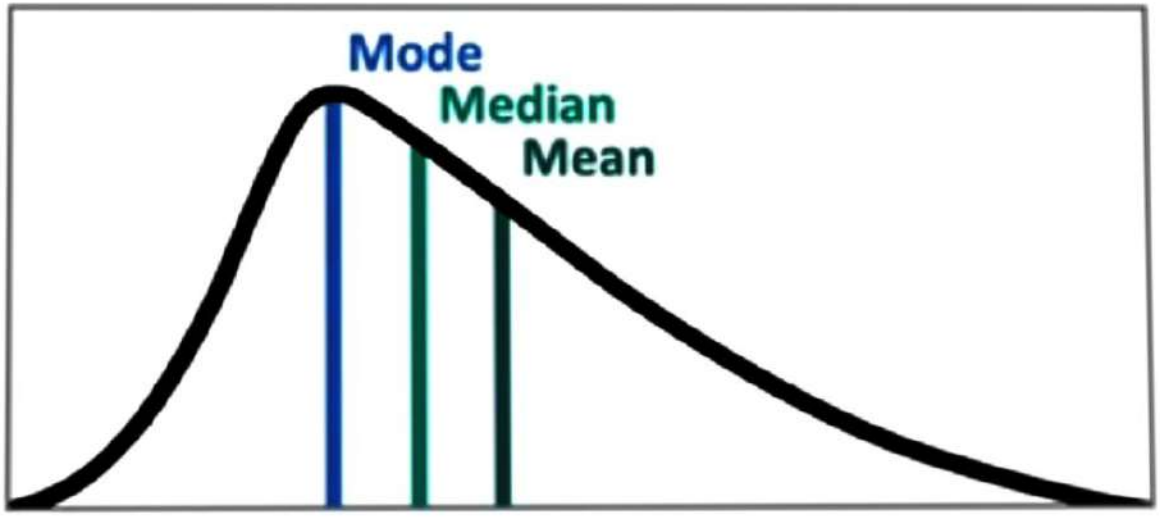
Measures of central tendency

- Measures of central tendency describe the most representative value of the data in a single number.
- The **mean, median** and **mode** are the most commonly used measures for describing the mid-point or most characteristic value.
- How are these measures calculated?



Steps of a data journey

Supported by a foundation of stewardship, metadata, standards and quality

The mean

The mean represents the average of all the values for one variable in a dataset.

Calculating the mean

Dataset = 3,4,8,5,7,3

1. Sum all the values → 3 + 4 + 8 + 5 + 7 + 3 = 30

2. Divide the sum by the number of values → 30 / 6 = 5

Calculating the mean

Dataset = 3,4,8,5,7,3

1. Sum all the values → 3 + 4 + 8 + 5 + 7 + 3 = 30

2. Divide the sum by the number of values → 30 / 6 = 5

3. Result is the mean → Mean = 5

Calculating the mean

- Be careful when working with data that have outliers.
- Extreme values are values that are extremely high or low compared with the rest of the values.
- Extreme values pull the mean up or down.

$3 + 4 + 8 + 5 + 7 + 33 = 60$

$60 / 6 = 10$

Mean

atistics 101: Exploring measures of central tendency

The median

Represents the middle value when all values are arranged in increasing order.

Calculating the median: For an odd number of values

The median splits the data values down the middle, so half are below it and half are above it.

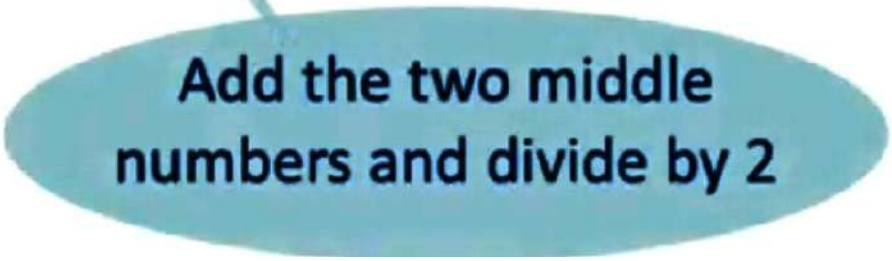
Example of dataset = **5, 6, 7, 8, 8, 9, 9, 9, 12, 15, 21, 28, 33**

1. Sort the values into increasing order
2. Count the values and find the "middle" value that separates the lower half from the higher half
3. The middle value is the median. In this example the median is **9**

Calculating the median: For an even number of values

Sample dataset = **5, 6, 7, 8, 8, 9, 9, 9, 12, 15, 21, 28, 33, 35**

1. Sort values into increasing order
2. Count the values (14) and locate the two middle numbers where there is an equal number of values on either side (9,9)
3. Add the two middle numbers (9+9)
4. Divide by 2 (median is **9**)



Extreme values and the median

Consider the following datasets:

Dataset A: 5, 6, 6, 7, 8, 9, **9**, 9, 12, 15, 21, 28, 33

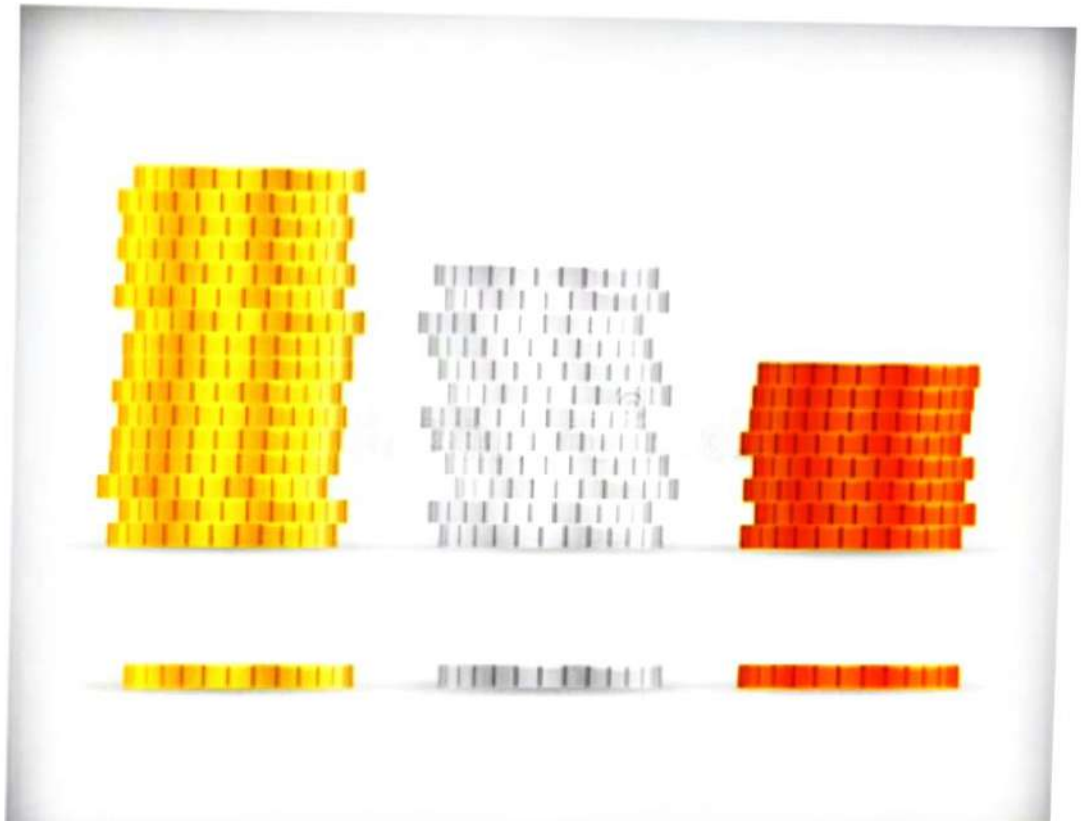
Versus

Dataset B: 5, 6, 6, 7, 8, 9, **9**, 9, 12, 15, 21, 28, 333

- The median is **9** in both cases, despite the inclusion of a much higher value in Dataset B.
- The median is not as influenced by extreme values as the mean is

The mode

- ✓ The mode is the most common value or range
- ✓ A descriptive measure based solely on counts
- ✓ There may be more than one mode

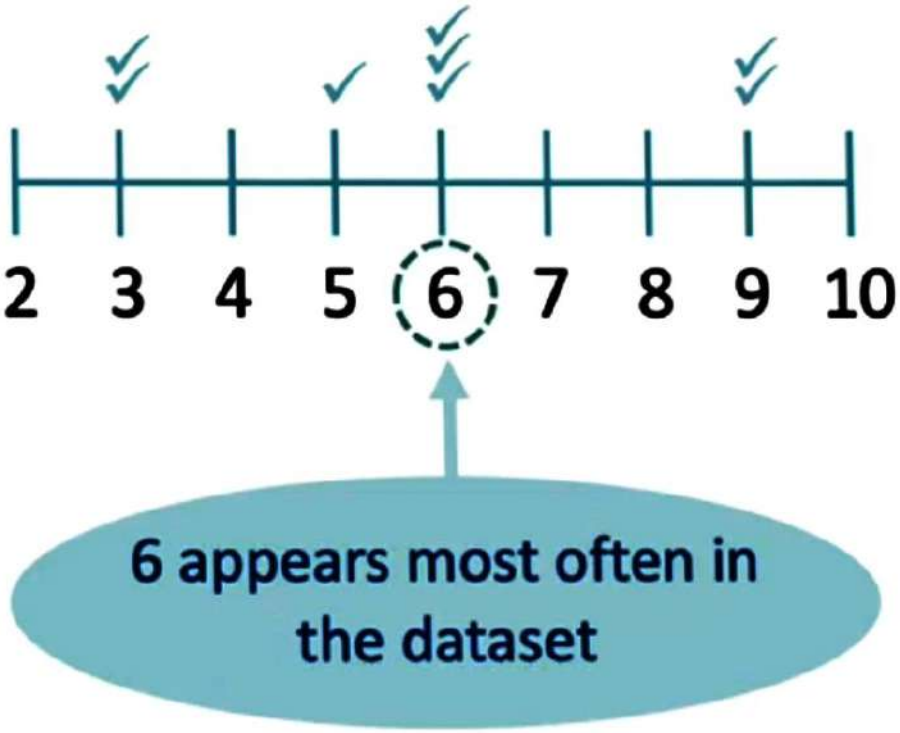


Calculating the mode

Example dataset = **6, 3, 9, 6, 6, 5, 9, 3**

1. Count the occurrence of each value
2. Value appearing most often is mode

- If all values have the same number of occurrences, there is no mode
- If the highest number of occurrences is found for more than one value, there is more than one mode



Question

Look at the following numbers

1, 1, 1, 1, 1, 4, 5

1. What is the mean?
2. What is the median?
3. What is the mode?

Answer

1, 1, 1, 1, 1, 4, 5

1. The mean is 14 divided by 7 (14/7), which equals **2**
2. The median is **1.**
3. The mode is **1.**

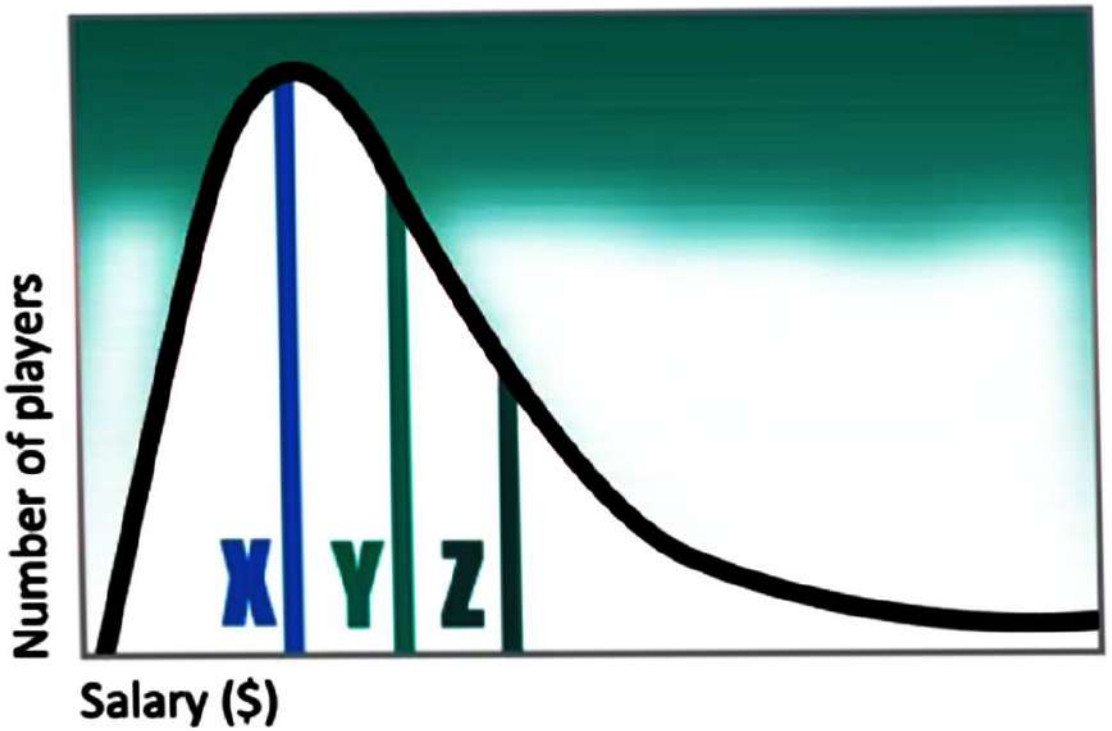
Top tips: mean, median and mode

- ✓ If the data are not numerical, the only measure of central tendency that can be used is the mode.
- ✓ If there are extremely low or extremely high values, the median usually provides a better measure of the central tendency.
- ✓ If the data have more than one mode, the mode may not be the best measure of central tendency.

It can be useful to look at more than one measure of central tendency.

Answer

- X is the mode.
- Z is the mean.
- Y is the median.



Statistics

Statistics is the study of how to collect, organize, analyze and interpret the information about one or more population(s) under investigation.

Frequency: Number of observations in a particular class interval is called frequency of that class.

Note: 1. After arranging the data in ascending or descending order, it is presented in tabular form consisting of columns headed by symbols x and f.

$x \equiv$ various observations recorded.

$f \equiv$ frequency.

Rules of tabulation

The class interval must be well defined.

e.g if class intervals are 0-10, 10-20, 20-30, 30-40, 40-50;

Note: ① value 20 should be put in 20-30 (not 10-20)

② $x = 50$ must be in 40-50

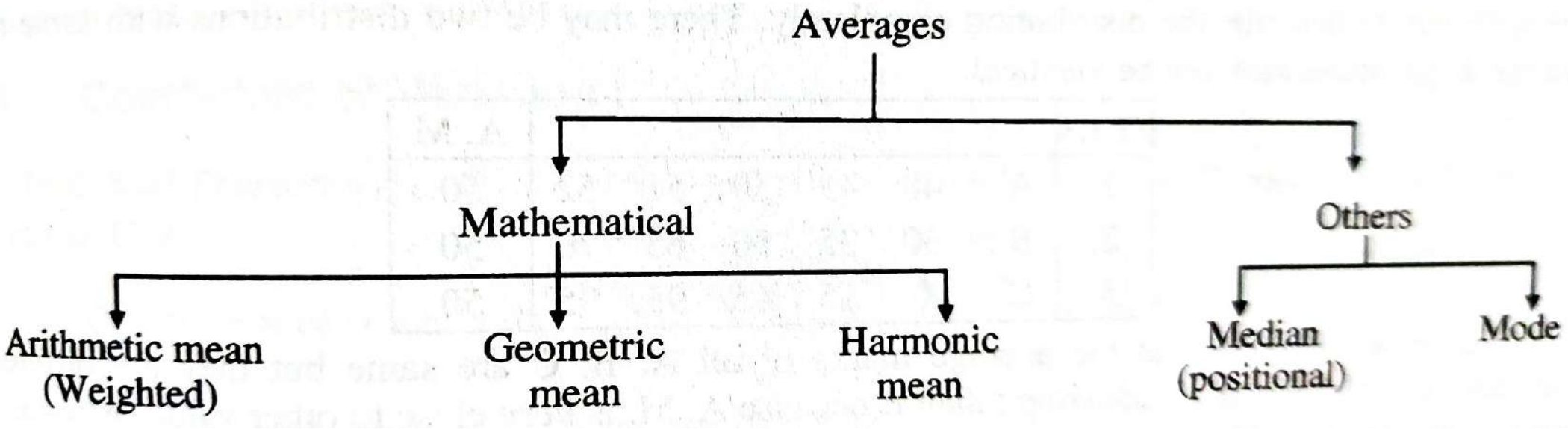
Note

* – The class interval should be uniform.
– No single observation left out.
– The number of class interval should be adequate.

Central Tendency:

Most of the observations get clustered in the central part of the data. This property of observations is described as Central tendency.

* Average is a value around which most of the observations are clustered, hence single value itself gives clear idea regarding the concept under study.


Averages
Mathematical
Others
Arithmetic mean
(Weighted)
Geometric
mean
Harmonic
mean
Median
(positional)
Mode

Arithmatic Mean (A.M):

We denote mean by letter $\bar{x}$.

(i) Let $x_1, x_2, x_3, \ldots x_n$ be the set of observations.

$$\text{Mean } (\bar{x}) = \frac{\text{Sum of observations}}{\text{total number of observations}}$$

$$\bar{x} = \frac{\sum x_i}{n}$$

(ii) For Grouped data (classified data)
(Discrete frequency distribution)

Let $x_1, x_2, \ldots x_n$ be n observations
and $f_1, f_2, \ldots f_n$ be the corresponding frequencies

$$\therefore \bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{\sum f_i x_i}{N}$$

(i) Direct Method

Here we take the mid point of every class interval as X, or m multiply these values of X or m with their respective frequencies. Take total and apply the formula ;

$$\bar{X} = \frac{\Sigma fX}{N} \quad \text{or} \quad \bar{X} = \frac{\Sigma fm}{N}$$

Where f is frequency; X or m the mid point of the class-interval and $N = \Sigma f$

Example 1. Use direct method to find $\bar{X}$.

Income :	10–20	20–30	30–40	40–50	50–60	60–70
No. of Persons :	4	7	16	20	15	8

Solution

Income	Mid value m	No. of Persons f	fm
10–20	15	4	60
20–30	25	7	175
30–40	35	16	560
40–50	45	20	900
50–60	55	15	825
60–70	65	8	520
		$N = 70$	$\Sigma fm = 3040$

As $\bar{X} = \frac{\Sigma fm}{N}$

$\therefore \quad \bar{X} = \frac{3040}{70} = 43.43.$

Date: YOUVA

(iii) For continous frequency distribution:

In this case, frequency is associated to the entire class and not any specific single value.

Note: for calculation purpose, we make a reasonable assumption that the frequency is associated to the midpoint of class.

Let $d_i = x_i - a$ is called deviation of x_i from a, Here a is any point from x_i.

$$\text{Mean}(\bar{x}) = a + \frac{\Sigma f_i d_i}{N}$$

(iv) If h be the length of interval (width) of class.

$$\text{Mean}(\bar{x}) = a + h\left[\frac{\Sigma f_i u_i}{N}\right]$$

where $u_i = \frac{x_i - a}{h}$

Effect of change of origin and scale:

Note: 1. If we add (or substract) a constant from each observations, then we say that the origin is changed.

2. If each observation is multiplied (or divided) by some constant then we say

Date: YOUVA

Dispersion:

Note: 1. We have seen that average value condenses (or merge) into a single value, But average value is not sufficient to describe the distribution completely

2. There may be two distributions with same means but the distribution may not be identical

Ex							A.M.
1	A:	48	49	50	51	52	50
2	B:	30	35	50	65	70	50
3	C:	0	15	45	95	95	50

Observation: 1. Average marks of all A, B, C are same but they are different in variation.

2. Clearly, A is more consistent than B because A.M is very close to all values whereas, B is more consistant than C.

3. Observations are scattered or dispersed from central value in the row of B & C
This is called dispersion.

4. Average remains good representative if dispersion is less.
(i.e when dispersion is small number, the average is good.)

To measure the scatteredness of data from mean, we use

(i) Range (ii) Quartile deviation
(iii) Mean deviation (iv) Standard deviation.

Standard deviation:

It is defined as the positive square root of the arithmatic mean of the squares of the deviations of given values from their A.M.

Notation: σ

(i) $S.D = \sigma = \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2}$

—— For individual observ.

OR

$\sigma = \sqrt{\frac{\sum x_i^2}{n} - (\bar{x})^2}$

~~(iii).~~

(ii) $\sigma = \sqrt{\frac{1}{N}\sum f_i(x_i - \bar{x})^2}$

—— For frequency distributio

OR

$\sigma = \sqrt{\frac{\sum f_i x_i^2}{N} - (\bar{x})^2}$

(iii) $\sigma = \sqrt{\left(\frac{\sum f_i d_i^2}{N}\right) - \left(\frac{\sum f_i d_i}{N}\right)^2}$

(iv) $\sigma = h\sqrt{\left(\frac{\Sigma f_i u_i^2}{N}\right) - \left(\frac{\Sigma f_i u_i}{N}\right)^2}$

where $u_i = \frac{d_i}{h}$, $d_i = x_i - a$, $N = \Sigma f_i$

* Variance: The square of standard deviation is called variance & it is denoted by σ^2.

Coefficient of variation:

Prof. Karl Pearson suggested the relative measure of standard deviation. It is also called coefficient of variation.

$$C.V = \frac{S.D}{|A.M|} \times 100 = \frac{\sigma}{|\bar{x}|} \times 100\ \%$$

C.V is always expressed in percentage.

Solved Examples :

Ex. 5.2.1 : The scores obtained by two batsman A and B in 10 matches are given below. Determine who is more consistent and who is better run getter ?

Batsman A :	30	44	66	62	60	34	80	46	20	38
Batsman B :	34	46	70	38	55	48	60	34	45	30

Soln. :

Step I : Since, Coefficient of Variation **(C.V.)** $= \dfrac{\sigma}{\bar{x}} \times 100$

Step II : $\text{S.D.} = \sigma = \sqrt{\dfrac{1}{n}\sum\left(x_i - \bar{x}\right)^2}$

(∵ Here, data is individual, so we use this formula)

Where, n = number of matches = 10 ; $\bar{x} = \dfrac{\sum x_i}{n}$

Step III :

(i) For Batsman A :

x_i	$x_i - \bar{x}$ $(x_i - 48)$	$\left(x_i - \bar{x}\right)^2$
30	– 18	324
44	– 4	16
66	18	324
62	14	196
60	12	144
34	– 14	196
80	32	1024
46	– 2	4
20	– 28	784
38	– 10	100
$\sum x_i = 480$	$\sum (x_i - \bar{x}) = 0$	$\sum (x_i - \bar{x})^2 = 3112$

Step IV : $\overline{x_A} = \dfrac{\sum x_i}{n} = \dfrac{480}{10} = 48$; $\sigma_A = \sqrt{\dfrac{3112}{10}} = 17.6409$

Step V : $(\text{C.V.})_A = \dfrac{\sigma_A}{\bar{x}_A} \times 100 = \left(\dfrac{17.6409}{48}\right) \times 100 = 36.7519$

Step VI :

(ii) For Batsman B :

x_i	$x_i - \bar{x} = x_i - 46$	$(x_i - \bar{x})^2$
34	– 12	144
46	0	0
70	24	576
38	– 8	64
55	9	81
48	2	4
60	14	196
34	– 12	144
45	– 1	1
30	– 16	256
$\sum x_i = 460$	$\sum (x_i - \bar{x}) = 0$	$\sum (x_i - \bar{x})^2 = 1466$

$$\bar{x}_B = \frac{460}{10} = 46$$

Step VII : S.D. $= \sigma_B = \sqrt{\frac{1}{10}(1466)} = 12.10785$

Step VIII : $(C.V.)_B = \frac{\sigma_B}{\bar{x}_B} \times 100 = \left(\frac{12.10785}{46}\right) \times 100 = 26.321$

By observation,

Step IX : $(C.V.)_A > (C.V.)_B$

This shows that Batsman B is more consistent and $\bar{x}_A > \bar{x}_B$, this shows that Batsman A is more run getter.

To measure the scatteredness of data from mean, we use

(i) Range (ii) Quartile deviation
(iii) Mean deviation (iv) Standard deviation.

Standard deviation:

It is defined as the positive square root of the arithmatic mean of the squares of the deviations of given values from their A.M.

Notation: σ.

(i) $S.D = \sigma = \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2}$

—— For individual obser

OR

$\sigma = \sqrt{\frac{\sum x_i^2}{n} - (\bar{x})^2}$

~~(iii).~~

(ii) $\sigma = \sqrt{\frac{1}{N}\sum f_i(x_i - \bar{x})^2}$

—— For frequency distributio

OR

$\sigma = \sqrt{\frac{\sum f_i x_i^2}{N} - (\bar{x})^2}$

(iii) $\sigma = \sqrt{\left(\frac{\sum f_i d_i^2}{N}\right) - \left(\frac{\sum f_i d_i}{N}\right)^2}$

(iv) $\sigma = h\sqrt{\left(\frac{\Sigma f_i u_i^2}{N}\right) - \left(\frac{\Sigma f_i u_i}{N}\right)^2}$

where $u_i = \frac{d_i}{h}$, $d_i = x_i - a$, $N = \Sigma f_i$

* Variance: The square of standard deviation is called variance & it is denoted by σ^2.

Coefficient of variation:

Prof. Karl Pearson suggested the relative measure of standard deviation. It is also called coefficient of variation.

$$C.V = \frac{S.D}{|A.M|} \times 100 = \frac{\sigma}{|\bar{x}|} \times 100\ \%$$

C.V is always expressed in percentage.

Working Rule to find C.V. for comparing two objects:

1. Write the formula of $C.V = \frac{\sigma}{\bar{x}} \times 100$
2. Write the suitable formula of S.D(σ) and mean($\bar{x}$)
3. Construct table according to formula of S.D(σ) and mean($\bar{x}$)

4. Find mean ($\bar{x}$) and S.D(σ)
5. Find C.V

Repeat above steps for second object.

6. Compare C.V of both objects.

If $(C.V)_A < (C.V)_B$ then

(i) Object A is more consistent than object B
(ii) object B is more variable than object A.

Ex: 1. The scores obtained by two batsman A and B in matches are given below. Determine who is more consistent and who is better run getter?

Soln:

Batsman A	30	44	66	62	60	34	80	46	20	38
Batsman B	34	46	70	38	55	48	60	34	45	30

Soln. Since, coefficient of variation $(C.V) = \frac{\sigma}{\bar{x}} \times 100$

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

(∵ Here, data is individual, so we use this formula)

where n = number of matches = 10

$$\bar{x} = \frac{\sum x_i}{n}$$

For Batsman A:

$$\bar{x}_A = \frac{\sum x_i}{n} = \frac{480}{10} = 48, \quad \sigma_A = \sqrt{\frac{3112}{10}} = 17.6409$$

	$x_i - 48$	
x_i	$x_i - \bar{x}_A$	$(x_i - \bar{x}_A)^2$
30	−18	324
44	−4	16
66	18	324
62	14	196
60	12	144
34	−14	196
80	32	1024
46	−2	4
20	−28	784
38	−10	100
$\sum x_i = 480$	$\sum (x_i - \bar{x}) = 0$	$\sum (x_i - \bar{x})^2 = 3112$

$$\sigma_A = \sqrt{\frac{3112}{10}} = 17.6409$$

$$(C.V)_A = \frac{\sigma_A}{\bar{x}_A} \times 100 = \left(\frac{17.6409}{48}\right) \times 100$$

$$= 36.7519$$

For Batsman B: $\bar{x}_B = \frac{460}{10} = 46$

x_i	$x_i - \bar{x}_B$	$(x_i - \bar{x}_B)^2$
34	−12	144
46	0	0
70	24	576
38	−8	64
55	9	81
48	2	4
60	14	196
34	−12	144
45	−1	1
30	−16	256
$\sum x_i = 460$	$\sum (x_i - \bar{x}_B) = 0$	$\sum (x_i - \bar{x}_B)^2 = 1466$

$$S.D = \sigma_B = \sqrt{\frac{1466}{10}} = 12.10785$$

$$(C.V)_B = \frac{\sigma_B}{\bar{x}_B} \times 100 = \left(\frac{12.10785}{46}\right) \times 100$$

$$= 26.321$$

By observation,

$$(C.V)_A > (C.V)_B$$

① This shows that batsman B is more consistent

② $\bar{x}_A > \bar{x}_B$ this shows that batsman A is more run getter (run scorer)

Ex: 2. Goals scored by two teams A and B in a football season were as follows:

Number of goals scored in a match	0	1	2	3	4
Number of matches A	27	9	8	5	4
B	17	9	6	5	3

Find out which team is more consistent.

Soln: Since $C.V = \frac{S.D}{\bar{x}} \times 100$

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

(∵ Here we use any formula of S.D)

M T W T F S S

Page No.: YOUVA

Date:

$$\bar{x} = a + \frac{\Sigma f_i d_i}{N}, \quad N = \Sigma f_i$$

where a is any value from x_i (generally central value)

For Team A: Let $a = 2$.

x_i Number of goals	f_i Number of matches	$d_i = x_i - a = x_i - 2$	$f_i d_i$	d_i^2	$f_i d_i^2$
0	27	−2	−54	4	108
1	9	−1	−9	1	9
2	8	0	0	0	0
3	5	1	5	1	5
4	4	2	8	4	16
	$\Sigma f_i = 53$		$\Sigma f_i d_i = -50$		$\Sigma f_i d_i^2 = 138$

$$\bar{x}_A = 2 + \frac{(-50)}{53} = 1.0566$$

$$\text{S.D} = \sigma_A = \sqrt{\frac{138}{53} - \left(\frac{-50}{53}\right)^2} = 1.30914$$

$$(\text{C.V})_A = \frac{\sigma_A}{\bar{x}_A} \times 100 = 1.2390 \times 100 = 123.90119$$

For Team B : Let $a = 2$

x_i	f_i	$d_i = x_i - A = x_i - 2$	$f_i d_i$	d_i^2	$f_i d_i^2$
0	17	−2	−34	4	68
1	9	−1	−9	1	9
2	6	0	0	0	0
3	5	1	5	1	5
4	3	2	6	4	12
	$N = \Sigma f_i = 40$		$\Sigma f_i d_i = -32$		$\Sigma f_i d_i^2 = 94$

$$\bar{x}_B = a + \frac{\Sigma f_i d_i}{N}$$

$$= 2 + \frac{(-32)}{40} = 1.2$$

$$S.D = \sigma_B = \sqrt{\frac{94}{40} - \left(\frac{-32}{40}\right)^2}$$

$$= 1.30766$$

$$(C.V)_B = \frac{\sigma_B}{\bar{x}_B} \times 100 = \frac{1.30766}{1.2} \times 100$$

$$= 108.9716$$

It is observed that $(C.V)_A > (C.V)_B$
This shows team B is more consistent.

Ex: 3. Calculate standard deviation for the following frequency distribution.
Decide whether A.M is good average.

Wages in Rs earned per day	0-10	10-20	20-30	30-40	40-50	50-60
Number of labourers	5	9	15	12	10	5

Soln:

Since $\sigma = h\sqrt{\left(\frac{\Sigma f_i u_i^2}{N} - \left(\frac{\Sigma f_i u_i}{N}\right)^2\right)}$

(∵ Here class width and frequency given so we use this formula).

Mean, $\bar{x} = a + h\frac{\Sigma f_i u_i}{N}$

where, $u_i = \frac{x_i - a}{h}$, $N = \Sigma f_i$

Here $h = 10$, let $a = 35$

class	x_i	f_i	$u_i = \frac{x_i - a}{h}$ $u_i = (x_i - 35)/10$	u_i^2	$f_i u_i$	$f_i u_i^2$
0-10	5	5	-3	9	-15	45
10-20	15	9	-2	4	-18	36
20-30	25	15	-1	1	-15	15
30-40	35	12	0	0	0	0
40-50	45	10	1	1	10	10
50-60	55	3	2	4	6	12
		$\Sigma f_i = 54$			$\Sigma f_i u_i = -32$	$\Sigma f_i u_i^2 = 118$

$$A.M(\bar{x}) = 35 + 10\left(\frac{-32}{54}\right) = 29.0740$$

$$S.D(\sigma) = 10\sqrt{\left(\frac{118}{54}\right) - \left(\frac{-32}{54}\right)^2}$$

$$= 13.5428$$

Note: Here, σ is very much deviated from (A.M) arithmatic mean;
∴ A.M is not good average.

Ex:4. The mean and standard deviation of 25 items is found to be 11 and 3 respectively. It was observed that one item 9 was incorrect. Calculate the mean and standard deviation if
(i) The wrong item is omitted.
(ii) It is replaced by 13.

Soln. Given: $n=25$, $\bar{x}=11$, $\sigma=3$.

Since, $\bar{x} = \frac{\Sigma x_i}{n} \Rightarrow 11 = \frac{\Sigma x_i}{25}$

$\Sigma x_i = 275$

(i) Wrong item 9 is omitted.
∴ $n=24$, $\Sigma x_i - 9 = 275 - 9 = 266$

∴ corrected mean $(\bar{x}) = \frac{\Sigma x_i}{n} = \frac{266}{24}$

$= 11.0833$

Now $\sigma = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n}}$ (∵ It is suitable formula as per data)

$3 = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{25}}$

$$\Rightarrow 9 = \frac{1}{25}\sum(x_i - \bar{x})^2$$

$$225 = \sum(x_i - \bar{x})^2 \quad \text{——— (1)}$$

$$\Rightarrow \sum x_i - \bar{x} = 15$$

* By omitting item 9 from data,

$$\therefore (x_i - \bar{x})^2 = (9 - 11)^2 = 4$$

Now, $\sum(x_i - \bar{x})^2 = 225 - 4$ from (1)

$$\sum(x_i - \bar{x})^2 = 221$$

$$\underline{\text{Corrected S.D} = \sigma = \sqrt{\frac{221}{74}} = 3.0345}$$

(ii). If item '9' replaced by 13

$\therefore n = 25$, $\sum x_i = \cancel{266 + 13}$

$= 275 - 9 + 13$

$= 266 + 13$

$= 279$

$$\text{corrected mean} = \frac{279}{25} = 11.16$$

For S.D $(x_i - \bar{x})^2 = (13 - 11.16)^2$

$= (1.84)^2$

$= 3.3856$

$\therefore \sum(x_i - \bar{x})^2 = 221 + 3.3856$

$= 224.3856$

$$\text{Corrected S.D}(\sigma) = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}} = \sqrt{\frac{224.3856}{25}}$$

$$\sigma = 2.9959$$

Moments, Skewness and Kurtosis

Moment is a familiar mechanical term which refers to the measure of a force w.r.t it's tendency to provide rotation.

Note: The strength of the tendency depends on the amount of force and the distance from the origin of the point at which force is exerted.

If the number of forces $f_1, f_2, \cdots f_n$ at distance $x_1, x_2, \cdots x_n$ are applied, the moment of the first force about the origin is f_1x_1, for the second is f_2x_2 ... and so on.

$\Sigma f_i x_i$ = is the total moment about the origin.

Note: If the total moment is divided by the total force, the quotient is termed as 'a moment'. ($= \frac{\Sigma f_i x_i}{N}$, $N = \Sigma f_i$ = total force)

Moment about mean:

The r^{th} moment of the variable x about the mean ($\bar{x}$) of a distribution is denoted by μ_r and is defined as —

$$\mu_r = \frac{1}{n} \sum (x_i - \bar{x})^r \quad \text{(without frequency)} \quad \text{---(1)}$$

OR

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

where $N = \sum f_i$

(without frequency)

where $r = 0, 1, 2, 3, \ldots$

$r = 0$, $\mu_0 = 1/N \times \sum f_i = \frac{N}{N} = 1$, $\boxed{\mu_0 = 1}$

$r = 1$, $\mu_1 = 0$

As $\mu_1 = \frac{1}{N} \sum f_i (x - \bar{x})$

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

$\boxed{\mu_1 = 0}$ [first moment of distribution about mean]

$r = 2$, $\boxed{\mu_2 = \sigma^2 = \text{Variance}}$ [second MOI about mean]

$\therefore \boxed{\text{S.D} = \sigma = \sqrt{\mu_2}}$

Moment about any value 'a':

The r^{th} moment of the variable x about any point 'a' of a distribution is denoted by μ'_r and is defined as -

$$\mu'_r = \frac{1}{N} \sum f_i (x_i - a)^r = \frac{1}{N} \sum f_i d_i^r \quad \text{---(2)}$$

(If frequency is given)

$$\mu_r' = \frac{1}{n}\sum_{r=0}^{r}(x_i - a)^r = \frac{1}{n}\sum d_i^r$$

OR

$$\mu_r' = h^r \frac{1}{N}\sum f_i u_i^r, \quad r = 0, 1, 2, 3, \ldots$$

(If class width is given)

where $u_i = \frac{x_i - a}{h}$

From eq^n (2).

$$\boxed{\mu_0' = 1}$$

$$\boxed{\mu_1' = \bar{x} - a}, \quad \boxed{\bar{x} = a + \mu_1'}$$

As, $\mu_1' = \frac{\sum(x_i - a)}{n}$

$= \frac{\sum x_i - a}{n}$

$= \bar{x} - a$

Relation between Moments about Mean (μ_r) in terms of Moments about any point (μ_r'):

Since $\mu_r = \frac{1}{N}\sum f_i (x_i - \bar{x})^r$

but $d_i = x_i - a \Rightarrow x_i = d_i + a$

$$\therefore \mu_r = \frac{1}{N}\sum f_i \left[(d_i + a) - \bar{x}\right]^r$$

$$= \frac{1}{N}\sum f_i \left[d_i - (\bar{x} - a)\right]^r$$

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

By putting $r = 1, 2, 3, \cdots$, we get

$$\mu_1 = 0$$

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

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

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

$$\vdots$$

$$\mu_r = \mu_r' - {}^rC_1\mu_{r-1}'\mu_1' + {}^rC_2\mu_{r-2}'(\mu_1')^2 - {}^rC_3\mu_{r-3}'(\mu_1')^3 + \cdots + (-1)^r(\mu_1')^r$$

$$\mu_1 = \frac{1}{N}\sum f_i(d_i - \mu_1') = \frac{\sum f_i d_i}{N} - \mu_1'$$

$$= \mu_1' - \mu_1'$$

$$= 0$$

$$\mu_2 = \frac{1}{N}\sum f_i(d_i - \mu_1')^2$$

$$= \frac{1}{N}\sum f_i\left[d_i^2 - 2d_i\mu_1' + (\mu_1')^2\right]$$

$$= \frac{\sum f_i d_i^2}{N} - 2\frac{\sum f_i d_i}{N}\mu_1' + (\mu_1')^2$$

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

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

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

Note that:

1. The sum of the coefficients of the various terms on R.H.S is zero.
2. For grouped data, we may use.

$$\mu_r = \frac{h^r}{N} \sum f_i (u_i - \bar{u})^r$$

where $\bar{u} = a + h \frac{\sum f_i u_i}{N}$

(If frequency and class width is given)

Skewness:

For symmetrical distribution, the frequencies are symmetrically distributed about the mean. In such distribution, the mean, mode and median are coincides.

Skewness means "lack of symmetry".

Note: Skewness measures the degree of symmetry or the departure from symmetry.

For positive skewness : Mode < Median < Mean

and

For Negative skewness : Mode > Median > Mean

Coefficient of skewness is denoted by β_1 and is defined as -

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

If $\beta_1 = 0$ then the distribution is symmetrical.

Kurtosis: Kurtosis signifies the peakness of the distribution curve.

Note: ① The coefficient of curtosis is denoted by β_2 and is defined as -

$$\beta_2 = \frac{\mu_4}{(\mu_2)^2}$$

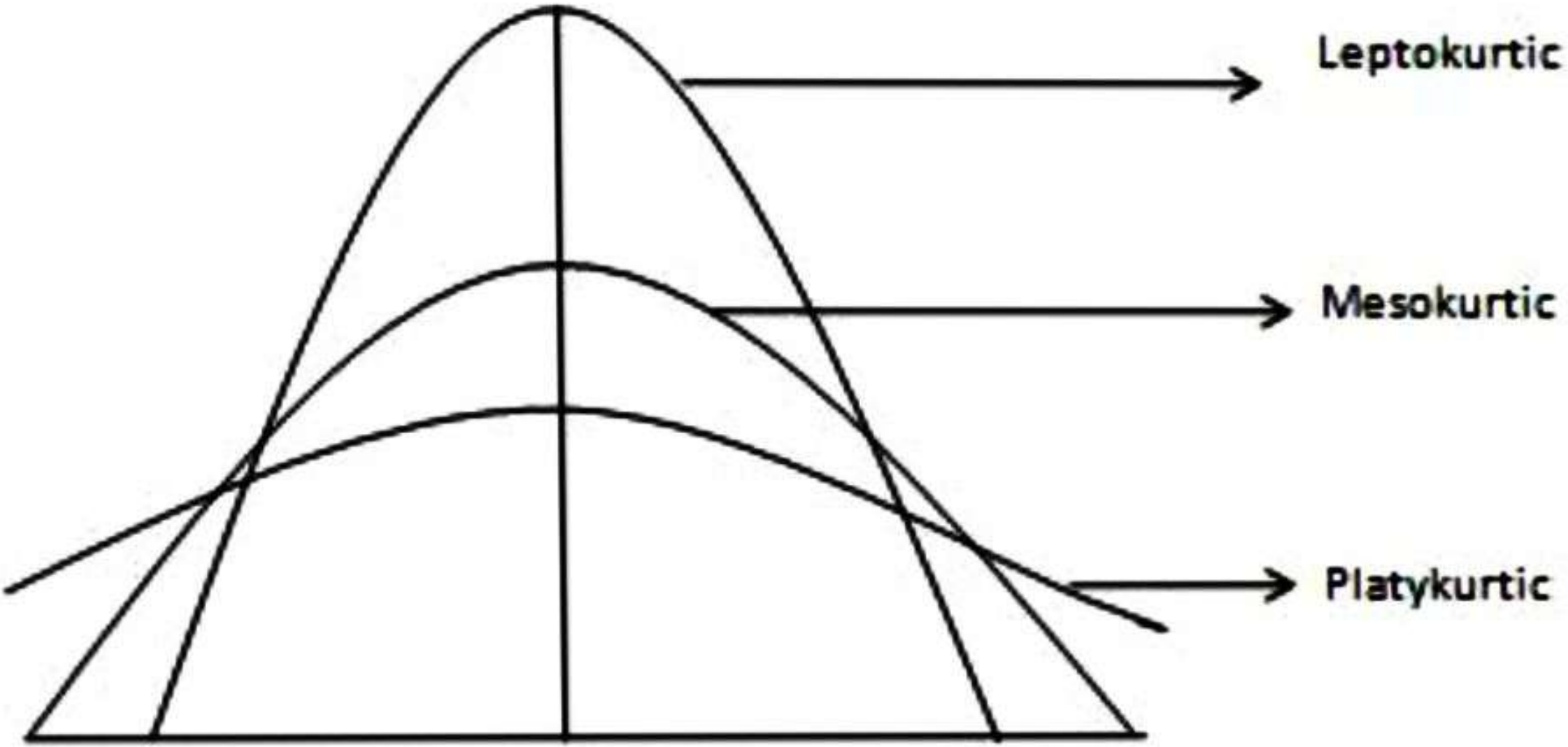
② Kurtosis gives the idea about the flatness or peakness of a curve

③ If the curve is neither flat nor peaked then it is called normal or Mesokurtic curve

If $\beta_2 < 3$, then distribution is
— platykurtic.

If $\beta_2 = 3$; then distribution is called
— Mesokurtic or normal.

If $\beta_2 > 3$; then the distribution is called
— Leptokurtic or more peaked.


Leptokurtic
Mesokurtic
Platykurtic

Note : (i) $\gamma_1 = \sqrt{\beta_1}$ is the simplest measure of skewness.

(ii) $\gamma_2 = \beta_2 - 3$ is the excess of Kurtosis.

Note : Sometimes we have to find moment from table value. In this case, first find the moment about any value by using the formula

$$\mu_r' = \frac{1}{N} \sum f_i (x_i - a)^r$$ and then find moment about mean by using formulae between μ_r and μ_r'.

Moments, Skewness and Kurtosis

Moment is a familiar mechanical term which refers to the measure of a force w.r.t it's tendency to provide rotation.

Note: The strength of the tendency depends on the amount of force and the distance from the origin of the point at which force is exerted.

If the number of forces $f_1, f_2, \cdots f_n$ at distance $x_1, x_2, \cdots x_n$ are applied, the moment of the first force about the origin is f_1x_1, for the second is f_2x_2 ... and so on.

$\sum f_ix_i$ = is the total moment about the origin.

Note: If the total moment is divided by the total force, the quotient is termed as 'a moment'. ($= \frac{\sum f_ix_i}{N}$, $N = \sum f_i$ = total force)

Moment about mean:

The r^{th} moment of the variable x about the mean ($\bar{x}$) of a distribution is denoted by μ_r and is defined as -

$$\mu_r = \frac{1}{n} \sum (x_i - \bar{x})^r \quad \text{(without frequency)} \quad \text{---(1)}$$

OR

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

where $N = \sum f_i$

(without frequency)

where $r = 0, 1, 2, 3, \ldots$

$r = 0$, $\mu_0 = 1/N \times \sum f_i = \frac{N}{N} = 1$, $\boxed{\mu_0 = 1}$

$r = 1$, $\mu_1 = 0$

As $\mu_1 = \frac{1}{N} \sum f_i (x - \bar{x})$

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

$\boxed{\mu_1 = 0}$ [first moment of distribution about mean]

$r = 2$, $\boxed{\mu_2 = \sigma^2 = \text{Variance}}$ [second MOD about mean]

$\therefore \boxed{\text{S.D} = \sigma = \sqrt{\mu_2}}$ ₹

Moment about any value 'a':

The r^{th} moment of the variable x about any point 'a' of a distribution is denoted by μ'_r and is defined as -

$$\mu'_r = \frac{1}{N} \sum f_i (x_i - a)^r = \frac{1}{N} \sum f_i d_i^r \quad \text{---(2)}$$

(If frequency is given)

$$\mu_r' = \frac{1}{n}\sum_{r=0}^{r}(x_i - a)^r = \frac{1}{n}\sum d_i^r$$

OR

$$\mu_r' = h^r \frac{1}{N}\sum f_i u_i^r, \quad r = 0, 1, 2, 3, \ldots$$

(If class width is given)

From eq^n ②.

$$\boxed{\mu_0' = 1}$$

$$\boxed{\mu_1' = \bar{x} - a}; \quad \boxed{\bar{x} = a + \mu_1'}$$

$$\text{As, } \mu_1' = \frac{\sum(x_i - a)}{n}$$

$$= \frac{\sum x_i - a}{n}$$

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

Relation between Moments about Mean (μ_r) in terms of Moments about any point (μ_r'):

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

$$\text{but } d_i = x_i - a \Rightarrow x_i = d_i + a$$

$$\therefore \mu_r = \frac{1}{N}\sum f_i \left[(d_i + a) - \bar{x}\right]^r$$

$$= \frac{1}{N}\sum f_i \left[d_i - (\bar{x} - a)\right]^r$$

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

By putting $r = 1, 2, 3, \cdots$, we get

$$\mu_1 = 0$$

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

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

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

$$\vdots$$

$$\mu_r = \mu_r' - {}^rC_1\mu_{r-1}'\mu_1' + {}^rC_2\mu_{r-2}'(\mu_1')^2 - {}^rC_3\mu_{r-3}'(\mu_1')^3 + \cdots + (-1)^r(\mu_1')^r$$

$$\mu_1 = \frac{1}{N}\sum f_i(d_i - \mu_1') = \frac{\sum f_i d_i}{N} - \mu_1'$$
$$= \mu_1' - \mu_1'$$
$$= 0$$

$$\mu_2 = \frac{1}{N}\sum f_i(d_i - \mu_1')^2$$

$$= \frac{1}{N}\sum f_i\left[d_i^2 - 2d_i\mu_1' + (\mu_1')^2\right]$$

$$= \frac{\sum f_i d_i^2}{N} - 2\frac{\sum f_i d_i}{N}\mu_1' + (\mu_1')^2$$

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

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

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

Note that:

1. The sum of the coefficients of the various terms on R.H.S is zero.

2. For grouped data, we may use.

$$\mu_r = \frac{h^r}{N} \sum f_i (u_i - \bar{u})^r$$

where $\bar{u} = a + h \frac{\sum f_i u_i}{N}$

(If frequency and class width is given)

Skewness:

For symmetrical distribution, the frequencies are symmetrically distributed about the mean. In such distribution, the mean, mode and median are coincides.

Skewness means "lack of symmetry".

Note: Skewness measures the degree of symmetry or the departure from symmetry.

For positive skewness : Mode < Median < Mean

and

For Negative skewness : Mode > Median > Mean

Coefficient of skewness is denoted by β_1 and is defined as –

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

If $\beta_1 = 0$ then the distribution is symmetrical.

Kurtosis: Kurtosis signifies the peakness of the distribution curve.

Note: ① The coefficient of curtosis is denoted by β_2 and is defined as –

$$\beta_2 = \frac{\mu_4}{(\mu_2)^2}$$

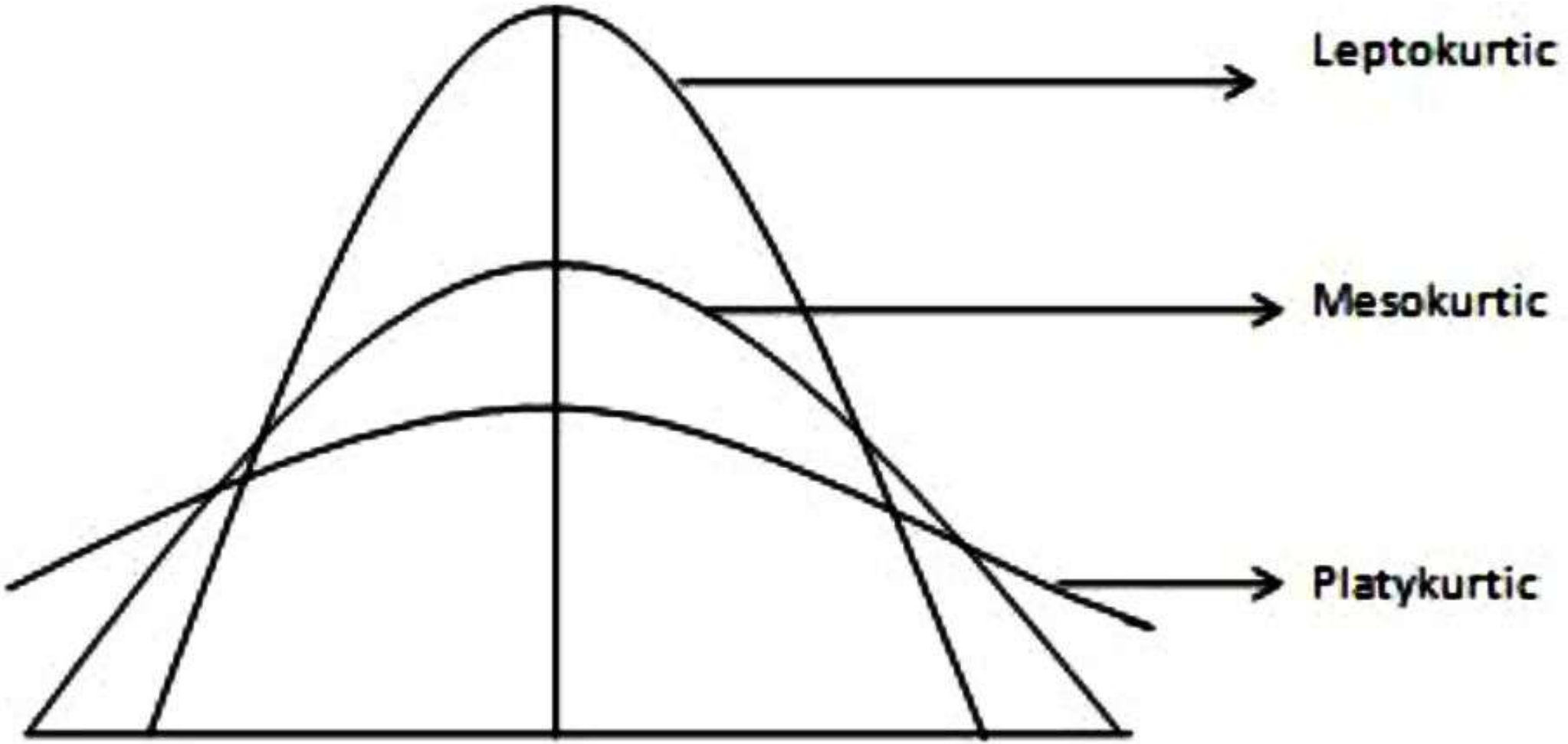
② Kurtosis gives the idea about the flatness or peakness of a curve

③ If the curve is neither flat nor peaked then it is called normal or Mesokurtic curve

If $\beta_2 < 3$, then distribution is
— platykurtic.

If $\beta_2 = 3$; then distribution is called
— Mesokurtic or normal.

If $\beta_2 > 3$; then the distribution is called
— Lepokurtic or more peaked.


Leptokurtic
Mesokurtic
Platykurtic

Note : (i) $\gamma_1 = \sqrt{\beta_1}$ is the simplest measure of skewness.

(ii) $\gamma_2 = \beta_2 - 3$ is the excess of Kurtosis.

Note : Sometimes we have to find moment from table value. In this case, first find the moment about any value by using the formula

$$\mu'_r = \frac{1}{N} \sum f_i (x_i - a)^r$$ and then find moment about mean by using formulae between μ_r and μ'_r.

M T W T F S S

Page No.: YOUVA

Date:

Ex: 1 : First four moments about the working mean 44.5 of a distribution are −0.4, 2.99, −0.08 and 27.63. Calculate the moments about the mean. Also calculate β_1 and β_2.

Soln: Given : $a = 44.5$

$\mu_1' = -0.4$, $\mu_2' = 2.99$, $\mu_3' = -0.08$, $\mu_4' = 27.63$.

Moments about mean : $\mu_1 = 0$.

$$\mu_2 = \mu_2' - (\mu_1')^2$$
$$= 2.99 - (-0.4)^2$$
$$= 2.99 - 0.16$$
$$= 2.83$$

$$\mu_3 = \mu_3' - 3(\mu_2')(\mu_1') + 2(\mu_1')^3$$
$$= (-0.08) - 3(2.99)(-0.4) + 2(-0.4)^3$$
$$= 3.38$$

$$\mu_4 = \mu_4' - 4(\mu_1')(\mu_3') + 6(\mu_1')^2(\mu_2') - 3(\mu_1')^4$$
$$= (27.63) - 4(-0.4)(-0.08) + 6(-0.4)^2(2.99) - 3(-0.08)^4$$

$\mu_4 = 27.63 - 0.128 + 2.8704 - 0.0768$

$= 30.2956$

$\therefore$ coefficient of skewness $(\beta_1) = \frac{\mu_3^2}{\mu_2^3} = \frac{(3.38)^2}{(2.83)^3}$

$= 0.504$

coefficient of kurtosis $(\beta_2) = \frac{\mu_4}{\mu_2^2} = \frac{30.2956}{(2.83)^2}$

$= 3.7827$

$\beta_2 > 3$, the distribution is leptokurtic.

Ex: 2 The first three moments about the value 2 of a distribution are 1, 16 and -40. Find mean, standard deviation and β_1 (coeffi. of skewness) of the distribution.

soln: Given: $a = 2$.

$\mu_1' = 1$; $\mu_2' = 16$ and $\mu_3' = -40$

Moments about mean,

$\mu_1 = 0$

$\mu_2 = \mu_2' - (\mu_1')^2 = 16 - 1 = 15$

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

$= (-40) - 3(16)(1) + 2(1)^3$

$= -40 - 48 + 2 = -86$

Mean, $(\bar{x}) = a + \mu_1' = 2 + 1 = 3$

Standard deviation $(\sigma) = \sqrt{\mu_2} = \sqrt{15}$
$= 3.8729$

$$\beta_1 = \frac{(\mu_3)^2}{(\mu_2)^3} = \frac{(-86)^2}{(15)^3} = \frac{7396}{3375} = 2.1914$$

Ex: 3. The first four moments of a distribution about the value 4 are $-1.5, 17, -30$ and 108. Obtain the first four central moments, mean, standard deviation and coefficient of skewness and kurtosis.

Soln: Given $a = 4$

$\mu_1' = -1.5$, $\mu_2' = 17$, $\mu_3' = -30$ and $\mu_4' = 108$

Moments about mean:

$\mu_1 = 0$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 17 - (-1.5)^2$$
$$= 17 - 2.25 = 14.75$$

$$\mu_3 = \mu_3' - 3(\mu_2')(\mu_1') + 2(\mu_1')^3$$
$$= (-30) - 3(17)(-1.5) + 2(-1.5)^3$$
$$= -30 + 76.5 - 6.75$$
$$= 39.75$$

$$\mu_4 = \mu_4' - 4(\mu_1')(\mu_3') + 6(\mu_1')^2(\mu_2') - 3(\mu_1')^4$$
$$= 108 - 4(-1.5)(-30) + 6(-1.5)^2(17) - 3(-1.5)^4$$
$$= 108 - 180 + 3901.5 - 15.1875$$
$$= 3814.3125$$

Co-efficient of skewness $(\beta_1) = \frac{\mu_3}{\mu_2^3}$

$= \frac{(39.75)^2}{(14.45)^3}$

$= 0.4923$

Co-efficient of kurtosis $(\beta_2) = \frac{\mu_4}{\mu_2^2}$

$= \frac{3814.3125}{(14.75)^2}$

$= 17.5320$

Mean $(\bar{x}) = a + \mu_1' = 4 + (-1.5) = 2.5$

standard deviation $(\sigma) = \sqrt{\mu_2}$

$= \sqrt{14.75}$

$= 3.84$

Ex:4. The first four moments of a distribution about the values are 2, 20, 40 and 50. From the given information, obtain the first four central moments, mean, standard deviation and coefficient of skewness and kurtosis.

$\beta_1 = 1$, $\beta_2 = 0.6328 < 3 \Rightarrow$ platy-kurtic

Ex:5. The first four moments about the working mean 30.2 of a distribution are 0.255, 6.222, 30.211 and 400.25. Calculate the moments about the mean. Also evaluate β_1, β_2 and comments upon the skewness and kurtosis of the distribution.

$\beta_1 = 2.78255$, $\beta_2 = 9.80916$

Given curve is Leptokurtic curve

Ex: 6. Calculate the first four moments of the following distribution about the mean and hence find β_1 and β_2.

x_i	0	1	2	3	4	5	6	7	8
f_i	1	8	28	56	70	56	28	8	1

Soln: We know the moments about any value is,

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

$N = \sum f_i$, Let $a = 4$.

x_i	f_i	$d_i = x_i - a = x_i - 4$	$f_i d_i$	$f_i d_i^2$	$f_i d_i^3$	$f_i d_i^4$
0	1	-4	-4	16	-64	256
1	8	-3	-24	72	-216	648
2	28	-2	-56	112	-224	448
3	56	-1	-56	56	-56	56
4	70	0	0	0	0	0
5	56	1	56	56	56	56
6	28	2	56	112	224	448
7	8	3	24	72	216	648
8	1	4	4	16	64	256
	$\sum f_i = 256$		$\sum f_i d_i = 0$	$\sum f_i d_i^2 = 512$	$\sum f_i d_i^3 = 0$	$\sum f_i d_i^4 = 2816$

Put $r = 1, 2, 3, 4$ in eq^n (1).

$$\mu_1' = \frac{\sum f_i d_i}{N} = 0, \quad \mu_2' = \frac{\sum f_i d_i^2}{N} = \frac{512}{256} = 2$$

$$\mu_3' = \frac{\sum f_i d_i^3}{N} = 0, \quad \mu_4' = \frac{\sum f_i d_i^4}{N} = \frac{2816}{256} = 11$$

Moments about mean are -

$\mu_1 = 0$ $\quad \mu_2 = \mu_2' - (\mu_1')^2 = 2$

$\mu_3 = \mu_3' - 3\mu_2'(\mu_1') + 2(\mu_1')^3 = 0$

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

$= 11 - 4 \times 0 \times 0 + 6 \times 2 \times 0 - 3 \times 0$

$= 11$

Now

$\beta_1 =$ coefficient of skewness

$= \frac{\mu_3^2}{\mu_2^3} = \frac{0}{(2)^3} = 0$

$\beta_2 =$ coefficient of kurtosis

$= \frac{\mu_4}{\mu_2^2} = \frac{11}{4} = 2.75 < 3$

The given distribution is platykurtic

Ex: 7. Calculate the first four moments about the mean of the given distribution.
Also find β_1 and β_2.

x_i	2.0	2.5	3.0	3.5	4.0	4.5	5.0
f_i	4	36	60	90	70	40	10

$\beta_1 = 1.7605 \times 10^{-4}$

$\beta_2 = 1.194984$

Ex: 8. Calculate the first four moments about mean for the following data.

x	1	2	3	4	5	6	7	8	9
f	1	6	13	25	30	22	9	5	2

$\mu_1 = 0$, $\mu_2 = 2.4878$, $\mu_3 = 0.6789$,
$\mu_4 = 18.3359$.

Ex:7 Bold: The moments about any value is

$$\mu_r' = h^r \frac{\sum f_i u_i^r}{\sum f_i}$$ ——— (1) we may use

(We may use the formula.

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

$$= \frac{1}{N} \sum f_i d_i^r$$)

take & $a = 3.5$, $h = 0.5$, $u_i = \frac{x_i - a}{h}$

$$= \frac{x_i - 3.5}{0.5}$$

x_i	$u_i = \frac{x_i - a}{h} = \frac{x_i - 3.5}{0.5}$	f_i	$f_i u_i$	$f_i u_i^2$	$f_i u_i^3$	$f_i u_i^4$
2.0	−3	4	−12	36	−108	324
2.5	−2	36	−72	144	−288	576
3.0	−1	60	−60	60	−60	60
3.5	0	90	0	0	0	0
4.0	1	70	70	70	70	70
4.5	2	40	80	160	320	640
5.0	3	10	30	90	270	810
		$\Sigma f_i = 310$	$\Sigma f_i u_i = 36$	$\Sigma f_i u_i^2 = 560$	$\Sigma f_i u_i^3 = 204$	$\Sigma f_i u_i^4 = 2480$

Put $r = 1, 2, 3, 4$ in eq^n (1).

$$\mu_1' = \frac{h \Sigma f_i u_i}{N} = 0.5 \times \frac{36}{310} = 0.058065$$

$$\mu_2' = \frac{h^2 \Sigma f_i u_i^2}{N} = (0.5)^2 \frac{560}{310} = 0.451613$$

$$\mu_3' = \frac{h^3 \Sigma f_i u_i^3}{N} = (0.5)^3 \times \frac{204}{310} = 0.082259$$

$$\mu_4' = \frac{h^4 \Sigma f_i u_i^4}{N} = (0.5)^4 \frac{2480}{310} = 0.5$$

Moments about mean:

$\mu_1 = 0$

1 −3 2

1 −4 6

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

$$= 0.451613 - (0.058065)^2$$

$$= 0.448242$$

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

$$= (0.082259) - 3(0.451613)(0.058065) + 2(0.058065)^3$$

$$= 0.0039819$$

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

$$= (0.5) - 4(0.082259)(0.058065) + 6(0.451613)(0.058065)^2 - 3(0.058065)^4$$

$$\mu_4 = 0.489997$$

(i) β_1 = coefficient of skewness

$$= \frac{\mu_3^2}{\mu_2^3} = \frac{(0.0039819)^2}{(0.448242)^3}$$

$$= 1.7605 \times 10^{-4}$$

(ii) β_2 = coefficient of kurtosis

$$= \frac{\mu_4}{\mu_2^2} = \frac{0.489997}{(0.448242)^2}$$

$$= 1.194984$$

Correlation and Regression

Many times we come across situations where two variables are inter-related. If two variables (i.e bivariate distribution) x and y changes such that the change in one variable affects a change in the other variables, then these variables are said to be correlated, and this relationship is called correlation

e.g. Relation between income and expenditure
Relation between price and demand.

Types of correlation:

(i) Positive correlation: If an increase (or decrease) in one variable corresponds to an increase (or decrease) in other variables, then correlation is said to be positive or directive correlation.

say $x \propto y$

(ii) Negative correlation: If an increase (or decrease) in one variable corresponds to an decrease (~~or increase~~ (or increase) in other variables then correlation is called negative or inverse relation.

$x \propto \frac{1}{y}$

Note: If there is no relation between two variables, then these variables are said to be independent.

(iii) Perfect correlation: If two variables, vary

in such a way that their ratio is ~~always~~ constant, then the correlation is called perfect correlation.

Covariance :

Let us consider n pair of observations of two variables X and Y such as (x_1, y_1), (x_2, y_2), (x_3, y_3) $\cdots$ (x_n, y_n)
where $x_1, x_2, \cdots x_n$ are observed values of X and $y_1, y_2, \cdots y_n$ are observed values of Y

Note: 1. The deviation of X values from their mean $\bar{x}$ are $(x_1 - \bar{x})$, $(x_2 - \bar{x})$, $(x_3 - \bar{x})$ $\cdots$ $\cdots (x_n - \bar{x})$.

2. The deviation of Y values from their mean $\bar{y}$ are $(y_1 - \bar{y})$, $(y_2 - \bar{y})$, $(y_3 - \bar{y})$ $\cdots$ $\cdots (y_n - \bar{y})$

Then,

$$Cov(x, y) = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots (x_n - \bar{x})(y_n - \bar{y})}{n}$$

$$= \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{n} \left[\sum (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y} \right]$$

$$= \frac{1}{n} \sum x_i y_i - \bar{y} \frac{\sum x_i}{n} - \bar{x} \frac{\sum y_i}{n} + \bar{x} \bar{y}$$

$$= \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y}$$

$$\boxed{Cov(x,y) = \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}$$

Properties of covariance

(i) $Cov(x,x) = Var(x)$

(ii) Effect of change of origin:

$Cov(x-a, y-b) = Cov(x,y)$, where a, b are constants.

(iii) Effect of change of scale:

$$Cov\left(\frac{x-a}{h}, \frac{y-b}{k}\right) = \frac{1}{hk} Cov(x,y),$$

where a, b, h, k are constants.

(iv) $Cov(aX+b, cY+d) = ac\, Cov(X,Y)$

(v) $Cov(X+Y, Z) = Cov(X,Z) + Cov(Y,Z)$

(vi) If x and y are independent variables then $Cov(X,Y) = 0$.

(vii) $Cov(X,Y) = Cov(Y,X)$

(viii) $Cov(X,Y)$ may be negative value.

(ix) Covariance is a joint central moment of order $(1,1)$ of (X,Y)

$\therefore \mu_{11} = Cov(X,Y)$

Karl-Pearson's Co-efficient of Correlation:

Correlation co-efficient is calculated to study the extent or the degree of relationship between variables x and y.

It is usually denoted by $\gamma(x,y)$ or γ_{xy} or $corr(x,y)$.

$\therefore$ Karl Pearson's co-efficient of correlation is defined as -

$$\gamma(x,y) = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$$

$$\sigma_x^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 = \frac{1}{n}\left[\sum(x_i^2 - 2x_i\bar{x} + \bar{x}^2)\right]$$

$$= \frac{1}{n}\sum x_i^2 - \frac{2\bar{x}}{n}\sum x_i + \frac{\sum n\bar{x}^2}{n}$$

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

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

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

since $\sum(\bar{x})^2$
$= (\bar{x})^2 \sum_{n=1}^{n}(1)$
$= (\bar{x})^2 n$
$= n(\bar{x})^2$

$$\gamma(x,y) = \frac{\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum(x_i-\bar{x})^2 \times \frac{1}{n}\sum(y_i-\bar{y})^2}}$$

$$\gamma(x,y) = \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{\sqrt{\left[\frac{1}{n}\sum x_i^2 - (\bar{x})^2\right]\left[\frac{1}{n}\sum y_i^2 - (\bar{y})^2\right]}}$$

Note: (1) The denominator of $\gamma(x,y)$ is always positive, so the value of $\gamma(x,y)$ is positive or negative according to covariance.

(2). Correlation coefficient $\gamma(x,y)$ does not change in magnitude under the change of origin and scale.

i.e $\gamma\left(\frac{x-a}{h}, \frac{y-b}{k}\right) = \gamma(x,y)$; if both h & k have same signs

$= -\gamma(x,y)$, if h & k have opposite signs.

Also $\gamma(x,y)$ lies between -1 and 1 i.e $-1 \le \gamma \le 1$

Let $u_i = \frac{x_i-a}{h}$, $v_i = \frac{y_i-b}{k}$

Note If h & k have same algebraic sign and let $h=k=1$

$\therefore u_i = x_i - a, \quad v_i = y_i - b$

$Cov(x_i - a, y_i - b) = Cov(u_i, v_i)$ —— by property

$= \frac{1}{n} \sum u_i v_i - \bar{u}\bar{v}$

$\sigma_u^2 = \frac{1}{n} \sum u_i^2 - (\bar{u})^2, \quad \sigma_v^2 = \frac{1}{n} \sum v_i^2 - (\bar{v})^2$

$\therefore \gamma(u,v) = \frac{Cov(u,v)}{\sigma_u \cdot \sigma_v}$

Note ① The probable error of coefficient of correlation (γ) is $\pm 0.6745 \left(\frac{1-\gamma^2}{\sqrt{N}}\right)$

② If h & k have same sign then $hk > 0$, hence $\gamma(x,y) = \gamma(u,v)$

For frequency distribution, i.e if frequency is given then,

(i) $Cov(x,y) = \frac{1}{N} \sum f_i (x_i - \bar{x})(y_i - \bar{y})$

$\sigma_x = \sqrt{\frac{1}{N} \sum f_i (x_i - \bar{x})^2}$

$\sigma_y = \sqrt{\frac{1}{N} \sum f_i (y_i - \bar{y})^2}$

Also in terms of u_i and v_i

$\longrightarrow$

(ii) $Cov(u,v) = \frac{1}{N}\sum f_i u_i v_i - \bar{u}\bar{v}$

$$\sigma_u = \sqrt{\frac{1}{N}\left(\sum f_i u_i^2\right) - (\bar{u})^2}$$

$$\sigma_v = \sqrt{\frac{1}{N}\left(\sum f_i v_i\right)^2 - (\bar{v})^2}$$

Properties : (i) If $\gamma(x,y) > 0$, then correlation is positive.

(ii) If $\gamma(x,y) < 0$, then correlation is negative.

(iii) If $\gamma(x,y) = 0$, then there is no relation.

Ex: 1 : Find the correlation coefficient between x & y. Given that : $n = 50$, $\sum(x_i - 40) = 30$, $\sum(y_i - 20) = 70$, $\sum(x_i - 40)^2 = 170$, $\sum(y_i - 20)^2 = 165$, $\sum(x_i - 40)(y_i - 20) = 140$.

Soln : Given : $n = 50$, $\sum(x_i - 40) = 30$, $\sum(y_i - 20) = 70$

$\sum(x_i - 40)^2 = 170$, $\sum(y_i - 20)^2 = 165$

$\sum(x_i - 40)(y_i - 20) = 140$

Let $u_i = x_i - 40$, $v_i = y_i - 20$

Hence, $\bar{u} = \frac{\sum u_i}{n} = \frac{30}{50} = 0.6$

$\bar{v} = \frac{\sum v_i}{n} = \frac{70}{50} = 1.4$

coefficient of correlation

$$r(x,y) = \frac{\sum u_i v_i - n\bar{u}\bar{v}}{\sqrt{\left[\sum u_i^2 - n(\bar{u})^2\right] \times \left[\sum v_i^2 - n(\bar{v})^2\right]}}$$

(since frequency is not given.)

$$\therefore r(x,y) = \frac{140 - 50 \times (0.6)(1.4)}{\sqrt{\left[(170) - 50(0.6)^2\right]\left[(165) - 50(1.4)^2\right]}}$$

Date:

$$\gamma(x,y) = \frac{98}{\sqrt{152 \times 67}} = \frac{98}{100.9158}$$

$$\boxed{\gamma(x,y) = 0.9711}$$

Ex: 3, Find the coefficient of correlation for the following table.

x_i	10	14	18	22	26	30
y_i	18	12	24	6	30	36

Soln: We know, Karl Pearson's coefficient of correlation $\gamma(x,y)$

$$\gamma(x,y) = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n}(\Sigma x_i y_i) - \bar{x}\bar{y}}{\sqrt{\frac{1}{n}\Sigma x_i^2 - \bar{x}^2}\sqrt{\frac{1}{n}\Sigma y_i^2 - \bar{y}^2}}$$

and $\bar{x} = \frac{\Sigma x_i}{n}$, $\bar{y} = \frac{\Sigma y_i}{n}$, Here $n = 6$

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
10	18	180	100	324
14	12	168	196	144
18	24	432	324	576
22	6	132	484	36
26	30	780	676	900
30	36	1080	900	1296
$\Sigma x_i^2 = 120$	$\Sigma y_i^2 = 126$	$\Sigma x_i y_i = 2772$	$\Sigma x_i^2 = 2680$	$\Sigma y_i^2 = 3276$

$\therefore \bar{x} = \frac{120}{6} = 20 \; ; \quad \bar{y} = \frac{126}{6} = 21.$

$$\gamma(x,y) = \frac{\frac{1}{6}(2772) - (20)(21)}{\sqrt{\frac{1}{6}(2680) - (20)^2}\sqrt{\frac{1}{6}(3276) - (21)^2}}$$

$$= \frac{462 - 420}{\sqrt{46.66} \times \sqrt{105}}$$

$$= \frac{42}{69.9994}$$

$$\boxed{\gamma(x,y) = 0.60}$$

Ex: 4 Calculate the coefficient of correlation for the following distribution.

x_i	5	9	15	19	24	28	32
y_i	7	9	14	21	23	29	30
f_i	6	9	13	20	16	11	7

Soln: Here frequency is given,
So we use the formula,

$$\gamma(x,y) = \frac{Cov(u,v)}{\sigma_u \cdot \sigma_v} = \frac{\frac{1}{N}(\Sigma f_i u_i v_i) - \bar{u}\bar{v}}{\sqrt{\frac{1}{N}\Sigma f_i u_i^2 - (\bar{u})^2}\sqrt{\frac{1}{N}(\Sigma f_i v_i^2 - (\bar{v})^2}}$$

$$= \frac{\Sigma f_i u_i v_i - N\bar{u}\bar{v}}{\sqrt{\Sigma f_i u_i^2 - N(\bar{u})^2}\sqrt{\Sigma f_i v_i^2 - N(\bar{v})^2}}$$

$$\bar{u} = \frac{\Sigma f_i u_i}{N}, \quad \bar{v} = \frac{\Sigma f_i v_i}{N}, \quad N = \Sigma f_i$$

$u_i = x_i - a$ and $v_i = y_i - b$

~~x_i y_i f_i $u_i = x_i - a$ $v_i = y_i - b$ $u_i v_i$ $f_i u_i$ $f_i v_i$ $f_i u_i^2$~~
$= x_i - 19$ $= y_i - 21$

x_i	y_i	f_i	$u_i = x_i - a$ $= x_i - 19$	$v_i = y_i - b$ $= y_i - 21$	$u_i v_i$	$f_i u_i$	$f_i v_i$	$f_i u_i^2$	$f_i v_i^2$	$f_i u_i v_i$
5	7	6	-14	-14	196	-84	-84	1176	1176	1176
9	9	9	-10	-12	120	-90	-108	900	1296	1080
15	14	13	-4	-7	28	-52	-91	208	637	364
19	21	20	0	0	0	0	0	0	0	0
24	23	16	5	2	10	80	32	400	64	160
28	29	11	9	8	72	99	88	891	704	792
32	30	7	13	9	117	91	63	1183	567	819
		Σf_i			$\Sigma f_i u_i$		$\Sigma f_i v_i$	$\Sigma f_i u_i^2$	$\Sigma f_i v_i^2$	$\Sigma f_i u_i v_i$
		$=82$			$=44$		$=-100$	$=4758$	$=4444$	$=4391$

Put these values in the formulae, we get,

$$\bar{u} = \frac{44}{82} = 0.5366, \quad \bar{v} = \frac{-100}{82} = -1.2195$$

$$\rho(x,y) = \frac{(4391) - (82)(0.5366)(-1.2195)}{\sqrt{(4758) - (82)(0.5366)^2}\sqrt{(4444) - 82(-1.2195)^2}}$$

$$= \frac{4391 - [illegible] 53.6595}{68.8069 \times 65.7423}$$

$$= \frac{4444.6595}{4523.5239}$$

$$\rho(x,y) = 0.9826$$

Page No.: YOUVA
Date:

Ex: 5. Following are the values of import of raw material and export of finished product in suitable units.

Export material	10	11	14	14	20	22	16	12	15	13
Import material	12	14	15	16	21	26	21	15	16	14

Calculate the coefficient of correlation between the import and export values.

Soln:

$$\gamma(x,y) = \frac{Cov(u,v)}{\sigma_u \sigma_v} = \frac{\Sigma u_i v_i - n\bar{u}\bar{v}}{\sqrt{\Sigma u_i^2 - n(\bar{u})^2}\sqrt{\Sigma v_i^2 - n(\bar{v})^2}}$$

$$\bar{u} = \frac{\Sigma u_i}{n}, \quad \bar{v} = \frac{\Sigma v_i}{n}, \quad n = 10$$

$$u_i = x_i - a, \quad v_i = y_i - b$$
$$= x_i - 20, \quad = y_i - 21$$

x_i	y_i	$u_i = x_i - 20$	$v_i = y_i - 21$	$u_i v_i$	u_i^2	v_i^2
10	12	-10	-9	90	100	81
11	14	-9	-7	63	81	49
14	15	-6	-6	36	36	36
14	16	-6	-5	30	36	25
20	21	0	0	0	0	0
22	26	2	5	10	4	25
16	21	-4	0	0	16	0
12	15	-8	-6	48	64	36
15	16	-5	-5	25	25	25
13	14	-7	-7	49	49	49
		$\Sigma u_i = -53$	$\Sigma v_i = 40$	$\Sigma u_i v_i = 351$	$\Sigma u_i^2 = 411$	$\Sigma v_i^2 = 326$

$$\bar{u} = \frac{-53}{10} = -5.3, \qquad \bar{v} = \frac{-40}{10} = -4$$

$$\rho(x,y) = \frac{(351) - 10(-5.3)(-4)}{\sqrt{(411) - 10(-5.3)^2}\sqrt{(326) - 10(-4)^2}}$$

$$= \frac{(351) - 212}{\sqrt{130.1 \times 166}}$$

$$= \frac{139}{146.9578}$$

$$= 0.945849.$$

Ex: 6. Find the co-efficient of correlation between x and y from the table.

x	1	3	4	6	8	9	11	16
y	1	2	4	4	5	7	8	9

Use: $$\rho(x,y) = \frac{\text{cov}(x,y)}{\sigma_x \cdot \sigma_y}$$

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

$$\rho(x,y) = 0.9770$$

Ex: 7. Calculate the correlation coefficient for the following data.

x	6	2	10	4	8
y	9	11	5	8	7

$$\rho(x,y) = -0.7905$$

$r(x, y) = -0.7905$

Exercise

1. The following marks have been obtained by a class of students in two papers of economics :

Paper I	45	55	56	58	60	65	68	70	75	80	85
Paper II	56	50	48	60	62	64	65	70	74	82	90

Calculate the coefficient of correlation for the above data. **(May 10)**

2. Obtain the correlation coefficient between population density (per square miles) and death rate (per thousand persons) from the following data : **(May 05, Dec. 10, May 12)**

Population density	200	500	400	700	300
Death rate	12	18	16	21	10

3. Calculate the coefficient of correlation from the following data :

$n = 20, \sum x_i = 40, \sum x_i^2 = 190, \sum y_i^2 = 200,$

$\sum x_i y_i = 150, \quad \sum y_i = 40$

4. Find the number of items if x and y are deviations from arithmetic mean given :

$r(x, y) = 0.9$, $\sum xy = 70$; $\sigma_y = 3.5$, $\sum x_i^2 = 100$

5. Calculate the co - efficient of correlation for the following data : **(May 14)**

x	1	2	3	4	5	6	7	8	9
y	9	8	10	12	11	13	14	16	15

Date:

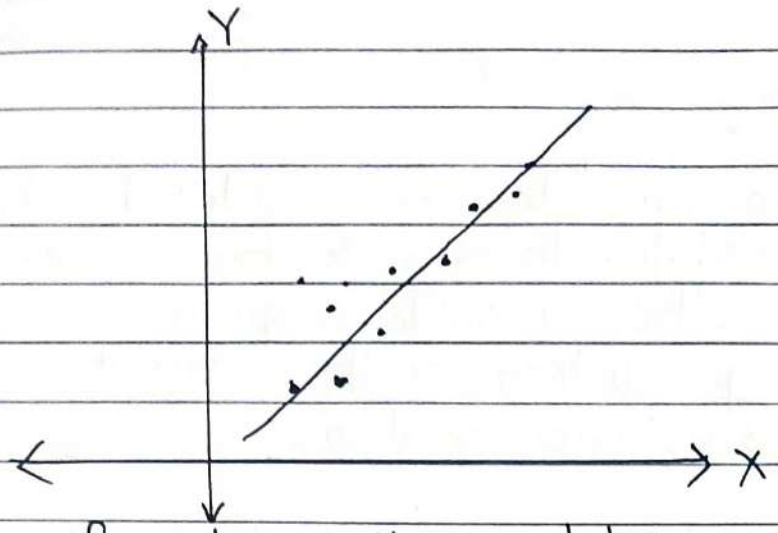
Regression :

When two variables are correlated, then we can use the correlation to estimate the one variable, if the value of other variable is given.
The method of prediction on the basis of correlation is known as regression analysis.

Note: 1. The correlation measures the linear relation between two variables, so we get linear equations of these variables. i.e we get the equation of straight line

2. By least square principle, a line which minimises a sum of squares of differences between true value and the value given by straight line is chosen.
The equation of line so obtained is called as least square regression line.

In this concept, we discuss linear regression using least square method.
This method helps us to find linear equation that best fits the data.



If we have these data points as shown the diagram.

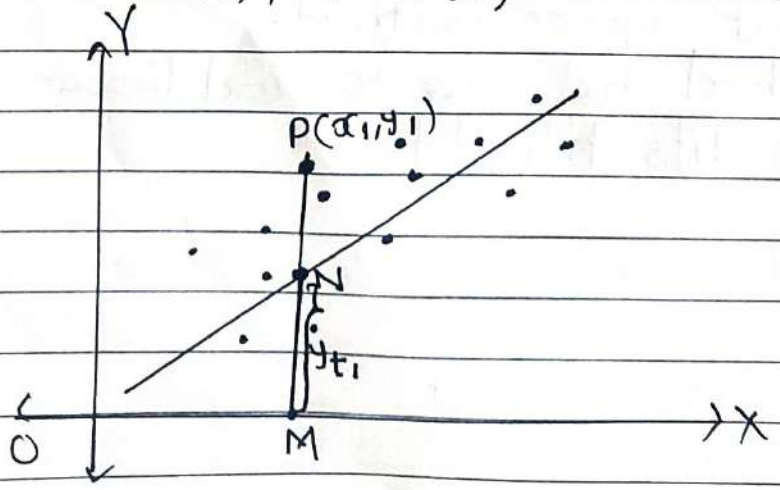
These data points don't fall on straight line. But we can find equation of straight line which best best fits the data where all data points are as close ~~to line~~ as possible to this straight line.

This straight line is found by least square method.

Method of least squares:

Let $y = a + bx$ ——— ①.

be the straight line to be fitted to given data points $(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)$



$PM = y_{t_1}$, $PM = y_1$

$PN = PM - NM = y_{t_1} - y_1$

$PN = e_1$

$\therefore e_1 = y_1 - y_{t_1}$

$= y_1 - (a + bx_1)$

$e_1^2 = (y_1 - a - bx_1)^2$

$S = e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2$

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

For S to be minimum,

$\frac{\partial S}{\partial a} = \sum_{i=1}^{n} 2(y_i - a - bx_i)(-1) = 0$

$\Rightarrow \Sigma(y_i - a - bx_i) = 0$ —— ①

or $\Sigma(y - a - bx) = 0$ —— ②

$\frac{\partial S}{\partial b} = \sum_{i=1}^{n} 2(y_i - a - bx_i)(-x_i) = 0$

$\Rightarrow \sum_{i=1}^{n} (x_i y_i - ax_i - bx_i^2) = 0$

$\Rightarrow \Sigma(xy - ax - bx^2)$ —— ③.

On simplification of ② & ③, we get,

$$\boxed{\begin{aligned} \Sigma y &= na + b\Sigma x \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 \end{aligned}} \quad \text{——} \quad ④$$

Equations in ④ are known as normal equations

On solving these equations, we get the values of a and b.
substituting these values of a and b in $y = a + bx$ we get the equation of straight line.

Note: To Remember:

Consider $y = a + bx$ ——— ①

(take Σ on both sides) $\Rightarrow \Sigma y = \Sigma a + b\Sigma x$

$\Rightarrow \Sigma y = a \cdot n + b\Sigma x$

$$\boxed{\Sigma y = na + b\Sigma x} \quad ②$$

(Now multiply eq^n ① by x & take Σ on both sides)

$$\therefore \boxed{\Sigma xy = a\Sigma x + b\Sigma x^2} \quad ③$$

Ex: 1. Find the best values of a and b so that $y = a + bx$ fits the data given in the table.

x	0	1	2	3	4
y	1·0	2·9	4·8	6·7	8·6

Soln. Equation of line: $y = a + bx$

x_i	y_i	$x_i y_i$	x_i^2
0	1.0	0	0
1	2.9	2.9	1
2	4.8	9.6	4
3	6.7	20.1	9
4	8.6	34.4	16
$\Sigma x_i = 10$	$\Sigma y_i = 24$	$\Sigma x_i y_i = 67$	$\Sigma x_i^2 = 30$

Normal equations are:

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

$$\Rightarrow \quad 24 = 5a + 10b \quad \text{——} (1)$$
$$67 = 10a + 30b \quad \text{——} (2)$$

$2 \times 5a + 10b = 24 \Rightarrow$ ~~10a +~~
$10a + 30b = 67$ ~~5a +~~
~~——~~
-2

$$10a + 20b = 48$$
$$10a + 30b = 67$$
$$-10b = -19$$
$$b = 1.9$$

put in (1)

$$5a + 10 \times 1.9 = 24$$
$$5a = 24 - 19$$
$$5a = 5$$
$$a = 1$$

Hence $y = a + bx = 1 + 1.9x$

Ex: 2 By the method of least squares, find the straight line that fits the following data.

x_i	1	2	3	4	5
y_i	14	27	40	55	68

Soln: Let the equation of straight line best fit be $y = a + bx$

x_i	y_i	$x_i y_i$	x_i^2
1	14	14	1
2	27	54	4
3	40	120	9
4	55	220	16
5	68	340	25
$\Sigma x_i = 15$	$\Sigma y_i = 204$	$\Sigma x_i y_i = 748$	$\Sigma x_i^2 = 55$

Here $n = 5$.

Normal equations are -

$$\left.\begin{aligned} \Sigma y_i &= na + b\Sigma x_i \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 \end{aligned}\right\} \text{---} (1)$$

put values from above table,

$\Rightarrow$

$$\left.\begin{aligned} 204 &= 5a + 15b \\ 748 &= 15a + 55b \end{aligned}\right\} \text{---} (2)$$

$$3 \times 5a + 15b = 204$$
$$15a + 55b = 748$$

$$15a + 45b = 612$$
$$15a + 55b = 748$$
$$- \quad - \quad -$$

$-10b = -136$

$b = 13.6$

put in $5a + 15b = 204$

$5a = 204 - 15 \times 13.6$

$= 204 - 204$

$5a = 0$

$a = 0$.

Hence the eqn of straight line is ~

$y = 13.6x$

Ex: 3 Use least square method to fit a curve of the form $y = ae^{bx}$ to the data.

x	1	2	3	4	5	6
y	7.209	5.265	3.846	2.809	2.052	1.499

Soln: $y = ae^{bx}$

taking log of both sides.

$\log_e y = \log_e a + bx \cdot \log e$

$\log_e y = \log_e a + bx$.

put $\log_e y = Y$, $\log_e a = C$

$\therefore Y = C + bx$

Normal eq^n(s) are ~

$\Sigma Y = nC + b\Sigma x$

$\Sigma xY = C\Sigma x + b\Sigma x^2$

x	y	$Y = \log_e y$	xY	x^2
1	7.209	1.97533	1.97533	1
2	5.265	1.66108	3.32216	4
3	3.846	1.34703	4.04109	9
4	2.809	1.03283	4.13132	16
5	2.052	0.71881	3.59405	25
6	1.499	0.40480	2.4288	36
$\Sigma x_i = 21$		$\Sigma Y = 7.13988$	$\Sigma xY = 19.49275$	$\Sigma x^2 = 91$

Normal eq^n(s) becomes –

$$6C + 21b = 7.13988$$
$$21C + 91b = 19.49275$$

solving –

$b = -0.3141$, $C = \log_e a = 2.28933$

$a = 9.86832$

then $y = 9.86832\, e^{-0.314x}$

To fit up the parabola :

Let $y = a + bx + cx^2$ be the equation of the parabola. —— (1)

The normal equations are

$$\Sigma y = na + b\Sigma x + c\Sigma x^2 \quad \text{——(2)}$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \quad \text{——(3)}$$
$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 \quad \text{——(4)}$$

On solving these three normal equations, we get the values of a, b and c.

After substituting values of a, b, c, we get required equation of parabola.

Ex: 1. Find the least square approximation of degree two to the data.

x	0	1	2	3	4
y	-4	-1	4	11	20

Solution: Let the equation of the polynomial be — $y = a + bx + cx^2$

x	y	xy	x^2	x^2y	x^3	x^4
0	-4	0	0	0	0	0
1	-1	-1	1	-1	1	1
2	4	8	4	16	8	16
3	11	33	9	99	27	81
4	20	80	16	320	64	256
$\Sigma x = 10$	$\Sigma y = 30$	$\Sigma xy = 120$	$\Sigma x^2 = 30$	$\Sigma x^2 y = 434$	$\Sigma x^3 = 100$	$\Sigma x^4 = 354$

Normal equations are -

$$\Sigma y = na + b\Sigma x + c\Sigma x^2 \quad \text{—②}$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \quad \text{—③}$$
$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 \quad \text{—④}$$

$$30 = 5a + 10b + 30c$$
$$120 = 10a + 30b + 100c$$
$$434 = 30a + 100b + 354c$$

On solving these equations,

$$6 = a + 2b + 6c$$
$$12 = a + 3b + 10c$$
$$217 = 15a + 50b + 177c$$

$$\begin{bmatrix} 1 & 2 & 6 \\ 1 & 3 & 10 \\ 15 & 50 & 177 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 6 \\ 12 \\ 217 \end{bmatrix}$$

$R_2 - R_1$, $R_3 - 15R_1$

$$\begin{bmatrix} 1 & 2 & 6 \\ 0 & 1 & 4 \\ 0 & 20 & 87 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 6 \\ 6 \\ 127 \end{bmatrix}$$

$50 - 15 \times 2 = 20$

$177 - 15 \times 6 = 177 - 90 = 87$

$217 - 15 \times 6 = 217 - 90 = 127$

$R_3 - 20R_2$

$$\begin{bmatrix} 1 & 2 & 6 \\ 0 & 1 & 4 \\ 0 & 0 & 7 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 6 \\ 6 \\ 7 \end{bmatrix}$$

$87 - 20 \times 4 = 87 - 80 = 7$

$127 - 20 \times 6 = 7$

$\therefore 7c = 7 \Rightarrow c = 1$

$b + 4c = 6$

$b + 4 \times 1 = 6 \Rightarrow b = 2$

$a + 2b + 6c = 6$

$a + 2\times 2 + 6\times 1 = 6$

$a = 6 - 10 = -4$

$\therefore a = -4,\ b = 2,\ c = 1$

Ex. 2. Employ the method of least squares to fit a parabola $y = a + bx + cx^2$ in the following data.

(x, y): $(-1, 2), (0, 0), (0, 1), (1, 2)$

Sol^n : Let the equation of parabola be

$$y = a + bx + cx^2 \quad \text{---} \quad (1)$$

Here $n = 4$;

x	y	x^2	x^3	x^4	xy	x^2y
-1	2	1	-1	1	-2	2
0	0	0	0	0	0	0
0	1	0	0	0	0	0
1	2	1	1	1	2	2
$\Sigma x = 0$	$\Sigma y = 5$	$\Sigma x^2 = 2$	$\Sigma x^3 = 0$	$\Sigma x^4 = 2$	$\Sigma xy = 0$	$\Sigma x^2y = 4$

Normal equations are –

$$\Sigma y = na + b\Sigma x + c\Sigma x^2$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$$
$$\Sigma x^2y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

Hence, $5 = 4a + 0\cdot b + 2c \Rightarrow 5 = 4a + 2c$

$0 = 0a + 2b + 0c \Rightarrow 0 = 2b$

$4 = 2a + 0\cdot b + 2c \Rightarrow 4 = 2a + 2c$

$b = 0$

$\Rightarrow$ ~~5~~ $4a + 2c = 5$

$2a + 2c = 4$

$2a = 1$

$a = 0.5$

Hence, $4 \times 0.5 + 2c = 5$

$2c = 5 - 2$

$2c = 3$

$c = 1.5$

Hence, $y = a + bx + cx^2$

$= 0.5 + 0x + 1.5x^2$

$\boxed{y = 0.5 + 1.5x^2}$

Ex: 3. Fit a parabola $y = ax^2 + bx + c$ to the following data taking x as independent variable:

x	1	2	3	5	7	11	13	17	19	23
y	2	3	5	7	11	13	17	19	23	29

Soln: Here, we have,

$y = ax^2 + bx + c$

Normal equations are -

$\Sigma y = na + b\Sigma x + c\Sigma x^2$

$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$

$\Sigma x^2 y = a\Sigma x + b\Sigma x^3 + c\Sigma x^4$

x	y	xy	x^2	x^2y	x^3	x^4
1	2	2	1	2	1	1
2	3	6	4	12	8	16
3	5	15	9	45	27	81
5	7	35	25	175	125	625
7	11	77	49	539	343	2401
11	13	143	121	1573	1331	14641
13	17	221	169	2873	2197	28561
17	19	323	289	5491	4913	83521
19	23	437	361	8303	6859	130321
23	29	667	529	15341	12167	279841
$\Sigma x =$	$\Sigma y =$	Σxy	Σx^2	Σx^2y	$\Sigma x^3 =$	Σx^4
101	129	$= 1926$	$= 1557$	$= 34354$	27971	$= 540009$

putting all values in normal equations.

$$129 = 10a + 101b + 1557c$$
$$1926 = 101a + 1557c + 27971c$$
$$34354 = 1557a + 27971b + 540009c$$

Solving these eq^n(s).

$$a = 1.41259297$$
$$b = 1.08901395 7$$
$$c = 0.00313658359 5$$

Hence eq^n of parabola is —

$$y = 1.41259297 + 1.08901395 7x + 0.00313658359 5x^2$$

Lines of regression of y on x:

$$(y-\bar{y}) = \frac{cov(x,y)}{\sigma_x^2}(x-\bar{x})$$

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

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

where $b_{yx} = \frac{r(x,y)\sigma_y}{\sigma_x} = \frac{cov(x,y)}{\sigma_x^2}$

= Regression coefficient of y on x

Lines of regression of x on Y.

$$(x-\bar{x}) = \frac{cov(x,y)}{\sigma_y^2}(y-\bar{y})$$

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

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

where $b_{xy} = \frac{r(x,y)\sigma_x}{\sigma_y} = \frac{cov(x,y)}{\sigma_y^2}$

= Regression coefficient of x on y

In share market, the term $\beta = b_{xy}$ (beta index) is used. If $\beta > 1$, then share is considered as aggressive / good otherwise defensive.

Note: (i) co-efficient of correlation and coefficient of regressions have same sign because all these terms have same numerator.

(ii) $r(x,y) = \sqrt{b_{xy} \cdot b_{yx}}$

i.e. co-efficient of correlation is a geometric mean of regression coefficients.

$$b_{xy} \cdot b_{yx} = \frac{cov(x,y)}{\sigma_y^2} \cdot \frac{cov(x,y)}{\sigma_x^2}$$

$$= \left[\frac{cov(x,y)}{\sigma_x \cdot \sigma_y}\right]^2$$

$$\boxed{b_{xy} \cdot b_{yx} = r^2}$$ is called the coefficient of determination.

(iii) $(b_{xy} \cdot b_{yx} = r^2) < 1$

(iv) $b_{yx} = r \frac{\sigma_y}{\sigma_x}$

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

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

& $b_{xy} = r \frac{\sigma_x}{\sigma_y}$

(v) As $r(x,y)$ is geometric mean of b_{yx} and b_{xy}, this means $r(x,y)$ lies between b_{xy} and b_{yx}.

(vi) If $r = \pm 1$, then $b_{xy} = \frac{1}{b_{yx}}$

or and $b_{yx} = \frac{1}{b_{xy}}$

(vii) If $\sigma_x = \sigma_y$ then $b_{yx} = b_{xy} = r(x,y)$

Observations:

(i) If $r = 0$, $\tan\theta \to \infty \Rightarrow \theta = \pi/2$

i.e if the variables are independent, then the two lines of regression are perpendicular to each other.

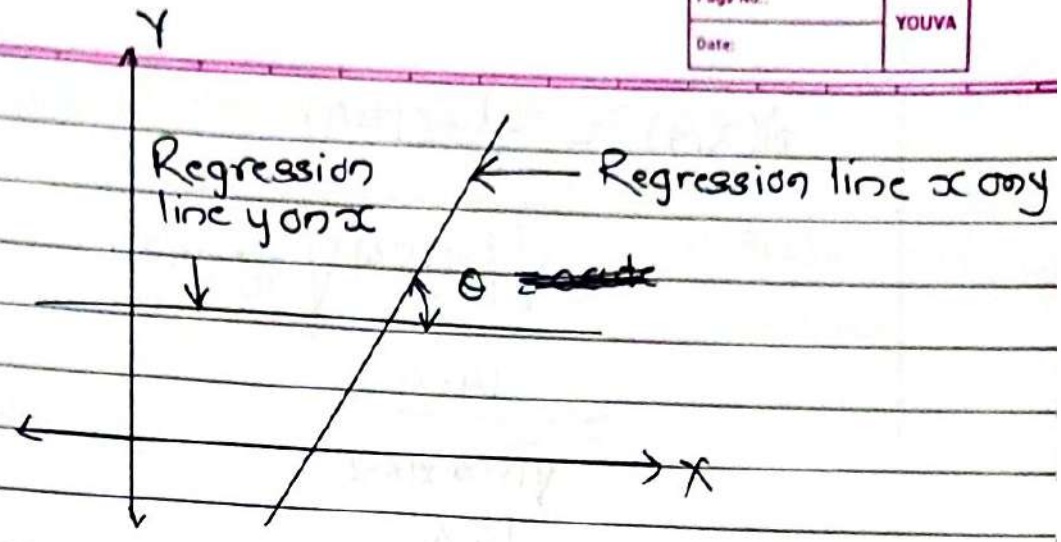
i.e variables are more spread means not correlated.

(ii) If $r = \pm 1$, then $\theta \to 0 \Rightarrow \theta = 0$ or π

i.e given two lines are coincides or parallel.

(iii) The lines of regression are intersect at $(\bar{x}, \bar{y})$.

(iv) If r^2 is larger, θ is smaller.



Ex: 1. Find the correlation coefficient from the following data

$n = 10$, $\Sigma x_i = 300$, $\Sigma y_i = 250$, $\Sigma (x_i - 30)^2 = 154$
$\Sigma (y_i - 25)^2 = 162$, $\Sigma (x_i - 30)(y_i - 50) = 144$

Also find the regression line Y on X and X on Y.

Soln: $n = 10$, $\Sigma x_i = 300$, $\Sigma y_i = 250$,

$$\Rightarrow \bar{x} = \frac{\Sigma x_i}{n} = \frac{300}{10} = 30$$

$$\& \quad \bar{y} = \frac{\Sigma y_i}{n} = \frac{250}{10} = 25$$

The correlation coefficient

$$\rho(x,y) = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n}\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\Sigma (x_i - \bar{x})^2}\sqrt{\frac{1}{n}\Sigma (y_i - \bar{y})^2}}$$

$$= \frac{\frac{1}{10}\Sigma (x_i - 30)(y_i - 25)}{\sqrt{\frac{1}{10}\Sigma (x_i - 30)^2}\sqrt{\frac{1}{10}\Sigma (y_i - 25)^2}}$$

$$r(x,y) = \frac{\frac{1}{10}(144)}{\sqrt{\frac{1}{10}(154)}\sqrt{\frac{1}{10}(162)}}$$

$$= \frac{14.4}{\sqrt{15.4 \times 16.2}}$$

$$= \frac{14.4}{15.79494}$$

$$= 0.911684$$

$$\sigma_x = \sqrt{\frac{1}{10}\Sigma(x_i - 30)^2} = \sqrt{15.4} = 3.9243$$

$$\sigma_y = \sqrt{\frac{1}{10}\Sigma(y_i - 25)^2} = \sqrt{16.4} = 4.0249$$

(i) Lines of regression of y on x is

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

$$(y - 25) = \frac{(0.911684)(4.0249)}{3.9243}(x - 30)$$

$$(y - 25) = 0.93505(x - 30)$$

$$y - 25 = 0.93505x - 0.93505 \times 30$$

$$y = 0.93505x - 3.0515$$

(ii) Lines of regressions of x on y

$$(x-\bar{x}) = r(x,y)\frac{\sigma_x}{\sigma_y}(y-\bar{y}) = \frac{(0.911684)(3.9423)}{4.0249}(y-25)$$

$$(x-30) = 0.89297\,(y-25)$$

$$x = 0.89297y + 7.67575$$

Ex: 2. The equation of two regression lines obtained in a correlation analysis are $4x-5y+33=0$ and $20x-9y-107=0$
If the variance of y is 16 then find
(i) The mean value of x and y
(ii) The correlation coefficient between x and y
(iii) The variance of x

Soln. Given: $\sigma_y^2 = 16 \Rightarrow \sigma_y = \pm 4$

Regression lines are
$4x-5y = -33$
$20x-9y = 107$

(i) Let $(\bar{x}, \bar{y})$ be the point of intersection of given regression lines.
put $x = \bar{x}$ & $y = \bar{y}$

$\therefore 4\bar{x} - 5\bar{y} = -33$ ——— (1)
$20\bar{x} - 9\bar{y} = 107$ ——— (2)

Solving these two equations.
$\bar{x} = 13, \quad \bar{y} = 17$
Mean values are $\bar{x} = 13, \ \bar{y} = 17$

Let us consider the line of regression of y on x is from eqn (1)

$$[4x - 5y = -33]$$

$$\Rightarrow y = 4/5 x + \frac{33}{5}$$

$$= 0.8x + \frac{33}{5}$$

$$\Rightarrow b_{yx} = 0.8$$

& the line of regression of x on y is from eqn (2) $20x - 9y = 107$

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

$$= 0.45y + \frac{107}{20}$$

$$\therefore b_{xy} = 0.45$$

$\therefore$ correlation co-efficient $r(x,y)$

$$r(x,y) = \sqrt{b_{yx} \cdot b_{xy}}$$

$$= \sqrt{0.8 \times 0.45}$$

$$= \sqrt{0.36}$$

$$= 0.6$$

Since $\sigma_y = 4$, As $b_{xy} = r(x,y)\frac{\sigma_x}{\sigma_y}$

$$\Rightarrow \sigma_x = \frac{b_{xy} \times \sigma_y}{r(x,y)} = \frac{0.45 \times 4}{0.6}$$

$$\sigma_x = 3$$

$\therefore$ Variance in $x = \sigma_x^2$

$$= 9$$

Ex: 3. Find the coefficient of correlation for distribution in which $\sigma_x = 4$, $\sigma_y = 1.8$ $b_{yx} = 0.32$

Soln: Given: $\sigma_x = 4$, $\sigma_y = 1.8$, $b_{yx} = 0.32$

Since $b_{yx} = r \frac{\sigma_y}{\sigma_x}$

$$r = \frac{b_{yx} \times \sigma_x}{\sigma_y}$$

$$= \frac{0.32 \times 4}{1.8}$$

$$= 0.711$$

Ex: 4. If $\bar{x} = 8.2$, $\bar{y} = 12.4$, $\sigma_x = 6.2$, $\sigma_y = 20$, $r(x,y) = 0.9$. Then find the lines of regression. Estimate the value of x for $y = 10$ and estimate y for $x = 10$.

Soln: Given $\bar{x} = 8.2$; $\bar{y} = 12.4$, $\sigma_x = 6.2$, $\sigma_y = 20$; $r(x,y) = 0.9$

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} = (0.9)\frac{(20)}{6.2} = 2.9032$$

and $b_{xy} = r \frac{\sigma_x}{\sigma_y} = (0.9)\frac{(6.2)}{20} = 0.279$

Regression lines of y on x is -

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

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

$$= 2.9032(x - \bar{x})$$

$$= 2.9032(x - 8.2)$$

$$y - 12.4 = 2.9032x - 23.80624$$

$$y = 2.903x - 11.40624$$

y at $x = 10$

$$y = 2.903 \times 10 - 11.40624$$

$$= 17.62376$$

The regression line of x on y is –

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

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

$$(x - 8.2) = 0.279(y - 12.4)$$

$$x - 8.2 = 0.279y - 3.4596$$

$$\Rightarrow x = 0.279y + 4.7404$$

x at $y = 10$

$$\Rightarrow x = (0.279) \times 10 + 4.7404$$

$$x = 7.5304$$

Ex: 5. The two regression equations of the variable x and y are —

$$x = 19.3 - 0.87y ;$$

$$y = 11.64 - 0.50x$$

Find (i) $\bar{x}, \bar{y}$

(ii) the correlation coefficient between x and y.

$\Rightarrow$ Ans: $\bar{x} = 16.2358$, $\bar{y} = 3.5222$

$r(x, y) = 0.6596$

Ex:6. Find the lines of regression for the data.

x	2	3	5	7	9	10	12	15
y	2	5	8	10	12	14	15	16

Soln: Since the regression lines are:

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

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

where, $b_{yx} = r\frac{\sigma_y}{\sigma_x}$, $b_{xy} = r\frac{\sigma_x}{\sigma_y}$

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

$$\sigma_x^2 = \frac{1}{n}(\Sigma x_i^2) - (\bar{x})^2, \quad \sigma_y^2 = \frac{1}{n}(\Sigma y_i^2) - (\bar{y})^2$$

$$\bar{x} = \frac{\Sigma x_i}{n}, \quad \bar{y} = \frac{\Sigma y_i}{n}$$

$$cov(x,y) = \frac{1}{n}\Sigma(x_i y_i) - \bar{x}\bar{y}, \quad n=8$$

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	2	2	4	4	4
	3	5	9	25	15
	5	8	25	64	40
	7	10	49	100	70
	9	12	81	144	108
	10	14	100	196	140
	12	15	144	225	180
	15	16	225	256	240
	$\Sigma x_i =$	$\Sigma y_i =$	$\Sigma x_i^2 =$	$\Sigma y_i^2 =$	$\Sigma x_i y_i =$
	63	82	637	1014	797

Hence, $\bar{x} = \frac{63}{8} = 7.825$, $\bar{y} = \frac{82}{8} = 10.25$

$$\sigma_x^2 = \frac{637}{8} - (7.825)^2 = 18.3944$$

$$\sigma_y^2 = \frac{1014}{8} - (10.25)^2 = 21.6875$$

$$\text{Cov}(x,y) = \frac{797}{8} - (7.825)(10.25)$$

$$= 19.4188$$

$$\therefore b_{xy} = \frac{19.4188}{21.6875} = 0.8954$$

$$b_{yx} = \frac{19.4188}{18.3944} = 1.0557$$

Hence,

(i) Line of regression of y on x is —

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

$$y - 10.25 = (1.0557)(x - 7.825)$$

(ii) Line of regression of x on y is

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

$$x - 7.825 = (0.8954)(y - 10.25)$$

Ex: 6. Obtain the regression line for the following data:

x	2	3	5	7	9	10
y	2	5	8	10	12	14