UNIT - I :

Measures of Central Tendency:

Class:

Each stated interval such as 0-10, 10-20 .... etc., is called a class.

Class limit:

There are two limits of every class, they are lower limit and upper limit

Class Interval / class width:

The difference between the boundaries is known as Class Interval. i.e., $(U-L)$ or (L_2-L_1) i.e., upper limit - lower limit

Mid values:

The values lying in the middle / centre of the two class limits of a group or a class is known as mid value.

i.e., $\frac{L_1+L_2}{2}$

Class frequency:

The number of observations to a particular class is known as class frequency.

Types of series:

There are 3 types of series:

1) Individual series :

Every observation is independent or separate is called Individual series.

Ex: 1, 2, 3, 4, ...

2) Discrete series:

The series dealing with the discrete

Discrete series means where frequencies are given but variable is without class interval.

Ex:

Marks (x)	No. of students (f)
10	2
20	3
30	4
40	5

3) Continuous series:

There are 5 types of continuous series:

i) Inclusive series:

In this inclusive series first class interval of its upper limit is included and immediate next class interval of its lower limit is also included.

Ex:

Marks (x)	No. of students (f)
10-19	2
20-29	3
30-39	4
40-49	5

ii) Exclusive series:

In this exclusive series first class interval of its upper limit is excluded and immediate next class interval of its lower limit is included

Ex:

Marks (x)	No. of students (f)
10-20	2
20-30	3
30-40	4
40-50	5

iii) Open-end series:

In this open-end series first class interval of its lower limit is not given and last class interval (highest class interval) of its upper limit is not given

Ex:

Marks (x)	No. of students (f)
below 20	2
20-30	3
30-40	4
Above 40	5

iv) Mid-value series:

In this series, difference between each mid point and dividing with two, therefore reducting and adding to each midpoint we get the interval.

Ex:

Midpoint		frequency	class Interval
6	} 10-6	2	(6-2 - 6+2)
10	=4	7	4 - 8
14	$\frac{4}{2}=2$	18	8 - 12
18		29	12 - 16
22		17	16 - 20
			20 - 24

v) Cumulative frequency series:

It is of 2 types:

i) less than cumulative frequency series:

(Lcf)

Ex:

C-I	f	less than C.F	C-I
below 40	3	3	30-40
" 50	11	11-3=8	40-50
" 60	34	34-11=23	50-60
" 70	59	59-34=25	60-70
" 80	72	72-59=13	70-80
" 90	80	80-72=8	80-90

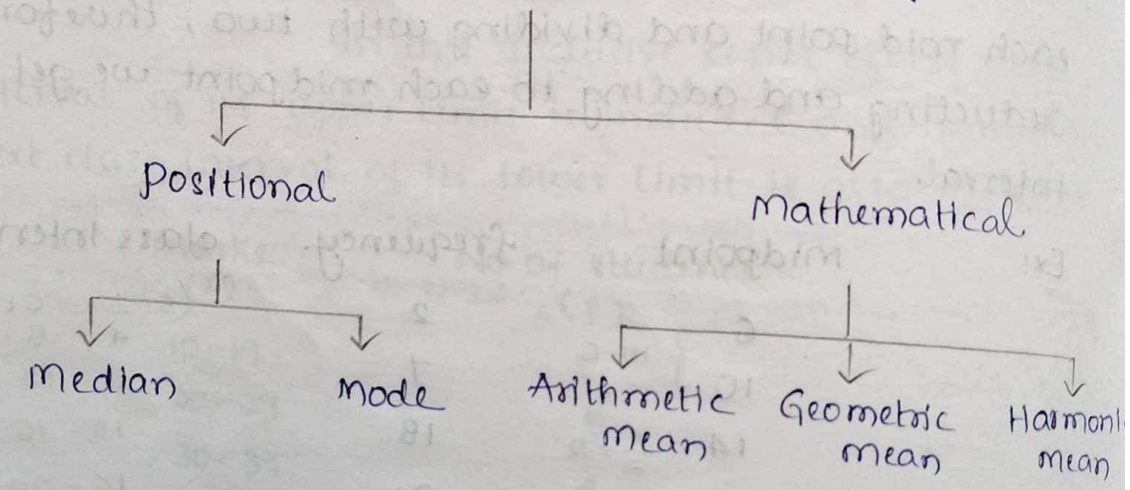
2) More than Cumulative frequency Series:

Ex:

C.I	f	more than C.F	C.I
Above 40	80	80-77 = 3	
" 50	77	77-69 = 8	
" 60	69	69-46 = 23	
" 70	46	46-21 = 25	
" 80	21	21-8 = 13	
" 90	8	8-0 = 8	
		N = 80	

Types of Averages:

There are 2 types of Averages:



Arithmetic mean:

It is denoted by $\bar{x}$

$$\therefore \bar{x} = \frac{\text{Sum of observations}}{\text{No of observations}}$$

- Individual series:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad ; n \rightarrow \text{no. of observations}$$

- Discrete series:

$$\bar{x} = \frac{\Sigma fx}{N} \quad ; N \rightarrow \text{Sum of frequencies}$$

Problems:

1) Find mean for the following data:

30, 41, 47, 54, 23, 34, 37, 51, 53, 47

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{417}{10} = 41.7$$

2) Calculate arithmetic mean from the following data:

Marks (x)	No. of students (f)	fx
4	6	24
8	12	96
12	18	216
16	15	240
20	9	180
	N=60	Σfx=756

$$\bar{x} = \frac{\Sigma fx}{N} = \frac{756}{60} = 12.6$$

- continuous series:

$$\bar{x} = \frac{\Sigma fm}{N}$$; m → mid value, N → sum of frequencies

3) Find average income of:

Exclusive series

Income	No. of persons (f)	m	fm
10-20	4	15	60
20-30	7	25	175
30-40	16	35	560
40-50	20	45	900
50-60	15	55	825
60-70	8	65	520
	N=70		Σfm=3040

$$\bar{x} = \frac{\Sigma fm}{N} = \frac{3040}{70} = 43.42$$

4) Find arithmetic mean for the following data:

Inclusive series	C.I	frequency	C.I	m	fxm
	4-6	1	3.5-6.5	5	5
	7-9	3	6.5-9.5	8	24
$\frac{7-6}{2}=0.5$	10-12	7	9.5-12.5	11	77
	13-15	15	12.5-15.5	14	210
$\frac{10-9}{2}=0.5$	16-18	11	15.5-18.5	17	187
	19-21	3	18.5-21.5	20	60
	22-24	2	21.5-24.5	23	46
		N=42			$\Sigma fm = 609$

$$\bar{x} = \frac{\Sigma fm}{N} = \frac{609}{42} = 14.5$$

Median:

Median is a positional measure. The position which is exactly in the centre, equal number of terms lie on either side of it, when terms are arranged in ascending order / descending order.

- Individual series:

case(i): when no. of obs'ns are odd

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

case(ii): when no. of obs'ns are even

Median = Avg. of size of $\left(\frac{n}{2}\right)^{th}$ & $\left(\frac{n}{2}+1\right)^{th}$ obs

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

Problems:

1) Find median of 4, 8, 16, 20, 2, 7, 13

2, 4, 7, 8, 13, 16, 20

$n = 7$

Median = Size of

2) Find median of 4, 8, 16, 20, 2, 7

2, 4, 7, 8, 16, 20 ; $n = 6$

$$\text{Median} = \frac{\text{size of } \left(\frac{6}{2}\right)^{th} + \left(\frac{6}{2}+1\right)^{th} \text{ obs}}{2} = \frac{\text{size of } 3^{rd} + 4^{th} \text{ obs}}{2}$$

$$= \frac{7+8}{2}$$

$$= 7.5$$

- Discrete series:

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

3) Compute median from the following data:

	X	f	Cf	
	10	4	4	
	20	7	11	
	30	21	32	
Median ←	40	34	66	→ Median class
	50	25	91	
	60	12	103	
	70	3	106	
		N = 106		

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

$$= \left(\frac{106+1}{2}\right)^{th} \text{ obs.}$$

$$= (53.5)^{th} \text{ obs.}$$

$\therefore$ median = 40

- Continuous series:

$$\text{Median} = L + \frac{\frac{N}{2} - Cf}{f} \times i ;$$

where, L → lower limit of the median class

Cf → cumulative frequency preceding the median class

f → frequency corresponding to the median class

N → Total of frequency.

i → common factor of median class. (upper limit - lower limit)

4) Find median from the following data:

Marks (C.I)	No. of students (f)	C.f	
0-10	20	20	
10-20	45	65	
20-30	85	150	
30-40	160	310	→ median class
40-50	70	380	
50-60	55	435	
60-70	35	470	
70-80	30	500	
	N = 500		

$\frac{N}{2} = 250$;

$Cf = 150$; $f = 160$; $L = 30$; $i = 10$

$$\text{Median} = 30 + \frac{250 - 150}{160} \times 10 = 30 + \frac{100}{16} = 36.25$$

Mode:

Mode is the most repeated frequency or value.

- Individual series:

Problems:

1) Find mode : 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 5

Mode = 5

- Discrete series:

2) Find mode from the following data:

X	f	
4	3	
7	9	
11	14	
16	21	→ Max. frequency
25	13	

Mode ←

∴ Mode = 16

3) Find mode from the following data:

Grouping Table

X	f	I	II (2)	III 1/2	IV 3	V 1/3	VI 2/3
5	1		4	leave 1st		leave 1st	leave 1st & 2nd
10	3			7	8		
15	4					16	
20	9			13			24
25	11		23	20	32		
30	12 (max)			15		26	17
35	3			5			
40	2				7		
45	2			4			

Add 2 | Add 3

Analysis Table

X	I	II	III	IV	V	VI	Total
5							
10							
15						✓	1
20			✓	✓		✓	3
25		✓	✓	✓	✓	✓	5
30	✓	✓		✓	✓		4
35					✓		1
40							
45							

∴ Mode = 25

- Continuous series:

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

where, L → lower limit corresponding to the modal class

f_1 → modal class - frequency

f_0 → preceeding the modal class "

f_2 → succeeding the modal class "

i → common factor corresponding to the modal class

4) Calculate mode for the following data:

Grouping table X	(I) f	Add 2 (II)	(1/2) (III)	Add 3 (IV)	(1/3) (V)	(2/3) (VI)
			leave 1st		leave 1st	leave 1st & 2nd, add 3
0-10	4					
		17				
10-20	13			38		
			34			
20-30	21				78	
		65				
30-40	(44)					98
			77			
40-50	33			99		
		55				
50-60	22					
			29		62	
60-70	7					

Analysis table

X	I	II	III	IV	V	VI	Total
0-10							
10-20					✓		1
20-30		✓			✓	✓	3
Modal class 30-40	✓	✓	✓	✓	✓	✓	6
40-50			✓	✓		✓	3
50-60				✓			1
60-70							

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$$= 30 + \frac{44 - 21}{2 \times 44 - 21 - 33} \times 10$$

$$= 30 + \frac{230}{34}$$

$$= 30 + 6.76$$

$\therefore z = \text{Mode} = 36.76$

Standard deviation: (σ)

{Introduced by Karl Pearson in 1893}

The idea of standard deviation was introduced by Karl Pearson in 1893. This measure is widely used for studying dispersion.

Standard deviation means the square root of the mean of the squared deviations from the actual mean

- individual series:

$$\sigma = \sqrt{\frac{\Sigma x_i^2}{n}} \quad ; \quad x_i = X_i - \bar{X} \quad ; \quad \bar{X} \rightarrow \text{sample mean}$$

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

(direct method)

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2} \quad ; \quad d = X - A \quad ; \quad A \rightarrow \text{Assumed mean}$$

(short cut method)

Problems:

1. Find standard deviation by using direct and short cut methods:

25, 27, 31, 32, 35

direct method

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{25+27+31+32+35}{5} = \frac{150}{5} = 30$$

X	$X_i - \bar{X} + 30$	$(x_i - \bar{X})^2$
25	-5	25
27	-3	9
31	1	1
32	2	4
35	5	25
		$\Sigma(x_i - \bar{X})^2 = 64$

$$\sigma = \sqrt{\frac{64}{5}}$$

$$\therefore \sigma = 3.56$$

Short cut method

X	$d = X - A \rightarrow 31$	d^2
25	-6	36
27	-4	16
31	0	0
32	1	1
35	4	16
	$\Sigma d = -5$	$\Sigma d^2 = 69$

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$$

$$= \sqrt{\frac{69}{5} - \left(\frac{-5}{5}\right)^2}$$

$$= \sqrt{13.8 - 1}$$

$$= \sqrt{12.8}$$

$$\sigma = 3.56$$

- Discrete series:

Direct Method:

$$\sigma = \sqrt{\frac{\Sigma f x^2}{N}} \quad ; \quad x = X - \bar{X} \quad \text{where } \bar{X} = \frac{\Sigma f x}{N}$$

Shortcut method:

$$\sigma = \sqrt{\frac{\Sigma f d^2}{N} - \left(\frac{\Sigma f d}{N}\right)^2} \quad ; \quad d = X - A \quad \text{where } A \rightarrow \text{Assumed mean}$$

2. Calculate standard deviation from the following data:

Direct method

X	f	fx	$X - \bar{X} \rightarrow 8.47$	$(X - \bar{X})^2$	$f(X - \bar{X})^2$
2	5	10	-6.47	41.8609	209.3045
4	15	60	-4.47	19.9809	299.7135
6	20	120	-2.47	6.1009	122.018
8	25	200	-0.47	0.2209	5.5225
10	25	250	1.53	2.3409	58.5225
12	20	240	3.53	12.4609	249.218
15	8	120	6.53	42.6409	341.1272
	$N = 118$	$\Sigma fx = 1000$			$\Sigma f(X - \bar{X})^2 = 1285.4262$

$$\bar{X} = \frac{\Sigma f x}{N} = \frac{1000}{118} = 8.47$$

$$\sigma = \sqrt{\frac{\Sigma f x^2}{N}} = \sqrt{\frac{1285.4262}{118}} = \sqrt{10.8934} = 3.3005$$

Short cut method

x	f	d = x − A	d^2	fd	fd^2
2	5	−6	36	−30	180
4	15	−4	16	−60	240
6	20	−2	4	−40	80
(8) A	25	0	0	0	0
10	25	2	4	50	100
12	20	4	16	80	320
15	8	7	49	56	392
	N = 118			Σfd = 56	$\Sigma fd^2 = 1312$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$$

$$= \sqrt{\frac{1312}{118} - \left(\frac{56}{118}\right)^2}$$

$$= \sqrt{11.11 - (0.47)^2}$$

$$= \sqrt{11.11 - 0.2209}$$

$$= \sqrt{10.8891}$$

$$= 3.2998$$

Variance:

Variance is the square of standard deviation. If a phenomenon is affected by a number of variables, variance helps in isolating the effects of different factors. This term was used by R.A. Fisher in 1913

$$\text{Variance} = (\sigma)^2 = (S.D)^2$$

- **Continuous series:**

Direct method:

$$\sigma = \sqrt{\frac{\Sigma f(m - \bar{x})^2}{N}}$$

Step-deviation method: (short cut method)

$$\sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times m \quad m = A$$

(if class intervals are not equal) $\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$

3. Compute standard deviation and variance from the following data:

Short cut method

Assumed mean ↑ = 5

C.I	f	m	$d' = \frac{m-A}{C}$	d'^2	fd'	fd'^2
0-10	1	5	0	0	0	0
10-20	4	15	1	1	4	4
20-30	17	25	2	4	34	68
30-40	45	35	3	9	135	405
40-50	26	45	4	16	104	416
50-60	5	55	5	25	25	125
60-70	2	65	6	36	12	72
	$N = 100$				$\Sigma fd' = 314$	$\Sigma fd'^2 = 1090$

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times C$$

$$= \sqrt{\frac{1090}{100} - \left(\frac{314}{100}\right)^2} \times 10$$

$$= \sqrt{10.9 - 9.85} \times 10$$

$$= \sqrt{1.05} \times 10$$

$$= 1.0246 \times 10$$

$\therefore \sigma = 10.246$ & $\sigma^2 = 104.04$

Mean

$$\bar{X} = A + \frac{\Sigma fd'}{N} \times C$$

$$= 5 + \frac{314}{100} \times 10$$

$$= 36.4$$

Direct Method

C.I	f	m	fm	$m-\bar{x}$	$(m-\bar{x})^2$	$f(m-\bar{x})^2$
0-10	1	5	5	-31.4	985.96	985.96
10-20	4	15	60	-21.4	457.96	1831.84
20-30	17	25	425	-11.4	129.96	2209.32
30-40	45	35	1575	-1.4	1.96	88.2
40-50	26	45	1170	8.6	73.96	1922.96
50-60	5	55	275	18.6	345.96	1729.8
60-70	2	65	130	28.6	817.96	1635.92
	$N=100$		$\Sigma fm=3640$			$\Sigma f(m-\bar{x})^2=10404$

$$\bar{x} = \frac{\Sigma fm}{N} = \frac{3640}{100} = 36.4$$

$$\sigma = \sqrt{\frac{\Sigma f(m-\bar{x})^2}{N}}$$

$$= \sqrt{\frac{10404}{100}}$$

$$= \sqrt{104.04}$$

$$\therefore \sigma = 10.2$$

Moments:

There are 2 types of Moments:

i) Non-Central Moments / Raw Moments / Moments about a point

ii) Central Moments / Moments about mean

Non-central Moments:

These moments are calculated from any value 'A' (or) any point 'A'. The r^{th} non-central moments is denoted by μ_r'

i.e., $\mu_r' = E(X-A)^r$ (or) $\frac{1}{N}\Sigma f_i (x_i - A)^r$ [Discrete series]

$\mu_r' = E(X-A)^r = \frac{1}{n}\Sigma (x_i - A)^r$ [Individual series]

If $r=0$,

$\mu_0' = \frac{1}{N}\Sigma f_i (x_i - A)^0$ ($\Sigma f_i = N$, $(x_i - A)^0 = 1$)

$\therefore \mu_0' = 1$

If $r=1$,

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

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

$= \bar{x} - A$

$\therefore \mu_1' = \bar{x} - A$

Hy, $r = 2, 3, 4 \ldots$ we get 1st 4 Non-Central Moments

$\mu_2' = \frac{1}{N}\Sigma f_i (x_i - A)^2$ (or) $\mu_2' = \frac{\Sigma f_i d_i^2}{N}$, $d_i = x_i - A$

$\mu_3' = \frac{1}{N}\Sigma f_i (x_i - A)^3$ (or) $\mu_3' = \frac{\Sigma f_i d_i^3}{N}$

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

Central Moments:

These moments are calculated from actual mean $\bar{x}$. The r^{th} central moments is defined as

i.e., $\mu_r = E(X-\bar{x})^r$ (or) $\frac{1}{N}\Sigma f_i(x_i-\bar{x})^r$ [Discrete series]

$\mu_r = E(X-\bar{x})^r = \frac{1}{n}\Sigma(x_i-\bar{x})^r$ [Individual series]

If $r=0$,

$$\mu_0 = \frac{1}{N}\Sigma f_i (x_i-\bar{x})^0$$

$$\therefore \mu_0 = 1$$

If $r=1$,

$$\mu_1 = \frac{1}{N}\Sigma f_i (x_i-\bar{x})^1$$

$$= \frac{\Sigma f_i x_i}{N} - \frac{1}{N}\Sigma f_i \bar{x}$$

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

$$\therefore \mu_1 = 0$$

Wkg $r = 2,3,4$ we get 1st 4 central moments

$$\mu_2 = \frac{1}{N}\Sigma f_i(x_i-\bar{x})^2 \text{ (or) } \mu_2 = \frac{\Sigma f_i d_i^2}{N}; \; d_i = x_i - \bar{x}$$

$$\mu_3 = \frac{1}{N}\Sigma f_i (x_i-\bar{x})^3 \text{ (or) } \mu_3 = \frac{\Sigma f_i d_i^3}{N}$$

$$\mu_4 = \frac{1}{N}\Sigma f_i (x_i-\bar{x})^4$$

Properties of Moments (Central Moments):

1. The first moment about mean is always zero
   i.e., $\mu_1 = 0$
2. The second moment about mean indicates variance
   i.e., $\mu_2 = \sigma^2$
3. The third moment about mean measures of skewness of a given distribution.
4. The fourth moment about mean measures Kurtosis of a frequency distribution.

Relation between central and Non-central Moments:

Central Moments:

$\mu_0 = 1$

$\mu_1 = 0$

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

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

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

Non-central Moments:

$\mu_0' = 1$

$\mu_1' = \bar{x} - A$

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

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

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

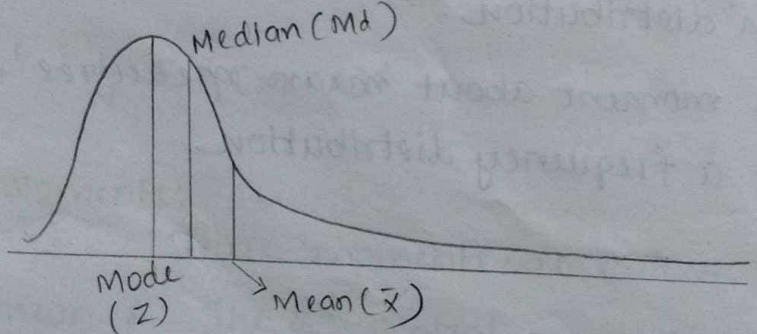
Skewness:

Skewness is a measure of symmetry (or more precisely the lack of symmetry.

We study skewness to have an idea about the shape of the curve which we can draw with the help of given data

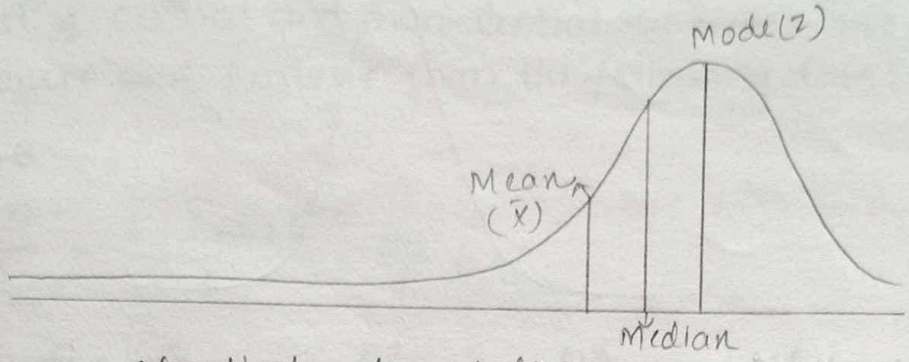
The third moment about mean measure of skewness of given frequency distribution

1. If $\mu_3 > 0$, that is if the frequency curve has longer tail to right, then the distribution is positively skewed.



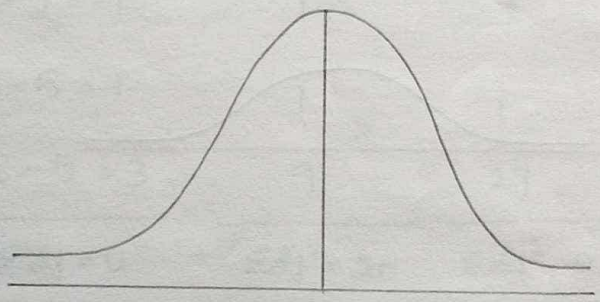
$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$

2. If $\mu_3 < 0$, i.e., if the frequency curve has longer tail to left, then the distribution is negatively skewed.



Negatively skewed distribution

3. If $\mu_3 = 0$, then the distribution is symmetrical.



Kurtosis:

Kurtosis measures the relative peakedness or flatedness of the distribution.

Coefficient of kurtosis based on second and fourth central moments, and it is denoted by β_2

limits are -3 to $+3$

$$\therefore \beta_2 = \frac{\mu_4}{\mu_2^2} \quad (\text{or}) \quad \gamma_2 = \beta_2 - 3$$

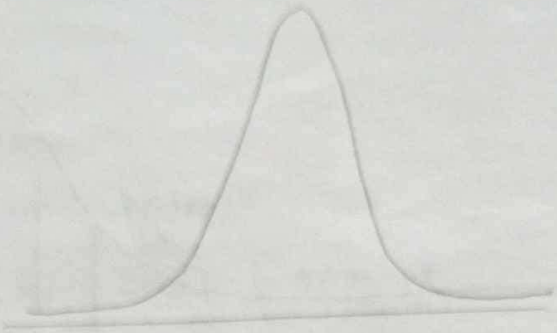
(γ gamma 2)

Meso Kurtic (or) Normal Curve:

The curve is Normal or Symmetrical is called Meso kurtic

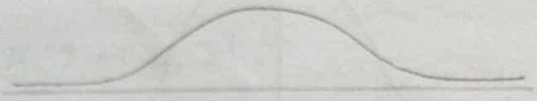
Lepto Kurtic Curve:

A curve which is more peaked than the normal (or) Meso Kurtic curve, then it is called Lepto Kurtic curve.



Platy Kurtic Curve:

The curve which is less peaked than the normal (or) Meso Kurtic curve, is called Platy Kurtic Curve



Note:

1) If $\beta_2 > 3$, it is Lepto Kurtic

2) If $\beta_2 < 3$, it is Platy Kurtic

3) If $\beta_2 = 3$, it is Meso Kurtic / Normal curve

Coefficient of Skewness:

The coefficient of Skewness based on second and third central moments, and it is denoted by β_1

$$\therefore \beta_1 = \frac{\mu_3^2}{\mu_2^3} \quad (\text{or}) \quad \gamma_1 = \pm\sqrt{\beta_1}$$

- The sign of Skewness depends upon the sign of μ_3

Note:

1) If $\beta_1 > 0$ / $\gamma_1 > 0$, then the distribution is positively skewed.

3) If $\beta_1 = 0 / \gamma_1 = 0$, then the distribution is normal or symmetrical

Problems:

1. Find first 4 central and non-central moments and also find skewness and kurtosis from the following data:

2, 4, 6, 8

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2+4+6+8}{4} = \frac{20}{4} = 5$$

X	$d_i = X - \bar{X}$ (5)	d_i^2	d_i^3	d_i^4
2	2-5 = -3	9	-27	81
4	4-5 = -1	1	-1	1
6	6-5 = 1	1	1	1
8	8-5 = 3	9	27	81
	$\Sigma d_i = 0$	$\Sigma d_i^2 = 20$	$\Sigma d_i^3 = 0$	$\Sigma d_i^4 = 164$

Central Moments:

$$\mu_r = \frac{1}{n} \Sigma (x_i - \bar{x})^r$$

if $r=1$, $\mu_1 = \frac{1}{n} \Sigma d_i^1 = 0$

if $r=2$, $\mu_2 = \frac{1}{n} \Sigma d_i^2 = \frac{20}{4} = 5$

if $r=3$, $\mu_3 = \frac{1}{n} \Sigma d_i^3 = 0$

if $r=4$, $\mu_4 = \frac{1}{n} \Sigma d_i^4 = \frac{164}{4} = 41$

Skewness $= \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0)^2}{(5)^3} = 0$

$\boxed{\beta_1 = 0}$ → Normal

Kurtosis $= \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{41}{25} = 1.64$

$\boxed{\beta_2 = 1.64}$ → Platy kurtic

x_i	$d_i = x_i - A = 4$	d_i^2	d_i^3	d_i^4
2	$2-4 = -2$	4	-8	16
4	$4-4 = 0$	0	0	0
6	$6-4 = 2$	4	8	16
8	$8-4 = 4$	16	64	256
	$\Sigma d_i = 4$	$\Sigma d_i^2 = 24$	$\Sigma d_i^3 = 64$	$\Sigma d_i^4 = 288$

Non-central Moments:

$$\mu_r' = \frac{1}{n} \Sigma (x_i - A)^r$$

If $r=1$, $\mu_1' = \frac{1}{4}(4) = 1$

If $r=2$, $\mu_2' = \frac{1}{4}(24) = 6$

If $r=3$, $\mu_3' = \frac{1}{4}(64) = 16$

If $r=4$, $\mu_4' = \frac{1}{4}(288) = 72$

$\mu_1 = 0$

$\mu_2 = \mu_2' - (\mu_1')^2 = 6 - 1 = 5$

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

$= 16 - 3(6)(1) + 2(1)^3$

$= 0$

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

$= 72 - 4(16)(1) + 6(6)(1)^2 - 3(1)^4$

$= 41$

$$\text{Skewness} = \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0)^2}{(5)^3} = 0$$

$$\text{Kurtosis} = \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{41}{25} = 1.64$$

2. Calculate first 4 non-central moments for the following data: Also find coefficient of skewness & Kurtosis

x	f	$d_i = x - A = 4$	d_i^2	d_i^3	d_i^4	fd_i	fd_i^2	fd_i^3	fd_i^4
0	1	-4	16	-64	256	-4	16	-64	256
1	8	-3	9	-27	81	-24	72	-216	648
2	28	-2	4	-8	16	-56	112	-224	448
3	56	-1	1	-1	1	-56	56	-56	56
4	70	0	0	0	0	0	0	0	0
5	56	1	1	1	1	56	56	56	56
6	28	2	4	8	16	56	112	224	448
7	8	3	9	27	81	24	72	216	648

$\mu_r' = \frac{1}{N} \Sigma f_i (x_i - A)^r \quad ; x_i - A = d_i$

if $r=1$, $\mu_1' = \frac{1}{N} \Sigma f_i d_i' = 0$

if $r=2$, $\mu_2' = \frac{1}{N} \Sigma f_i d_i^2 = \frac{512}{256} = 2$

if $r=3$, $\mu_3' = \frac{1}{N} \Sigma f_i d_i^3 = 0$

if $r=4$, $\mu_4' = \frac{1}{N} \Sigma f_i d_i^4 = \frac{2816}{256} = 11$

$\mu_1 = 0$

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

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = 0 - 3(4)(0) + 2(0)^3 = 0$

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

$= 11 - 4(0)(0) + 6(2)(0)^2 - 3(0)^4 = 11$

Coefficient of skewness $= \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0)^2}{(2)^3} = 0$ → Normal

Kurtosis $= \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{11}{4} = 2.75$ → platy kurtic

3. Find skewness and kurtosis from the following data:

C-I	f	M_i	$d_i = M_i - A$	d_i^2	d_i^3	d_i^4	fd_i	fd_i^2	fd_i^3	fd_i^4
10-20	4	15	-20	400	-8000	160000	-80	1600	-32000	640000
20-30	7	25	-10	100	-1000	10000	-70	700	-7000	70000
30-40	16	35 A	0	0	0	0	0	0	0	0
40-50	20	45	10	100	1000	10000	200	2000	20000	200000
50-60	15	55	20	400	8000	160000	300	6000	120000	2400000
60-70	8	65	30	900	27000	810000	240	7200	216000	6480000
	N=70		30	1900	27000	1150000	590	17500		

$\mu_r' = \frac{1}{N} \Sigma f_i d_i^r$

If $r=1$, $\mu_1' = \frac{1}{70} \times 590 = 8.42$

If $r=2$, $\mu_2' = \frac{1}{70} \times 17500 = 250$

If $r=3$, $\mu_3' = \frac{1}{70} \times 317000 = 4528.57$

If $r=4$, $\mu_4' = \frac{1}{70} \times 9790000 = 139857.14$

$\mu_1 = 0$

$\mu_2 = \mu_2' - (\mu_1')^2 = 250 - (8.42)^2 = 250 - 70.89 = 179.11$

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

$= 4528.57 - 3(250)(8.42) + 2(8.42)^3$

$= 4528.57 - 6315 + 1193.89$

$= -592.54$

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

$= 139857.14 - 4(4528.57)(8.42) + 6(250)(8.42)^2 - 3(8.42)^4$

$= 139857.14 - 152522.23 + 106344.6 - 15078.89$

$= 78600.62$

$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{351103.65}{5745919.02} = 0.06 \rightarrow$ positively skewed

$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{78600.62}{32080.39} = 2.45 \rightarrow$ platy kurtosis

Curve fitting:

Suppose that a data is given in the two variables x and y, the problem of finding an analytical expression of the form $y = f(x)$ which fits the given data is called Curve fitting

(or)

Curve fitting means an exact relationship between two variables by algebraic equations. This relationship is equation of the curve. Curve fitting means to form an equation of the curve from the given data.

Method of Least Square:

principle

The sum of the squares of the deviation between observed values and expected values should be minimum is called Residual Error (or) Method of least square.

$$E = \Sigma [y - f(x)]^2 \text{ is minimum}$$

- Some curve fitting equations are given below:
  - Straight line : $y = a + bx$
  - parabola : $y = a + bx + cx^2$
  - power curve : $y = ax^b$
  - exponential curve : $y = a \cdot b^x$ (or) $y = a \cdot e^{bx}$

Straight line equation:

Let the straight line : $y = a + bx$ — ①

Taking summation on both sides of ①, we get

$$\Sigma y = na + b\Sigma x \text{ — ②}$$

Taking 'x' on both sides of ①, we get

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \text{ — ③}$$

Solve eq'n ② & ③, we get a & b

Problems:

1. Fit a straight line for the following data:

x	y	x^2	xy
1	1	1	1
3	2	9	6
4	4	16	16
6	4	36	24
8	5	64	40
9	7	81	63
11	8	121	88
14	9	196	126

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

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

$\Rightarrow 40 = 8a + 56b$

$364 = 56a + 524b$

$a = 0.5454,$

$b = 0.6363.$

Straight line is

2. Fit a straight line to the following data by using equation: $y = a_0 + a_1 x$

x	y	x^2	xy
0	-1	0	0
2	5	4	10
5	12	25	60
7	20	49	140
$\Sigma x = 14$	$\Sigma y = 36$	$\Sigma x^2 = 78$	$\Sigma xy = 210$

$\Sigma y = n a_0 + a_1 \Sigma x$

$\Sigma xy = a_0 \Sigma x + a_1 \Sigma x^2$

$36 = 4a_0 + a_1(14)$

$210 = a_0(14) + a_1(78)$

$a_0 = -1.137$

$a_1 = 2.89$

$y = a_0 + a_1 x$

$y = -1.137 + (2.89)x$

Fitting of a second degree parabola:

Let the second degree parabola equation:

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

* Normal equations:

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

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

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

Problems:

1. Fit a second degree parabola for the following data:

x	y	x^2	x^3	x^4	xy	x^2y
0	1	0	0	0	0	0
1	1.8	1	1	1	1.8	1.8
2	1.3	4	8	16	2.6	5.2
3	2.5	9	27	81	7.5	22.5
4	6.3	16	64	256	25.2	100.8
$\Sigma x = 10$	$\Sigma y = 12.9$	$\Sigma x^2 = 30$	$\Sigma x^3 = 100$	$\Sigma x^4 = 354$	$\Sigma xy = 37.1$	$\Sigma x^2 y = 130.3$

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

$$12.9 = 5a + b(10) + c(30)$$
$$\Rightarrow \quad 37.1 = a(10) + b(30) + c(100)$$
$$130.3 = a(30) + b(100) + c(354)$$

$a = 1.42$
$b = -1.07$
$c = 0.55$

Second degree parabola eq'n is:

$$y = a + bx + cx^2$$
$$y = 1.42 - 1.07(x) + 0.55(x^2)$$

2. Fit a second degree parabola for the following data:

x	y	x^2	x^3	x^4	xy	x^2y
20	5.5	400	8000	160000	110	2200
40	9.1	1600	64000	2560000	364	14560
60	14.9	3600	216000	12960000	894	53640
80	22.8	6400	512000	40960000	1824	145920
100	33.3	10000	1000000	100000000	3330	333000
120	46	14400	1728000	207360000	5520	662400
$\Sigma x = 420$	$\Sigma y = 131.6$	$\Sigma x^2 = 36400$	$\Sigma x^3 = 3528000$	$\Sigma x^4 = 364000000$	$\Sigma xy = 12042$	$\Sigma x^2y = 1211720$

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

$$131.6 = (6)a + b(420) + c(36400)$$
$$\Rightarrow \quad 12042 = a(420) + b(36400) + c(3528000)$$
$$1211720 = a(36400) + b(3528000) + c(364000000)$$

$$a = 4.35$$
$$b = 2.410714 \times 10^{-3}$$
$$c = 2.8705 \times 10^{-3}$$

Second degree parabola eq'n is:

$$y = a + bx + cx^2$$
$$= (2.410714 \times 10^{-3})x + (2.8705 \times 10^{-3})x^2$$

Power Curve:

The power curve $y = ax^b$

Taking 'log' on both sides

$$\log y = \log(ax^b)$$
$$\log y = \log a + \log x^b$$
$$\log y = \log a + b \log x$$

Here, $\log y = Y$

$\log a = A$, $a = \text{Antilog}(A)$

$\log x = X$

$$\boxed{Y = A + bX}$$

Normal Equations:

$$\Sigma Y = nA + b\Sigma X \quad \text{---} (1)$$
$$\Sigma XY = A\Sigma X + b\Sigma X^2 \quad \text{---} (2)$$

Solve eq'us (1) & (2), we get A & b

Problems:

1. Fit a power curve for the following data:

x	y	$X = \log x$	$Y = \log y$	x^2	XY
1	1200	0	3.0791	0	0
2	900	0.3010	2.9542	0.0906	0.8892
3	600	0.4771	2.7781	0.2276	1.3254
4	200	0.6020	2.3010	0.3624	1.3852
5	100	0.6989	2	0.4884	1.3978
6	50	0.7781	1.6989	0.6054	1.3219
$\Sigma x = 21$	$\Sigma y = 3050$	$\Sigma X = 2.8571$	$\Sigma Y = 14.81137$	$\Sigma x^2 = 1.7744$	$\Sigma XY = 6.3195$

$$\Sigma Y = nA + b\Sigma X \quad \Rightarrow \quad 14.8113 = 6A + 2.8571(b) \quad \text{---} (1)$$
$$\Sigma XY = A\Sigma X + b\Sigma X^2 \quad \Rightarrow \quad 6.3195 = 2.8571A + 1.7744(b) \quad \text{---} (2)$$

Solve (1) & (2), we get

$$A = 3.3123, \quad b = -1.77189$$

$a = \text{antilog}(A)$

$= 2052.57$

$y = ax^b$

2. Fit a power curve for the following data:

x	y	$X = \log x$	$Y = \log y$	X^2	XY
1	2.98	0	0.4742	0	0
2	4.26	0.3010	0.6294	0.0906	0.1894
3	5.21	0.4771	0.7168	0.2276	0.3419
4	6.1	0.6020	0.7853	0.3624	0.4727
5	6.8	0.6989	0.8325	0.4884	0.5818
6	7.5	0.7781	0.8750	0.6054	0.6808
		$\Sigma X = 2.8571$	$\Sigma Y = 4.3132$	$\Sigma X^2 = 1.7744$	$\Sigma XY = 2.2666$

$\Sigma Y = nA + b\Sigma X$

$\Sigma XY = A\Sigma X + b\Sigma X^2$

$\Rightarrow$ $4.3132 = (6)A + b(2.8571)$

$2.2666 = A(2.8571) + b(1.7744)$

$A = 0.4741$, $b = 0.5139$

$a = \text{Antilog}(A)$

$= 2.9792$

power curve: $y = ax^b$

$y = (2.9792)\, x^{(0.5139)}$

Exponential Curve:

Let the exponential curve $y = ab^x$

$\log y = \log(ab^x)$

$\log y = \log a + x \log b$

$Y = \log y$

$A = \log a$

$B = \log b$

$\boxed{Y = A + Bx}$

Normal Equations:

$\Sigma Y = nA + B\Sigma x$ ——— ①

$\Sigma xY = A\Sigma x + B\Sigma x^2$ ——— ②

Solve eq'n ① & ②, we get A & B

$a = \text{Antilog}(A)$, $b = \text{Antilog}(B)$

Problems:

1. Fit an exponential curve of the form: $y = ab^x$ to the following data:

x	y	x^2	$Y = \log y$	xY
1	1	1	0	0
2	1.2	4	0.0791	0.1582
3	1.8	9	0.2552	0.7656
4	2.5	16	0.3979	1.5916
5	3.6	25	0.5563	2.7815
6	4.7	36	0.6720	4.032
7	6.6	49	0.8195	5.7365
8	9.1	64	0.9590	7.672
$\Sigma x = 36$		$\Sigma x^2 = 204$	$\Sigma Y = 3.7390$	$\Sigma xY = 22.7374$

$$\Sigma Y = nA + B\Sigma x \quad \Rightarrow \quad 3.7390 = (8)A + B(36)$$
$$\Sigma xY = A\Sigma x + B\Sigma x^2 \quad \Rightarrow \quad 22.7374 = A(36) + B(204)$$

$$A = -0.1660, \quad B = 0.1407$$

$$a = \text{antilog}(A), \quad b = \text{antilog}(B)$$
$$a = 0.6823 \qquad b = 1.3826$$

Exponential curve: $y = ab^x$

$$y = (0.6823)(1.3826)^x$$

2. Fit an exponential curve of the form: $y = ab^x$ to the following data:

Year (x)	production (y)	x^2	$Y = \log y$	xY
1901	3.9	3613801	0.5910	1123.491
1911	5.3	3651921	0.7242	1383.9462
1921	7.3	3690241	0.8633	1658.3993
1931	9.6	3728761	0.9822	1896.6282
1941	12.9	3767481	1.1105	2155.4805
1951	17.1	3806401	1.2329	2405.3879
1961	23.2	3845521	1.3654	2677.5494
1971	30.5	3884841	1.4842	2925.3582
$\Sigma x = 15488$		$\Sigma x^2 = 29988968$	$\Sigma Y = 8.3537$	$\Sigma xY = 16226.2407$

$\Sigma Y = nA + B\Sigma x$ $\quad 8.3537 = (8)A + B(15488)$

$\Sigma xY = A\Sigma x + B\Sigma x^2$ $\Rightarrow$ $16226.2407 = A(15488) + B(29988968)$

$A = -23.6063$, $B = 0.0127$

$a = \text{antilog}(A)$, $b = \text{antilog}(B)$

$a = 2.4757 \times 10^{-24}$ $\quad b = 1.0296$

Exponential curve: $y = ab^x$

$$y = (2.4757 \times 10^{-24})(1.0296)^x$$

Exponential Curve:

Let the exponential curve $y = ae^{bx}$

Taking log on both sides

$$\log y = \log(a.e^{bx})$$
$$\log y = \log a + \log e^{bx}$$
$$\log y = \log a + bx \log_{10} e$$

$[\log_{10} e = 0.4342]$

$\log y = Y$
$\log a = A$
$b \log_{10} e = B$

$$\boxed{Y = A + Bx}$$

Normal equations:

$$\Sigma Y = nA + B\Sigma x \quad \text{—①}$$
$$\Sigma xY = A\Sigma x + B\Sigma x^2 \quad \text{—②}$$

Solve eq'n ① & ②, we get A & B

$a = \text{Antilog}(A)$ & $b = \frac{B}{\log_{10} e}$

$\Rightarrow b = \frac{B}{0.4342}$

Problems:

1. Fit an exponential curve of the form: $y = ae^{bx}$ to the following data:

x	y	x^2	$Y = \log y$	xY
1	1.6	1	0.2041	0.2041
2	4.5	4	0.6532	1.3064
3	13.8	9	1.1398	3.4194
4	40.2	16	1.6042	6.4168
5	12.5	25	1.0969	5.4845
6	30	36	1.4771	8.8626
$\Sigma x = 21$		$\Sigma x^2 = 91$	$\Sigma Y = 6.1753$	$\Sigma xY = 25.6938$

$$\Sigma Y = nA + B\Sigma x \qquad 6.1753 = 6(A) + B(21)$$
$$\Sigma xY = A\Sigma x + B\Sigma x^2 \Rightarrow 25.6938 = A(21) + B(91)$$

$$A = 0.2131, \quad B = 0.2331$$

$$a = \text{Antilog}(A), \quad b = \frac{0.2331}{0.4342}$$

$$a = 1.6334 \qquad b = 0.5368$$

Exponential curve: $y = ae^{bx}$

$$y = (1.6334)e^{(0.5368)x}$$

2. Fit an exponential curve of the form: $y = ae^{bx}$ to the following data:

x	y	x^2	$Y = \log y$	xY
0	20	0	1.3010	0
1	30	1	1.4771	1.4771
2	52	4	1.7160	3.432
3	77	9	1.8864	5.6592
4	135	16	2.1303	8.5212
5	211	25	2.3242	11.621
6	326	36	2.5132	15.0792
7	550	49	2.7403	19.1821
$\Sigma x = 28$		$\Sigma x^2 = 140$	$\Sigma Y = 16.0885$	$\Sigma xY = 64.9718$

$\Sigma Y = nA + B\Sigma x$ → $16.0885 = (8)A + B(28)$

$\Sigma xY = A\Sigma x + B\Sigma x^2$ → $64.9718 = A(28) + B(140)$

$A = 1.2892$, $B = 0.2062$

$a = \text{Antilog}(A)$, $b = \frac{0.2062}{0.4342}$

$a = 19.4625$

$b = 0.4748$

Exponential curve: $y = ae^{bx}$

$y = (19.4625)e^{(0.4748)x}$

Correlation:

In a bivariate distribution, the relationship between two variables is called correlation

(or)

If the change in one variable effects the change in the other variable, then the variables are said to be correlated.

Ex: Price and demand
income and expenditure
height and weight of a group of people

Types of correlation:

There are 5 types of correlation. They are as follows:

1. Positive correlation:

In a bivariate distribution, one variable increase (or) decrease, the other variable also increase (or) decrease, then the relation is called positive correlation.

i.e, If $r > 0$

2. Negative correlation:

In a bivariate distribution, one

Negative correlation

i.e, If $r<0$

3. Nonsense correlation / Zero correlation:

The correlation which is no meaning is called Nonsense correlation

(or)

In a bivariate distribution, the two variables x and y are independent, then it is called Nonsense correlation or Zero correlation

i.e, If $r=0$

4. Perfect positive correlation:

The two variables x and y are said to be perfect positively correlated, if they deviated same direction with constant ratio.

i.e, If $r=1$

Ex:

x	1	3	5	7
y	2	4	6	8

5. Perfect negative correlation:

The two variables x and y are said to be perfect negatively correlated, if they deviated opposite direction with constant ratio.

i.e, If $r=-1$

Ex:

x	1	3	5	7
y	8	6	4	2

Methods to measure of correlation:

There are different types of measures of correlation. Some of them are:

1. Karl Pearson's Coefficient of correlation ('r')

Karl pearson's coefficient of correlation:

The numerical measure of intensity or degree of linear relationship between the random variable x and y is called coefficient of correlation between the variables usually denoted by r (or) $r(x,y)$ (or) r_{xy}

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

Here,

$X = x - \bar{x}$

$Y = y - \bar{y}$

$\bar{x}$ = Mean in terms of 'x' & $\bar{y}$ = Mean in terms of 'y'

$$\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i \qquad \bar{y} = \frac{1}{n_2}\sum_{j=1}^{n_2} y_j$$

Properties of correlation:

- The limits of the Karl pearson's coefficient of correlation : ± 1

  i.e, $-1 \le r \le 1$

- The coefficient of correlation is independent of change of origin and scale.
- Two independent random variables are uncorrelated but converse is not true

Problems:

1. Find Karl pearson's coefficient of correlation from the following data:

Height (x)	weight (y)	$X = x - \bar{x}$	$Y = y - \bar{y}$	X^2	XY	Y^2
100	98	1	3	1	3	9
101	99	2	4	4	8	16
102	99	3	4	9	12	16
102	97	3	2	9	6	4
100	95	1	0	1	0	0
99	92	0	-3	0	0	9
97	95	-2	0	4	0	0
98	94	-1	-1	1	1	1
96	90	-3	-5	9	15	25
95	91	-4	-4	16	16	16
$\Sigma x = 990$	$\Sigma y = 950$	$\Sigma X = 0$		$\Sigma X^2 = 54$	$\Sigma XY = 61$	$\Sigma Y^2 = 96$

$$\bar{x} = \frac{990}{10} = 99$$

$$\bar{y} = \frac{950}{10} = 95$$

$$r = \frac{\Sigma XY}{\sqrt{\Sigma x^2}\sqrt{\Sigma Y^2}} = \frac{61}{\sqrt{54}\sqrt{96}} = 0.84$$

2. Calculate correlation coefficient from the following data:

x (height in inches)	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	X^2	XY	Y^2
65	67	-3	-2	9	6	4
66	68	-2	-1	4	2	1
67	65	-1	-4	1	4	16
67	68	-1	-1	1	1	1
68	72	0	3	0	0	9
69	72	1	3	1	3	9
70	69	2	0	4	0	0
72	71	4	2	16	8	4
$\Sigma x = 544$	$\Sigma y = 552$			$\Sigma X^2 = 36$	$\Sigma XY = 24$	$\Sigma Y^2 = 44$

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

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

$$\frac{\Sigma XY}{}$$

Spearman's Rank coefficient of correlation:

This method is based on rank and is useful in dealing with qualitative characteristic such as mortality, intelligence, blindness, kindness, goodness etc.,

The formula for spearman Rank coefficient of correlation:

1. observations are not repeated / Ranks are not repeated:

$$\rho = 1 - \left[\frac{6\Sigma D^2}{N^3 - N}\right]$$; where N = pair of observation
D = difference between two ranks
i.e., $D = \rho(x) - \rho(y)$

2. When Ranks are repeated:

$$\rho = 1 - \frac{6\left(\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \cdots\right)}{N^3 - N}$$;

Here N = pair of observations
D = difference between two ranks
m = repeated observation value

Problems:

1. From the ranks obtained by 10 students in 2 subjects statistics and mathematics, to what extent the knowledge of the students in two subjects related:

(X) Statistics	(Y) Mathematics	$D = \rho(X) - \rho(Y)$	D^2
1	2	-1	1
2	4	-2	4
3	1	2	4
4	5	-1	1
5	3	2	4
6	9	-3	9
7	7	0	0
8	10	-2	4
		3	9

Spearman's Rank coeff. of correlation: $\rho = 1 - \frac{6\Sigma D^2}{N^3 - N}$

$= 1 - \frac{6 \times 40}{10^3 - 10} = 1 - \frac{240}{990} = 0.75$

∴ Mathematics and statistics are positively correlated.

2. 10 competitors in a musical test were ranked by the 3 judges A, B and C in the following order

Ranks by A	B	C	$[R(A)-R(B)]$ D_1	D_1^2	$[R(B)-R(C)]$ D_2	D_2^2	$[R(A)-R(C)]$ D_3	D_3^2
1	3	6	-2	4	-3	9	-5	25
6	5	4	+1	1	1	1	2	4
5	8	9	-3	9	-1	1	-4	16
10	4	8	6	36	-4	16	2	4
3	7	1	-4	16	6	36	2	4
2	10	2	-8	64	8	64	0	0
4	2	3	2	4	-1	1	1	1
9	1	10	8	64	-9	81	-1	1
7	6	5	1	1	1	1	2	4
8	9	7	-1	1	2	4	1	1
				$\Sigma D_1^2 = 200$		$\Sigma D_2^2 = 214$		$\Sigma D_3^2 = 60$

Using Rank coeff. of correlation, discuss which pair of judge approach a common liking in music

$\rho_1 = 1 - \frac{6\Sigma D_1^2}{N^3 - N} = 1 - \frac{6 \times 200}{990} = 1 - \frac{120}{99} = 1 - 1.21 = -0.21$

$\rho_2 = 1 - \frac{6\Sigma D_2^2}{N^3 - N} = 1 - \frac{6 \times 214}{990} = 1 - \frac{1284}{990} = 1 - 1.29 = -0.29$

$\rho_3 = 1 - \frac{6\Sigma D_3^2}{N^3 - N} = 1 - \frac{6 \times 60}{990} = 1 - \frac{36}{99} = 1 - 0.36 = 0.64$

3. Find Rank coefficient of correlation to the following data:

X		Y		$D = P(X) - P(Y)$	D^2
85	2	93	1	1	1
60	4	75	3	1	1
73	3	65	4	-1	1
40	5	50	5	0	0
90	1	80	2	-1	1
					$\Sigma D^2 = 4$

$$\rho = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 4}{125 - 5}$$

$$= 1 - 0.2$$

$$\rho = 0.8$$

4. Obtain the rank coefficient of correlation for the following data:

X	Y	$P(X)$	$P(Y)$	$P(X) - P(Y)$ D	D^2
68	62	4	5	-1	1
64	58	6	7	-1	1
75	68	2.5	3.5	-1	1
50	45	9	10	-1	1
64	81	6	1	5	25
80	60	1	6	-5	25
75	68	2.5	3.5	-1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
					$\Sigma D^2 = 72$

In X-series, 75 is repeated 2-times $\therefore m = 2$

" , 64 " 3 " $\therefore m = 3$

In Y-series, 68 " 2 " $\therefore m = 2$

$$\rho = 1 - \frac{6\left(\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m)\right)}{N^3 - N}$$

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

$$\frac{6(72 + 0.5 + 2 + 0.5)}{} \qquad \frac{6(75)}{}$$

5. Obtain the rank coefficient of correlation for the following data:

X	Y	R(X)	R(Y)	R(X) − R(Y) = D	D^2
48	13	2	6.5	−4.5	20.25
33	13	4	6.5	−2.5	6.25
40	24	3	1	2	4
9	6	10	9	1	1
16	15	8	5	3	9
25	20	5	2	3	9
24	9	6	8	−2	4
16	16	8	4	4	16
51	19	1	3	−2	4
16	4	8	10	−2	4

$\frac{7+8+9}{3} = 8$ $\quad \frac{6+7}{2} = 6.5$ $\quad \Sigma D^2 = 77.5$

In X-series, 16 repeated 3 times $\Rightarrow \therefore m = 3$

In Y-series, 13 " 2 times $\Rightarrow \therefore m = 2$

$$\rho = 1 - \frac{6\left(\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m)\right)}{N^3 - N}$$

$$= 1 - \frac{6\left(77.5 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2)\right)}{10^3 - 10}$$

$$= 1 - \frac{6(77.5 + 2 + 0.5)}{990}$$

$$= 1 - 0.080$$

$$= 0.92$$

Regression Analysis:

Regression Analysis in general sense the estimation of prediction of the unknown value of one variable from the known value of other variable. It is very important statistical tool which is extensively used in almost all sciences - natural, social and physical science

Definition:

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

In Regression Analysis, there are two types of variables. The variable whose value is influenced / is to be predicted is called "dependent variable" and the variable which influences the values is used for prediction is called "Independent variable".

Lines of Regression:

Lines of Regression (or) are the line which gives the best estimate of the value of one variable for any given value of the other variable. In case of two variables 'X' and 'Y' we shall have two Lines of Regression. They are

1. Regression Line X on Y
2. Regression Line Y on X

1. Regression Line X on Y $= X - \bar{X} = r \frac{\sigma x}{\sigma y} (Y - \bar{Y})$

(or)

$X - \bar{X} = b_{xy} (Y - \bar{Y})$

(or)

Here $x = x - \bar{x}$ → mean in terms of x

$Y = y - \bar{Y}$ → " " Y

$r \frac{\sigma x}{\sigma y}$ (or) b_{xy} (or) $\frac{\Sigma xY}{\Sigma Y^2}$ = Regression coefficient of x on Y

2. Regression line Y on X = $Y - \bar{Y} = r \cdot \frac{\sigma y}{\sigma x}(x - \bar{x})$

$r \cdot \frac{\sigma y}{\sigma x}$ (or) b_{yx} (or) $\frac{\Sigma xy}{\Sigma x^2}$ = Regression coefficient of Y on x

3. Coefficient of correlation: $r = \sqrt{b_{xy} \cdot b_{yx}}$ (or) $\sqrt{\frac{\Sigma xy}{\Sigma Y^2} \times \frac{\Sigma xy}{\Sigma x^2}}$

4. Probable error = $P.E = 0.6745 \times \frac{1 - r^2}{\sqrt{N}}$

Properties of Regression Coefficient:

1. The correlation coefficient is geometric mean of two regression coefficients.

$$r = \sqrt{b_{xy} \cdot b_{yx}}$$

2. The arithmetic mean of two regression coefficients is greater than or equal to correlation coefficient

$$\frac{b_{xy} + b_{yx}}{2} \geq r$$

3. ffll the coefficient of correlation cannot excued 1 Incase of regression, one of the regression coefficient is greater than 1, then the other must be less than 1.

4. Both the regression coefficient will have the same sign either positive or negative

5. If one regression coefficient is '+ve' then the other should be '+' and vice versa

Uses of Regression Analysis:

1) The Regression Analysis technique is very useful in predicting the probable value of an unknown variable in response to some known related variable

2) The Regression Analysis is useful in establishing the nature of relationship between 2 variables

3) Regression Analysis is extensively used for measurement and estimating the relationship among variables.

4) Regression Analysis provides regression coefficient which are generally used in calculation of correlation coefficient

Problems:

1. Calculate the Regression coefficient of X on Y and Y on x and also find probable error from the following data:

price (x)	Demand (y)	$X = x - \bar{x}$	X^2	$Y = Y - \bar{Y}$	Y^2	XY
10	40	-3	9	-1	1	3
12	38	-1	1	-3	9	3
13	43	0	0	2	4	0
12	45	-1	1	4	16	-4
16	37	3	9	-4	16	-12
15	43	2	4	2	4	4
			$\Sigma X^2 = 24$		$\Sigma Y^2 = 50$	$\Sigma XY = -6$

Estimate the likely demand when the price is ₹ 20

$$\bar{X} = \frac{10+12+13+12+16+15}{6} \quad ; \quad \bar{Y} = \frac{40+38+43+45+37+43}{6}$$

$$\bar{X} = 13 \qquad \bar{Y} = 41$$

Regression line X on Y:

Regression line Y on x: