Basic Statistics

Measure of central tendancy (Average)

1) Arithmetic Mean
2) Median
3) Node
4) Geometric Mean
5) H.M (Harmonic Mean)

Measure of central tendancy or an average is a single significant figure which sums up the characteristics of a group of figures. Various measures of central tendancy are arithmetic mean, Median mean, Node, Geometric Mean, Harmonic mean

Arithmetic Mean :-

Case I - Individual Series

let $x_1, x_2 \ldots x_n$ are n values then arithmetic mean of these numbers is given by $A.M = \frac{x_1 + x_2 \ldots x_n}{n}$

$$\boxed{Am = \frac{\Sigma x}{n}}$$

Q Find mean of 10, 12, 8, 14, 16, 19 ....

$$\frac{10+12+8+14+16+19}{6} = 13.167$$

Q Find mean of 2.6, 8.4, 1.5, 10.9, 4.5, 6.8, 7.5, 5.9, 6.7, 8.8

$$\frac{2.6+8.4+1.5+10.9+4.5+6.8+7.5+5.9+6.7+8.8}{10} = 6.36$$

Discreat frequency

let $x_1, x_2 \ldots x_n$ are n values with corresponding frequencies $f_1, f_2 \ldots f_n$ then $Am = \frac{f_1x_1 + f_2x_2 + \ldots + f_nx_n}{f_1 + f_2 + \ldots f_n}$

Q Calculate mean from the following data

value	5	15	25	35	45	55	65	75
freq	15	20	25	24	12	31	71	52

x	f	fx
5	15	75
15	20	300
25	25	625
35	24	840
45	12	540
55	31	1705
65	71	4615
75	52	3900
	250	12600

The following data related to the destance travelled by 520 villagers to buy their weekly requirment

x	f	fx
2	38	56
4	104	416
6	140	840
8	78	624
10	48	480
12	42	504
14	28	392
16	24	384
18	16	288
20	2	40
	520	4024

$$\frac{4024}{520} = 7.7769$$

Case III mean in contenues Freqvency distribution

In the case of contenues Freqvency distribution we can take mid value of the class x then $Am = \frac{\Sigma fx}{N}$

Q Find the Arithemetic mean from the following data

Age :	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of :	15	30	53	75	100	110	115	125

Class	Frequency(f)	Midvalue (x)	fx
0-10	15	5	75
10-20	30	15	450
20-30	53	25	1325
30-40	75	35	2625
40-50	100	45	4500
50-60	110	55	6050
60-70	115	65	7475
70-80	125	75	9375
	623		31,875

$$\frac{31875}{623} = 51.16$$

Median

Median is the value of that item which occupies the central position when the items are arranged in the ascending or descending order of their magnitude. Therefore median is the value of that item which has eavual number of items above & below that is, number of items greater than median and number of items less than median are eavual.

(a) Median in Individual Series

Arrange the values in the data in the ascending or descending order of their magnitude and find out the value of middle item. It is the median

It is also called size of $\left(\frac{n+1}{2}\right)$th item.

Q Find the median for the following values?

4, 45, 60, 20, 83, 19, 26, 11, 27, 12, 52

- writing the values in the ascending order

4, 11, 12, 19, 20, 26, 27, 45, 52, 60, 83

Median = Size of $\left(\frac{n+1}{2}\right)^{th}$ item = 6^{th} item = 26

when 'n' is even, there are two middle items. Take the average of them.

Q calculate median

35, 23, 45, 50, 80, 61, 92, 40, 52, 61

- writing the values in the ascending order

23, 35, 40, 45, 50, 52, 61, 61, 80, 92

median = size of $\left(\frac{n+1}{2}\right)^{th}$ item = size of $\left(\frac{10+1}{2}\right)^{th}$

item = size of 5. 5^{th} item = $\left(\frac{50+52}{2}\right) = 51$

(b) Median in Discrete frequency distribution

Formula : Median = Size of $\left(\frac{N+1}{2}\right)^{th}$ item,

where $N = \Sigma f$

Q calculate Median

Size	5	8	10	15	20	25
Frequency	3	12	8	7	5	4

Size	Frequency (f)	Cumulative Frequency
5	3	3
8	12	3 + 12 = 15
10	8	15 + 8 = 23
15	7	23 + 7 = 30
20	5	30 + 5 = 35
25	4	35 + 4 = 39

$N = 39$

$\frac{N+1}{2}$

Median = size of $\left(\frac{N+1}{2}\right)^{th}$ item = size of $\left(\frac{40}{2}\right)^{th}$ item = size of 20th item

Median corresponds to 20th item of the series

The first cumulative frequency which includes 20 is 33.

The size of item for which cumulative frequency is 23, is 10 ∴ median = 10 //

(a) Median in continuous frequency distribution

In the case of continuous frequency distribution, median class corresponds to the cumulative frequency which includes $N/2$. After getting median class, find median by using the following interpolation formula

$$\text{Median} = l_1 + \frac{\left(\frac{N}{2} - cf\right)}{f} \times c$$

where l_1 is the lower limit of median class. 'cf' is the cumulative frequency of the class just proceeding the median class. 'f' is the frequency of the median class. 'c' is the interval of median class.

Q Compute median

class	0-10	10-20	20-30	30-40	40-50	50-60	60-70
frequency	8	12	20	23	18	7	2

class	frequency	cumulative frequency
0-10	8	8
10-20	12	20
20-30	20	40
30-40	23	63 $\frac{N}{2}$
40-50	18	81
50-60	7	88
60-70	2	90
	90	

Median = Size of $(\frac{N}{2})^{th}$ item = Size of $(\frac{90}{2})^{th}$ item = Size of 45^{th} item 45 is included in the cumulative frequency 63. The class having cumulative frequency 63 is 30-40

∴ 30-40 is the median class.

Applying interpolation formula,

$$\text{Median} = l_1 + \frac{(\frac{N}{2} - cf)}{f} \times c$$

Here $l_1 = 30, N/2 = 45, cf = 40, f = 23, c = 10$

$$\text{median} = 30 + \frac{(45-40)}{23} \times 10 = 30 + \frac{50}{23} = 30 + 2.7$$

$$= 32.17$$

Q Find ~~the~~ median of the values

17, 32, 35, 33, 15, 21, 41, 32, 11, 10, 20

- 10, 11, 15, 17, 20, 21, 32, 32, 33, 35, 41

$$\text{median} = \text{size of } \left(\frac{N+1}{2}\right)^{th} \text{ item} = 6^{th} \text{ item} = 21$$

Q The following table shows age of 8 student find median

Sl No :	1	2	3	4	5	6	7	8
Age :	18	16	14	11	13	10	9	20

class	Frequency	Cumulative frequency
1	18	18
2	16	34
3	14	48
4	11	59
5	13	72
6	10	82
7	9	91
8	20	101

~~N = 101~~

9, 10, 11, 13, 14, 16, 18, 20

$$\frac{N+1}{2} = \frac{8+1}{2} = \frac{9}{2} = 4.5 \qquad \frac{13+14}{2}$$

$$= 13.5$$

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

$= \frac{101+1}{2} = 51^{th}$ item

The 1st cumulative frequency which include, 51 is 59 The size of item for which cumulative frequency is 5 ||

$\therefore$ Median = 10

Q Find the average no. of person per house using median

No. of person in house	1	2	3	4	5	6	7	8	9
No. of house	26	113	120	95	60	42	21	14	54

class	Frequency	Cumulative frequency
1	26	26
2	113	139
3	120	259
4	95	354
5	60	414
6	42	456
7	21	477
8	14	491
9	54	545

N = 545

Mode

Median = size of $\left(\frac{N+1}{2}\right)^{th}$ item = size of $\left(\frac{546}{2}\right)^{th}$

item = size of 273th item.

The 1st C.F which includes 273 is 354, 4

∴ Median = 4

Q Find the median from the following data

Marks :	15-25	25-35	35-45	45-55	55-65	65-75	75-85
no. of std :	3	5	12	15	9	9	7

Class	F	Comulative frequency
15-25	3	3
25-35	5	8
35-45	12	20
45-55	15	35 ← N/2
55-65	9	44
65-75	9	53
75-85	7	60
	60	

Median = size of $\left(\frac{N}{2}\right)^{th}$ item = size of $\left(\frac{60}{2}\right)^{th}$ item

= size of 30th item 30 is included in the C.F 35. The class having C.F 35 is 45-55

∴ 45-55 is the median class applying interpolation formula:

$$\text{Median} = l_1 + \frac{\left(\frac{N}{2} - CF\right) \times C}{f}$$

$l_1 = 45, \frac{N}{2} = 30, CF = 20, f = 15, c = 10$

$$45 + \left(\frac{30 - 20}{15}\right) \times 10$$

$$= \frac{155}{3}$$

$$= 51.6667$$

Q Calculate median from the following data

class :	10-20	20-40	40-70	70-120	120-140
F :	4	10	26	8	2

Class	F	CF
10-20	4	4
20-40	10	14
40-70	26	40 ← N/2
70-120	8	4
120-140	2	50
	50	

Median = size of $(N/2)^{th}$ item = size of $\left(\frac{50}{2}\right)^{th}$ item

$= 25^{th}$ item. 25 is included in the C.F 40. The class having C.F 40 is 40-70

$\therefore$ 40-70 is the median class interpolation formula

$$\text{media} = l_1 + \frac{\left(\frac{N}{2} - C.F\right)}{f} \times C$$

$$40 + \frac{(25-14)}{26} \times 30$$

$$= 50.69$$

$l_1 = 40$

$N/2 = 25$

$C.F = 14$

$f = 26$

$C = 30$

Mode

The value of the variable which occur most frequently in a distribution is called mode.

Mode in Individual Series :-

In the case of individual Series the value which occures more no. of times is mode

Q find mode 23 35 28 42 62 53 35 28 42 35 23 42 35

Ans. 35 is the mode

NOTE :

when no item appears more no.of times than other we say mode is ill-defined in that case mode is obtain by the formula mode = 3 median - 2 mean

Q Find mode from the values

Ans 40, 25, 60, 35, 81, 75, 90, 10

Here all items appear eavual no.of times so mode is ill-defined mode is obtain by the formula 3 meadian - 2 mean

Mean $\frac{416}{8} = 52$ 10, 25, 35, 40, 60, 75, 81, 90

median $= \frac{n+1}{2} = \frac{8+1}{2} = 9/2 = 4.5$

$\frac{(40+60)}{2} = 50$

mode = 3median − 2mean

$= 3 \times 50 - 2 \times 52 = 46$

Mode is discreate frequency distribution

In the case of discreate frequancy distribution the value having highest frequancy is taken as mode.

Find mode

Size :	5	8	10	12	29	35	40	46
no of item :	3	12	25	40	31	20	18	7

Ans. Mode is 12

The value has the highest frequancy so mode is equal to 12

Mode in continuos frequancy distribution. In the case of continuos frequancy distribution mode lays in the class having the highest frequancy this class is known as mode class from the model class mode is calculated by the formula

$$l + \frac{(f_1 - f_0)c}{2f_1 - f_0 - f_2}$$

where l = lower limit of model class

f_0 = frequancy of class just preceeding model class

f_2 = frequancy of the class just saceeding model class

f_1 = frequancy of model class

c = Interval of the model class

Q calculate mode from the following data

Size :	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50
freq :	4	8	18	30	20	10	5	2

25-30 is the model class then

$l = 25$

$f_0 = 18$

$f_1 = 30$

$f_2 = 20$

$c = 5$

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

$$= l + \frac{(f_1 - f_0)c}{2f_1 - f_0 - f_2}$$

$$= 25 + \frac{(30 - 18)5}{2 \times 30 - 18 - 20}$$

$$= 47.73$$

Essential Properties (or characterestics) of Good Average

1) An average is regarded as a good average if it has the following properties.

1) It should be clearly defined.
2) It should be based on all the observations of the data.
3) It should be easy to calculate and simple to follow
4) It should not be influenced by sampling fluctuations
5) It should be amenable to further algebraic treatment.
6) It should not be affected by extreme items

MERITS and DEMERITS of measurement of central Tendency merits of Arithmetic mean

1) Arithmetic mean is simple to understand.
2) Arithmetic mean can be easily calculated
3) Arithmetic mean can be determined in most of the cases.
4) It is based on all the observations of the series.
5) It is capable of more algebraic treatment.
6) Arithmetic mean is stable. It does not differ from sample to sample.

Demerits of Arithmetic Mean :-

1) Arithmetic mean is affected by extreme values.
2) Usually mean does not coincide with any of the observed values.
3) It is not suitable for averaging ratios and percentages.
4) It cannot be calculated for qualitative data which cannot be measured numerically.
5) It may offer misleading and absurd results in some cases. For example average number of children per family in a village may work out as 3.2 which is impossible.
6) In the case of frequency distributions with open end classes, the mid values of all classes cannot be obtained. Therefore in such distributions, mean cannot be calculated accurately.

Merits of Median :-

1) It is very simple measure
2) Some times it can be located even by a mere glance
3) It is not affected by extreme items
4) It is suitable for even such data which are not capable of numerical expression
5) It can be determined graphically also.
6) Median is the suitable average in the case of

Open and class series.

Demerits of Median :-

1) Median is not based on all the observations.
2) It is not capable of algebraic treatments.
3) It requires arraying (ie writing in the ascending or descending order) also.
4) In the case of continuous series interpolation formula is to be used. The value thus obtained may give only an approximate value.

Merits of Mode :-

1) Mode is a very simple among measures of central tendancy. Even a glance at the series is enough to locate the modal value. It is a popular average.
2) Mode is less affected by extreme values in the series.
3) Mode can be located graphically also.
4) Usually mode coincides with one of the values of the series.
5) Mode is the value which occurs most frequently. So mode is the best representative.

~~Demerits of Mode :-~~

Q Why is Arithmetic mean considered to be the best average?

- Arithmetic mean can be easily calculated. It is simple to understand. It is well defined. It can be determined in most of the cases. It is capable of more algebraic treatment. It is based on all observations of the data. It can be located even without arraying data. It is stable as it does not differ from sample to sample when the sample selected is sufficiently large. Thus mean satisfies many of the properties of a good average. That is why Arithmetic mean is considered to be the best among commonly used averages. It has become an average of every day life. It is used in the study of many social and economic problems.

Compare Mean, Median and Mode :-

Mean is a mathematical average while Median and mode are positional averages. Mean is based on all observations while median is the middle item and mode is commonly found item. Mean is capable of farther mathematical treatment while median and mode are not capable of more mathematical treatment. Median and mode are values found in the series while mean is not necessarily be such a value. Median and mode can be graphically located.

But mean cannot generally be located graphically. Median and Mean can be obtained by mere inspection while mean can be obtained by calculation. Mean is affected by extreme items while median and mode are not much affected by extreme items. Mode is ill-defined in some cases but median and mean are well defined in all cases.

Demerits of Mode :-

1) mode is uncertain and vague measure of central tendancy. It is illdefined in some cases.

2) mode is not capable of further algebraic treatment

3) Some times grouping becomes necessary to identify the mode value.

4) mode is not based on all the values of the series

5) when the extreme value appears most frequently the extreme value becomes mode. for eg:- If zero appears more number of times then mode is zero This is not a representative value.

Geometric Mean :-

If there are n value in a series then their G.M is defined as the nth root of the product of those n values G.M is a mathematical average.

Geometric Mean in Individual Series :-

If $x_1, x_2 \ldots x_n$ are n values then

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

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

$$\text{Log } G.M = \log (x_1 \times x_2 \times \ldots \times x_n)^{1/n}$$

$$= \frac{1}{n} \log (x_1 \times x_2 \times \ldots \times x_n)$$

$$= \frac{1}{n} [\log x_1 + \log x_2 + \ldots \log x_n)$$

$$= \frac{\log x_1 + \log x_2 \ldots + \log x_n}{n}$$

$$\log G.M = \frac{\Sigma \cdot \log x}{n}$$

$\log x^n = n \log x$

$\sqrt{x} = x^{1/2}$

$\sqrt[3]{x} = x^{1/3}$

$\sqrt[n]{x} = x^{1/n}$

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

Q calcalate G.M of the following figure

57.5, 87.75, 53.5, 73.5, 81.75

x	log x
57.5	1.7597
87.75	1.9432
53.5	1.7283
73.5	1.8663
81.75	1.9125
	9.21

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

$$= \text{Antilog}\left[\frac{9.21}{5}\right]$$

$$= \text{Antilog}\,[1.842]$$

$$= 69.51$$

G.M in frequency distribution :-

For Discrete frequency distribution

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

Q Find G.M from the following data

Size : 5 8 10 12

Freq : 2 3 4 1

x	f	log x	f log x
5	2	.6980	1.396
8	3	.9031	2.7093
10	4	1	4
12	1	1.0792	1.0792
	N = 10		9.1845

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

$$= \text{Antilog}\,[0.91845]$$

$$G.M = 8.287$$

G.M in Continous frequency distribution

In Continous frequency distribution mid value of the class is taken as x

Find G.M

Mark	0-10	10-20	20-30	30-40	40-50
no. of stud	5	7	15	25	8

class	freq	mid value x	$\log x$	$f \log x$
0-10	5	5	.6990	3.495
10-20	7	15	1.1761	8.2327
20-30	15	25	1.3980	20.97
30-40	25	35	1.5441	38.6025
40-50	8	45	1.6532	13.2256
	60			84.5258

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

$$= \text{Antilog}\left[\frac{84.5258}{60}\right]$$

$$= \text{Antilog}\,[1.4088]$$

$$= 25.63$$

Q calculate G.M for the following data

Data	100-104	105-109	110-114	115-119	120-124	125-129
no. of items	24	30	45	65	72	84

130-134	135-139
124	58

class	frequency	midvalue x	log x	f log x
100 – 104	24	102	2.0086	48.2064
105 – 109	30	107	2.0294	60.882
110 – 114	45	112	2.0492	92.214
115 – 119	65	117	2.0682	134.433
120 – 124	72	122	2.0863	150.2136
125 – 129	84	127	2.1038	176.7192
130 – 134	124	132	2.1206	262.9544
135 – 139	58	137	2.1367	123.9286
	N = 502			1049.55

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

$$= \text{Antilog}\left[\frac{1049.55}{502}\right]$$

$$= \text{Antilog}\left[2.091\right]$$

$$= 123.3$$

Harmonic Mean (H.M)

Harmonic mean of a set of n values is defined as the reciprocal of the mean of reciprocal of the values

H.M in Individual Series

Let $x_1, x_2 \ldots x_n$ are n values then

$$H.M = \frac{n}{1/x_1 + 1/x_2 \cdots 1/x_n}$$

i.e., $$\boxed{HM = \frac{n}{\Sigma 1/x}}$$

Q Find H.M 2, 3, 4, 5

$$1/x = 1/2 + 1/3 + 1/4 + 1/5$$

$$\Sigma 1/x = 77/60 = 1.283$$

$$H.M = \frac{4}{\Sigma 1/x} = \frac{4}{1.283} = 3.125$$

Q Find H.M 10, 20, 40, 25

x	$1/x$
10	0.1
20	0.05
40	0.025
25	0.04
$n=4$	0.215

$$H.M = \frac{n}{\Sigma 1/x} = \frac{4}{0.215} = 18.604$$

H.M in frequency distribution

In the case of Descrete and continous frequency distribution

$$H.M = \frac{N}{\Sigma f \cdot 1/x}$$

Q Find H.M

Size	6	10	14	18
freq	20	40	30	10

x	f	$1/x$	$f \times 1/x$
6	20	0.167	3.34
10	40	0.1	4
14	30	0.071	2.13
18	10	0.056	.56
	100		10.03

$$H.M = \frac{N}{\Sigma f 1/x} = \frac{100}{10.03} = 9.97$$

Q Find H.M from the following data

Class :	10-20	20-30	30-40	40-50	50-60
f :	4	6	10	7	3

class	f	midvalue x	$1/x$	$f \times 1/x$
10-20	4	15	.67	.268
20-30	6	25	.04	.24
30-40	10	35	.028	.28
40-50	7	45	.022	.154
50-60	3	55	.018	.054
	30			6.996

$$H.M = \frac{N}{\Sigma f 1/x} = \frac{30}{0.996} = 29.88$$

Q Find H.M of

x	2	4	8	10	12
f	1	3	5	4	2

x	f	1/x	f 1/x
2	1	.5	.5
4	3	.25	.75
8	5	.125	.625
10	4	.1	.4
12	2	.083	.166
	15		2.441

$$H.M = \frac{N}{\Sigma f 1/x} = \frac{15}{2.441} = 6.145$$

Q Find H.M

Size	0-10	10-20	20-30	30-40	40-50
freq	2	3	7	5	2

class	freq	midvalue x	1/x	f x 1/x
0-10	2	5	0.2	.4
10-20	3	15	.067	.201
20-30	7	25	.04	.28
30-40	5	35	.028	.14
40-50	2	45	.022	.044
	19			1.065

$$H.M = \frac{N}{\Sigma f \frac{1}{x}} = \frac{19}{1.065} = 17.84$$

Measure of Dispersion

Dispersion refers to the variability in the size of items

Measure of dispersion are classified into:

1) Absolute measures
2) Relative measures

Absolute Measures of dispersion are

1) Range
2) Quartile Deviation (QD)
3) Mean Deviation (MD)
4) Standard Deviation (SD)

The various relative measures of dispersion are

1) Coefficent of Range
2) Coefficent of quartile deviation
3) Coefficent of standard deviation
4) Coefficent of variation

Range

- It is the simplest possible measure of dispersi
- It is the differents between highest & lowest value in a series

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

$$\text{Coefficent of Range} = \frac{H-L}{H+L}$$

Q Find range and coefficent of Range for the followin values

25, 32, 85, 32, 42, 10, 20, 18, 28

- Here $H = 85 \quad L = 10$

$$\text{Range} = H - L = 75$$

$$\text{co of Range} = \frac{H-L}{H+L} = \frac{75}{95} = .789$$

Q Find Range

Wage	10	15	18	20	25
no. of employee	3	5	12	8	6

- $H = 25 \quad L = 10$

$$\text{Range} = H - L = 15$$

$$\text{C. of Range} = \frac{H-L}{H+L} = \frac{15}{35} = 0.428$$

Q Five students obtained the following marks in statistics 20, 35, 25, 30, 15. Find out Range and coefficent of Range of the marks

- Here $H = 35$ & $L = 15$

Range $= 35 - 15 = 20$

C. of Range $= \frac{H-L}{H+L} = \frac{20}{50} = 0.4$

Q Find out Range and coefficent of Range of the following series

Marks :	20-29	30-39	40-49	50-59	60-69
no. of stud :	8	12	20	7	3

[Hint : Read the problem as exclusive series]

- we have to make series exclusive

∴ It is inclusive

1st find interval b/n upper limit of a class and lower limit of next class and find its average and substract from every lower limit and add to every upper limit

Marks	19.5-29.5	29.5-39.5	39.5-49.5	49.5-59.5	59.5-69.5
No. of std	8	12	20	7	3

$H = 69.5 \quad L = 19.5$

$\text{Range} = H - L = 50$

$\text{C. of Range} = \frac{H-L}{H+L} = \frac{50}{89} = 0.562$

Quartile Deviation (semi interquartile Range)

It is defined as half the distance between 3rd and 1st quartiles

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

$$\text{Coefficent of } Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Q.D in Individual Series

$$Q_1 = \text{size of } \left(\frac{n+1}{4}\right)^{th} \text{ item}$$

$$Q_3 = \text{size of } \left(3 \times \frac{(n+1)}{4}\right)^{th} \text{ item}$$

Q Compute Q.D and coefficient of Q.D

23, 25, 8, 10, 9, 29, 45, 85, 10, 16, 24

Arrange values in Ascending Order

9, 8, 10, 10, 16, 23, 24, 25, 29, 45, 85

$Q_1 = \left(\frac{n+1}{4}\right)^{th}$ item $= \left(\frac{11+1}{4}\right) = 3^{rd}$ item $= 10$

$Q_3 = \frac{3\times(n+1)}{4}$ item $= 3\times 3 = 9^{th}$ item $= 9.5$

Coefficient of Q.D $= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{29-10}{29+10} = \frac{19}{39} = 0.48$

Q Find Q.D and coefficient Q.D

170, 82, 110, 100, 150, 120, 200, 116, 250

Ascending order

82, 100, 110, 116, 120, 150, 170, 200, 250

$Q_1 = \left(\frac{n+1}{4}\right)^{th}$ item $= \left(\frac{9+1}{4}\right)^{th} = \left(\frac{10}{4}\right)^{th} = (2.5)^{th}$ item

2^{th} item $+ 0.5\,(3^{rd}$ item $- 2^{th}$ item$) = 100 + .5(110-100) = 105$

$Q_3 = 3\times\frac{(n+1)^{th}}{4}$ item $= \left(\frac{3\times 10}{4}\right)^{th}$ item $= \left(\frac{30}{4}\right)^{th} = (7.5)^{th}$ item

7^{th} item $+ 0.5\times(8^{th}$ item $- 7^{th}$ item$) = 170 + .5(200-170)$

$Q_3 = 185$

$$Q.D = \frac{Q_3 - Q_1}{2} = \frac{185 - 105}{2} = \frac{80}{2} = 40$$

$$\text{Coefficen of } Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{185 - 105}{185 + 105} = \frac{80}{290} = 0.275$$

Q.D in Descrete Frequency distribution

In the case of Discrete frequency distribution

Q_1 = Size of $\left(\frac{N+1}{4}\right)^{th}$ item

Q_3 = Size of $\left(\frac{N+1}{4} \times 3\right)^{th}$ item, where N = cf

This formula give in which the quentiles lie

Q Find Q.D and coefficent of Q.D

Size : 5 8 10 12 19 20 32

Frequency : 3 10 15 20 8 7 6

Size	frequency	Cumilative frequency
5	3	3
8	10	13
10	15	28
12	20	48

19	8	56
20	7	63
32	6	69
	N = 69	

$$\frac{N+1}{4} = 17.5$$

$$Q_1 = \left(\frac{N+1}{4}\right)^{th} = \left(\frac{70}{4}\right)^{th} = 17.5 \Rightarrow 10$$

$$Q_3 = \left(\frac{N+1}{4} \times 3\right)^{th} = 3 \times 17.5 = 52.5 \Rightarrow 19$$

$$Q.D = \frac{Q_3 - Q_1}{2} = \frac{19-10}{2} = \frac{9}{2} = 4.5$$

$$\text{Coe of } Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{9}{29} = 0.3103$$

The specified value of the quartiles are obtained by the interpolation formula

$$Q_1 = l_1 + \frac{\left(\frac{N}{4} - cf\right)}{f} \times c, \quad Q_3 = l_1 + \frac{\left(\frac{3N}{4} - cf\right)}{f} \times c$$

where l_1 is the lower limit of the quartile class
f is the frequency of the class

cf is the camulative frequency of the preceding class

c is the class ~~method~~ interval of the quartile class

Q obtain the quartile measure measure of dispersion and its coefficent for the data given below

Age	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of person	15	30	53	75	100	110	115

70-80

125

class	frequency	cf	
0-10	15	15	
10-20	30	45	
20-30	53	98	
30-40	75	173	$(N/4)^{th}$ class
40-50	100	273	
50-60	110	383	
60-70	115	498	$(3N/4)^{th}$ class
70-80	125	623	
	623		

Q_1 = size of $(N/4)^{th}$ item $= \left(\frac{623}{4}\right)^{th}$ item $= 155.75$

Q_3 = size of $(3N/4)^{th}$ item $= \left(\frac{3 \times 623}{4}\right)^{th}$ item $= 467.25$

interpolation formula

$$Q_1 = l_1 + \frac{\left(\frac{N}{4} - cf\right)}{f} \times c$$

$l_1 = 30$

$\frac{N}{4} = 155.75$

$cf = 98$

$c = 10$

$f = 75$

$$= 30 + \frac{(155.75 - 98)}{75} \times 10$$

$$= 30 + \frac{577.5}{75}$$

$$= 37.7$$

$$Q_3 = l_1 + \frac{\left(\frac{3N}{4} - cf\right)}{f} \times c$$

$l_1 = 60$

$cf = 383$

$c = 10$

$\frac{3N}{4} = 467.25$

$f = 115$

$$= 60 + \frac{(467.25 - 383)}{115} \times 10$$

$$= 60 + \frac{842.5}{115}$$

$$= 67.33$$

$$Q.D = \frac{Q_3 - Q_1}{2} = \frac{67.33 - 37.7}{2} = 14.815$$

$$\text{Coef of } Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{67.33 - 37.7}{67.33 + 37.7} = 0.28$$

Mean Deviation (M.D)

It is defined as the arithmetic mean of deviation of all the values in a series from their average counting all such deviations as positive. The average may be mean, median or mode

$$\text{The coefficent of } M.D = \frac{M.D}{\text{Average from which } M.D \text{ is calculated}}$$

M.D in Individual Series

If $|d|$ represent deviation from an average then $M.D = \frac{\Sigma |d|}{n}$

Q Find M.D from mean and its coefficent from the following data

25, 63, 85, 75, 62, 70, 83, 28, 30, 12

X	$\|d\|$ $\|x - 53.3\|$
25	28.3
63	9.7
85	31.7
75	21.7

62	8.7
70	16.7
83	29.7
28	25.3
30	23.3
12	41.3
533	236.4

$$\text{Mean} = \frac{\Sigma x}{n} = \frac{533}{10} = 53.3$$

$$M.D = \frac{\Sigma |d|}{n} = \frac{236.4}{10} = 23.64$$

$$\text{Coefficent of } M.D = \frac{M.D}{\text{mean}} = \frac{23.64}{53.3} = 0.443$$

M.D is Discrete frequency distribution

In the case of discrete frequency distribution

$$M.D = \frac{\Sigma f|d|}{N}$$

Q Compute the mean deviation about mean and its coefficent for the following data

no. of children	: 0	1	2	3	4	5	6
no. of families	: 171	82	50	25	13	7	2

x	F	Fx	|d| (x-1.02)	f|d|
0	171	0	1.02	174.42
1	82	82	.02	1.64
2	50	100	.98	49
3	25	75	1.98	49.5
4	13	52	2.98	58.74
5	7	35	3.98	27.86
6	2	12	4.98	9.96
	350	356		351.12

$$\text{mean} = \frac{\Sigma Fx}{N} = \frac{356}{350} = 1.02$$

$$M.D = \frac{\Sigma f|d|}{N} = \frac{351.12}{350} = 1.0032$$

Q calculate mean deviation from median and its coefficent for the following values.

Values: 5, 28, 33, 44, 83, 87, 96, 99, 25, 35, 82

Asc order 5, 25, 28, 33, 35, 44, 82, 83, 87, 96, 99

Asc order

x	$\lvert d \rvert [x - 441]$
5	39
25	19
28	16
33	11
35	9
44	0
82	38
83	39
87	43
96	52
99	55
	321

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{item} = \left(\frac{11+1}{2}\right)^{th} \text{item} = 6^{th} \text{item} = 44$$

$$M.D = \frac{\Sigma |d|}{n}$$

$$= \frac{321}{11} = 29.18$$

$$\text{Coefficent of } M.D = \frac{M.D}{\text{median}} = \frac{29.18}{44} = 0.66$$

M.D is continuous frequency distribution

In continous frequency distribution =

$$M.D = \frac{\Sigma f|d|}{N}$$

here midvalue of the class is choosen as x

Q calculate mean deviation about medium for the following data and coefficent of mean deviation

Class	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
freq	18	16	15	12	10	5	2	2

class	f	cf	midvalue	|d|	f|d|
0-10	18	18	5	19	342
10-20	16	34	15	9	144
20-30	15	49	25	1	15
30-40	12	61	35	11	132
40-50	10	71	45	21	210
50-60	5	76	55	31	155
60-70	2	78	65	41	82
70-80	2	80	75	51	102
	80				1182

$$\text{Median} = \left(l_1 + \frac{\left(\frac{N}{2} - cf\right)}{f} \times c \right)$$

$l_1 = 20$

$\frac{N}{2} = 40$

$cf = 34$

$f = 15$

$c = 10$

$$= 20 + \frac{40 - 34}{15} \times 10$$

$$= 20 + \frac{6}{15} \times 10$$

$$= 20 + \frac{60}{15}$$

$$= \underline{24}$$

$$M.D = \frac{\Sigma f|d|}{N}$$

$$= \frac{1182}{80}$$

$$= 14.775$$

$$\text{Coefficent of } M.D = \frac{M.D}{\text{median}} = \frac{14.775}{24} = \underline{0.615}$$

Q calculate M.D about mean of the following data

21, 29, 35, 10, 42, 75, 50, 30, 18, 80

X	|d|
21	18
29	10
35	4
10	29
42	3
75	36
50	11
30	9
18	21
80	41
390	182

$$\text{Mean} = \frac{\Sigma x}{n}$$

$$= \frac{390}{10} = 39$$

$$M.D = \frac{\Sigma |d|}{n} = \frac{182}{10} = 18.2$$

Q Find mean deviation about the median

Sl. No	1	2	3	4	5	6	7	8	9	10
Marks	50	27	38	40	90	42	70	60	15	40

[Hint This is an individual series]

→ 15, 27, 38, 40, 40, 42, 50, 60, 70, 90

X	|d| (x − 41)
15	26
27	14
38	3
40	1
40	1
42	1
50	9
60	19
70	29
90	49
	152

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

$(11/2)^{th}$ item $= (5.5)^{th}$ item

$= \frac{40+42}{2} = \frac{82}{2}$

median = 41

$M.D = \frac{\Sigma|d|}{n}$

$= \frac{152}{10}$

$= 15.2$

Q calculate M.D about mean from the following data

X	:	10	11	12	13	14
F	:	3	12	18	12	3

X	F	FX	|d|	F|d|
10	3	30	2	6
11	12	132	1	12
12	18	216	0	0
13	12	156	1	12
14	3	42	2	6
	48	576		36

$$\text{Mean} = \frac{\Sigma Fx}{N}$$

$$= \frac{576}{48} = 12$$

$$M.D = \frac{\Sigma F|d|}{N} = \frac{36}{48} = .75$$

Q Compute mean deviation about mean from the following frequency

class : 0-10 10-20 20-30 30-40 40-50

freq : 3 8 12 9 8

class	freq	midvalue x	Fx	|d| x−27.75	F|d|
0-10	3	5	15	22.75	68.25
10-20	8	15	120	12.75	102
20-30	12	25	300	2.75	33
30-40	9	35	315	7.25	65.25
40-50	8	45	360	17.25	138
	40		1110		406.5

$$\text{Mean} = \frac{\Sigma fx}{N}$$

$$= \frac{1110}{40}$$

$$= 27.75$$

$$M.D = \frac{\Sigma f|d|}{N}$$

$$= \frac{406.5}{40} = 10.16$$

Q calculate M.D from mean and from median for the following distribution of the scores of 50 college student is

Scores : 140 150 160 170 180 190

Freq : 4 6 10 18 9 3

[Hint : Read close as 140, 150, 160, 160 ...]

M.D about Mean

Class	f	midvalue x	Fx	|d| x-171.2	F|d|
140-150	4	145	580	26.2	104.8
150-160	6	155	930	16.2	97.2
160-170	10	165	1650	6.2	62
170-180	18	175	3150	3.8	68.4
180-190	9	185	1665	13.8	124.2
190-200	3	195	585	23.8	71.4
	50		8560		528

$$\text{Mean} = \frac{\Sigma Fx}{N} = \frac{8560}{50} = 171.2$$

$$M.D = \frac{\Sigma f|d|}{N}$$

$$= \frac{528}{50} = 10.56$$

M.D about Mean

class	f	cf	midvalue x	$\|d\|$ $x-171.78$	$f\|d\|$
140-150	4	4	145	27.78	111.12
150-160	6	10	155	17.78	106.68
160-170	10	20	165	7.78	77.8
$N/3$ 170-180	18	38	175	2.22	39.96
180-190	9	47	185	12.22	109.98
190-200	3	50	195	22.22	66.66
	50				512.2

$$\text{Median} = l_1 + \frac{N/2 - cf}{f} \times c$$

$l_1 = 170$

$N/2 = 25$

$cf = 20$

$f = 18$

$c = 10$

$$\text{Median} = 170 + \frac{25-20}{18} \times 10$$

$$= 170 + \frac{50}{18}$$

$$= 172.78$$

$$M.D = \frac{\Sigma F|d|}{N}$$

$$= \frac{512.2}{50}$$

$$= 10.24$$

Standard Deviation

It is the square root of mean of squares of the deviation of all values of a series from their arithmetic mean

S.D in individual series

$$\text{S.D } (\sigma) = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$$

$$\text{Variance } \sigma^2 = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2$$

$$\text{Coefficent of variation } CV = \frac{SD}{\text{mean}} \times 100$$

Q for the following values find S.D, variance coefficent of variance

5, 8, 7, 11, 9, 10, 8, 2, 4, 6

X	X^2
5	25
8	64
7	49
11	121
9	81
10	100
8	64
2	4
4	16
6	36
70	560

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

$$= \sqrt{\frac{560}{10} - \left(\frac{70}{10}\right)^2}$$

$$= \sqrt{56 - 49}$$

$$= \sqrt{7}$$

$$= 2.64$$

$$V \rightarrow \sigma^2 = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2$$

$$= 7$$

$$CV = \frac{S.D}{mean} \times 100$$

$$= \frac{2.64}{7} \times 100$$

$$= 37.71$$

S.D in continous and Discrete frequency distribution

$$S.D = \sqrt{\frac{\Sigma F x^2}{N} - \left(\frac{\Sigma F x}{n}\right)^2}$$

Q for the following data calculate S.D variance and C.V

Mark	2	4	6	8	10
no. of std	8	10	6	9	7

X	F	FX	X^2	FX^2
2	8	16	4	32
4	10	40	16	160
6	16	96	36	576
8	9	72	64	576
10	7	70	100	700
	50	294		2044

$$S.D = \sqrt{\frac{\Sigma FX^2}{N} - \left(\frac{\Sigma FX}{N}\right)^2}$$

$$= \sqrt{\frac{2044}{50} - \left(\frac{294}{50}\right)^2}$$

$$= \sqrt{40.88 - 35.28}$$

$$= \sqrt{5.6}$$

$$= 2.36$$

$$V = \frac{\Sigma Fx^2}{N} - \left(\frac{\Sigma Fx}{N}\right)^2$$

$$= 5.6$$

$$CV = \frac{S.D}{mean} \times 100$$

$$= \frac{2.36}{35.88} \times 100$$

$$= \frac{236}{5.88}$$

$$= 40.14$$

Q Calculate SD, Variance & CV for the following data

Class	0-2	2-4	4-6	6-8	8-10	10-12
Freq	2	4	6	4	2	6

Class	Freq	midvalue x	Fx	x^2	Fx^2
0-2	2	1	2	1	2
2-4	4	3	12	9	36
4-6	6	5	30	25	150
6-8	4	7	28	49	196
8-10	2	9	18	81	162
10-12	6	11	66	121	726
	24		156		1272

$$S.D = \sqrt{\frac{\Sigma FX^2}{N} - \left(\frac{\Sigma FX}{N}\right)^2}$$

$$S.D = \sqrt{\frac{1272}{24} - \left(\frac{156}{24}\right)^2}$$

$$= \sqrt{53 - 42.25}$$

$$= \sqrt{10.75}$$

$$= 3.28$$

$$V = \frac{\Sigma FX^2}{N} - \left(\frac{\Sigma FX}{N}\right)^2 = 10.75$$

$$CV = \frac{S.D}{mean} \times 100$$

$$= \frac{3.28}{6.5} \times 100 = 50.46$$

Merits and Demerits of Measure of Dispersion

Merits of Range

1) Range is the simplest measure of dispersion
2) It can be easily calculated
3) It can be understand even by a layman.

Demerits of Range

1) Range is not based on all items of the series
2) It is highly affected by sampling fluctuations
3) It cannot be computed on the case of open end distribution

Uses of Range

Range is used in certain fields to measure variability particularly those data where variation is not much. For example, doctors are generally interested in the range of the fluctuating temperatures of patients. It is also used in Quality control. Range is used to study variation in prices of shares and interest rates. In weather forecats. the minimum and the maximum temperature of every day are studied from this, they can forecast the range within which the temperature may likely vary.

Merits of Quartile Deviation

1) It is easy to understand and to calculate
2) It is not affected by extreme values.

Demerits of Quartile Deviation

1) It ignores extreme items
2) It is not capable of more algebraic treatments.

Merits of Mean Deviation

1) Mean Deviation is a very simple and on easy measure of dispersion it is easily understood.
2) It is based on all the items of the series. so it is more representative.
3) Mean deviation is less affected by extreme values

Demerits of Mean Deviation

1) Mean Deviation suffers from inaccuracy because '+' or '−' signs are ignored
2) mean deviation is not capable of any further algebraic treatment.

3) Mean deviation is not reliable measure when calculated from mode as the mode is uncertain in some cases.

Uses of Mean deviation

Mean Deviation is significantly used for measuring variability of the series. relating to Economic and social phenomena. variability in the distribution of wealth and income is generally measured in terms of mean deviation.

Merits of Standard Deviation

1) standard deviation is based on all the values of a series it does not ignore any value
2) standard deviation is a clear and definite measure of dispersion so that it can be measured from all series.
3) It is not very much affected by sampling fluctuators.

4) It is capable of further algebraic treatment.

Demerits of Standard Deviation

1) standard is difficult to calculate
2) It gives more weight to extreme values

Features of Standard Deviation

1) Deviations are measured from Arithmetic mean
2) Signs of the deviations are not ignored.

Q Why is Standard Deviation considered to be the best measure of dispersion.

- Standard Deviations possess most of the important characteristics which ~~can~~ an ideal measure of dispersion should have

I Standard Deviation is rigidly defined
II It is based on all the observations of the data
III It is amentable to more algebraic treatment
IV It possesses many mathematical properties
V It is not much affected by sampling fluctuations
VI It does not ignore the signs of the deviations
VII It is possible to find out S.D of two or more groups

VIII Coefficient of variation is based on standard Deviation and it can be used to compare variability of 2 series.

IX Standard Deviation is used to find out statistical measures like coefficent of skewness, coefficent of correlation Regression equations etc...

So standard Deviation is best measure of dispersion

Difference between Standard Deviation and mean Deviation

1) In standard Deviation, deviations are taken only from the mean of the values of the series. But in mean Deviation, deviations may be taken from mean or median or mode

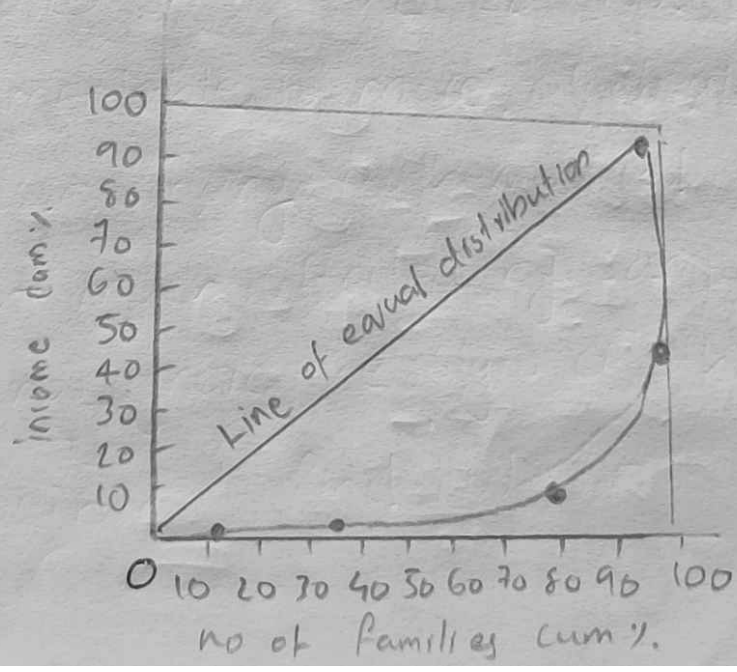
2) In standard deviation, signs of deviations are not ignored But in mean deviation, signs of deviations (+ or −) are ignored.

Lorenz Curve

Lorenz curve is a graphic method of studying dispersion in a series It is used in business to study the disparities of the distribution of wages, turnover, production, population etc.... In Economics it is useful to measure inequalities in the distribution of income between different countries or between different periods of time

Lorenz curve is a graph drawn to frequency distribution, taking the cumulated percentage values of frequencies along the x-axis and cumulated percentage values of the variable along the Y-axis

Given below is an example of Lorenz curve



Q How to measure variability (or inequality) from a Lorenz presentation

- If there is no inequality in the distribution, the Lorenz curve will coincide with the line of equal distribution. The more the Lorenz curve is away from the line of equal distribution the greater is the inequality (or variability)

Lorenz curve is useful to

1) study the variability in a distribution.
2) compare the variability relating to a phenomenon for two regions
3) study the changes in variability over a period.

unit IV

Curve Fitting

Fitting of a straight line of the form $y = a+bx$

Suppose we want to find a straight line as a best approximation to the actual curve $y=f(x)$ passing through n given points $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$ this line will be reffered as the line of best fit and we take its equation as $y = a+bx$ — ① where a & b are the parameters to be determined. values of a and b are obtained by solving two equations known as Normal equations

$$\Sigma y = na + b\Sigma x$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

After determine the values of a & b put in these values in ① we get the equation of the line of best fit for the given data.

Q Fit a straight line to the following data

x	1	2	3	4	5
y	14	13	4	5	2

Estimate the value of y at $x = 3.8$

x	y	xy	x^2
1	14	14	1
2	13	26	4
3	4	12	9
4	5	20	16
5	2	10	25

$\Sigma x = 15 \quad \Sigma y = 38 \quad \Sigma xy = 82 \quad \Sigma x^2 = 55$

The normal equation are

$$\Sigma y = na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

ie, $\Sigma y = na + b\Sigma x$

$$38 = 5a + b15$$

$$82 = a15 + b55$$

$$5a + 15b = 38 \quad -(1)$$

$$15a + 55b = 82 \quad -(2)$$

$3 \times (1) \Rightarrow 15a + 45b = 114 \quad -(3)$

$(2) - (3) \Rightarrow \quad 10b = -32$

$$b = \frac{-32}{10} = -3.2$$

put $b = -3.2$ in eqn (1)

$$5a + 15x - 3.2 = 38$$

$$5a - 48 = 38$$

$$5a = 38 + 48$$

$$5a = 86$$

$$a = 86/5$$

$$= 17.5$$

$$y = a + bx$$

$$y = 17.2 - 3.2x$$

at $x = 3.5$

$$y = 17.2 - 3.2 \times 3.5$$

$$y = 17.2 - 11.2$$

$$y = 6$$

Q Fit a straight line to the following data

x	1	2	3	4	6	8
y	2.4	3	3.6	4	5	6

x	y	xy	x^2
1	2.4	2.4	1
2	3	6	4
3	3.6	10.8	9
4	4	16	16
6	5	30	36
8	6	48	64

$\Sigma x = 24$ $\Sigma y = 24$ $\Sigma xy = 113.2$ $\Sigma x^2 = 130$

The normal equation are

$$\Sigma y = na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

i.e,

$$24 = 6a + b\,24$$

$$113.2 = a\,24 + b\,130$$

$$6a + 24b = 24 \quad -①$$

$$24a + 130b = 113.2 \quad -②$$

$$4 \times ① \Rightarrow 24a + 98b = 96 \quad -③$$

$$② - ③ \Rightarrow 34b = 17.2$$

$$b = \frac{17.2}{34}$$

$$= 0.50$$

put $b = 0.50$ in eqn ①

$$6a + 24 \times 0.50 = 24$$

$$6a + 12 = 24$$

$$6a = 24 - 12$$

$$6a = 12$$

$$a = 12/6$$

$$a = 2$$

$$y = a + bx$$

$$y = 2 + 0.50x$$

Correlation And Regression

Correlation

It is a statistical measure for finding out degree of association between two or more variables.

-> Suppose x & y are two variables if the movement of x & y be in the same direction ie, either both x & y increase or both decreases

Then we say that x & y are in positive correlation

-> If the movement are in the opposite direction ie, if x increase then y decrease or viseversa then we say that there is a negative correlation between x & y.

-> If Y is unaffected by any change in X then X & Y are said to be uncorrelated or zero Correlation

Determination of correlation

Correlation between two variables can be determined by the following methods

1 -> Scatter diagram

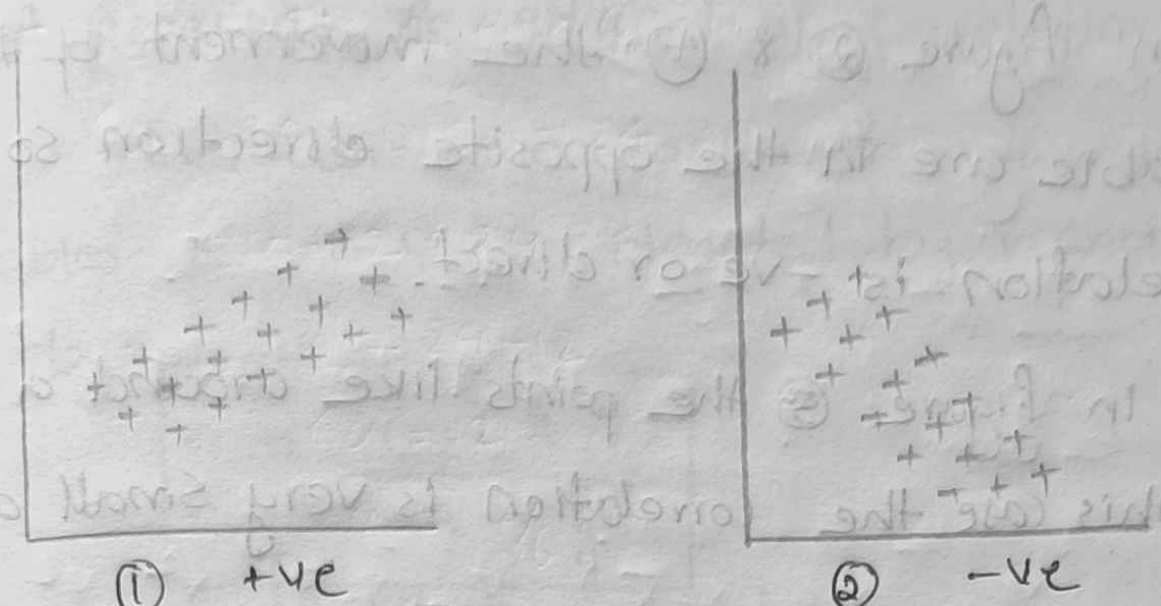
2 -> Karl pearson's method

3 -> Rank method

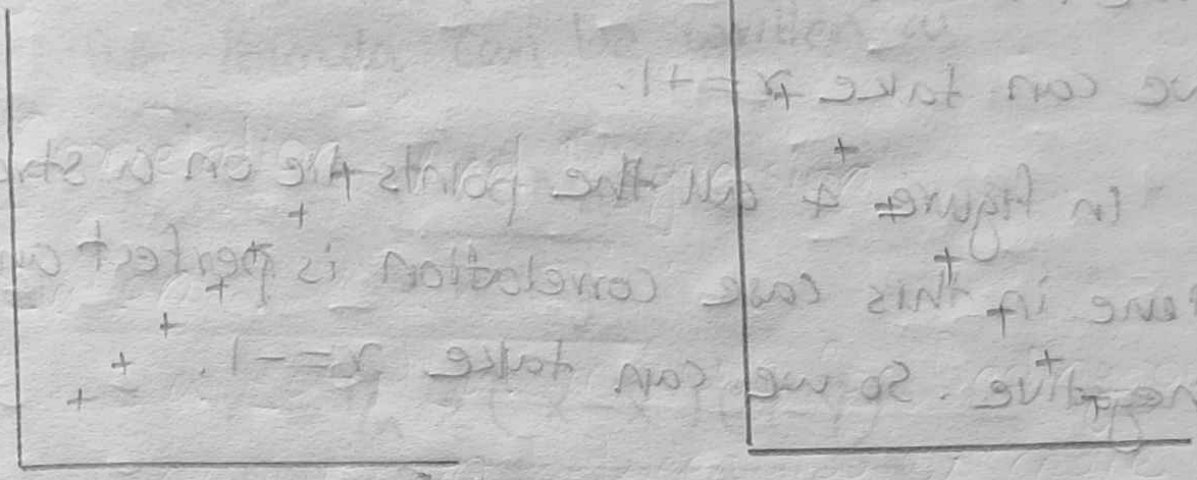
1) Scatter Diagram

The existance of correlation can be shown graphically by means of a scatter diagram

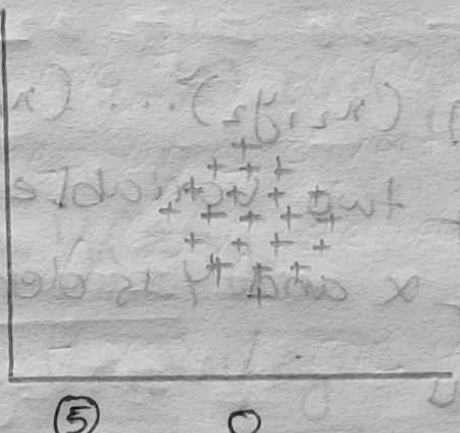
The data relating to two variables can be graphically represented by points taking one of the variables x along horizontal axis and second variable Y along vertical axis. All the pair of value x and y are shown by points on the graph paper. This diagramatic representation of bivariate the data is known as scatter diagram.



① +ve ② -ve



③ $r=1$ ④ $r=-1$



⑤ 0

In Figure ① & ③ the movement of the 2 variable are in the same direction in thes the correlation +ve or direct.

In figure ② & ④ the movement of the 2 variable are in the opposite direction so. the correlation is -ve or direct.

In figure ⑤ the points like around a curve in this case the correlation is very small and we take $r=0$

In figure 3 all points are lie in a straight line. In this case the correlation is perfect. so we can take $r=+1$.

In figure 4 all the points lie on a straight line in this case correlation is perfect and also negative. So we can take $r=-1$.

2. Karl pearson's method [co-variance method]

Co-variance

If (x_1,y_1) $(x_2,y_2)\dots(x_n,y_n)$ be n pair of observation of two variable x & y. then the co-variance of x and y is denoted as $Cov(x,y)$ and defined as

$$Cov(x,y) = \frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y})$$

Karl pearson's co-efficent of correlation

The correlation coefficent between two variables x and y are denoted by r and it is give by

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

where σ_x and σ_y are standard deviation of x & y respectively.

The formula can be written as

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

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

take $x = (x-\bar{x})$ and $y = (y-\bar{y})$

then $$r = \frac{1/n \sum xy}{\frac{\sqrt{\sum x^2}}{\sqrt{n}} \frac{\sqrt{\sum y^2}}{\sqrt{n}}}$$

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

Q Find the correlation co-efficent from the following data

x =	1	2	3	4	5	6	7
y	6	8	11	9	12	10	14

x	y	$x = (x-\bar{x})$	$y = (y-\bar{y})$	x^2	y^2	xy
1	6	-3	-4	9	16	12
2	8	-2	-2	4	4	4
3	11	-1	1	1	1	-1
4	9	0	-1	0	1	0
5	12	1	2	1	4	2
6	10	2	0	4	0	0
7	14	3	4	9	16	12

$\Sigma x^2 = 28 \quad \Sigma y^2 = 42 \quad \Sigma xy = 29$

$\bar{x} = \frac{\Sigma x}{n} = 4$

$\bar{y} = \frac{\Sigma y}{n} = 10$

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

$$= \frac{29}{\sqrt{28}\cdot\sqrt{42}}$$

$$= 0.8456$$

Q calculate the karl pearson's correlation between advertisement cost & sales as per the data given below.

cost	39	65	62	90	82	75	25	98	36	78
sales	47	53	58	86	62	68	60	91	51	84

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

$\bar{x} = \frac{\Sigma x}{n}$ $\bar{y} = \frac{\Sigma y}{n}$ $\Sigma x^2 = 5398$ $\Sigma y^2 = 2224$ $\Sigma xy = 2704$

$= 65$ $= 66$

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

$$= \frac{2704}{\sqrt{5398}\sqrt{2224}} = 0.7804$$

Rank Correlation Coefficient

Simple correlation coefficent is based on the magnitude of the variables. But in some situation it is not possible to find magnitude of all the variables.

For example we cannot measure beauty or intelligents qualitative But in this case it is possible to rank the intellegents in some order. Rank correlation is based on the order of the rank, spherman formula for rank coefficent correlation

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

where d -> the difference between the rank.

Q Calculate rank correlation coefficent

mark in math	78	36	98	25	75	82	90	62	65	69
marks in stati	84	51	91	60	68	62	86	58	53	47

x	y	R(x)	R(y)	d	d^2
78	84	4	3	1	1
36	51	9	9	0	0
98	91	1	1	0	0
25	60	10	6	4	16
75	68	5	4	1	1
82	62	3	5	-2	4
90	86	2	2	0	0
62	58	8	7	1	1
65	53	7	8	-1	1
69	47	6	10	-4	16

$\Sigma d^2 = 40$

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

$$= 1 - \frac{6 \times 40}{10(100-1)}$$

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

$$R = 0.7575$$

Q Find the rank co-efficent correlation

X	28	45	40	35	33	41	32	36	64
Y	23	34	33	36	36	26	28	31	37

X	Y	R(x)	R(Y)	d	d^2
28	23	9	9	0	0
45	34	1	3	-2	4
40	33	3	4	-1	1
35	36	5	2	3	9
33	30	7	6	1	1
41	26	2	8	-6	36
32	28	8	7	1	1
36	31	4	5	-1	1
34	37	6	7	5	25

$$\Sigma d^2 = 78$$

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

$$= \frac{1 - 6 \times 78}{9(81-1)}$$

$$= 1 - \frac{468}{720}$$

$$= 0.35$$

Spearman's formula for Repeated Ranks

If in a series 2 or more individual or items having the same score then we find the average of the rank of these individuals and allowed this rank to each of them. In such a case Spearman's modified formula is

$$R = 1 - \frac{6\left(\Sigma d^2 + \Sigma (t^3 - t)\right)}{n(n^2-1)}$$

where t is the no. of individuals involved in a tie in the 1st or 2nd series.

Q Find the rank correlation coefficent of the following data

Series A	115	109	112	87	98	120	98	100	98	118
Series B	75	73	85	70	76	82	65	73	68	80

X	Y	R(X)	R(Y)	d	d^2
115	75	3	5	-2	4
109	73	5	6.5	-1.5	2.25
112	85	4	1	3	9
87	70	10	8	2	4
98	76	8	4	4	16
120	82	1	2	-1	1
98	65	8	10	-2	4
100	73	6	6.5	-0.5	0.25
98	68	8	9	-1	1
118	80	2	3	-1	1

$\Sigma d^2 = 42.5$

$$R = 1 - \frac{6\left(\Sigma d^2 + \frac{\Sigma(t^3 - t)}{12}\right)}{n(n^2-1)}$$

$$= 1 - \frac{6\left(42.5 + \frac{30}{12}\right)}{990}$$

$$= 1 - \frac{6(42.5 + 2.5)}{990}$$

$$= 1 - \frac{270}{990}$$

$= 1 - 0.272$

$= \underline{\underline{0.728}}$

Q Find the co-efficent of correlation of the marks of some students in subjects

Paper 1 80 45 55 56 58 60 65 68 70 75 85

paper 2 82 56 50 48 60 62 64 65 70 64 90

X	Y	R(x)	R(Y)	d	d^2
80	82	2	2	0	0
45	56	11	9	2	4
55	50	10	10	0	0
56	48	9	11	-2	4
58	60	8	8	0	0
60	62	7	7	0	0
65	64	6	5.5	0.5	0.25
68	65	5	4	1	1
70	70	4	3	1	1
75	64	3	5.5	-2.5	6.25
85	90	1	1	0	0

$\Sigma d^2 = 16.5$

$$R = 1 - \frac{6\left(\Sigma d^2 + \frac{\Sigma(t^3 - t)}{12}\right)}{n(n^2-1)}$$

$$= 1 - \frac{6\left(16.5 + \frac{6}{12}\right)}{11 \times 10^2}$$

$$= 1 - \frac{6\,(16.5 + 0.5)}{1320}$$

$$= 1 - \frac{6 \times 17}{1320}$$

$$= 0.9227$$

Regression Analysis

Suppose we are given n pair of values (x_1, y_1) $(x_2, y_2) \ldots (x_n, y_n)$ of two variables x and y if we fit a straight line to this data by taking x as independent variable and y as dependent variable. Then the straight line obtained is called regression line of y on x. Its slope is called the regression co-efficent of y on x similarly if we fit a straight line to the data by taking y as independant variable and x as dependent variable Then line obtained is called regression line of x on y The reciprocal of its slope is called regresston co-efficent of x on y.

Equation for regression line

Regression equation of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

where b_{yx} is the regression co-efficent y on x

Regression equation of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

where b_{xy} is the regression co-efficent x on y.

Here $b_{yx} = \frac{cov(x,y)}{\sigma^2 x}$

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

$$\boxed{b_{yx} = \frac{\Sigma xy}{\Sigma x^2}}$$

$$b_{xy} = \frac{cov(x,y)}{\sigma^2 y}$$

$$\boxed{b_{xy} = \frac{\Sigma xy}{\Sigma y^2}}$$

Q You are given the data relating to purchase and sale obtain the a regression equation by the method of least squares and estimates the value of the sales when the purchase = 100

Purchase	62	72	98	76	81	56	76	92	88	49
Sales	112	124	131	117	132	96	120	136	97	85

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	y^2	xy
62	112	-13	-3	169	9	39
72	124	-3	9	9	81	-27
98	131	23	16	529	256	368
76	117	1	2	1	4	2
81	132	6	17	36	289	102
56	96	-19	-19	361	361	361
76	120	1	5	1	25	5
92	136	17	21	289	441	357
88	97	13	-18	169	324	-234
49	85	-26	-30	676	900	780

$\bar{x} = 75 \quad \bar{y} = 115 \quad \Sigma xy = 1753$

$\Sigma y^2 = 2690 \quad \Sigma x^2 = 2240$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{1753}{2240} = 0.78$$

$$b_{xy} = \frac{\Sigma xy}{\Sigma xy^2} = \frac{1753}{2690} = 0.651$$

Regression equation of y on x,

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 115 = 0.78(x - 75)$$

$$y - 115 = 0.78x - 58.5$$

$$y = 0.78x - 58.5 + 115$$

$$y = 0.78x + 56.5$$

Regression equation of x on y

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 75 = 0.65(y - 115$$

$$x - 75 = 0.65y - 74.75$$

$$x = 0.65y - 74.75 + 75$$

$$x = 0.65y + 0.25$$

when purchase $x = 100$, then sales

$$y = 0.78 \times 100 + 56.5$$

$$y = 134.5$$

Q For the given data obtain the two regression lines

x	8	6	4	7	5
y	9	8	5	6	2

	x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	X^2	Y^2	XY
	8						