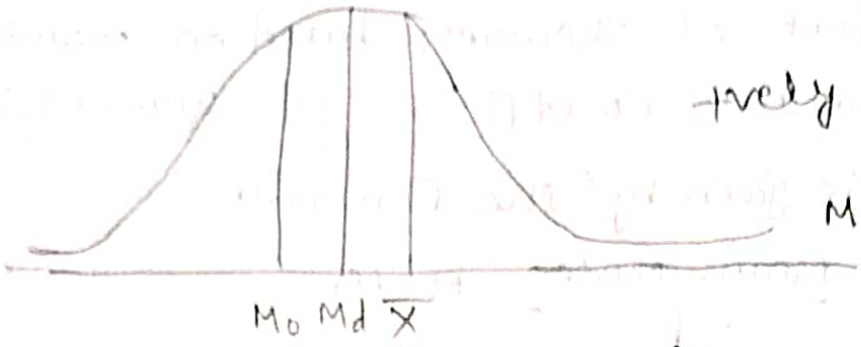
$M = M_d = M_o$

(b) Asymmetrical Distribution:—

A distribution which is not symmetrical is called a skewed distribution and such a distribution could either positively skewed or negatively skewed distribution.

(c) Positively skewed Distribution:—

In the positively skewed distribution, the values of the mean is maximum, mode is least and median lies in between the two. A frequency distribution for which the curve has a longer tail towards the right is said to be tvely skewed distribution.

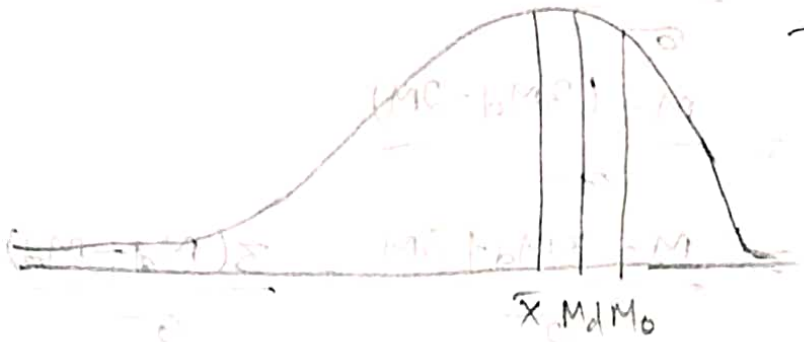


-ively skewed distribution

$M > M_d > M_o$

(d) Negatively skewed Distribution:—

In a negatively skewed distribution the value of mode is maximum, mean is least and median is lies in between the two. In negatively skewed distribution if the longer tail towards the left.



-vely skewed distribution

$M_o > M_d > M$

Measures of Skewness:—

Various measures of skewness are—

(i) $S_K = M - M_d$ (ii) $S_K = M - M_o$ (iii) $(Q_3 - M_d) - (M_d - Q_1)$

→ There are the absolute measures of skewness. As in dispersion for compairing two series. So, do not calculate these absolute measures but we calculate the relative measures called the co-efficient of skewness.

→ There are three important co-efficient of skewness are—

(i) Prof Karl pearson's co-efficient of skewness

(ii) Prof Bowley's co-efficient of skewness

(iii) Co-efficient of skewness based on moment

Karl Pearson's co-efficient of skewness:—

This is given by the formula

$$S_k = \frac{\text{Mean} - \text{Mode}}{\text{S.d}} = \frac{M - M_0}{\sigma}$$

where, σ is the S.D of the distribution.

→ IF mode is ill-defined, then using the empirical relating $M_0 = 3M_d - 2M$, for a moderately asymetrical distribution we get,

$$\text{MODE} = 3M_d - 2M$$

$$S_k = \frac{M - M_0}{\sigma}$$

$$= \frac{M - (3M_d - 2M)}{\sigma}$$

$$= \frac{M - 3M_d + 2M}{\sigma} = \frac{3(M_d - M_d)}{\sigma}$$

we observed data, $S_k = 0$, if $M = M_0 = M_d$

Hence for a symmetrical distribution, mean, median, and mode coincide.

Skewness is positive if $M > M_0$ or $M > M_d$

Skewness is negative if $M < M_0$ or $M < M_d$

The limits of Karl Pearson's co-efficient of skewness are $-3 \leq S_k \leq 3$.

Bowley's co-efficient of skewness:—

$$S_k = \frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)}$$

$$= \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

2) If $S_K = 0$, $Q_3 - M_d = M_d - Q_1$

This implies that for a symmetrical distribution ($S_K = 0$), Median is equidistant from the upper and lower quartiles.

Skewness is positive if

$$Q_3 - M_d > M_d - Q_1 \Rightarrow Q_3 + Q_1 > 2M_d$$

Skewness is negative if

$$Q_3 - M_d < M_d - Q_1 \Rightarrow Q_3 + Q_1 < 2M_d$$

The limits of Bowley's co-efficient of skewness is $-1 \leq S_K \leq 1$.

3) Based upon moments, co-efficient of skewness is

$$S_K = \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$$

where, $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ is the co-efficient of skewness.

$\beta_2 = \mu_4/\mu_2^2$ is the co-efficient of skewness.

Pearson's β and γ co-efficients:—

Karl pearson's defined the following four co-efficient, based upon the first moment about mean:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}, \quad \gamma_1 = +\sqrt{\beta_1}$$

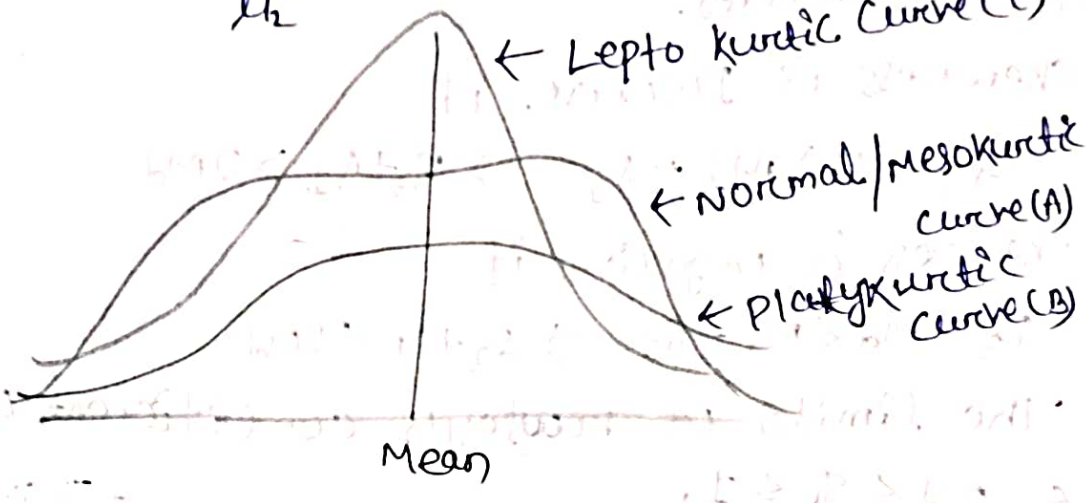
$$\beta_2 = \mu_4/\mu_2^2, \quad \gamma_2 = \beta_2 - 3$$

Kurtosis:—

Kurtosis gives us to have an idea about the "flatness of peakedness" of the frequency curve.

It is measured by the co-efficient β_2 or its deviation γ_2 is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2}, \quad \gamma_2 = \beta_2 - 3$$



→ Curve of the type 'A' which is neither flat nor packed is called the normal curve or mesokurtic curve and for such as curve $\beta_2 = 3$ i.e $\gamma_2 = 0$.

→ Curve of the type 'B' which is flatter than the normal curve is known as platykurtic curve and for such a curve, $\beta_2 < 3$ i.e $\gamma_2 < 0$.

→ Curve of the type 'C' which is more packed then the normal curve is called leptokurtic curve and for such a curve $\beta_2 > 3$ i.e $\gamma_2 > 0$.

UNIT-3

Correlation:—

In a bivariate distribution we have studied two variables. If we change in one variable affects the change in other variables. so, we say that this two variables are said to be correlated.

→ correlation is the linear relationship between two variables

Types of Correlation:—

There are 3 types of correlation.

1. Positive correlation
2. Negative correlation
3. Zero correlation

Positive correlation:—

If the two variables deviate in same direction i.e increase or decrease of one variables result in increase or decrease of the other variables. The correlation is said to be Positive.

Ex:— Income and Expenditure
Heights and weights of a group of person

Negative correlation:—

If the two variables deviate in opposite direction i.e increase or decrease in one variables result in increase or decrease of another variables. The correlation is said to be Negative.

Ex:— Price and demand of a commadity
sales of woolen garments and temperature
volume and pressure of a Perfect gas

Zero correlation:—

If there is no relationship between two variables is known as zero correlation.

Ex:— The rate of marriage and agricultural production.

Methods of studying correlation:—

There are 3 methods for studying correlation.

1. Karl. Pearson's Co-efficient of correlation
2. Scatter Diagram
3. Spearman's Rank correlation

Karl Pearson's Co-efficient of correlation:—

Correlation co-efficient between two variables X and Y usually denoted by $R(X,Y)$ or r_{xy} or 'r'.

→ It is a numerical measure of linear relationship between two variables.

→ The Karl pearson's developed a formula called co-efficient of correlation.

→ It is defined as

$$r = \frac{\frac{1}{n}\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\frac{1}{n}\Sigma(X-\bar{X})^2}\sqrt{\frac{1}{n}\Sigma(Y-\bar{Y})^2}}$$

or

$$r = \frac{\frac{1}{n}\Sigma XY - \bar{X}\bar{Y}}{\sqrt{\frac{1}{n}\Sigma X^2 - (\bar{X})^2}\sqrt{\frac{1}{n}\Sigma Y^2 - (\bar{Y})^2}}$$

→ The limits of correlation co-efficient is -1 to $+1$. i.e $-1 \leq r \leq 1$.

→ It mean the correlation co-efficient can not exceed unity.

Interpretation of 'r':—

→ If $r=0$; No correlation

→ If $r>0$, the correlation is positive.

→ If $r<0$, the correlation is Negative

→ If $r=+1$, the correlation is perfect positive.

→ If $r=-1$, the correlation is perfect Negative.

EX:—

Calculating the correlation co-efficient for the following heights of fathers (x) and their sons (y).

x	y	xy	x^2	y^2
65	67	4355	4225	4489
66	68	4488	4356	4624
67	65	4355	4489	4225
67	68	4556	4489	4624
68	72	4896	4624	5184
69	72	4968	4761	5184
70	69	4830	4900	4761
72	71	5112	5184	5041
$\Sigma x = 554$	$\Sigma y = 552$	$\Sigma xy = 38,560$	$\Sigma x^2 = 38,028$	$\Sigma y^2 = 38132$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{544}{8} = 68$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{552}{8} = 69$$

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

$$= \frac{\frac{1}{8} \times 37560 - 4692}{\sqrt{\frac{1}{8} \times 37028 - 4624}\sqrt{\frac{1}{8} \times 38132 - 4761}}$$

$$= \frac{4965 - 4692}{\sqrt{4628.5 - 4624}\sqrt{4766.5 - 4761}}$$

$$= \frac{3}{\sqrt{4.5}\sqrt{5.5}} = \frac{3}{2.12 \times 2.34} = 0.60$$

$\Rightarrow$ The correlation of co-efficient is 0.60.

Properties of co-efficient correlation:—

(i) The correlation co-efficient lies in between -1 to +1, i.e $-1 \leq r \leq 1$.

(ii) The correlation co-efficient is independent of the change of origin and scale.

(iii) Two variables are ~~one~~ uncorrelated.

If $r=0$, then the variables are independent

(iv) $r(aX+b, cY+d) = \frac{ac}{|ac|} r(X,Y)$

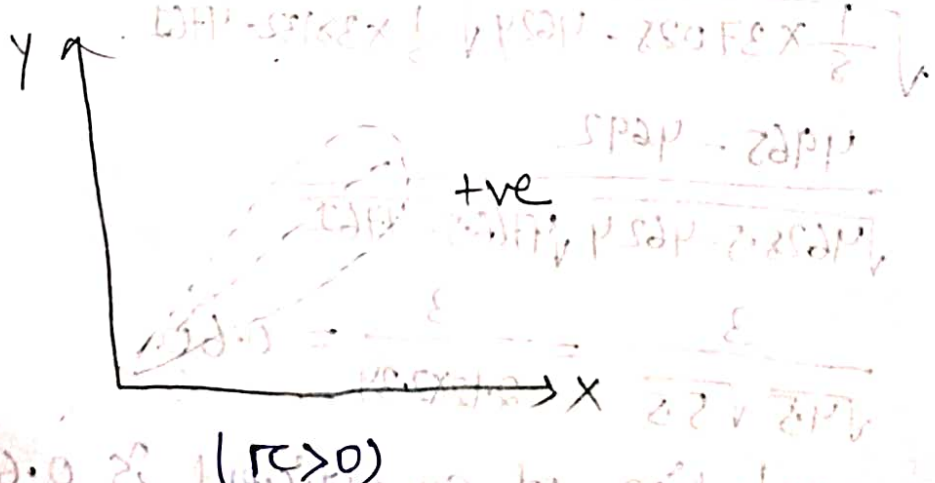
Scatter Diagram method:—

Scatter Diagram is one of the simplest way of diagramatic representation of bivariate distribution.

Suppose we are given n pairs of values $(X_1, Y_1)(X_2, Y_2) \ldots\ldots (X_n, Y_n)$ of two variables 'x' & 'y'. The values of x and y are plotted as dots on the x-axis and y-axis in xy plane. The diagram of dots are obtain. so, it is known as scatter Diagram.

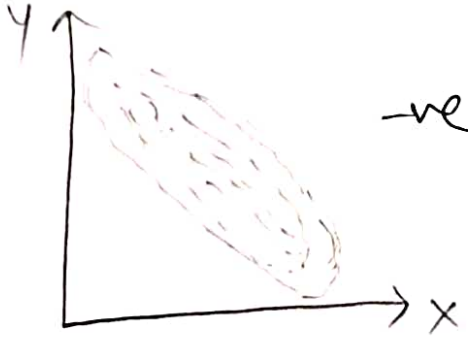
Positive correlation:—

If there is an upward trend raising from lower left hand corner and going upward to the upper right hand corner the correlation is positive.



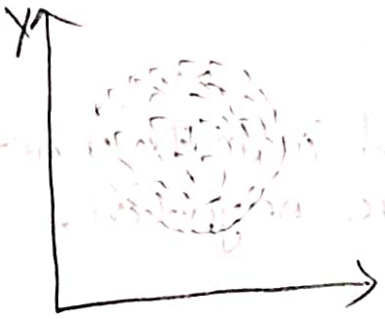
Negative Correlation :—

If there is a downward trend from the upper left hand corner to the lower right corner then it is negative correlation.



$(r<0)$

NO correlation :—

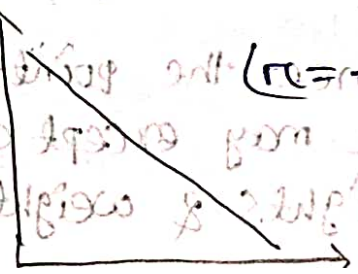


Perfect Positive Correlation :—

If all the points lie on a straight line starting from the left bottom and going upwards the right top the correlation is Perfect Positive.

Perfect negative correlation :—

If all the points lie on a straight line starting From left up and coming down to right bottom the correlation is perfect negative.

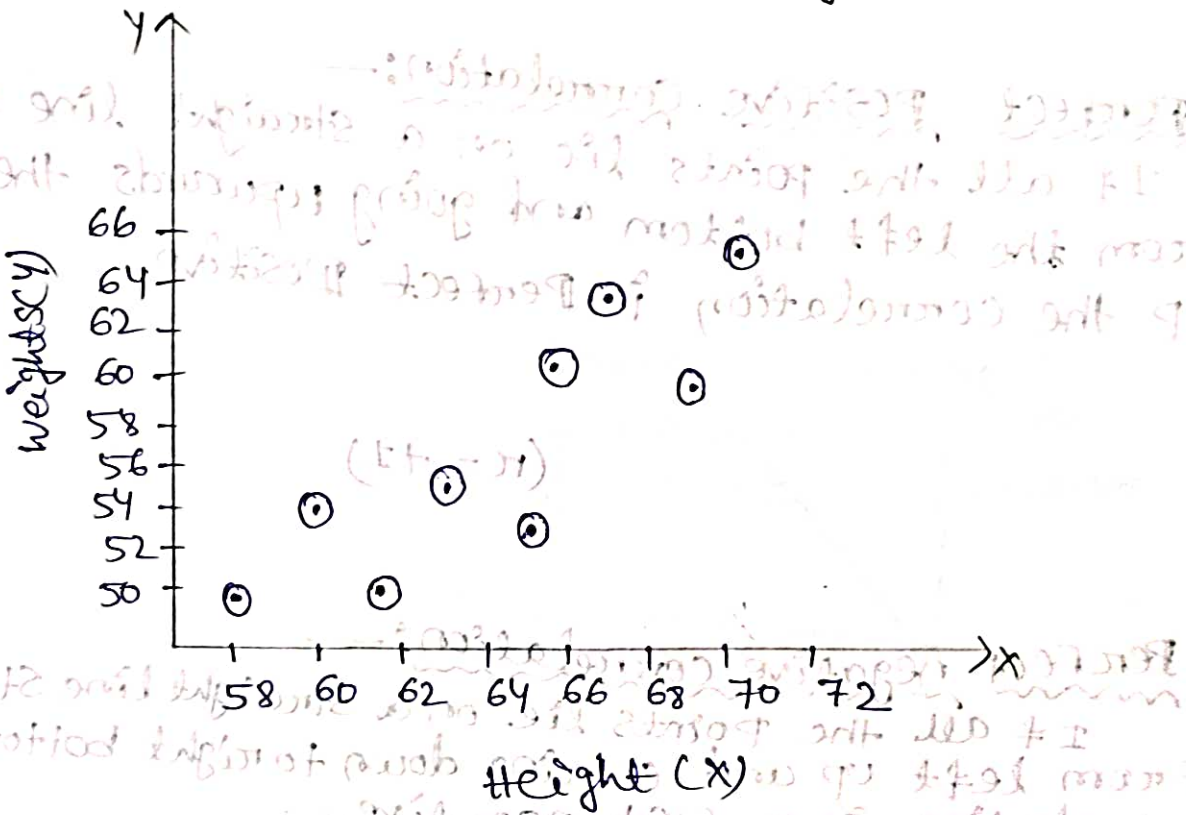


Ex:—

The following are the heights and weights of 10 students of a BCom class.

Heights(X)	Weights(Y)
62	50
72	65
68	63
58	50
65	54
70	60
66	61
63	55
60	54
72	65

Draw a scatter diagram and indicate whether the correlation is positive or negative.



(Since) the points are dense i.e close to each other. So, we may except a high degree of correlation betn the heights & weights further the points are relevant upward

end: starting from the left bottom going up towards the right top. The correlation is positive. So, hence we may except a high degree of +ve. correlation between the series of heights and weights in the class of BCom Student.

Rank Correlation:—

→ It is used to measure the degree of association between two sets of qualitative observations which can be rank graded through scores.

→ The spearman's Rank correlation co-efficient usually denoted by (ρ).

→ It is given by $\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$

(where, d_i = Rank of x − Rank of y
n = no. of observation)

Computation of correlation Co-efficient:—

(i) When actual ranks are given
(ii) When ranks are not given

Case-1 :— When Actual Ranks are Given

The following steps are:—

(i) Compute d, the difference of ranks.
(ii) Compute d^2
(iii) Obtain the sum Σd^2
(iv) Use the formula to get the value of ρ.

EX:—

The Ranks of the 15 students in two subjects A and B are given below. The two numbers within the brackets denoting the ranks of the same student in A and B respectively.

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

Ans:—

Rank(X)	Rank(Y)	$d = x - y$	d^2
1	10	-9	81
2	7	-5	25
3	2	1	1
4	6	-2	4
5	4	1	1
6	8	-2	4
7	3	4	16
8	1	7	49
9	11	-2	4
10	15	5	25
11	9	2	4
12	5	7	49
13	14	-1	1
14	12	-2	4
15	13	2	4
		$\Sigma d = 0$	$\Sigma d^2 = 272$

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 272}{15(15^2-1)}$$

$$= 1 - \frac{6 \times 272}{15 \times 225}$$

$$= 1 - \frac{17}{35} = \frac{18}{35} = 0.51$$

Case-2 When Ranks are not Given

→ When the actual data but not ranks relating to two variables are given. In such a case we shall convert data in to ranks.

→ The highest (smallest) observation is given the rank 1. Then next highest (next lower) observation is given in rank 2 and so on.

→ In case of repeated ranks the rank correlation co-efficient is given by

$$\rho = 1 - \frac{6}{n(n^2-1)}\left[\Sigma d^2 + \Sigma\frac{m(m^2-1)}{12}\right]$$

(where n = no. of times an item is repeated)

EX:—

A psycologist wanted to compare two methods of teaching, he selected a random sample of 20 students, he grouped them in to 11 pairs. so that the students in a pair have approximately equal scores on an intelligence sites in each pairs. one student was taught by method A and method B and examine after the course. The marks often by them are tabulated below. Then find the correlation co-efficient.

Ans:

Pairs	A	B	Rank in A (x)	Rank in B (y)	d=x-y	d^2
1	24	37	6	1	5	25
2	29	35	3	2	1	1
3	19	16	8.5	9.5	-1	1
4	14	26	10	4	6	36
5	30	23	1.5	5	-3.5	12.25
6	19	27	8.5	3	5.5	30.25

7	27	19	5	8	−3	9
8	30	20	15	7	−5·5	30.25
9	20	16	7	9·5	−2·5	6·25
10	28	11	4	11	−7	49
11	11	21	11	6	5	25

$\Sigma d^2 = 225$

We see that in series A, the item 19 is repeated in two times. The correction factor is given by $\frac{m(m^2-1)}{12}$

Where $m = 2$

$C.f = \frac{2(2^2-1)}{12} = \frac{1}{2}$

Similarly 30, $m=2$, $CF = \frac{1}{2}$

In series B, the item 16 is repeated in two times. The correction factor is given by $\frac{m(m^2-1)}{12}$

For 16, ~~where~~ $m=2$

$C.f = \frac{2(2^2-1)}{12} = \frac{1}{2}$

So, the rank correlation is given by,

$$\rho = 1 - \frac{6}{n(n^2-1)}\left[\Sigma d^2 + \frac{m(m^2-1)}{12}\right]$$

$$= 1 - \frac{6}{11(11^2-1)}\left(225 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}\right)$$

$$= 1 - \frac{6 \times 226.5}{11 \times 120}$$

$$= 1 - \frac{226.5}{220}$$

$$= -\frac{26.5}{220} = -0.0295$$

Probable Error:—

→ If r is the correlation co-efficient in a sample of n pairs of observations. The standard error is usually denoted as $S.E(r)$.

→ It is given by

$$S.E(r) = \frac{1-r^2}{\sqrt{n}}$$

→ The probable error of correlation co-efficient is given by

$$P.E(r) = \frac{1-r^2}{\sqrt{n}} \times 0.6745$$

Multiple Correlation:—

→ The multiple correlation is the study of combine effect of two or more variables on a single variable.

Ex:- yield of crop say (X_1) depends upon quality of seed (X_2). fertility of soil (X_3) and so on.

→ The multiple correlation co-efficient is denoted as $R_{1.23}$.

→ It is given by

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1-r_{23}^2}}$$

(r_{12} is the total correlation co-efficient the total correlation co-efficient between two variable x_1 & x_2)

where, $R_{1.23}$ is the multiple correlation co-efficient. we study the multiple impacts of 2nd & 3rd independent variable on 1st dependent variable.

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{21}r_{23}r_{13}}{1-r_{13}^2}}$$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{13}r_{23}r_{12}}{1-r_{12}^2}}$$

→ The limits of multiple correlation co-efficient is 0 to 1 i.e. $0 \leq R_{1.23} \leq 1$.

→ $R_{1.23}$ is always greater than equal to total correlation co-efficient. i.e $R_{1.23} \geq r_{12} r_{13} r_{23}$.

Partial correlation:—

The correlation between two variables x_1 and x_2 after eliminating the linear effect of other variable x_3 is known as partial correlation.

→ It is denoted by $r_{12.3}$.

→ It is given by,

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23})^2}}$$

Where $R_{12.3}$ is the partial correlation between x_1 and x_2. we study the partial impact of 2nd & 1st variable keeping 3rd independent variable constant.

$$r_{13.2} = \frac{r_{13} - r_{12} r_{32}}{\sqrt{(1-r_{12}^2)(1-r_{32}^2)}}$$

$$r_{23.1} = \frac{r_{23} - r_{21} r_{31}}{\sqrt{(1-r_{21}^2)(1-r_{31}^2)}}$$

Ex:—

From the data relating to the yield of dry bark (x_1), height (x_2) and girth (x_3). For 18 cinchona plants. The following correlation co-efficient were obtained.

$r_{12} = 0.77$, $r_{13} = 0.72$, $r_{23} = 0.52$

Find the partial correlation and multiple correlation co-efficient.

Ans:—

Partial Correlation $(r_{12.3}) = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$

$= \frac{0.77 - 0.72 \times 0.52}{\sqrt{(1-(0.72)^2)(1-(0.52)^2)}}$

$= \frac{0.77 - 0.37}{\sqrt{(1-0.51)(1-0.27)}}$

$= \frac{0.4}{\sqrt{0.49 \times 0.73}}$

$= \frac{0.4}{0.59} = 0.67$

Multiple correlation co-efficient $(R_{1.23}) =$

$$\sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1-r_{23}^2}}$$

$$= \sqrt{\frac{(0.77)^2 + (0.72)^2 - 2 \times 0.77 \times 0.72 \times 0.52}{1-(0.52)^2}}$$

$$= \sqrt{\frac{0.59 + 0.51 - 0.57}{1-0.27}}$$

$$= \sqrt{\frac{1.1 - 0.57}{0.73}}$$

$$= \sqrt{\frac{0.53}{0.73}}$$

$$= \sqrt{0.726}$$

$$= 0.85$$

Simple regression:—

The regression analysis confined to the study of only two variable at a time is known as simple regression.

Regression Analysis :—

→ Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of original units of data.

or

→ Regression analysis is the technique for estimating the relationship among variables.

→ In Regression analysis there are two types of variable.

(i) Dependent variable

(ii) Independent Variable

Dependent variable:—

The variable whose value is to be predicted or implemented is called as dependent variable.

Independent variable:—

The variable whose value is used for prediction is called as Independent variable.

Multiple regression :—

The regression analysis study more than two variables at a time is known as multiple regression.

EX: The cost of a product depends on the production and advertising expenditure.

Linear regression:—

If the regression curve is a straight line we say that there is linear regression between the variables under study.

Lines of regression :—

In regression there are always two lines of regression

⇒ lines of regression of Y on X:

It is given by

$$y - \bar{y} = r \cdot \frac{\sigma y}{\sigma x} (x - \bar{x}) \qquad y = ax + b$$

→ Lines of regression of x on y

It is given by

$$x - \bar{x} = r \cdot \frac{\sigma x}{\sigma y} (y - \bar{y}) \qquad x = ay + b$$

Note:—

The two lines of regression passes through the point $\bar{x} \cdot \bar{y}$.

Why there are two lines of regression?

→ There are always two lines of regression, one is regression line of y on x another is regression line of x on y.

→ The line of regression of y on x is used to predict the value of y for any given values of x, where x is independent variable and y is dependent variable.

→ similarly the line of regression of x on y is used to predict the value of x for any given values of y. where x is dependent variable y is independent variable.

Ex:—

Obtain the equations of two lines of regression for the following data also obtain the estimate of x for $y = 70$.

x	y	xy	x^2	y^2
65	67	4355	4225	4489
66	68	4488	4356	4624
67	65	4355	4489	4225
67	68	4556	4489	4624
68	72	4896	4624	5184
69	72	4968	4761	5184
70	69	4830	4900	4761
72	71	5112	5184	5041

$\Sigma x = 544$ $\Sigma y = 552$ $\Sigma xy = 37,560$ $\Sigma x^2 = 37,028$ $\Sigma y^2 = 38,132$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{552}{8} = 69 \qquad \bar{x} = \frac{\Sigma x}{n} = \frac{544}{8} = 68$$

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

$$= \sqrt{\frac{1}{8} \times 38132 - (69)^2}$$

$$= \sqrt{4766.5 - 4761} = \sqrt{5.5} = 2.345$$

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

$$= \sqrt{\frac{1}{8} \times 37028 - (68)^2}$$

$$= \sqrt{4628.5 - 4624}$$

$$= \sqrt{4.5} = 2.121$$

$$r = \frac{\frac{1}{n}\Sigma xy - \bar{x}\,\bar{y}}{\sigma_y \sigma_x} = \frac{\frac{1}{8} \times 37560 - 4692}{2.345 \times 2.121} = \frac{4695 - 4692}{4.973}$$

$= \frac{3}{4.973}$

$= 0.60$

Eq^n of Regression line y on $x = y - \bar{y} = r \cdot \frac{\sigma_y}{\sigma_x}(x - \bar{x})$

$\Rightarrow y - 69 = 0.6 \times \frac{2.35}{2.12}(x - 68)$

$\Rightarrow y = 0.66(x - 68) + 69$

$= 0.66x - 44.88 + 69$

$= 0.66x + 24.12$

Eq^n of line of regression x on y is

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

$\Rightarrow x - 68 = 0.6 \times \frac{2.12}{2.35}(y - 69)$

$\Rightarrow x = 054(y - 69) + 68$

$= 0.54y - 37.26 + 68$

$= 0.54y + 30.74$

To estimate x for given y, we use the line of regression of x on y.

If $y = 70$, estimated value of x is given by

$\hat{x} = 0.54 \times 70 + 30.74$

$= 68.54$

Regression Co-efficient:—

The regression co-efficient of y on x represent the increment in the value of dependent variable y corresponding to a unit change in the value of independent variable x.

b_{yx} = Regression co-efficient y on x

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

The regression co-efficient of x on y represent the increment in the value of dependent variable x corresponding to a unit change in the value of independent variable y.

$$b_{xy} = \text{Regression co-efficient } x \text{ on } y$$
$$= r \cdot \frac{\sigma_x}{\sigma_y}$$

Properties of regression variable:—

(1) Correlation co-efficient (r) is the geometric mean between the regression co-efficient (b_{yx}, b_{xy}).

i.e $r = \pm\sqrt{b_{yx} * b_{xy}}$

(2) If one of the regression co-efficient is greater than unity other must be less than unity.

$b_{yx} > 1$, $b_{xy} < 1$

(3) The regression co-efficient are independent of change of origin but not scale.

(4) The Modulus value of A.M of the regression co-efficient is not less than the modulus value of correlation co-efficient.

i.e $\left|\frac{b_{yx} + b_{xy}}{2}\right| \geq |r|$

NOTE:—

→ If $r = 0$, the two variables are uncorrelated. The lines of regression perpendicular to each other.

→ If $r = \pm 1$, the two variables are coincide. The lines of regression parallel to each other.

Ex:—

In a partially destroyed laboratory, record of an analysis of correlation data, the following results are obtained. The variance of $X = 9$. Regression equations $8X - 10Y + 66 = 0$, $40X - 18Y = 214$.

find (a) Mean value of X & Y

(b) the correlation co-efficient bet^n X & Y

(c) the standard deviation of Y.

Ans:—

We know that both the line of regression passes through $(\bar{X}, \bar{Y})$, we have $8\bar{X} - 10\bar{Y} + 66 = 0$ &

$40\bar{X} - 18\bar{Y} - 214 = 0$

$$8\bar{X} - 10\bar{Y} + 66 = 0 \quad \text{---(1)}$$
$$40\bar{X} - 18\bar{Y} + 214 = 0 \quad \text{---(2)}$$

eq^n(1) $\times 5$ $= 40\bar{X} - 50\bar{Y} + 330 = 0$

$$40\bar{X} - 18\bar{Y} + 214 = 0$$

eq^n(1) − (2) $\quad -32\bar{Y} + 544 = 0$

$\Rightarrow -32\bar{Y} = -544$

$\Rightarrow \bar{Y} = \frac{-544}{-32} = 17$

Putting the value of $\bar{Y}$ in eq^n (1)

$8\bar{X} - 10\bar{Y} + 66 = 0$

$\Rightarrow 8\bar{X} - 10 \times 17 + 66 = 0$

$\Rightarrow 8\bar{X} - 170 + 66 = 0$

$\Rightarrow 8\bar{X} = 104$

$\Rightarrow \bar{X} = \frac{104}{8} = 13$

Solving above eq^n, we get $\bar{X} = 13$, $\bar{Y} = 17$

Let $8x - 10y + 66 = 0$ & $40x - 18y = 214$ be the line of eqn of y on x and x on y.

So, These eqn can be put in the form

$$y = \frac{8}{10}x + \frac{66}{10}$$

$$x = \frac{18}{40}y + \frac{214}{40}$$

So, b_{yx} will be regression co-efficient of y on x.

b_{yx} = regression co-efficient of y on $x = \frac{8}{10}$

b_{xy} = regression co-efficient of x on $y = \frac{18}{40}$

$$r = \pm\sqrt{b_{yx} * b_{xy}}$$

$$= \pm\sqrt{\frac{8}{10} \times \frac{18}{40}}$$

$$= \pm\sqrt{\frac{9}{25}} = \pm\frac{3}{5} = \pm 0.6$$

Since both the regression co-efficient are positive, we take $r = \pm 0.6$

$$b_{yx} = r \cdot \frac{\sigma y}{\sigma x}$$

$$\Rightarrow \frac{4}{5} = \frac{3}{5} \times \frac{\sigma y}{3}$$

$$\Rightarrow \frac{4}{5} \times \frac{5}{3} \times 3 = \sigma y$$

$$\Rightarrow \sigma y = 4$$

Co-efficient of Determination:—

→ The square of the correlation co-efficient is called as co-efficient of determination.

Co-efficient of determination $= r^2 =$	explained variance / Total variance

Ex:—

If the value of $r = 0.9$ then $r^2 = 0.81$ it means that 80% of the variation in the dependent variable has been explained by the independent variable. It gives the percentages variation in the dependent variable that is accounted by the independent variable.

$$Y = ax + b$$

Y ↙ Dependent variable

x ↘ Independent variable

Difference bet^n correlation and regression:—

Correlation	Regression
(1) Correlation co-efficient are used to measure the relationship between two variables.	(1) Regression analysis is to study the nature of relationship between two variables.
(2) In correlation analysis we cannot say that one variable is the cause and the other is the effect.	(2) In regression analysis it is possible to study cause and effect relationship.
(3) In correlation analysis both r_{xy} and r_{yx} are symmetric i.e ($r_{xy} = r_{yx}$)	(3) In regression analysis b_{yx} and b_{xy} are not symmetric. i.e. ($b_{yx} \neq b_{xy}$)
(4) Correlation co-efficient is independent of change of origin & scale.	(4) Regression co-efficient are independent of change of origin but not scale.

UNIT - 4

Principle of least square :—

The Principle of least square states that by getting the sum of square of the errors of a minimum value, the most probable value of a system of unknown quantity can be obtain.

→ The principle of least square states that, the based value of $a_1, a_2 \ldots\ldots a_m$ in the relation

$$Y_i = f(x_i, a_1, a_2, \ldots\ldots a_m) + \epsilon_i^2$$

where $i = 1, 2, 3 \ldots\ldots n$

are those values of $a_1, a_2 \ldots\ldots a_m$ values which is minimize the sum of square of the errors. Hence, to find the values of $a_1, a_2 \ldots a_m$ such that

$$E = \sum_{i=1}^{n} \epsilon_i^2$$

$$= \sum_{i=1}^{n} [Y - f(x_i, a_1, a_2 \ldots\ldots a_m)]^2$$

So, this can be done by solving the normal eq?

$$\frac{dE}{da_1} = 0, \frac{dE}{da_2} = 0 \ldots\ldots \frac{dE}{da_m} = 0,$$

$$Y = ax + b + \epsilon$$

$$\epsilon = \sum_{i=1}^{n} \epsilon$$

$E = \sum (Y - ax - b)^2$ is minimum

The Partial differentiation of E with respect to a and b and equating to zero.

$\frac{dE}{da} = -2\Sigma x(y-ax-b)=0$ and $\frac{dE}{db} = -2\Sigma(y-ax-b)=0$

$\Rightarrow \Sigma xy - a\Sigma x^2 - b\Sigma x = 0$ $\Rightarrow \Sigma y - a\Sigma x - nb = 0$

$\Rightarrow \Sigma xy = a\Sigma x^2 + b\Sigma x$ —(i) $\Rightarrow \Sigma y = a\Sigma x + nb$ —(ii)

solving eq^n(i) and eq^n(ii) we get $\hat{a}$ and $\hat{b}$, the best fitting line is $y = \hat{a}x + \hat{b}$

Fitting of Straight line eq^n :—

The Straight line eq^n is given by $y = ax+b$ using the Principle of list square, we have to determine the constant a and b. The

$E = \Sigma(y-ax-b)^2$ is minimum

The Partial differentiation of E with respect to a and b and equating to zero.

$\frac{dE}{da} = -2\Sigma x(y-ax-b)=0$ $\frac{dE}{db} = -2\Sigma(y-ax+b)=0$

$\Rightarrow \Sigma xy - a\Sigma x^2 - b\Sigma x = 0$ $\Rightarrow \Sigma y - a\Sigma x - nb = 0$

$\Rightarrow \Sigma xy = a\Sigma x^2 + b\Sigma x$ —(i) $\Rightarrow \Sigma y = a\Sigma x + nb$ —(ii)

Solving eq^n (i) and eq^n (ii) we get $\hat{a}$ and $\hat{b}$, the best fitting straight line is $y = \hat{a}x + \hat{b}$.

Ex:— fit the straight line of the forem $y = ax+b$.

Sol^n:—

x	y	xy	x^2
1	8	8	1
3	12	36	9
5	15	75	25
7	17	119	49
8	18	144	64
10	20	200	100
$\Sigma y = 34$	$\Sigma y = 90$	$\Sigma xy = 582$	$\Sigma x^2 = 248$

The normal eq^n are $582 = 248a + 34b$ —— (i)

$90 = 34a + 6b$ —— (ii)

Solving eq^n (i) & (ii)

eq^n (i) $\times 3$ = $1746 = 744a + 102b$

eq^n (ii) $\times 17$ = $1530 = 578a + 102b$

eq^n (i) - (ii) = $216 = 166a$

$= \frac{216}{166} = a$

$= 1.301 = a$

Putting value of a in eq^n (ii), we get

$90 = 34a + 6b$

$\Rightarrow 90 = 34 \times 1.301 + 6b$

$\Rightarrow 90 - 44.2 = 6b$

$\Rightarrow \frac{45.8}{6} = b$

$\Rightarrow 7.63 = b$

Solving eq^n (i) & (ii) we get $\hat{a}$ & $\hat{b}$, $Y = 1.301x + 7.63$.

Fitting of 2nd degree polynomial eq^n:—

The 2nd degree polynomial eq^n is given by,

$$Y = a + bx + cx^2$$

Using the principle of least square, we have to determine the constant a, b and c.

$$\epsilon = \sum_{i=1} (Y - a - bx - cx^2)^2 \text{ is minimum.}$$

The partial difference of ϵ with respect to a, b & c and equating to zero.

$$\frac{d\epsilon}{da} = 2\sum (Y - a - bx - cx^2)^2 = 0 \text{ is minimum}$$

$$\Rightarrow \sum Y - na - b\sum x - c\sum x^2 = 0 \quad \text{—— (i)}$$

$$\frac{d\epsilon}{db} = -2\sum x (Y - a - bx - cx^2) = 0$$

$$\Rightarrow \sum xY - a\sum x - b\sum x^2 - c\sum x^3 = 0$$

$\Rightarrow \Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$ ——(ii)

$\frac{dE}{dc} = -2\Sigma x^2(y - a - bx - c - cx^2)$

$\Rightarrow \Sigma x^2 y - a\Sigma x^2 - b\Sigma x^3 - c\Sigma x^4$

$\Rightarrow \Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$ ——(iii)

Solving eqn (i), (ii) and (iii) we get $\hat{a}, \hat{b}$ & $\hat{c}$ the best fitting line eqn is $y = \hat{a} + \hat{b}x + \hat{c}x^2$.

Fitting of exponential curve:——

(i) $y = ab^x$ ——(i)

Taking logarithms both the side,

$\log y = \log(ab^x)$

$\log y = \log a + x \log b$

Let $U = \log y$, $A = \log a$ & $B = \log b$

$U = A + Bx$

The normal eqns for estimating A and B are

$\Sigma U = nA + B\Sigma x$ ——(ii)

$\Sigma Ux = A\Sigma x + B\Sigma x^2$ ——(iii)

Solving eqn (ii) and (iii), we get

$a = \text{Antilog}(A)$

$B = \text{Antilog}(B)$

With these value of a and b in eqn (i), we get the curve of best fit.

(ii) $y = ae^{bx}$

Taking logarithms both the side,

$\log y = \log(ke^{bx})$ ——(i)

$\log y = \log a + bx \log e$

Let $U = \log y$, $A = \log a$ & $B = b \log e$

$U = A + BX$

The normal eq^n for estimating A and B are

$$\Sigma U = nA + B\Sigma X \quad \text{——(ii)}$$

$$\Sigma UX = A\Sigma X + B\Sigma X^2 \quad \text{——(iii)}$$

Solving eq^n (ii) and (iii), we get

$a = \text{Antilog}(A)$

$B = \frac{B}{\log e}$

With these value of a and b in eq^n (i) we get the curve of best fit.

EX:— Fit an exponential curve of the form $Y = ab^x$ to the following data.

x	y	$u = \log x$	ux	x^2
1	1.0	0.000	0	1
2	1.2	0.0792	0.1584	4
3	1.8	0.253	0.7659	9
4	2.5	0.3979	1.5916	16
5	3.6	0.5536	2.7815	25
6	4.7	0.6721	4.0326	36
7	6.6	0.8195	5.7365	49
8	9.1	0.9590	7.6720	64
$\Sigma x = 36$		$\Sigma U = 3.7393$	$\Sigma Ux = 22.7385$	$\Sigma x^2 = 204$

The normal eq^n for estimating A and B are

$\Sigma U = nA + B\Sigma x$

$3.7393 = 8A + 36B$ ——(i)

$\Sigma Ux = A\Sigma X + B\Sigma X^2$

$\Rightarrow 22.7385 = 36A + 204B$ ——(ii)

Solving eq^n (i) and (ii), we get.

eq^n (i) $\times 9 = 33.65 = 72A + 324B$

(ii) $\times 2 = 45.47 = 72A + 408B$

$\Rightarrow -11.82 = -84B$

$\Rightarrow \frac{-11.82}{-84} = B$

$\Rightarrow 0.1407 = B \Rightarrow b = \text{Antilog}(0.1407)$

$= 1.380$

Putting the value of B in eq^n (i) we get, $3.7393 = 8A + 36 \times 0.1407$.

$\Rightarrow 3.7393 = 8A + 5.0652$

$\Rightarrow 3.7393 - 5.0652 = 8A$

$\Rightarrow \frac{-1.3259}{8} = A$

$\Rightarrow 0.1657 = A \rightarrow a = \text{Antilog}(0.1657) = 0.6821$

The exponential curve eq^n is

$$Y = (0.6821 \cdot 1.380)^x$$

Theory of attributes:—

→ Theory of attributes deals with qualitative characteristics calculated by using quantitative measurement.

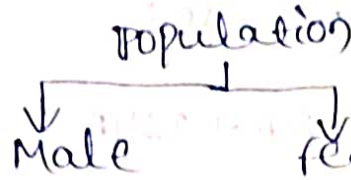
e.g - Honesty, beauty, habit of smoker, intelligence.

→ In the study of attributes the objects are classified according to the absence or presence of the attribute in them.

Dichotomous classification:—

If the universe is divided or to two sub class are known as Dichotomous classification

e.g →

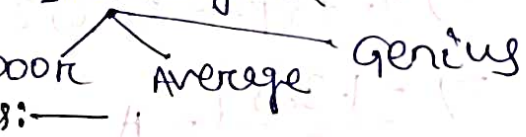


Manifold classification:—

→ If the universe is divided in to more than two classes are known as manifold classification.

e.g → The attribute intelligent is divided in to 4 types.

1. Genius
2. Average intelligent
3. very intelligent
4. Before average intelligent



Types of attributes:—

There are two types of attributes.

(1) Positive attributes (Presence of attributes)

(2) Negative attribute (Absence of attributes)

Symbols and Notations:—

→ The presence of attributes are represented by capital letters : i.e A, B, C ..... Z.

→ The absence of attributes are represented by greek letters i.e α, β, γ etc.

Number of attributes:—

1 - Attributes - 3 classes - N, A, α

2 - Attributes - 9 classes - N, A, B, α, β, AB, αB, Aβ, $\alpha\beta$

⋮ ⋮ ⋮

n - Attributes - 3^n classes -

class frequency:—

Class frequency means no. of frequency assign in each other.

e.g → (A), (AB), (BC), (AC) → +ve frequencies

(α), (β), (γ), (αβ) → -ve frequencies

(ABγ), (αβγ), (Aβ) → Contrary frequencies

Some algebric Expression:—

$(A) = (AB) + (A\beta)$

$(B) = (AB) + (\alpha B)$

$(\alpha) = (\alpha B) + (\alpha\beta)$

$(\beta) = (A\beta) + (\alpha\beta)$

formula for any frequency classes:—

No. of classes of n^{th} order frequency $= {}^nC_n \times 2^n$ classes

n = no. of attributes

n = order

$\boxed{{}^nC_n = 1,\ 0! = 1,\ 1! = 1}$

Q// Find the no. of order 2 classes if the no. of attributes is 2.

No. of 2 order class $= {}^nC_n \times 2^n$

$= {}^2C_2 \times 2^2$

Contingency Table:—

Contingency matrix shows the relation between two variables or attributes.

→ It is first used by Karl pearson.

Attributes	A	α	Total
B	(AB)	(αB)	(B)
β	(Aβ)	(αβ)	(β)
Total	(A)	(α)	N

(AB), (A) (B) → +ve frequency

(Aβ), (α), (β) → -ve frequency

(αβ), (Aβ) → contrary Frequency

Conditions for consistency of data:—

Rule:—(i) Frequency of every class is greater than equals to zero. i.e $(A) \geq 0$.

(ii) Class frequency is less than equals to N (population) i.e $(A) \leq N$.

Ex:— Determine the consistency in the given data.

$(AB) = 70, (\alpha B) = 40, \alpha\beta = 60, (B) = 100$

Attributes	A	α	Total
B	AB 60	αB 40	B 100
β	$A\beta$ 70	$\alpha\beta$ 60	β 130
Total	A 130	α 100	N=230

$AB = B - \alpha B = 100 - 40 = 60$, $\beta = \alpha B + A\beta = 70 + 60 = 130$

$B = 100$, $\beta = 130$, $A = 130$, $\alpha = 100$, $N = 230$.

No frequency is negative and all class frequency is less than N. so the data is consistency

Independent of attributes:—

→ The attributes are said to be independent because their does not exist any relationship between them.

ex:— beauty and intelligence, gender and success

→ If the attributes A & B are said to be independent then the proportion of AB in the population is equal to product of the proportion of A & B in the population. $\frac{(AB)}{N} = \frac{(A)}{N} \cdot \frac{(B)}{N}$

Methodology to Check association and independence attributes:

(1) Proportion Method:—

(i) $\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)}$ → Two attributes are independent

(ii) $\frac{(AB)}{(B)} > \frac{(A\beta)}{(\beta)}$ → Positive association

(iii) $\frac{(A\beta)}{(B)} < \frac{(A\beta)}{\beta}$ → Negative Association

EX:— $(AB) = 100$, $(B) = 10$, $(A\beta) = 150$, $(\beta) = 15$.

Find out how A and B are associated.

Ans:—

$$\frac{(AB)}{(B)} = \frac{100}{10} = 10$$

$$\frac{(A\beta)}{(\beta)} = \frac{150}{15} = 10$$

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)}$$

⇒ Two attributes are independent.

(2) Comparision Method:—

(i) $(AB) = \frac{(A)(B)}{N}$ → Two attributes are independent

(ii) $(AB) > \frac{(A)(B)}{N}$ → Positive association

(iii) $(AB) < \frac{(A)(B)}{N}$ → Negative association

EX:— Find out the association in the given data.

$N = 106$, $(A) = 70$, $(B) = 36$, $(AB) = 20$

Ans:— $(AB) = 20$

$$\frac{(A)(B)}{N} = \frac{70 \times 36}{106} = 23.77$$

$(AB) < \frac{(A)(B)}{N}$ ⇒ Negative association

③ Yule's co-efficient of association Method:—

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

→ Yule's co-efficient of association is lies in between −1 to +1.

Representation of values of Q:—

→ If Q lies bet^n 0 to 1 → positive association
→ If Q lies bet^n −1 to 0 → Negative association
→ If Q = 0 → Independent association
→ If Q = +1 → Perfectly positive association
→ If Q = −1 → Perfectly Negative association

Q// (AB) = 60, (αB) = 20, (Aβ) = 80, (αβ) = 30. find the association.

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} = \frac{60 \times 20 - 80 \times 30}{60 \times 20 + 80 \times 30}$$

$$= \frac{1200 - 2400}{1200 + 2400} = \frac{-1200}{3600} = \frac{-1}{3} = -0.33$$

⇒ Negatively association

Co-efficient of colligation:—

It is another measure of association suggested by Prof. yule's. It is denoted by 'Y'.

$$Y = \frac{1 - \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}{1 + \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}$$

Relation bet^n Q & Y:—

$$Q = \frac{2Y}{1+Y^2}$$

limits of Y is −1 to +1.

UNIT-1

Def^n of statistics:—

Statistics is a science of collection, organization, presentation, analysis and interpretation of numerical data.

→ Father of Morden statistics is Prof. R. A fisher.

→ Father of Indian statistics is Prof. Prasanta Chandra Mahalanobis.

Scope of statistics:—

→ Statistics are numerical statement of facts capable of analysis and interpretation as well as study of method used in collection, organization, presentation, analysis and interpretation of numerical data.

→ Statistics are used in different sectors such as statistics in Economics, in education, in planning, in commerce, in business etc.

Limitation of statistics:—

→ Statistic does not study for qualitative phenomenon.

→ Statistics does not study for individuals.

→ Statistics laws are not exact.

→ Statistic is liable to misused.

Function of statistics:—

(1) Collection of data:—

→ The 1st step of an investigation is a collection of data.

→ Carefully collection is needed because further analysis is based on this.

→ Data must be reliable. If the data is not reliable the collected data are faulty. Therefore the good investigator must taken special care in collection of data.

(2) Organization of Data :—

→ Organization of data refers to the arrangement of failures. In such a form of comparisons of the mass similar data may be facilated and further analysis may be possible.

(3) Presentation of Data :—

→ The collection of data are generally in raw form and need to be classified and tabulated before they can analysed.

→ Therefore the collected data are to be presented in tabular or diagramatic or graphical form.

(4) Analysis of Data :—

→ After the presentation of data the next step is to analysis the presented data.

→ Analysis includes some condensation, summerization and conclusion through measures of central tendency, dispersion, skewness etc.

(5) Interpretation of Data :—

→ Interpretation of Data takes the result of analysis and makes interference to a perticular research studies and draw conclusions about these relations.

What do you mean by Data :—

→ Data is a collection of information or facts on a certain enquiry is called as Data.

→ There are two types of Data.

(i) Primary Data

(ii) Secondary Data

(i) Primary Data :—

→ Primary Data refers to the first hand data gather by researcher himself.

→ The primary Data are usually in a row or bulky form (wide, heavy, large)

(ii) Secondary Data :—

→ Secondary Data is a data that has already been gathered or collected form others.

→ Secondary data is accessible in the form of data collected form different sources such as goverment, publication, census, books, journals, articles, websites etc.

Methods of collecting primary Data :—

(1) Direct personal investigation

(2) Indirect oral interview

(3) Information received through local agencies

(4) Mailed Questionnaire Method

(5) Schedules sent through enumerators

Direct Personal Investigation :—

Direct Personal Investigation or personal interview is a method of collecting primary data through

which the investigator contacts the informant directly to collect data by conducting on the spot enquiry.

(II) Indirect oral interview:—

Indirect oral investigation is a method of collecting primary data through which the investigator approaches third parties who are in possession of required information about the subject of enquiry.

(III) Information received through local agencies:—

It is a method of collecting primary data under which the investigator appoints local persons or correspondents at different places. These correspondents collect information in their own way and provide it to the investigator.

(IV) Method Questionnaire Method:—

The Questionnaires are the fundamental instrument for gathering information in review research. Fundamentally, it is a bunch of standardized questions, frequently called items which follow a decent plan together individual information around at least one explicit theme.

(V) Schedule sent through enumerators:—

In this method, list of questions or schedules are sent to the informants through the enumerators.

Classification:—

The arrangement of data in to different classes, which are to be determined depending upon the nature, objectives and scope of the enquiry.

EX:— The number of students registered in UNIC college in the academic year 2022-23 may be classified on the basis of any of the following Criteria.

(i) Sex
(ii) Age
(iii) Religion
(iv) state which they belong.
(v) Height
(vi) Weight

→ The data are classified Primarily depending upon the objective and the Purpose of the enquiry. The data can be classified in to following 4 basis.

(i) Geographical Classification
(ii) Chronological Classification
(iii) Qualitative Classification
(iv) Quantitative Classification

(i) Geographical classification:—

In this Classification the basis of classification is the geographical or locational differences between the various items in the data like:—states, cities, regions, areas etc.

EX:— The yield of agricultural output for hector for different countries in given period/Density of Population is different cities of India

(ii) Chronological classification:—

Chronological Classification is one of the data are classified on the basis of differences in time.

Ex:— The population of any country for different years.

(iii) Qualitative classification:—

When the data are classified according to some qualitative phenomenon which are not capable of quantitative measurement like:- honesty, beauty, intelligence etc.

(iv) Quantitative classification:—

When the data are classified on the basis of Quantitative Phenomenon which are capable of qualitative measurement like:— height, weight, income, expenditure etc.

Variable:—

The qualitative phenomenon under study like marks in a test, height, weight, number of student in a class etc. is known as variables.

→ There are two types of variable:

① Discrete variable
② Continuous variable

Discrete variable:—

Those variables which can not take all possible values within a given specified range is known as Discrete variable.

Ex:- Marks in a test, number of student in a class

Continuous variable:—

Those variables which can take all possible values within a given specified range is known as Continous Variable.

Ex:— Height, weight, Age etc.

Class interval:—

Class interval in a numerical width in a class and frequency distribution each class is specified with two extreme value is called as class interval.

→ The smaller one is Lower limit and the bigger one is upper limit.

→ There are two type of class interval

(i) Inclusive
(ii) Exclusive

Inclusive class interval:—

The class of type like: 1-9, 10-19, 20-29 etc. in which both upper limit and lower limit includes in a class is known as Inclusive class interval.

Exclusive class interval:—

The class of type like- 0-10, 10-20, 20-30 etc. in which lower limit is include and the upper limit is excluded from the respective class and immediately included in the next class is known as Exclusive class interval.

Mid value:—

Mid value of any class can be obtained by dividing the sum of lower limit and upper limit.

$$\text{Midvalue} = \frac{\text{Lower limit} + \text{Upper limit}}{2}$$

Ex:—

Age	Midvalue
10-20	15
20-30	25
30-40	35

Class Boundaries:—

→ In inclusive type classification or ungrouped frequency distribution there are the gap between the upper limit of any class and lower limit of succedding class. So, there is need to convert the data in to a continuous freqeuncy distribution.

→ The upper and lower limit of new enclusive type class are called as class boundaries.

→ If d is the gap between upper limit of any class and lower limit of succeeding class, so, the class Boundaries of any class is given by

* Upper class Boundaries = upper limit + $\frac{d}{2}$
* Lower class Boundaries = Lower limit - $\frac{d}{2}$

OPEN END Class:—

If the lower limit of first class and upper limit of last class are not specified and this classes one of the limit is missing is called as open End classes.

Ex:— Less than 30, age above 60.

Tabulation of Data:—

After classification the next step in the purpose of summerisation of data is to put the classification data in rows and columns having special Characteristics on a piece of paper. such a representation of data is known as Tabulation.

Silent features of a Table:—

→ A table should have a table number, table title and subtitle for rows and columns in the body of the table.

→ A good table should be simple to understand.
→ The numerical figures represented in the table should be compare the other related figures without include waiting for further calculation.
→ A good table should be proper size and shape with proper specing for rows and columns and should have attractive get up.
→ The data in co-operated in the table should be order alphabetically, geographically or depending upon the necessity of facilitate ready for comparison.

Types of Table:—

Classification of tables according to purpose-

(1) General purpose table:—

It also known as information table and reference table. It providing a symmetric manner usually in chronological order. All information connected with enquiry. This table are also called as ~~General purpose table~~. Master table.

(2) Summary Table:—

This table are constructed and derived from the master table with certain purpose and are useful form analytical and comparating studies, involving the study relationships among variables. Calculation of analytical statistics like ratio, percentage, etc. are in co-operated to this tables.

Classification of tables according to nature of data:-

There are two types.

① Primary table:—

The primary tables represence data in original form without arithmetical rounding up.

② Derived table:—

This table Condensed figures and entries.

Eg→ ratio, percentages etc. From the primary data

Classification of table according to characteristics:-

① Simple table:—

A simple table is the one way classification dealing with only one characteristics.

② Complex table:—

A Complex table is based on more than one Characteristics.

Graphical Representation of data:—

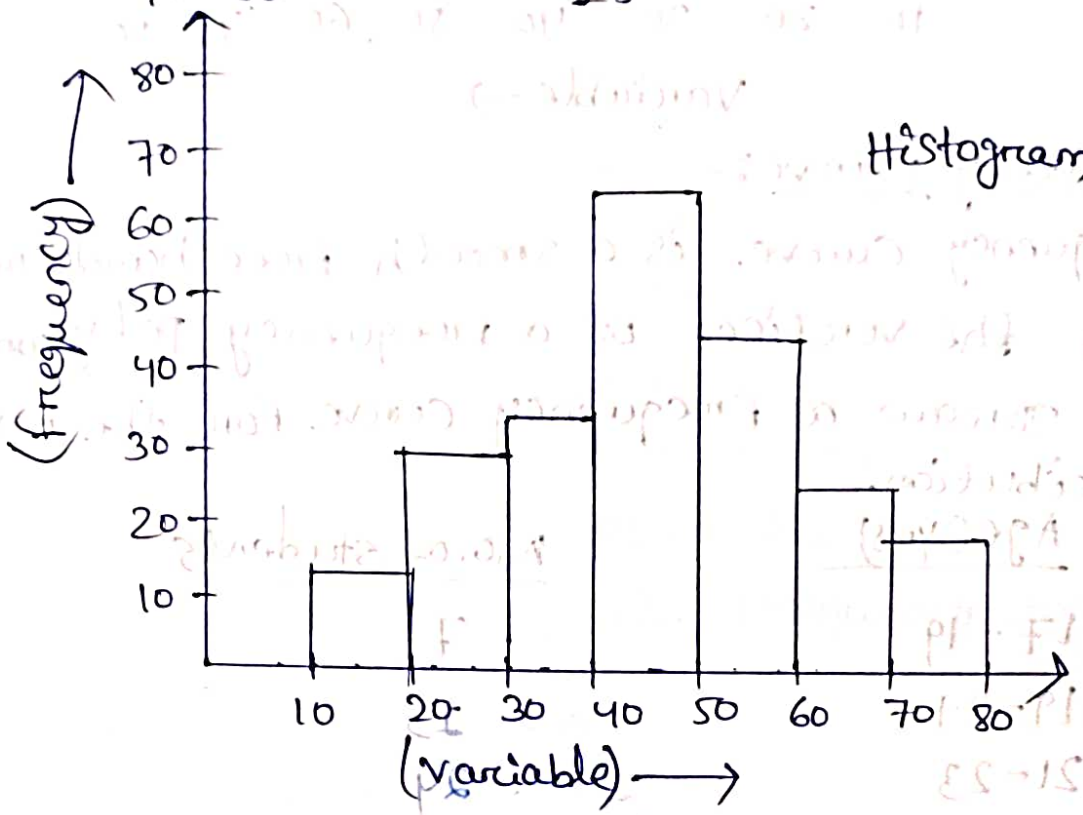
(1) Histogram:—

Histogram is a graphical representation of continuous frequency distribution. It consists of series of adjacent vertical rectangle on the sections of the horizontal axis (x-axis) with bases section equal to width of the corresponding class interval and heights are taken the areas of a rectangles are equal to the frequencies of the corresponding class.

The variable values are taken along the x-axis and the frequency along the y-axis.

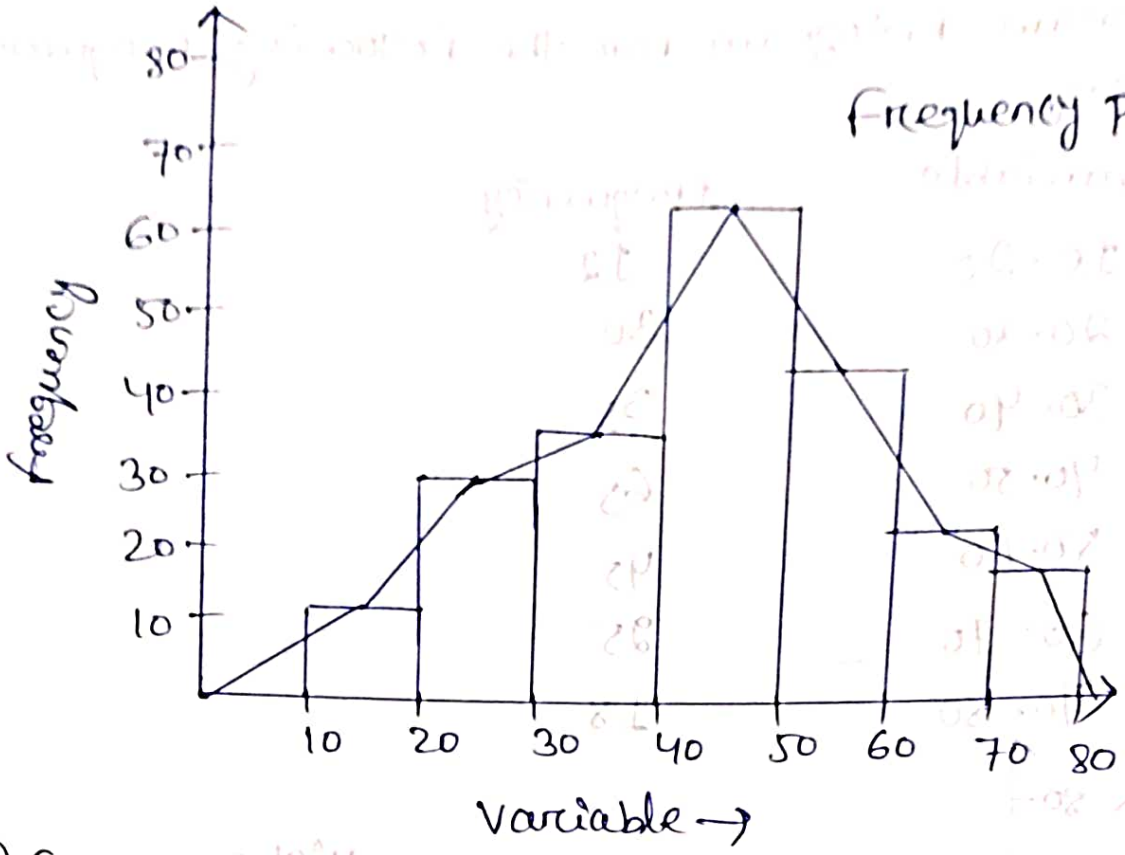
Ex:— Draw histogram for the following frequency distribution.

Variable	Frequency
10-20	12
20-30	30
30-40	35
40-50	65
50-60	45
60-70	25
70-80	18



(2) Frequency Polygon:—

Frequency polygon is obtain on plotting the frequencies on the vertical axis (Y-axis) against the corresponding values of the variable on the horizontal axis (X-axis) and joining the midpoints obtained by the straight line.

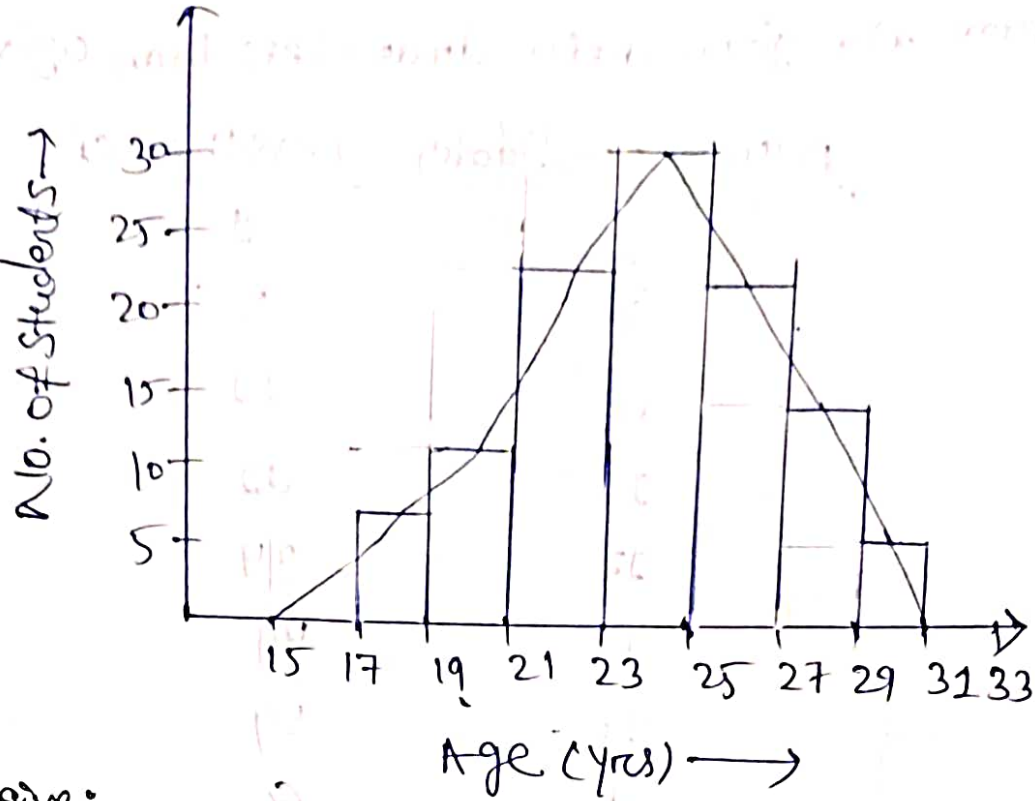


③ Frequency Curve:—

Frequency curve is a smooth free hand curve through the vertices of a frequency polygon.

Ex:— Draw a frequency curve for the following distribution.

Age (yrs)	No. of students
17 - 19	7
19 - 21	13
21 - 23	24
23 - 25	20
25 - 27	29
27 - 29	15
29 - 31	6



Ogive:—

→ The graphical representation of the partition value is known as ogive. In other words, ogive is a graphical representation of cumulative frequency distribution of continuous variable. It consist plotting the cumulative frequency (along the y axis) against the class boundaries of the x-axis. So, there are two types of cumulative frequency distribution.

(i) Less than ogive
(ii) More than ogive

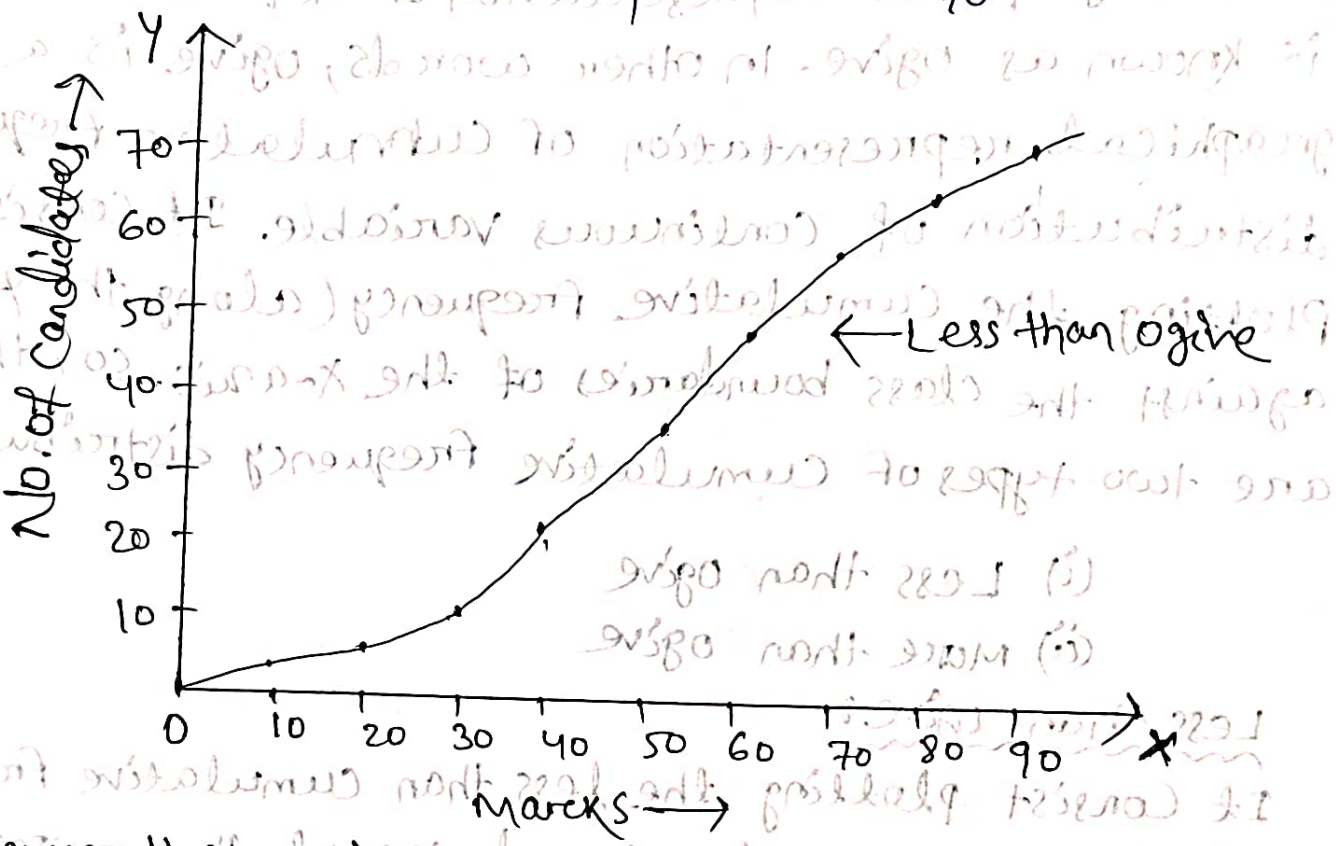
Less than ogive:—

It consist plotting the less than cumulative frequency against the upper class boundaries of their respective classes. The points are obtained and are joined by smooth hand curve.

→ Less than ogive an increasing curve, sloping upward from left to right and the shape is 'S' type.

Ex:— From the given data draw less than ogive.

Marks	No. of candidates	Less than cf
0-10	2	2
10-20	3	5
20-30	6	11
30-40	11	22
40-50	12	34
50-60	15	49
60-70	10	59
70-80	7	66
80-90	4	70



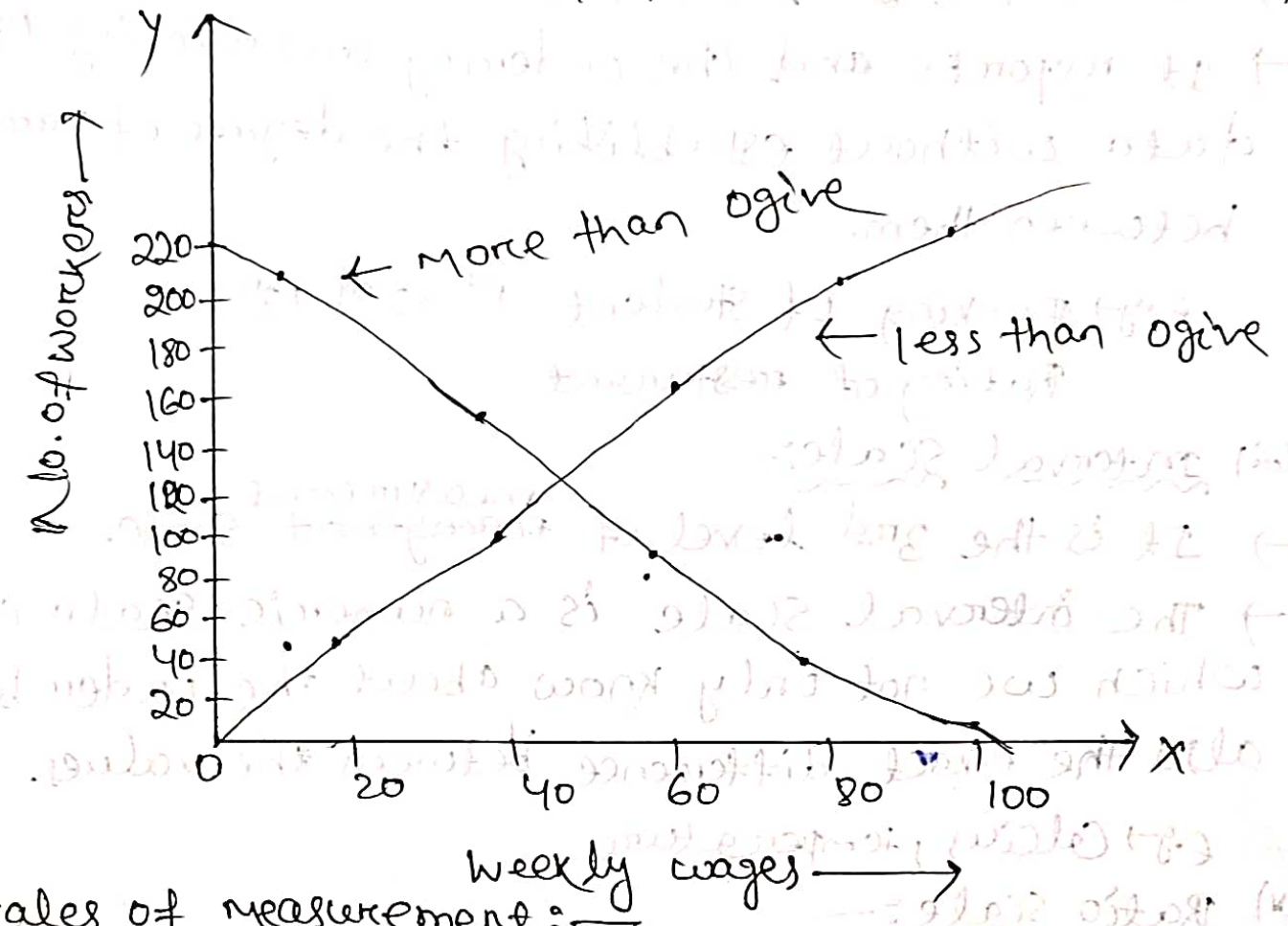
(2) More than ogive:—

In more than ogive or more than cumulative frequency are plotted against the lower class boundaries of that respective classes. The points are obtained and are joined by smooth hand curve.

→ More than ogive is a decreasing curve and slopes downward from left to right.

EX:— Draw less than and more than ogive and find the value of median.

Weekly wages	No. of workers	less than cf	More than cf
0-20	41	41	201
20-40	92-41=51	92	201-41=160
40-60	156-92=64	156	160-51=109
60-80	194-156=38	194	109-64=45
80-100	201-194=7	201	45-38=7



Scales of measurement:—

The scales of measurement is a classification of that describe the nature of information within the number assign to the variable.

→ There are four types of scale.

(i) Nominal scale
(ii) ordinal scale
(iii) Interval scale
(iv) Ratio scale

(i) Nominal Scale:—

→ It is the 1st level of measurement scale.
→ It deals with non-numeric variable or the number that do not have any value.
e.g → sex, Region, Cast etc.

(ii) Ordinal scale:—

→ It is the 2nd level of measurement scale.
→ It reports and the ordering and ranking of data without establishing the degree of variation between them.
e.g → Ranking of student 1st, 2nd, 3rd
Rating of resturant

(iii) Interval Scale:—

→ It is the 3rd level of ~~management~~ measurement scale.
→ The interval scale is a numeric scale in which we not only know about the order but also the exact difference between the values.
eg → Celcius, temperature

(iv) Ratio Scale:—

→ It is the 4th level of measurement scale which is quantitative.
→ It allow the researcher to compare the difference between unit.

→ one of the most important property that is zero position indicates the absence of quantity being measure.

eg) Hight, weight etc.

Concepts of statistical Population & sample:—

Population:— Group of individual under study is known as population.

EX:— All people living in India indicates the population of India

Sample:— It indicates one or more observations that are drawn from the population.

EX:— All SIKhs people living in India is the sample of population.

Tabulation:—

After classification, the next step in the purpose of summerization of data is to put the classified data in rows and columns having special characteristics on a piece of paper. such representation of data is called as tabulation.

Graphical:—

The graphs provide an alternative method of representing data in a condensed & summary form. A graph is a scale-dependent geometrical figures and provides a visual representation of statistical data.