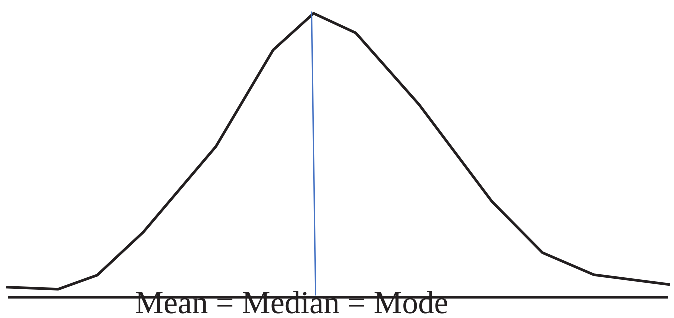
Descriptive Statistics
Unit IV
MEASURES OF SKEWNESS

Meaning of Skewness:

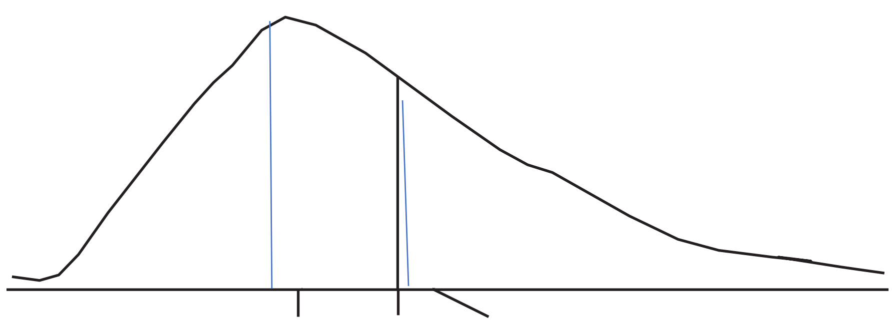
- Skewness means **lack of symmetry** .
- We study skewness to have an idea about the shape of the curve which we can draw with the help of the given data.
- If, in a distribution**, Mean = Median = Mode**, then that distribution is known as **Symmetrical Distribution**.
- If, in a distribution, **Mean ≠ Median ≠ Mode**, then it is not a symmetrical distribution and it is called a **Skewed Distribution** and such a distribution could either be **positively skewed or negatively skewed.**

a) Symmetrical distribution:



It is clear from the above diagram that in a symmetrical distribution the values of mean,median and mode coincide. The spread of the frequencies is the same on both sides of the centerpoint of the curve.

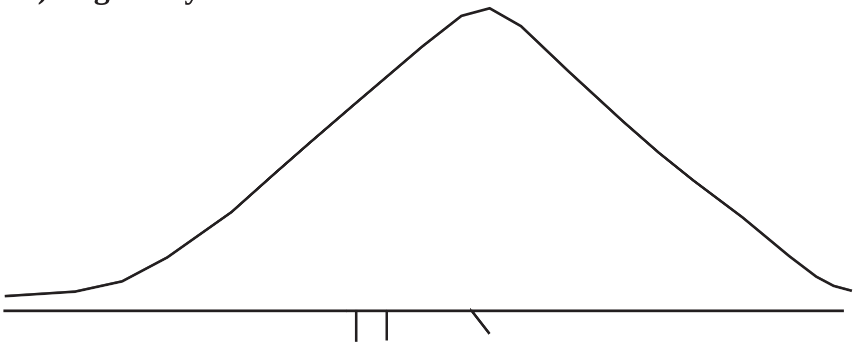
b) Positively skewed distribution:



Mode < Median < Mean

It is clear from the above diagram, in a positively skewed distribution, the value of the mean is maximum and that of the mode is least, the median lies in between the two. In the positively skewed distribution, the frequencies are spread out over a greater range of values onthe right-hand side than they are on the left hand side.

c) Negatively skewed distribution:



Mean < Median < Mode

It is clear from the above diagram, in a negatively skewed distribution, the value of themode is maximum and that of the mean is least. The median lies in between the two. In the negatively skewed distribution the frequencies are spread out over a greater range of values onthe left hand side than they are on the right hand side.

Measures of skewness:

The important measures of skewness are

[1] Karl – Pearson' s coefficient of skewness

[2] Bowley' s coefficient of skewness

[3] Measure of skewness based on moments.

We are interested in studying only the **first two** methods only.

1. **Karl-Pearson's Coefficient of skewness:**

According to Karl – Pearson, the absolute measure of skewness = mean – mode. This measure is not suitable for making valid comparison of the skewness in two or more distributionsbecause the unit of measurement may be different in different series. To avoid this difficulty, we use relative measure of skewness, called Karl-Pearson' s coefficient of skewness given by:

$$\text{Karl-Pearson's Coefficient of Skewness} = \frac{\text{Mean - Mode}}{\text{S.D.}}$$

In case of **mode is ill- defined**, the coefficient can be determined by

$$\text{Coefficient of skewness} = \frac{3(\text{Mean - Median})}{\text{S.D.}}$$

2. Bowley's coefficient of Skewness (Sk_B)

$$Sk_B = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

Karl Pearson's coefficient of Skewness (Sk_p)

1. From the marks secured by 120 students in Section A and B of a class, the following measures are obtained:
   Section A: $\overline{X}$ = 46.83; S.D = 14.8; Mode = 51.67
   Section B: $\overline{X}$ = 47.83; S.D = 14.8; Mode = 47.07
   Determine which distribution of marks is more skewed.
   Solution: Karl Pearson's coefficient of Skewness
   For Section A: $Sk_p = \frac{\overline{X} - Z}{\sigma} = \frac{46.83 - 51.67}{14.8} = \frac{-4.84}{14.8}$ = -0.3270
   For Section B: $Sk_p = \frac{\overline{X} - Z}{\sigma} = \frac{47.83 - 47.07}{14.8} = \frac{0.76}{14.8}$ = 0.05135

Marks of Section A is more Skewed. But marks of Section A is negatively Skewed. Marks of Section B are Positively skewed.

2. From a moderately skewed distribution of retail prices for men's shoes, it is found that the mean price is Rs. 20 and the median price is Rs. 17. If the coefficient of variation is 20%, find the Pearsonian coefficient of skewness of the distribution.

Solution: Given: C.V. = 20, $\overline{X}$ = 20, M = 17

C. V. = $\frac{\sigma}{\overline{X}} X100$

20= $\frac{\sigma}{20} X100$= 20 X 20 /100 = 400/100 = 4

σ = 4

$Sk_p = \frac{3(\overline{X} - M)}{\sigma} = \frac{3(20-17)}{4}$ = 9/4 = 2.25

3. Calculate Karl Pearson's coefficient of Skewness for the following data.
   X 25, 15, 23, 40, 27, 25, 23, 25,20

X	X^2
25	625
15	225
23	529
40	1600
27	729
25	625
23	529
25	625
20	400
$\sum X$ =223	$\sum X^2$ = 5887

$\overline{X} = \frac{\sum X}{N} = \frac{223}{9}$ = 24.78

σ = $\sqrt{\frac{\sum X^2}{N} - \left[\frac{\sum X}{N}\right]^2} = \sqrt{\frac{5887}{9} - (24.78)^2}$

= $\sqrt{654.1111 - 614.0484}$ = $\sqrt{40.06}$ = 6.33

Z = 25

$Sk_p = \frac{\overline{X} - Z}{\sigma} = \frac{24.78 - 25}{6.33} = \frac{-0.22}{6.33}$ = - 0.0348

4. Calculate Karl Pearson's coefficient of Skewness for the following data.

Wage per Item Rs.(x)	Number of items f	fx	x^2	fx^2
12	10	120	144	1440
15	25	375	225	5625
20	40	800	400	16000
25	70	1750	625	43750
30	32	960	900	28800
40	13	520	1600	20800
50	10	500	2500	25000
	$\sum f = 200$	$\sum fx = 5025$		$\sum fX^2 = 141415$

$$\overline{X} = \frac{\sum fX}{\sum f} = \frac{5025}{200} = 25.13$$

$$\sigma = \sqrt{\frac{\sum fX^2}{\sum f} - \left[\frac{\sum fX}{\sum f}\right]^2} = \sqrt{\frac{141415}{200} - (25.13)^2} = \sqrt{707.075 - 631.5169} = \sqrt{75.5581} = 8.69$$

Greatest frequency = 70, Z = 25

$$Sk_p = \frac{\overline{X} - Z}{\sigma} = \frac{25.13 - 25}{8.69} = 0.13/8.69 = 0.0149$$

5. Calculate Karl Pearson's coefficient of Skewness for the following data.

Profit (Rs.Lakhs)	No of Companies (f)	m	fm	m^2	fm^2
10-20	18	15	270	225	4050
20-30	$20 = f_0$	25	500	625	12500
30-40	$30 = f_1$	35	1050	1225	36750
40-50	$22 = f_2$	45	990	2025	44550
50-60	10	55	550	3025	30250
	$\sum f = 100$		$\sum fm = 3360$		$\sum fm^2 = 128100$

$$\overline{X} = \frac{\sum fm}{\sum f} = 3360/100 = 33.6$$

$$\sigma = \sqrt{\frac{\sum fm^2}{\sum f} - \left[\frac{\sum fm}{\sum f}\right]^2} = \sqrt{\frac{128100}{100} - (33.6)^2} = \sqrt{1281 - 1128.96} = \sqrt{152.04} = 12.33$$

$D_1 = f_1 - f_0 = 30-20 = 10$: $D_2 = f_1 - f_2 = 30-22 = 8$: L = 30: i=10

$$Z = L + \left[\frac{D_1}{D_1 + D_2}\right] i = 30 + \left[\frac{10}{10+8}\right] 10 = 30 + \left[\frac{10}{18}\right] 10 = 30+5.56 = 35.56$$

$$Sk_p = \frac{\overline{X} - Z}{\sigma} = \frac{33.6 - 35.56}{12.33} = -1.96/12.33 = -0.1590$$

6. Calculate Karl Pearson's coefficient of Skewness for the following data.

Weight (lbs)	No of Students(f)	m	fm	m^2	fm^2	c.f
90-100	4	95	380	9025	36100	4
100-110	2	105	210	11025	22050	6
110-120	18	115	2070	13225	238050	24
120-130	22	125	2750	15625	343750	46
130-140	21	135	2835	18225	382725	67
140-150	19	145	2755	21025	399475	86
150-160	10	155	1550	24025	240250	96
160-170	3	165	495	27225	81675	99
170-180	2	175	350	30625	61250	101
	$\sum f = 101$		$\sum fm =$ 13395		$\sum fm^2 =$ 1805325	

$$\overline{X} = \frac{\sum fm}{\sum f} = 13395/101 = 132.62$$

$$\sigma = \sqrt{\frac{\sum fm^2}{\sum f} - \left[\frac{\sum fm}{\sum f}\right]^2} = = \sqrt{\frac{1805325}{101} - (132.62)^2} = \sqrt{17874.51 - 17588.06}$$

$$= \sqrt{286.45} = 16.9$$

$\frac{\sum f}{2}$ = 101/2 = 50.5, Median Class = 130-140,L = 130,p.c.f =46, f = 21,i = 10

$$M = L + \left[\frac{\sum f/2 - p.c.f}{f}\right] i = 130 + \left[\frac{50.5-46}{21}\right]10 = 130 + \left[\frac{4.5}{21}\right]10 = 130+2.14 = 132.14$$

$$Sk_p = \frac{3(\overline{X} - M)}{\sigma} = \frac{3(132.62 - 132.14)}{16.9} = \frac{3(0.48)}{16.9} = 1.44/16.9 = 0.0852$$

BOWLEY'S COEFFICIENT OF SKEWNESS

7. Compare the Skewness of A and B

	Q_1		M	Q_3
Series A	40		60	80
Series B	62.85	65.25	72.15	

Series A

$$Sk_B = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} = \frac{80 + 40 - 2(60)}{80 - 40} = \frac{120 - 120}{40} = 0$$

Series B

$$Sk_B = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} = \frac{72.15 + 62.85 - 2(65.25)}{72.15 - 62.85} = \frac{135 - 130.5}{9.3} = 4.5/9.3 = 0.4839$$

In series A there is no skewness, In Series B there is moderate positive skewness.

8. Calculate Bowley's coefficient of Skewness.

No of child per family x	No of Families f	Cf
0	7	7
1	10	17
2	16	33
3	25	58
4	18	76
5	11	87
6	8	95
	$\sum f = 95$	

Solution: Position of $Q_1 = \frac{\sum f + 1}{4}$ = 95+1/4 = 96/4 =24

$Q_1 = 2$

Position of $Q_3 = 3(\frac{\sum f + 1}{4}) = 3(24) = 72$

$Q_3 = 4$

Position M = $\frac{\sum f + 1}{2}$ = 95+1/2 = 96/2 = 48

M = 3

$$Sk_B = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} = \frac{4 + 2 - 2(3)}{4 - 2} = \frac{6 - 6}{2} = 0$$

9. Calculate Bowley's coefficient of Skewness.

Weekly Wages (Rs.)	No of Workers(f)	cf
Below 200	10	10
200-250	25	35
250-300	145	180

300-350	220	400
350-400	70	470
400 & above	30	500
	$\sum f = 500$	

$\frac{\sum f}{4} = \frac{500}{4}$ = 125, Q_1 Class = 250-300, L_1 = 250, p.c.f_1 = 35, f_1 = 145, i_1 = 50

$$Q_1 = L_1 + \left[\frac{\sum f/4 - p.c.f_1}{f_1}\right] i_1 = 250 + \left[\frac{125-35}{145}\right]50 = 250 + \left[\frac{90}{145}\right]50$$

Q_1 = 250+31.03 = Rs.281.03

$3\left(\frac{\sum f}{4}\right)$ = 3(125) = 375, Q_3 Class = 300-350, L_3 = 300, p.c.f_3 = 180, f_3 = 220, i_3 = 50

$$Q_3 = L_3 + \left[\frac{3(\sum f/4) - p.c.f_3}{f_3}\right] i_3 = 300 + \left[\frac{375-180}{220}\right]50 = 300 + \left[\frac{195}{220}\right]50$$

Q_3 = 300+44.32 = Rs. 344.32

Median:

$\frac{\sum f}{2}$ = 500/2 = 250, Median Class = 300-350, L = 300, p.c.f = 180,

f = 220, i = 50

$$M = L + \left[\frac{\sum f/2 - p.c.f}{f}\right] i = 300 + \left[\frac{250-180}{220}\right]50 = 300 + \left[\frac{70}{220}\right]50 = 300+15.91$$

M = Rs. 315.91

$$Sk_B = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} = -0.1022$$

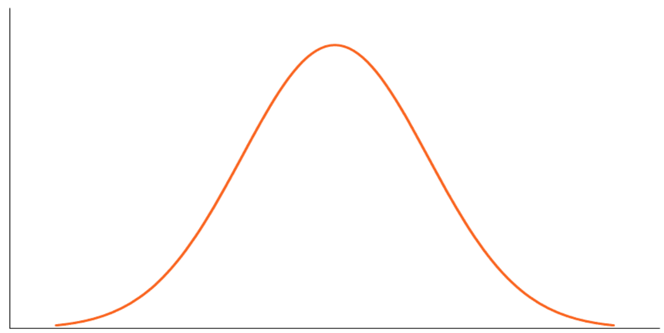
KURTOSIS

The Measure of location, dispersion and skewness alone cannot give a complete idea of a distribution.

Even if the distributions are symmetrical about the mean. But the frequency curves have different flatness or peakness of a distribution.

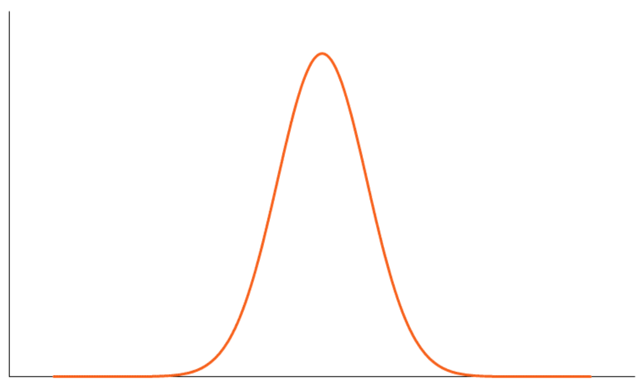
types of Kurtosis: The types of kurtosis are determined by the excess kurtosis of a particular distribution. The excess kurtosis can take positive or negative values, as well as values close to zero.

1. Mesokurtic: Data that follows a mesokurtic distribution shows an excess kurtosis of zero or close to zero. This means that if the data follows a normal distribution, it follows a mesokurtic distribution.



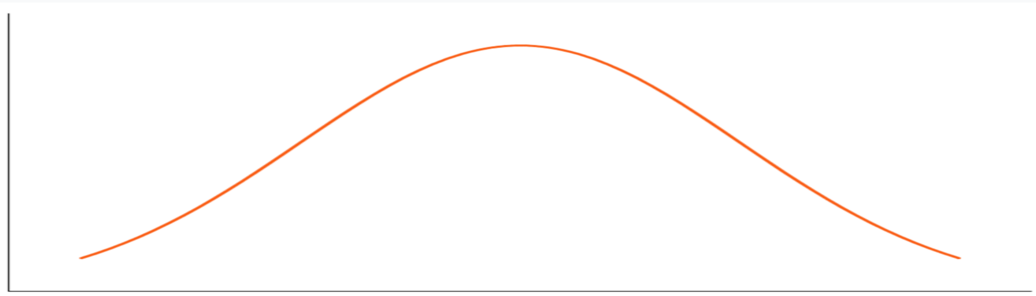
2. Leptokurtic

Leptokurtic indicates a positive excess kurtosis. The leptokurtic distribution shows heavy tails on either side, indicating large outliers. In finance, a leptokurtic distribution shows that the investment returns may be prone to extreme values on either side. Therefore, an investment whose returns follow a leptokurtic distribution is considered to be risky.



3. Platykurtic

A platykurtic distribution shows a negative excess kurtosis. The kurtosis reveals a distribution with flat tails. The flat tails indicate the small outliers in a distribution. In the finance context, the platykurtic distribution of the investment returns is desirable for investors because there is a small probability that the investment would experience extreme returns.



The **moment coefficient of kurtosis** of a data set is computed almost the same way as the coefficient of skewness: just change the exponent 3 to 4 in the formulas:

kurtosis: $a_4 = m_4 / m_2^2$ and excess kurtosis: $g_2 = a_4 - 3$

where

$$m_4 = \sum(x-\bar{x})^4 / n \quad \text{and} \quad m_2 = \sum(x-\bar{x})^2 / n$$

CORRELATION

Definition:

The term correlation refers to the relationship between two or more Variables. Correlation studies the extent of relationship between two or more variables.

Types of Correlation

Various types of correlation are considered under the following three heads. They are

(i) Positive or negative correlation
(ii) Simple or Partial or Multiple correlation
(iii) Linear or Non-linear or No correlation

Methods of measuring Correlation

Four methods of correlation are

(i) Scatter Diagram
(ii) Karl Pearson's correlation coefficient(r)
(iii) Spearman's rank correlation coefficient(ρ)
(iv) Correlation coefficient by concurrent deviation method(r_c)

KARL PEARSON'S COEFFICIENT OF CORRELATION (R)

This is also called product moment correlation coefficient. This is denoted by r. This is covariance between the two variables divided by the product of their standard deviations.

Formula

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

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

$$r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}, \text{ Where } \sum x = 0, \sum y = 0$$

$$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma (u)^2 - (\Sigma u)^2} x \sqrt{n\Sigma (v)^2 - (\Sigma v)^2}}$$,where u = x-A, v = y-B

1.Compute the coefficient of correlation between X – Advertisement Expenditure and Y – Sales.

X	10	12	18	8	13	20	22	15	5	17
Y	88	90	94	86	87	92	96	94	88	85

Solution

X	Y	X^2	Y^2	XY
10	88	100	7744	880
12	90	144	8100	1080
18	94	324	8836	1692
8	86	64	7396	688
13	87	169	7569	1131
20	92	400	8464	1840
22	96	484	9216	2112
15	94	225	8836	1410
5	88	25	7744	440
17	85	289	7225	1445
$\sum X = 140$	$\sum Y = 900$	$\sum X^2 = 2224$	$\sum Y^2 = 81130$	$\sum XY = 12718$

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

$$r = \frac{10(12718) - (140)(900)}{\sqrt{10(2224) - (140)^2}\sqrt{10(81130) - (900)^2}} \qquad r = \frac{127180 - 126000}{\sqrt{22240 - 19600}\sqrt{811300 - 810000}}$$

$$r = \frac{1180}{\sqrt{2640}\sqrt{1300}} \qquad r = \frac{1180}{51.3809 X 36.0555} \qquad r = \frac{1180}{1852.56404} \qquad r = 0.6370$$

2. The following table gives aptitude test scores and productivity indices of 8 randomly selected workers. Calculate the correlation coefficient between the aptitude score and productivity index.

Let X = Aptitude Score, Y = Productivity Index

Apptitude Score X	Productivity Index Y	X^2	Y^2	XY
57	67	3249	4489	3819
58	68	3364	4624	3944
59	65	3481	4225	3835
59	68	3481	4624	4012
60	72	3600	5184	4320
61	72	3721	5184	4392
62	69	3844	4761	4278
64	71	4096	5041	4544
$\sum X = 480$	$\sum Y = 552$	$\sum X^2 = 28836$	$\sum Y^2 = 38132$	$\sum XY = 33144$

$$r=\frac{N\sum XY-\left(\sum X\right)\left(\sum Y\right)}{\sqrt{N\sum X^2-\left(\sum X\right)^2}\sqrt{N\sum Y^2-\left(\sum Y\right)^2}}$$

$$r=\frac{8(33144)-(480)(552)}{\sqrt{8(28836)-(480)^2}\sqrt{8(38132)-(552)^2}} \qquad r=\frac{265152-264960}{\sqrt{230688-230400}\sqrt{305056-304704}}$$

$$r=\frac{192}{\sqrt{288}\sqrt{352}} \qquad r=\frac{192}{16.9706 X 18.7617} \qquad r=\frac{192}{318.3974}$$

r = 0.6030

3. Calculate the coefficient of correlation between Expenditure on Advertisement in Rs.'000 (X) and Sales in Rs. Lakhs (Y) after allowing the time lag of two months.

Months	X	Y	X	Y	X^2	Y^2	XY
Jan	40	75	40	65	1600	4225	2600
Feb	45	69	45	64	2025	4096	2880
Mar	47	65	47	70	2209	4900	3290
Apr	50	64	50	71	2500	5041	3550
May	53	70	53	75	2809	5625	3975
June	60	71	60	83	3600	6889	4980
July	57	75	57	90	3249	8100	5130
Aug	51	83	51	92	2601	8464	4692
Sep	48	90					
Oct	45	92					
			$\sum X=403$	$\sum Y=610$	$\sum X^2=20593$	$\sum Y^2=47340$	$\sum XY=31097$

$$r=\frac{N\sum XY-\left(\sum X\right)\left(\sum Y\right)}{\sqrt{N\sum X^2-\left(\sum X\right)^2}\sqrt{N\sum Y^2-\left(\sum Y\right)^2}}$$

$$r=\frac{8(31097)-(403)(610)}{\sqrt{8(20593)-(403)^2}\sqrt{8(47340)-(610)^2}} \qquad r=\frac{248776-245830}{\sqrt{164744-162409}\sqrt{378720-372100}}$$

$$r=\frac{2946}{\sqrt{2335}\sqrt{6620}} \qquad r=\frac{2946}{48.3218 X 81.3634} \qquad r=\frac{2946}{3931.6259}$$

r = 0.7493

4. From the following data, compute the coefficient of correlation between X and Y

	X	Y
Sum of squares of deviations from the arithmetic mean	8250	724
Sum of products of deviations of X and Y from respective means	2350	
No of pairs of observations	10	

Solution:

Given

$\sum x^2 = \sum (X - \overline{X})^2 = 8250$

$\sum y^2 = \sum (Y - \overline{Y})^2 = 724$

$\sum xy = \sum (X - \overline{X})(Y - \overline{Y}) = 2350$

N = 10

$$r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}} = r = \frac{2350}{\sqrt{8250}\sqrt{724}} = r = \frac{2350}{90.8295 X\, 26.9072}$$

$$r = \frac{2350}{2443.9675} = 0.9615$$

5. Calculate Karl Pearson's co-efficient of correlation from the following data using 44 and 26 respectively as the origin of x and y

X	43	44	46	40	44	42	45	42	38	40	42	57
y	29	31	19	18	19	27	27	29	41	30	26	10

$$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma (u)^2 - (\Sigma u)^2}\, x \sqrt{n\Sigma (v)^2 - (\Sigma v)^2}}$$,where u = x-A, v = y-B

X	y	u = x – A = x- 44	v = y – B = y - 26	uv	u^2	v^2
43	29	-1	3	-3	1	9
44	31	0	5	0	0	25
46	19	2	-7	-14	4	49
40	18	-4	-8	32	16	64
44	19	0	-7	0	0	49
42	27	-2	1	-2	4	1
45	27	1	1	1	1	1
42	29	-2	3	-6	4	9
38	41	-6	15	-90	36	225
40	30	-4	4	-16	16	16
42	26	-2	0	0	4	0
57	10	13	-16	-208	169	256
		Σu = -5	Σ v = - 6	Σ(uv) = -306	$\Sigma(u^2)$ = 255	$\Sigma(v^2)$ =704

$$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma(u)^2} - (\Sigma u)^2 x \sqrt{n\Sigma(v)^2} - (\Sigma v)^2}$$

$$r = \frac{[12\ \ (-306)] - [(-5)(-6)]}{\sqrt{12(-255)-}\,(-5)^2\ \sqrt{12(704)} - (-6)^2}$$

$$r = \frac{[-3672] - [30]}{\sqrt{3060-}\,25\quad \sqrt{8448} - (36)} \qquad r = \frac{-3702}{3035\,\sqrt{8412}}$$

r = -0.7327

Note:

1. Correlation coefficient (r) lies between $-1 \leq r \leq +1$
2. If r = 0, absence of linear correlation
3. If r = +1, perfect positive correlation
4. If r = -1, perfect negative correlation
5. r = Coefficient of correlation
6. r^2 = Coefficient of Determination
7. $K^2 = 1-r^2$ = Coefficient of Non – Determination
8. $K = \pm\sqrt{1-r^2}$ = Coefficient of Alienation
9. S.E(r) = $\frac{1-r^2}{\sqrt{N}}$

SPEARMAN'S RANK CORRELATION (ρ)

(i) When there is no tie and actual ranks are given

$$\rho = 1 - \left[\frac{6\sum d^2}{N(N^2-1)}\right]$$, Where d = difference between Rank of X and rank of Y

(ii) When one value occurs m times

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

(iii) When more than one value is repeated

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

1. Rankings of 10 trainees at the beginning (X) and at the end (Y) of a certain course are given below. Calculate the rank correlation.

Trainees	A	B	C	D	E	F	G	H	I	J
X	1	6	3	9	5	2	7	10	8	4
Y	6	8	3	7	2	1	5	9	4	10

Trainees	R_X	R_Y	$d = R_X - R_Y$	d^2
A	1	6	-5	25
B	6	8	-2	4
C	3	3	0	0
D	9	7	2	4
E	5	2	3	9
F	2	1	1	1
G	7	5	2	4
H	10	9	1	1
I	8	4	4	16
J	4	10	-6	36
			$\sum d = 0$	$\sum d^2 = 100$

$$\rho = 1 - \left[\frac{6\sum d^2}{N(N^2-1)}\right]$$

$$\rho = 1 - \left[\frac{6X100}{10(10^2-1)}\right] = 1 - \left[\frac{6X100}{10(100-1)}\right] = 1 - \left[\frac{6X100}{10X99}\right] = 1 - \left[\frac{600}{990}\right] = 1 - 0.6061$$

$\rho = 0.3939$

2. For the data given below, calculate the rank correlation coefficient.

X	21	36	42	37	25
Y	47	40	37	42	43

X	Y	R_X	R_y	$d = R_X - R_y$	d^2
21	47	5	1	4	16
36	40	3	4	-1	1
42	37	1	5	-4	16
37	42	2	3	-1	1
25	43	4	2	2	4
				$\sum d = 0$	$\sum d^2 = 38$

$$\rho = 1 - \left[\frac{6\sum d^2}{N(N^2-1)} \right]$$

$$\rho = 1 - \left[\frac{6X38}{5(5^2-1)} \right] = 1 - \left[\frac{6X38}{5(25-1)} \right] = 1 - \left[\frac{6X38}{5x24} \right] = 1 - \left[\frac{228}{120} \right] = 1\text{-}1.9$$

ρ = -0.9

3. Find the rank correlation coefficient for the percentage of marks secured by a group of 8 students in Economics and Statistics.

X	50	60	65	70	75	40	70	80
Y	80	71	60	75	90	82	70	50

X	Y	R_X	R_Y	$d = R_X - R_y$	d^2
50	80	7	3	4	16
60	71	6	5	1	1
65	60	5	7	-2	4
70	75	3.5	4	-0.5	0.25
75	90	2	1	1	1
40	82	8	2	6	36
70	70	3.5	6	-2.5	6.25
80	50	1	8	-7	49
				$\sum d = 0$	$\sum d^2 = 113.5$

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

when m = 2, $\frac{m(m^2-1)}{12} = \frac{2(2^2-1)}{12} = \frac{2(4-1)}{12} = \frac{2X3}{12} = \frac{6}{12} = \frac{1}{2} = 0.5$

$$1 - \left[\frac{6\{113.5+0.5\}}{8(8^2-1)} \right] = 1 - \left[\frac{6X114}{8X63} \right] = 1 - \left[\frac{684}{504} \right] = 1\text{-}1.3571 = -0.3571$$

5. Ten competitors in a musical test were ranked by three judges A,B and C in the following order:

Using rank correlation method, discuss which pair of judges has the nearest approach to common likings in music.

Rank of A	1	6	5	10	3	2	4	9	7	8
Rank of B	3	5	8	4	7	10	2	1	6	9
Rank of C	6	4	9	8	1	2	3	10	5	7

Ranks								
A	B	C	d_{AB}	d^2_{AB}	d_{AC}	d^2_{AC}	d_{BC}	d^2_{BC}
1	3	6	-2	4	-5	25	-3	9
6	5	4	1	1	2	4	1	1
5	8	9	-3	9	-4	16	-1	1
10	4	8	6	36	2	4	-4	16
3	7	1	-4	16	2	4	6	36
2	10	2	-8	64	0	0	8	64
4	2	3	2	4	1	1	-1	1
9	1	10	8	64	-1	1	-9	81
7	6	5	1	1	2	4	1	1
8	9	7	-1	1	1	1	2	4
			$\sum d_{AB} = 0$	$\sum d^2_{AB} = 200$	$\sum d_{AC} = 0$	$\sum d^2_{AC} = 60$	$\sum d_{BC} = 0$	$\sum d^2_{BC} = 214$

$$\rho_{AB} = 1 - \left[\frac{6\sum d^2}{N(N^2-1)}\right] = 1 - \left[\frac{6x200}{10(10^2-1)}\right] = 1 - \left[\frac{6X200}{10(100-1)}\right] = 1 - \left[\frac{1200}{990}\right] = 1 - 1.2121 = -0.2121$$

$$\rho_{AC} = 1 - \left[\frac{6\sum d^2}{N(N^2-1)}\right] = 1 - \left[\frac{6X60}{10(10^2-1)}\right] = 1 - \left[\frac{6X60}{10X99}\right] = 1 - \left[\frac{360}{990}\right] = 1 - 0.3636 = 0.6364$$

$$\rho_{BC} = 1 - \left[\frac{6\sum d^2}{N(N^2-1)}\right] = 1 - \left[\frac{6X214}{10(10^2-1)}\right] = 1 - \left[\frac{6X214}{10X99}\right] = 1 - \left[\frac{1284}{990}\right] = 1 - 1.2970 = -0.2970$$

ρ_{AC} is greater, so the pair A and C of judges has the nearest approach to common likings in music.

5.Marks obtained by 8 students in Accountancy (X) and Statistics (Y) are given below. Compute rank correlation coefficient.

X	15	20	28	12	40	60	20	80
Y	40	30	50	30	20	10	30	60

X	Y	R_X	R_Y	d	d^2
15	40	7	3	4	16
20	30	5.5	5	0.5	0.25
28	50	4	2	2	4
12	30	8	5	3	9
40	20	3	7	-4	16
60	10	2	8	-6	36
20	30	5.5	5	0.5	0.25
80	60	1	1	0	0
				$\sum d = 0$	$\sum d^2 = 81.5$

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

when m = 2, $\frac{m(m^2-1)}{12} = \frac{2(2^2-1)}{12} = \frac{2(4-1)}{12} = \frac{2X3}{12} = \frac{6}{12} = \frac{1}{2} = 0.5$

when m = 3, $\frac{m(m^2-1)}{12} = \frac{3(3^2-1)}{12} = \frac{3(9-1)}{12} = \frac{3X8}{12} = \frac{24}{12} = 2$

$$\rho = 1 - \left[\frac{6\{81.5+0.5+2\}}{8(8^2-1)} \right] = 1 - \left[\frac{6X84}{8X63} \right] = 1 - \left[\frac{504}{504} \right] = 1\text{-}1 = 0$$

METHOD OF CONCURRENT DEVIATIONS

This method requires only a direction of change (+ive to – ive or – ive to +ive) in the successive values of the variable x and in variable y. The co-efficient of correlation is given by the formula

$r_C = \pm\sqrt{\pm\frac{(2C-N)}{N}}$ where r_C is the coefficient of concurrent deviations, C is the number of concurrent deviations and N is the number of pairs of deviations compared. The sign of r_C is given as follows. If 2C – N is negative then ' – ' sign is taken both inside and outside the square root. In this case r_C is negative. If 2C – N is positive then '+' sign is taken both inside and outside the square root. In this case r_C is Positive. The value of 'r_C' will always lie between – 1 and +1, i.e., - 1 ≤ 1

Calculate co-efficient of correlation by the method of concurrent deviation from the following

X	84	85	62	48	84	95	103	100	85	115
Y	20	23	19	21	25	25	28	27	26	30

Solution:

X	Change in direction of variable X (D_x)	Y	Change in direction of variable Y(D_y)	D_x x D_y
84		20		
85	+	23	+	+
62	-	19	-	+
48	-	21	+	-
84	+	25	+	+
95	+	25	No change	-
103	+	28	+	+
100	-	27	-	+
85	-	26	-	+
115	+	30	+	+
No of concurrent deviation				7
Disagreement				2

$$r_C = \pm\sqrt{\pm\frac{(2C-N)}{N}}$$

$$= \pm\sqrt{\pm\frac{(2x7-9)}{9}} \quad = +\sqrt{+0.55} \quad = 0.74$$

1. Calculate co-efficient of correlation by the method of concurrent deviation from the following data.

X	60	59	72	51	55	54	65
Y	23	36	10	38	33	44	33

X	Change in direction of variable X (D_x)	Y	Change in direction of variable Y(D_y)	D_x x D_y
60		23		
59	-	36	+	-
72	+	10	-	-
51	-	38	+	-
55	+	33	-	-
54	-	44	+	-
65	+	33	-	-
No of concurrent deviation				0
Disagreement				6

$$r_C = \pm\sqrt{\pm\frac{(2C-N)}{N}}$$

$$= \pm\sqrt{\pm\frac{(2x0-6)}{6}} \quad = -\sqrt{-\frac{-6}{6}} \quad = -\sqrt{1} = -1$$

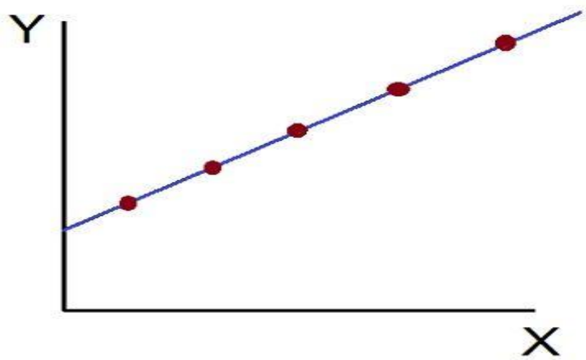
Scatter Diagram Method

Definition: The **Scatter Diagram Method** is the simplest method to study the correlation between two variables wherein the values for each pair of a variable is plotted on a graph in the form of dots thereby obtaining as many points as the number of observations. Then by looking at the scatter of several points, the degree of correlation is ascertained.

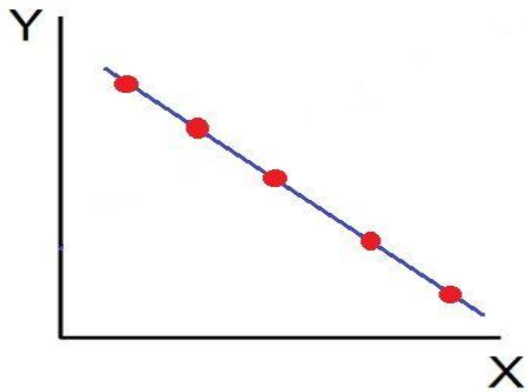
he degree to which the variables are related to each other depends on the manner in which the points are scattered over the chart. The more the points plotted are scattered over the chart, the lesser is the degree of correlation between the variables. The more the points plotted are closer to the line, the higher is the degree of correlation. The degree of correlation is denoted by **"r".**

The following types of scatter diagrams tell about the degree of correlation between variable X and variable Y.

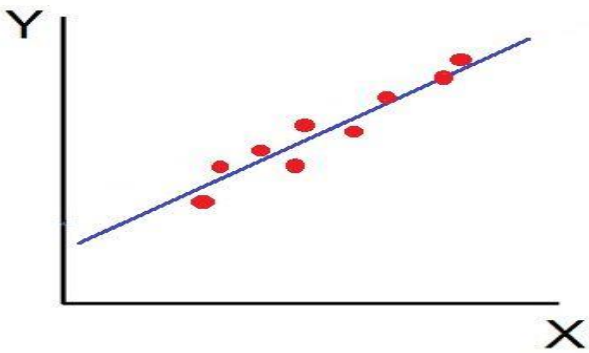
1. **Perfect Positive Correlation (r=+1):** The correlation is said to be perfectly positive when all the points lie on the straight line rising from the lower left-hand corner to the upper right-hand corner.



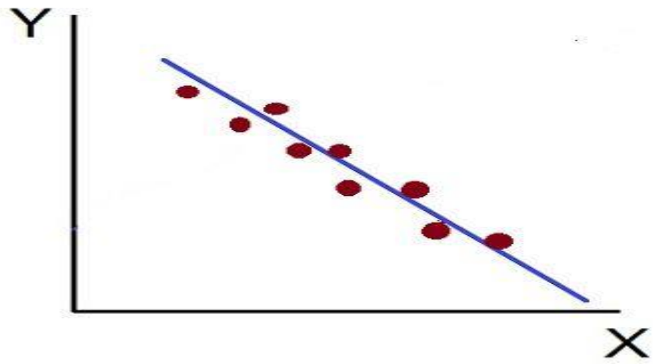
2. **Perfect Negative Correlation (r=-1):** When all the points lie on a straight line falling from the upper left-hand corner to the lower right-hand corner, the variables are said to be negatively correlated.



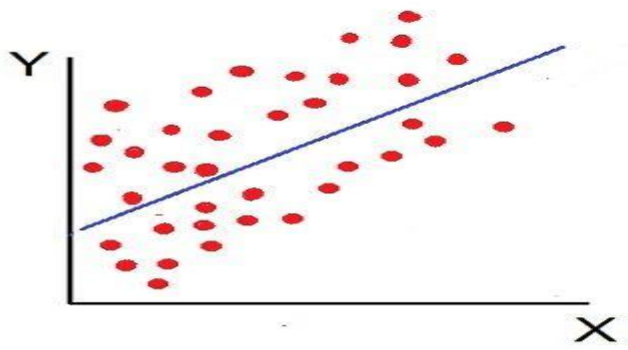
3. **High Degree of +Ve Correlation (r= + High):** The degree of correlation is high when the points plotted fall under the narrow band and is said to be positive when these show the rising tendency from the lower left-hand corner to the upper right-hand corner.



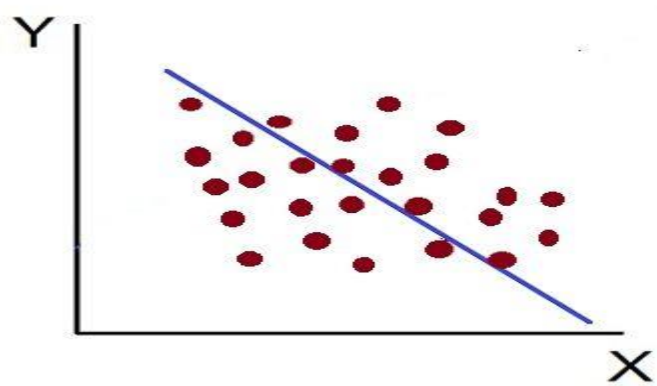
4. **High Degree of –Ve Correlation (r= – High):** The degree of negative correlation is high when the point plotted fall in the narrow band and show the declining tendency from the upper left-hand corner to the lower right-hand corner.



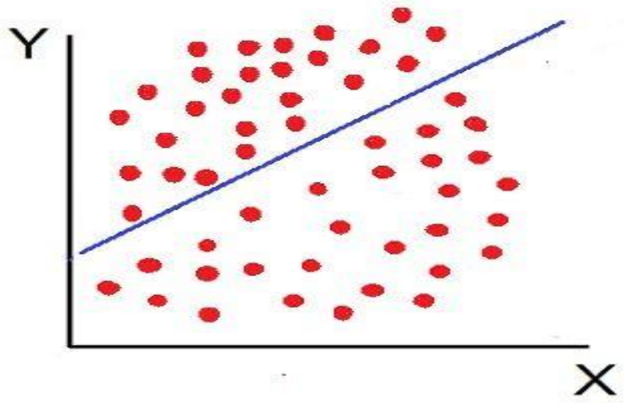
5. **Low degree of +Ve Correlation (r= + Low):** The correlation between the variables is said to be low but positive when the points are highly scattered over the graph and show a rising tendency left-hand corner to the upper right-hand corner.



6. **Low Degree of –Ve Correlation (r= + Low):** The degree of correlation is low and negative when the points are scattered over the graph and the show the falling tendency from the upper left-hand corner to the lower right-hand corner.



7. **No Correlation (r= 0):** The variable is said to be unrelated when the points are haphazardly scattered over the graph and do not show any specific pattern. Here the correlation is absent and



hence **r = 0**.

Thus, the scatter diagram method is the simplest device to study the degree of relationship between the variables by plotting the dots for each pair of variable values given. The chart on which the dots are plotted is also called as a **Dotogram**.