Solution :

Here the discrete frequency distribution is given. Take a = 8

x	f	u = x - a	fu	fu^2
8	2	0	0	0
11	3	3	9	27
17	4	9	36	324
20	1	12	12	144
25	5	17	85	1445
30	7	22	154	3388
35	3	27	81	2187
	N = 25		$\sum fu = 377$	$\sum fu^2 = 7515$

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

$$= \sqrt{\frac{7515}{25} - \left(\frac{377}{25}\right)^2}$$

$$= \sqrt{300.6 - 227.41}$$

$$= \sqrt{73.19}$$

$$= \mathbf{8.55}$$

Variance = $\sigma^2 = (8.55)^2 =$ **73.1025**

Example : 23.

The number of faults on the surface of each of 1000 tiles was distributed as follows.

No. of faults	0	1	2	3	4	5
No. of tiles	760	138	67	25	8	2

Calculate SD.

Solution : Here discrete frequency distribution is given.

x	f	fx	fx^2
0	760	0	0
1	138	138	138
2	67	134	268
3	25	75	225
4	8	32	128
5	2	10	50
	N = 1000	$\sum fx = 389$	$\sum fx^2 = 809$

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

$$= \sqrt{\frac{809}{1000} - \left(\frac{389}{1000}\right)^2}$$

$$= \sqrt{0.809 - 0.1513}$$

$$= \sqrt{0.6577}$$

$$= \mathbf{0.81}$$

Example : 24.

Calculate the SD for the following frequency distribution of heights of 30 persons by direct and step deviation methods.

Height (cm)	155 - 160	160 - 165	165 - 170	170 – 175	175 - 180	180 - 185	185 - 190
No. of persons	1	6	6	6	6	3	2

Solution : Continuous frequency distribution is given.

Direct method :

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

C - I	f	x	fx	fx^2
155 - 160	1	157.5	157.5	24806.25
160 - 165	6	162.5	975	158437.5
165 - 170	6	167.5	1005	168337.5
170 - 175	6	172.5	1035	178537.5
175 - 180	6	177.5	1065	189037.5
180 - 185	3	182.5	547.5	99918.75
185 - 190	2	187.5	375	70312.5
	N = 30		$\sum fx = 5160$	$\sum fx^2 =$ 889387.5

$$\sigma = \sqrt{\frac{889387.5}{30} - \left(\frac{5160}{30}\right)^2}$$

$$= \sqrt{29646.25 - 29584}$$

$$= \sqrt{62.25}$$

$$= \mathbf{7.89\ cm}$$

Step deviation method:

$$\sigma = \left[\sqrt{\frac{\sum f(u')^2}{N} - \left(\frac{\sum fu'}{N}\right)^2}\ \right] \times c\ ,\ \ u' = \frac{x-a}{c}$$

C - I	f	x	u = x-a	$u' = \frac{x-a}{c}$	fu'	$f(u')^2$
155 - 160	1	157.5	0	0	0	0
160 - 165	6	162.5	5	1	6	6
165 - 170	6	167.5	10	2	12	24
170 - 175	6	172.5	15	3	18	54
175 - 180	6	177.5	20	4	24	96
180 - 185	3	182.5	25	5	15	75
185 - 190	2	187.5	30	6	12	72
	N = 30				$\sum fu' = 87$	$\sum f(u')^2 = 327$

$$\sigma = \left[\sqrt{\frac{327}{30} - \left(\frac{87}{30}\right)^2}\ \right] \times 5$$

$$= \left[\sqrt{10.9 - 8.41}\ \right] \times 5$$

$$= \left[\sqrt{2.49}\ \right] \times 5$$

$$= 1.5779 \times 5$$

$$= \mathbf{7.89\ cm}$$

Example : 25.

The mean and standard deviation of a distribution of 100 and 150 items are 50, 5 and 40, 6 respectively. Find the standard deviation of all the 250 items taken together.

Solution : Given

I - Group	II - Group
$n_1 = 100$	$n_2 = 150$
$\bar{x}_1 = 50$	$\bar{x}_2 = 40$
$\sigma_1 = 5$	$\sigma_2 = 6$

Here,

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{(100 \times 50) + (150 \times 40)}{100 + 150} = \frac{5000 + 6000}{250} = \mathbf{44}$$

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

$$\mathbf{d_1 = 50 - 44 = 6}$$

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

$$\mathbf{d_2 = 40 - 44 = -4}$$

$$\sigma_c = \sqrt{\frac{n_1\left(\sigma_1{}^2 + d_1{}^2\right) + n_2\left(\sigma_2{}^2 + d_2{}^2\right)}{n_1 + n_2}}$$

$$= \sqrt{\frac{100(5^2 + 6^2) + 150(6^2 + (-4)^2)}{100 + 150}}$$

$$= \sqrt{\frac{(100 \times 61) + (150 \times 52)}{250}}$$

$$= \sqrt{\frac{6100 + 7800}{250}}$$

$$= \sqrt{\frac{13900}{250}}$$

$$= \sqrt{55.6}$$

$$= \mathbf{7.46}$$

Example : 26. Given the following data, find the combined SD of the groups together.

	I -Group (Males)	II - Group (Females)
Size	50	40
Mean	63	54
Variance	81	36

Solution: Given,

$n_1 = 50 \quad n_2 = 40$

$\bar{x}_1 = 63 \quad \bar{x}_2 = 54$

$\sigma_1{}^2 = 81 \quad \sigma_2{}^2 = 36$

Here,

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

$$= \frac{(50 \times 63) + (40 \times 54)}{50 + 90} = \frac{3150 + 2160}{90} = \mathbf{59}$$

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

$$\mathbf{d_1 = 63 - 59 = 4}$$

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

$$\mathbf{d_2 = 54 - 59 = -5}$$

$$\sigma_c = \sqrt{\frac{n_1(\sigma_1{}^2 + d_1{}^2) + n_2(\sigma_2{}^2 + d_2{}^2)}{n_1 + n_2}}$$

$$= \sqrt{\frac{50(81 + 16) + 40(36 + 25)}{50 + 40}} = \sqrt{\frac{(50 \times 97) + (40 \times 61)}{90}}$$

$$= \sqrt{\frac{4850 + 2440}{90}} = \sqrt{\frac{7290}{90}} = \sqrt{81} = \mathbf{9}$$

Example : 27.

If mean and S.D. of a distribution are 20 and 5 respectively. Find the coefficient of variation.

Solution:

$$CV = \frac{\sigma}{\bar{x}} \times 100 = \frac{5}{20} \times 100 = \mathbf{25\%}$$

Example : 28.

If mean and C.V of a distribution are 56 and 75% respectively. Find the S.D.

Solution :

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

$$75 = \frac{\sigma}{56} \times 100$$

$$\therefore \sigma = \mathbf{42}$$

Example : 29.

If C.V. and S.D. of a distribution are 41% and 22 respectively. Find the mean.

Solution :

$$\mathrm{CV} = \frac{\sigma}{\bar{x}} \times 100$$

$$41 = \frac{22}{\bar{x}} \times 100$$

$$\therefore \ \bar{x} = \mathbf{53.66}$$

Example : 30.

The following data refers to the dividend (%) paid by 2 companies A and B over the last five years.

Company A	2	3	5	8	4
Company B	6	3	8	10	7

Calculate the coefficient of variation and comment

Solution:

C.V. for Company A

$$\mathrm{CV} = \frac{\sigma}{\bar{x}} \times 100$$

$$\bar{x} = \frac{\sum x}{n} = \frac{22}{5} = 4.4$$

x	2	3	5	8	4	Σx= 22
x^2	4	9	25	64	16	$\Sigma x^2 = 118$

$$75 = \frac{\sigma}{56} \times 100$$

$$\therefore \sigma = \mathbf{42}$$

Example : 29.

If C.V. and S.D. of a distribution are 41% and 22 respectively. Find the mean.

Solution :

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

$$41 = \frac{22}{\bar{x}} \times 100$$

$$\therefore \bar{x} = \mathbf{53.66}$$

Example : 30.

The following data refers to the dividend (%) paid by 2 companies A and B over the last five years.

Company A	2	3	5	8	4
Company B	6	3	8	10	7

Calculate the coefficient of variation and comment

Solution:

C.V. for Company A

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

$$\bar{x} = \frac{\sum x}{n} = \frac{22}{5} = 4.4$$

x	2	3	5	8	4	$\Sigma x = 22$
x^2	4	9	25	64	16	$\Sigma x^2 = 118$

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

$$= \sqrt{\frac{118}{5} - \left(\frac{22}{5}\right)^2}$$

$$= \sqrt{23.6 - 19.36}$$

$$= \sqrt{4.24}$$

$$= 2.0591$$

$$\therefore \text{CV(A)} = \frac{\sigma}{\bar{x}} \times 100$$

$$= \frac{2.0591}{4.4} \times 100$$

$$= \mathbf{46.8\%}$$

C.V. for Company B

$$\bar{x} = \frac{\sum x}{n} = \frac{34}{5} = 6.8$$

x	6	3	8	10	7	$\Sigma x = 34$
x^2	36	9	64	100	49	$\Sigma x^2 = 258$

$$\sigma = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} = \sqrt{\frac{258}{5} - (6.8)^2}$$

$$= \sqrt{51.6 - 46.24} = \sqrt{5.36} = 2.32$$

$$\therefore \text{CV(B)} = \frac{2.32}{6.8} \times 100 = \mathbf{34.12\%}$$

As CV(A) > CV(B), company B is more consistent in payment of dividend than company A.

Example : 31.

Goals scored by two teams A and B in foot ball season are as follows:

No. of goals scored in a match (x)	No. of Matches	
	Team ' A'	Team ' B'
0	22	11
1	8	10
2	7	8
3	8	7
4	3	4

Find which team is more consistent in scoring.

Solution :

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

x	f	fx	fx^2
0	22	0	0
1	8	8	8
2	7	14	28
3	8	24	72
4	3	12	48
	N = 48	$\Sigma fx=58$	$\Sigma fx^2 = 156$

$$\bar{x} = \frac{\sum fx}{N} = \frac{58}{48} = 1.2083$$

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

$$= \sqrt{\frac{156}{48} - (1.2083)^2}$$

$$= \sqrt{3.25 - 1.4601} = \sqrt{1.7899} = \mathbf{1.3379}$$

$$\therefore CV(A) = \frac{1.3379}{1.2083} \times 100 = \mathbf{110.72\%}$$

To compute CV for Team B:

x	f	fx	fx²
0	11	0	0
1	10	10	10
2	8	16	32
3	7	21	63
4	4	16	64
	N = 40	Σfx=63	Σfx² = 169

$$\bar{x} = \frac{63}{40} = 1.575$$

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

$$= \sqrt{\frac{169}{40} - (1.575)^2}$$

$$= \sqrt{4.225 - 2.4806}$$

$$= \sqrt{1.7444}$$

$$= \mathbf{1.3207}$$

$$\therefore \text{CV(B)} = \frac{1.3207}{1.175} \times 100 = \mathbf{83.85\%}$$

As CV(B) < CV(A). Team B is considered to be more consistent in scoring.

Example: 32. Compare the variations and averages for the following distribution regarding expenditure on food of families in two different places.

Expenditure per month	No. of families	
	Place 'A'	Place 'B'
600 – 800	25	32
800 – 1000	42	65
1000 – 1200	68	84
1200 – 1400	152	124
1400 – 1600	53	30

Solution :

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

To compute CV for place ' A'

Take, a = 1100 and C = 200

C.I.	f	x	$u' = \frac{x-a}{c}$	fu'	$f(u')^2$
600-800	25	700	-2	-50	100
800-1000	42	900	-1	-42	42
1000-1200	68	1100	0	0	0
1200-1400	152	1300	1	152	152
1400-1600	53	1500	2	106	212
	N=340			$\Sigma fu'$=166	$\Sigma f(u')^2$ =506

$$\bar{x} = a + \left(\frac{\sum fu'}{N}\right) \times c$$

$$= 1100 + \left(\frac{166}{340}\right) \times 200$$

$$= 1100 + 97.65$$

$$= \text{Rs.}1197.65$$

$$\sigma = \sqrt{\frac{\sum f(u')^2}{N} - \left(\frac{\sum fu'}{N}\right)^2} \times c$$

$$= \sqrt{\frac{506}{340} - \left(\frac{166}{340}\right)^2} \times 200$$

$$= \sqrt{1.489 - 0.238} \times 200$$

$$= \sqrt{1.251} \times 200$$

$$= \text{Rs. } 223.696$$

$$CV(A) = \frac{223.696}{1197.65} \times 100$$

$$= \mathbf{18.67\%}$$

To compute CV for place' B'

Take, a = 1100 and C = 200

x	f	$u' = \frac{x-a}{c}$	fu'	$f(u')^2$
700	32	-2	-64	128
900	65	-1	-65	65
1100	84	0	0	0
1300	124	1	124	124
1500	30	2	60	120
	N=335		$\Sigma fu' = 55$	$\Sigma f(u')^2 = 437$

$$\bar{x} = a + \left(\frac{\sum fu'}{N}\right) \times c$$

$$= 1100 + \left(\frac{55}{335}\right) \times 200 = \text{Rs.}\,1132.84$$

$$\sigma = \sqrt{\frac{437}{335} - \left(\frac{55}{335}\right)^2} \times 200$$

$$= \sqrt{1.3045 - 0.0269} \times 200$$

$$= \sqrt{1.2776} \times 200 = \text{Rs.}\,226.062$$

$$\text{CV(B)} = \frac{226.062}{1132.84} \times 100 = \mathbf{19.95\%}$$

As CV(A) < CV(B). Expenditure pattern is consistent in place A. Also average expenditure of place A is more than place B.

Example : 33.

Calculate coefficient of variation from the following data.

Income (Rs)	Less than 10000	Less than 15000	Less than 20000	Less than 25000	Less than 30000
No. of families	8	20	38	50	60

Solution:

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

Given LCF distribution, initially form a simple frequency distribution.

Take, a = 7500, C = 5000

C.I.	LCF	f	x	$u' = \frac{x - A}{c}$	fu'	$f(u')^2$
5000-10000	8	8	7500	0	0	0
10000-15000	20	12	12500	1	12	12
15000-20000	38	18	17500	2	36	72
20000-25000	50	12	22500	3	36	108
25000-30000	60	10	27500	4	40	160
		N=60			124	352

$$\bar{x} = a + \left(\frac{\sum fu'}{N}\right) \times c = 7500 + \left(\frac{124}{60}\right) \times 5000 = \text{Rs. } 17833.33$$

$$\sigma = \sqrt{\frac{\sum f(u')^2}{N} - \left(\frac{\sum fu'}{N}\right)^2} \times c = \sqrt{\frac{352}{69} - \left(\frac{124}{60}\right)^2} \times 5000$$

$$= \sqrt{5.8667 - 4.2711} \times 5000$$

$$= \sqrt{1.5956} \times 5000$$

$$= \text{Rs. } 6315.85$$

$$\therefore CV = \frac{6315.85}{17833.33} \times 100 = \mathbf{35.42\%}$$

Example : 34. An analysis of monthly wages paid to workers, gave the following results.

	Firm 'A'	Firm 'B'
Number of wage earners	500	600
Average monthly wage (Rs.)	5600	6500
S.D. of wage (Rs.)	223.5	231.3

i) Which firm A or B pays a larger amount of monthly wage ?

ii) In which firm A or B is there greater variability in wages ?

Solution :

Firm 'A'	Firm 'B'
$\bar{x} = 5600$	$\bar{x} = 6500$
$n = 500$	$n = 600$
$\sigma = 223.5$	$\sigma = 231.3$
$\therefore$ Total wage $= n\bar{x}$	$\therefore$ Total wage $= n\bar{x}$
$(\Sigma x) \quad = 2800000$	$(\Sigma x) \quad = 3900000$

$$CV(A) = \frac{\sigma}{\bar{x}} \times 100 = \frac{223.5}{5600} \times 100 = \mathbf{3.99\%} \qquad CV(B) = \frac{\sigma}{\bar{x}} \times 100 = \frac{231.3}{6500} \times 100 = \mathbf{3.56\%}$$

i) Firm B pays larger amount of monthly wage.

ii) As CV(A) > CV(B), there is greater variability in wages of firm A.

Example : 35.

Following is data regarding marks obtained by two students X and Y

	Student 'X'	Student 'Y'
Mean marks	80	50
S.D. of Marks	4	3

i) Who is better?

ii) Who is more consistent?

Solution:

Student 'X'	Student 'Y'
$\bar{x} = 80$	$\bar{y} = 50$
$\sigma_x = 4$	$\sigma_y = 3$

$$CV(X) = \frac{\sigma_x}{\bar{x}} \times 100 = \frac{4}{80} \times 100 = \mathbf{5\%} \qquad CV(Y) = \frac{\sigma_y}{\bar{y}} \times 100 = \frac{3}{50} \times 100 = \mathbf{6\%}$$

i) As $\bar{x} > \bar{y}$, Therefore X is better.

ii) As CV(X) < CV(Y), Therefore student X is more consistent.

Questions

1. What is dispersion?
2. Mention the different measures of dispersion.
3. What are the desired qualities of a good measure of dispersion ?
4. Define Range, Q.D, M.D and S.D.
5. What are absolute and relative measures of dispersion ?
6. Mention two objectives of measuring variation.
7. State a merit of Range.
8. State a demerit of Range.
9. Why Q.D. is unaffected by abnormal extreme values ?
10. Name the measure of dispersion which can be easily calculated for distributions with open end class intervals.
11. Standard deviation is the least of all root mean square deviations. Give reason.
12. State the differences between mean deviation and standard deviation.
13. What is coefficient of variation ?
14. State two merits of Q.D, M.D and S.D.
15. State two demerits of Q.D, M.D and S.D.
16. What is variance?
17. If S.D. = 4 cm, find variance.
18. If variance = 16 Sq. feet, find S.D.
19. State all the properties of standard deviation.

Exercise problems

1. Calculate Range and Coefficient of range for the following data

   139, 150, 151, 151, 157, 158, 160, 161, 162, 165, 167, 175

   Ans : 36, 0.115

2. Calculate Inter quartile range and semi inter quartile range for the following data 11, 15, 16, 9, 14, 19, 10, 12, 8, 17, 20, 23, 22

   Ans : 9, 4.5

3.Compute coefficient of Q.D. from the data given below.

x	2	4	6	8	10	12
f	3	5	10	12	6	4

Ans : 0.25

4. Calculate an appropriate measure of dispersion from the following data.

Wages (Rs)/day	Less than 340	340 – 370	370 – 400	400 – 430	430 & above
No. of workers	14	62	99	18	7

Ans : Q.D = 17.5

5. Find semi-inter quartile range for the following distribution.

Age (years)	Less than 25	Less than 30	Less than 35	Less than 40	Less than 45	Less than 50	Less than 55
No. of Employees	10	25	75	130	170	189	200

Ans : 5

6. If lower quartile and inter quartile range of data are 31.75 and 30.88, find the upper quartile and Q.D. Ans : 62.63, 15.44

7. Calculate the mean deviation from mean from the following data.

Variable	10	11	12	13
f	3	12	18	12

Ans : M.D. = 0.71

8. Calculate the mean deviation from the mean for the following distribution.

C - I	2 – 4	4 – 6	6 – 8	8 – 10	10 – 12
f	3	5	8	4	2

Ans : 1.801

9. Calculate coefficient of mean deviation from median for the following distribution.

Age (Years)	16	17	18	19	20	21	22	23	24
No. of persons	4	5	7	12	20	13	5	0	4

Ans: 0.0685

10. Calculate mean deviation from median from the following data.

Marks	More than 0	More than 10	More than 20	More than 30	More than 40	More than 50	More than 60	More than 70
No. of students	100	95	87	80	68	40	20	10

Ans : 19.72

11. Find out mean deviation from mean and median for the data given below.

C-I	200-250	250-300	300-350	350-400	400-450	450-500	500-550	550-600	600-650	650-700	700-750	750-800
f	7	13	15	24	36	50	25	10	8	6	4	2

Ans : 86.4, 86.25

12. Find standard deviation of the following data.

25, 50, 45, 30, 70, 42, 36, 48, 34, 60 Ans : 13.1

13. Calculate standard deviation for the following distribution.

x	4	5	6	7	8	9	10
f	6	12	15	28	29	14	15

Ans : 1.54

14. The means of 2 samples of sizes 50 and 100 are 54.1 and 50.3 respectively and the standard deviations are 8 and 7. Obtain the standard deviation of the combined sample. Ans : 7.56

15. Find standard deviation and variance from the following data.

C.I.	0-6	6 -12	12-18	18-24	24-30	30-36	36-42
f	19	25	36	72	51	43	28

Ans : 9.9

16. Calculate standard deviation for the following distribution.

C.I.	30-39	40-49	50-59	60-69	70-79	80-89	90-99
f	1	4	14	20	22	12	2

Ans : 12.51

17. The number of runs scored by two batsmen A and B in differentinnings is as follows :

A	12	115	6	73	7	19	119	36	84	29
B	47	12	76	42	4	51	37	48	13	0

Who is the better run scorer? Who is more consistent?

Ans : Batsman A is better run scorer, Batsman B is more consistent.

18. If the mean and standard deviation of a set of 100 values are 50 and 4 respectively. Find the coefficient of variation. Ans : 8

19. Find standard deviation, variance and coefficient of variation from the following data.

Wage (Rs)	Less than 10	Less than 20	Less than 30	Less than 40	Less than 50	Less than 60	Less than 70	Less than 80
No. of persons	12	30	65	107	157	202	222	230

Ans: $\sigma = 17.26, \sigma^2 = 297.91$, CV = 42.7%

20. Following is the distribution of weights of students. Compare their coefficient of variations.

Weights (kg)		20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No. of Students	Class A	7	10	20	18	7
	Class B	5	9	21	15	6

Ans : CV(A) = 25%, CV(B) = 23.5%

21. If coefficient of variation and standard deviation of a distribution are 75% and 15 respectively, find its mean. Ans : $\bar{x} = 20$

22. The arithmetic mean of marks scored by 3 students A, B & C in a examination are 50, 44, 20 respectively. The standard deviations of marks are respectively 15, 11 and 3. Who is the most consistent scorer? Ans : Student C is a consistent scorer.

Moments

Moments are popularly used to describe the various characteristics of a frequency distribution such as average, dispersion, skewness and kurtosis.

Types of moments : Moments are classified as,

1. Raw moments
2. Central moments.

Raw Moments :

Moments about any arbitrary point 'a' are known as raw moments. They are denoted by μ_r'. It is given by,

$$\mu_r' = \frac{\sum (x-a)^r}{n}, r = 1, 2, \dots\dots\dots$$

Central moments :

Moments about the mean are known as central moments. They are denoted by μ_r. It is given by,

$$\mu_r = \frac{\sum (x-\bar{x})^r}{n}, r = 1, 2, \dots\dots\dots$$

The first moment about mean is zero.

i. e., $\mu_1 = \frac{\sum (x-\bar{x})}{n} = 0$

The first moment about zero is mean.

i. e., $\mu_1 = \frac{\sum (x-0)}{n} = \bar{x}$

The second moment about mean is variance

i. e., $\mu_2 = \frac{\sum (x-\bar{x})^2}{n} = \sigma^2$

The third moment about mean is a measure of skewness.

i. e., $\mu_3 = \frac{\sum (x-\bar{x})^3}{n}$

The fourth moment about mean is measure of kurtosis.

i. e., $$\mu_4 = \frac{\sum (x - \bar{x})^4}{n}$$

Karl-Pearson' s (β) and Gamma (γ) coefficients based on moments :

Prof. Karl Pearson defined the following coefficients based on the first four central moments,

$$\beta_1 = \frac{{\mu_3}^2}{{\mu_2}^3}, \qquad \beta_2 = \frac{\mu_4}{{\mu_2}^2}$$

Also, $\gamma_1 = \sqrt{\beta_1}$ and $\gamma_2 = \beta_2 - 3$

β_1 tells us whether a distribution is skewed or whether it is symmetrical and β_2 tells us the difference between a symmetrical curve and a normal curve. In other words, β_1 is a measure of skewness based on moments and β_2 is a measure of Kurtosis based on moments.

SKEWNESS

Introduction :

In the previous topics of measures of central tendency and measures of dispersion we have learnt to find a single value that represents the entire mass of data and the extent of spread of values in a data from that single value. But we may come across frequency distributions which differ widely in their nature and yet may have the same measure of central tendency and measure of dispersion

Eg : 1st frequency distribution 2nd frequency distribution.

x	f
3	10
9	30
15	60
21	60
27	30
33	10

x	f
3	10
9	40
15	30
21	90
27	20
33	10

For the above two frequency distributions we see that $\overline{x} = 18$ and $\sigma = 7.22$, but the distributions vary very much in their structure. Thus measures of central tendency and measures of dispersion are inadequate to characterise a frequency distribution.

In the first frequency distribution we see that frequencies are equal on either side of the central value (symmetric) whereas in the second frequency distribution frequencies are unequal on either side of the central value. Thus there arises a need to study whether the scatter of value around the central value are symmetric or not symmetric. The distribution which is not symmetric is called as a skewed distribution.

Definition: Skewness refers to ' lack of symmetry' or ' asymmetry'.

Properties of Symmetric distribution :

In a symmetric distribution we have,

1) Mean, Median and Mode are equal.
2) Median is equidistant from the lower and upper quartiles.

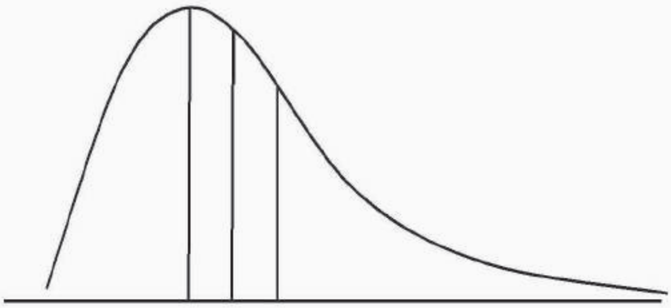
   i.e., $Q_3 - Q_2 = Q_2 - Q_1$ (here Q_2 is median)
3) When the symmetric distribution is plotted on a graph, we get a bell shaped curve.

Types of Skewness :

1) Positively skewed distribution
2) Negatively skewed distribution

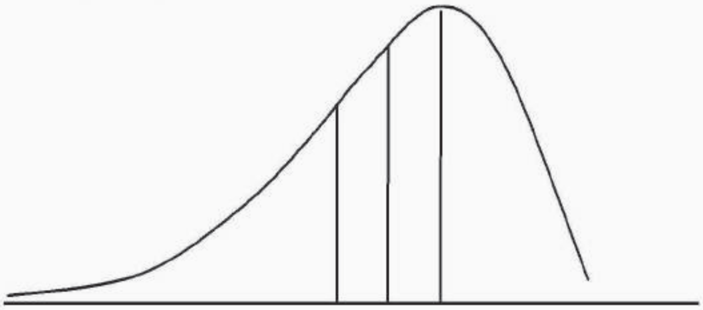
Diagrams showing types of skewness along with positions of Mean, Median and Mode :

1. Positively Skewed Curve



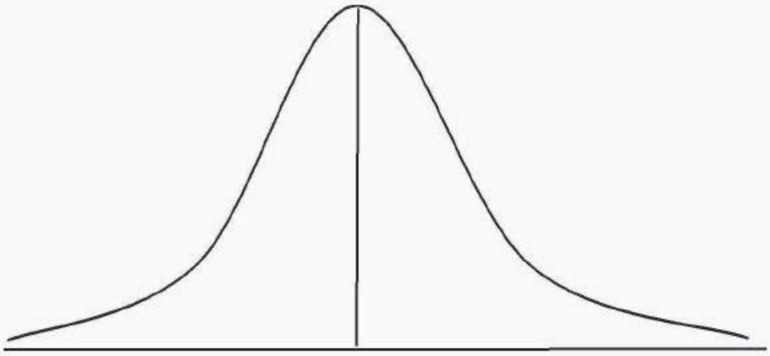
$Z < M < \overline{X}$

2. Negatively skewed curve



$\overline{X} < M < Z$

3. No Skewness (Symmetric)



In this case $\overline{X} = M = Z$

Measures of skewness :

Measures of skewness may be absolute or relative. As absolute measures are expressed in the units of the distribution, it cannot be used for comparison with another distribution with different units. Thus, for purposes of comparison it is necessary to have relative measures of skewness. These relative measures are known as coefficient of skewness. Some important measures of skewness are :

1) Karl Pearson' s coefficient of skewness.
2) Bowley' s Coefficient of skewness
3) Measures of skewness based on moments.

Karl – Pearson' s Coefficient of Skewness :

Karl–Pearson' s coefficient of skewness given by $\frac{\bar{x}-Z}{\sigma}$,

Where, $\bar{x}$ – mean, Z - mode and σ- standard deviation.

If mode is ill-defined, then skewness is estimated on the basis of the empirical relationship which exists among $\bar{x}$, M, Z.

Karl Pearson' s coefficient of Skewness = $\frac{3(\bar{x}-M)}{\sigma}$,

Where, M – median.

Bowley' s Coefficient of Skewness :

This measure of skewness is based on quartiles.

Bowley' s coefficient of skewness = $\frac{Q_3+Q_1-2Q_2}{Q_3-Q_1}$

Where, Q_1 = Lower Quartile (First Quartile)
Q_2 = Median (Second Quartile)
Q_3 = Upper Quartile (Third Quartile)

Measure of skewness based on moments :

It is calculated by using the third moment about the mean. It is given by

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

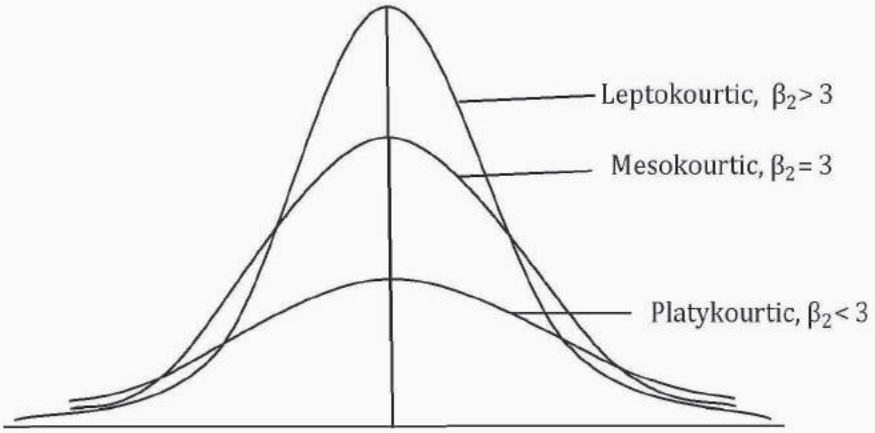
Where, μ_3 = Third central moment

μ_2= Second central moment

Also $\gamma_1 = \sqrt{\beta_1}$ is a measure of skewness.

KURTOSIS

So far we have studied there measures, viz, central tendency, dispersion and skewness to describe the characteristics of a frequency distribution. However these three measures are not enough to characterize a distribution completely. The following diagram will clear this point.


Leptokourtic, $\beta_2 > 3$
Mesokourtic, $\beta_2 = 3$
Platykourtic, $\beta_2 < 3$

All the three curves are symmetrical about mean and also have the same variation (range), but the shape and peakedness of the curves are different. This study of peakedness of the frequency curve is called kurtosis.

Definition : Kurtosis means peakedness in the region of mode of a frequency curve.

Types of Kurtosis :

In the above diagram, curve which is neither flat nor peaked is known as normal curve (mesokurtic) and is taken as standard to measure kurtosis. Curve which is more peaked than normal curve is called leptokurtic curve and curve which is less peaked than normal curve is known as platykurtic curve.

Measures of Kurtosis based on moments :

Karl pearson gave the coefficient β_2 or derivative γ_2 as measures of kurtosis. They are given as follows,

$\beta_2 = \frac{\mu_4}{{\mu_2}^2}$ where μ_4 is the fourth central moment and μ_2 is the second central moment

Also, $\gamma_2 = \beta_2 - 3$

For a mesokurtic distribution , $\beta_2 = 3$ and $\gamma_2 = 0$

For a leptokurtic distribution , $\beta_2 > 3$ and $\gamma_2 > 0$

And for a platykurtic distribution , $\beta_2 < 3$ and $\gamma_2 < 0$

Problems

Example : 1.

Calculate Karl Pearson's coefficient of skewness from the following data.

Value	6	12	18	24	30	36	42
Frequency	4	7	9	18	15	10	5

Solution : Discrete distribution is given

$$\text{Karl Pearson's coefficient of skewness } = \frac{\bar{x} - Z}{\sigma}$$

$$\text{Where, } \bar{x} = \frac{\sum fx}{N} \quad \text{and} \quad \sigma = \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2}$$

x	f	fx	$fx^2 = (fx)(x)$
6	4	24	144
12	7	84	1008
18	9	162	2916
24	18	432	10368
30	15	450	13500
36	10	360	12960
42	5	210	8820
	N=68	$\sum fx = 1722$	49716

Thus $\bar{x} = \frac{\sum fx}{N} = \frac{1722}{68} = 25.32$

$$\sigma = \sqrt{\frac{49716}{68} - \left(\frac{1722}{68}\right)^2}$$

$$= \sqrt{731.118 - 641.102} = \sqrt{90.016} = 9.488$$

Z = mode = value of x corresponding to the highest frequency = 24

$\therefore$ Karl Pearson's coefficient of skewness = $\frac{25.32-24}{9.488}$ = **0.139**

As coefficient of skewness is positive, the distribution is positively skewed.

Example : 2.

Calculate Pearson' s coefficient of skewness from the data given below.

Life (Hrs)	300 - 400	400 - 500	500 - 600	600 - 700	700 - 800	800 - 900	900 - 1000	1000-1100	1100-1200
No. of bulbs	14	46	58	76	68	62	48	22	6

Solution :

Continuous frequency distribution is given

Karl Pearson's coefficient of skewness $= \frac{\bar{x} - Z}{\sigma}$

Let a = 350, here, c = 100

C.I.	f	x	u=x-A	$u^{'} = \frac{u}{c}$	$fu^{'}$	$f(u^{'})^2$
300-400	14	350	0	0	0	0
400 – 500	46	450	100	1	46	46
500 – 600	58	550	200	2	116	232
600 – 700	76	650	300	3	228	684
700 – 800	68	750	400	4	272	1088
800 – 900	62	850	500	5	310	1550
900 – 1000	48	950	600	6	288	1728
1000 – 1100	22	1050	700	7	154	1078
1100 – 1200	6	1150	800	8	48	384
	N=400				$\sum fu^{'}=1462$	$\sum f(u^{'})^2= 6790$

Thus, $\bar{x} = a + \left(\frac{\sum fu^{'}}{N} \times c\right)$

$$= 350 + \left(\frac{1462}{400} \times 100\right)$$

= 350 +365.5 = 715.5 Hours.

$$\sigma = \sqrt{\frac{\sum f(u^{'})^2}{N} - \left(\frac{\sum fu^{'}}{N}\right)^2} \times c \quad = \sqrt{\frac{6790}{400} - \left(\frac{1462}{400}\right)^2} \times 100$$

$$= \sqrt{16.975 - 13.359} \times 100 = \sqrt{3.616} \times 100 = \text{Rs. } 190.15$$

To find mode:

Step 1 : Modal class = class corresponding to highest frequency

= (600 – 700)

Step 2: $Z = l + \left[\frac{(f_m - f_1)}{2f_m - f_1 - f_2} \times C\right]$

$$= 600 + \left[\frac{(76 - 58)}{2 \times 76 - 58 - 68} \times 100\right]$$

$$= 600 + \left[\frac{1800}{26}\right] = 600 + 69.231 = \mathbf{669.231 hours}$$

$$\text{Karl Pearson's coefficient of skewness} = \frac{715.5-669.231}{190.15} = \mathbf{0.243.}$$

As the coefficient of skewness is positive, the distribution is positively skewed.

Example 3:

Calculate Pearson's coefficient of skewness from the following.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of students	10	40	20	0	10	40	16	14

Solution: The distribution given is bimodal, Thus, KP coefficient of skewness is found by using the formula, $\frac{3(\bar{x}-M)}{\sigma}$

Take, a =35 and c = 10

C.I.	f	x	u =x-a	$u' = \frac{u}{c}$	fu'	$f(u')^2$
0 – 10	10	5	-30	-3	- 30	90
10 – 20	40	15	- 20	- 2	- 80	160
20 – 30	20	25	- 10	- 1	- 20	20
30 – 40	0	35	0	0	0	0
40 – 50	10	45	10	1	10	10
50 – 60	40	55	20	2	80	160
60 – 70	16	65	30	3	48	144
70 – 80	14	75	40	4	56	224
	N-150				$\sum fu' = 64$	$\sum f(u')^2 = 808$

$$\bar{x} = a + \left(\frac{\sum fu'}{N} \times c\right)$$

$$= 35 + \left(\frac{64}{150} \times 10\right)$$

$$= 35 + 4.267 = 39.267.$$

$$\sigma = \sqrt{\frac{\sum f(u')^2}{N} - \left(\frac{\sum fu'}{N}\right)^2} \times c$$

$$= \sqrt{\frac{808}{150} - \left(\frac{64}{150}\right)^2} \times 10$$

$$= \sqrt{5.387 - 0.182} \times 10$$

$$= \sqrt{5.205} \times 10$$

$$= 22.81$$

To find median

Step 1 : Find LCF

C - I	f	LCF
0-10	10	10
10-20	40	50
20-30	20	70
30-40	0	70
40-50	10	80
50-60	40	120
60-70	16	136
70-80	14	150

Step 2: Median class = class corresponding to $\left(\frac{N}{2}\right)^{th}$ observation.

= class corresponding to $(75)^{th}$ observation.

= (40 – 50)

Step 3: $M = l + \left[\frac{\left(\frac{N}{2} - cf\right)}{f} \times c\right] = 40 + \left[\frac{(75 - 70)}{10} \times 10\right] = 40 + 5 = 45$

∴ Karl Pearson' s coefficient of skewness is = $\frac{3(39.267-45)}{22.81}$ = - 0.754

As coefficient of skewness is negative, the distribution is negatively skewed.

Example : 4.

Calculate Bowley' s coefficient of skewness for the following data.

Capital (lakh Rs)	1 – 5	6 – 10	11–15	16–20	21–25	26–30	31–35
No. of companies	20	30	60	80	50	25	15

Solution:

Convert inclusive CI' s to exclusive type.

Bowley' s coefficient of skewness = $\frac{Q_3+Q_1-2Q_2}{Q_3-Q_1}$

Step 1 : Find LCF

C.I.	f	LCF	
0.5-5.5	20	20	
5.5-10.5	30	50	
10.5-15.5	60	110	Q_1 Class
15.5-20.5	80	190	Q_2 Class
20.5-25.5	50	240	Q_3 Class
25.5-30.5	25	265	
30.5-35.5	15	280	
	N=280		

Step 2 : Q_1 Class = Class corresponding to $\left(\frac{N}{4}\right)^{th}$ observation.

= Class corresponding to $(70)^{th}$ observation.

= (10.5 – 15.5)

Q_2 Class = Class corresponding to $\left(\frac{2N}{4}\right)^{th}$ observation.

= Class corresponding to $(140)^{th}$ observation.

= (15.5 – 20.5)

Q_3 class = Class corresponding to $\left(\frac{3N}{4}\right)^{th}$ observation.

= Class corresponding to $(210)^{th}$ observation

= (20.5 – 25.5)

Step 3: $Q_1 = l + \left[\frac{\left(\frac{N}{4} - cf\right)}{f} \times c\right]$

$= 10.5 + \left[\frac{(70 - 50)}{60} \times 5\right]$

$= 10.5 + 1.67$

= **Rs.12.17 lakh.**

$Q_2 = l + \left[\frac{\left(\frac{2N}{4} - cf\right)}{f} \times c\right]$

$= 15.5 + \left[\frac{(140 - 110)}{80} \times 5\right]$

$= 15.5 + 1.875$

= **Rs.17.375 lakh.**

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

$= 20.5 + \left[\frac{(210 - 190)}{50} \times 5\right]$

$= 20.5 + 2$

= **Rs.22. 25 lakh.**

Thus Bowley' s Coefficient of skewness = $\frac{22.5+12.17-2(17.375)}{22.5-12.17}$

$= \frac{34.67-34.75}{10.33}$

$= \frac{-0.08}{10.33}$

$= -0.0077$

The distribution is negatively skewed.

Example : 5. Find a measure of skewness based on quartiles for the following data.

Age (years)	Below 20	20-25	25-30	30-35	35-40	40-45	45-50	50 & above
No. of employees	13	29	46	60	112	94	45	21

Solution:

Step 1 : Find LCF

Age (years)	f	LCF	
Below 20	13	13	
20 – 25	29	42	
25 – 30	46	88	
30 – 35	60	148	Q_1 Class
35 – 40	112	260	Q_2 Class
40 – 45	94	354	Q_3 Class
45 – 50	45	399	
50 & above	21	420	
	N = 420		

Step 2:

Q_1 Class = Class corresponding to $\left(\frac{N}{4}\right)^{th}$ observation

= Class corresponding to $(105)^{th}$ observation

= (30 – 35)

Q_2 Class = Class corresponding to $\left(\frac{2N}{4}\right)^{th}$ observation

= Class corresponding to $(210)^{th}$ observation

= (35 – 40)

Q_3 Class = Class corresponding to $\left(\frac{3N}{4}\right)^{th}$ observation

= Class corresponding to $(315)^{th}$ observation

= (40 – 45)

Step 3:

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

$$= 30 + \left[\frac{(105 - 88)}{60} \times 5\right] = \mathbf{31.417}$$

$$Q_2 = l + \left[\frac{\left(\frac{2N}{4} - cf\right)}{f} \times c\right]$$

$$= 35 + \left[\frac{(210 - 148)}{112} \times 5\right] = \mathbf{37.768}$$

$$Q_3 = l + \left[\frac{\left(\frac{3N}{4} - cf\right)}{f} \times c\right] = 40 + \left[\frac{(315 - 260)}{94} \times 5\right] = \mathbf{42.926}$$

$$\therefore \text{Bowley' s Coefficient of Skewness} = \frac{Q_3+Q_1-2Q_2}{Q_3-Q_1}$$

$$= \frac{42.926+31.417-2(37.768)}{42.926-31.417}$$

$$= \frac{-1.193}{11.509}$$

$$= \mathbf{-0.104}$$

As coefficient of skewness is negative, the distribution is negatively skewed.

Example : 6.

Compute Bowley' s coefficient of skewness for the following data.

No. of Children / Couple	0	1	2	3	4	5
No. of families	6	15	25	8	4	1

Solution:

Discrete frequency distribution is given

Bowley' s Coefficient of skewness = $\frac{Q_3+Q_1-2Q_2}{Q_3-Q_1}$

Step 1: Find LCF

x	f	LCF
0	6	6
1	15	21
2	25	46
3	8	54
4	4	58
5	1	59
Total	N=59	

Step 2:

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

$=(15)^{th}$ observation = **1**

$Q_2 = \left(\frac{2(N+1)}{4}\right)^{th}$ observation

$= (30)^{th}$ observation = 2

$Q_3 = \left(\frac{3(N+1)}{4}\right)^{th}$ observation

$=(45)^{th}$ observation = 2

$\therefore$ Bowley' s coefficient of skewness = $\frac{2+1-2(2)}{2-1}$

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

= -1

The distribution is negatively skewed.

Example : 7.

The mean and mode of weekly wages of a group of workers are Rs. 45 and Rs. 36 respectively. If SD is Rs.18, then find the coefficient of skewness.

Solution :

$$\text{Karl Pearson}'\text{scoefficient of skewness} = \frac{\bar{x} - Z}{\sigma}$$

$$= \frac{45-36}{18} = \frac{9}{18}$$

$$= \mathbf{0.5}$$

Example : 8.

Calculate K.P. coefficient of skewness if Mean = 23, Median = 25 and σ = 10.

Solution :

$$\text{Karl Pearson}'\text{scoefficient of skewness} = \frac{3(\bar{x} - M)}{\sigma}$$

$$= \frac{3(23-25)}{10} = \frac{3(-2)}{10}$$

$$\mathbf{= -0.6}$$

Example : 9.

The mean of a distribution is 50. Its SD is 15 and the coefficient of skewness is -1. Find median.

Solution :

$$\text{Karl Pearson}'\text{scoefficient of skewness} = \frac{3(\bar{x} - M)}{\sigma}$$

$$-1 = \frac{3(50 - M)}{15}$$

$$-15 = 150 - 3M$$

$$3M = 165$$

$$\therefore M = \frac{165}{3} = \mathbf{55}$$

Example : 10.

In a distribution, sum of lower and upper quartiles is 40 and their difference is 16. If median is 18, then find the coefficient of skewness.

Solution : Here , M = Q_2

$$\text{Bowley's coefficient of skewness} = \frac{Q_3+Q_1-2M}{Q_3-Q_1} = \frac{40-2(18)}{16} = \mathbf{0.25}$$

Example : 11.

In a distribution, the first quartile is 142 and semi-inter quartile range is 18. If the distribution is symmetric, then find second quartile.

Solution: Given $Q_1 = 142$

$$\frac{Q_3-Q_1}{2} = 18 \Rightarrow Q_3 - Q_1 = 36$$

$$\text{i.e., } Q_3 = 36 + 142 = 178$$

For a symmetric distribution, $Q_3 - Q_2 = Q_2 - Q_1$

$$\Rightarrow Q_2 = \frac{Q_3+Q_1}{2} = \frac{178+142}{2} = \mathbf{160}$$

Example : 12.

For a distribution, Bowley' s coefficient of skewness = -0.36, lower quartile = 8.6 and median= 12.3. Find the upper quartile

Solution :

$$\text{Bowley's coefficient of skewness} = \frac{Q_3+Q_1-2Q_2}{Q_3-Q_1}$$

$$-0.36 = \frac{Q_3+8.6-2(12.3)}{Q_3-8.6}$$

$$\Rightarrow -0.36\, Q_3 + 3.096 = Q_3 + 8.6 - 24.6$$

$$\Rightarrow 1.36\, Q_3 = 19.096$$

$$\therefore Q_3 = \mathbf{14.04}$$

The upper quartile is 14.04

Example : 13.

The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Test the skewness and kurtosis of the distribution.

Solution :

Skewness :

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

$$\beta_1 = \frac{(0.7)^2}{(2.5)^3} = 0.031$$

As $\beta_1 = +0.03$, the distribution is slightly positively skewed.

Kurtosis:

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

$$\beta_2 = \frac{18.75}{2.5^2} = \frac{18.75}{6.25} = 3$$

As, $\beta_2 = 3$, the distribution is mesokurtic (normal).

Example : 14.

If $\mu_1 = 0$, $\mu_2 = 16$, μ_3 $(\sigma^2) = -36$ and $\mu_4 = 120$ then comment on the skewness and kurtosis of the distribution.

Solution :

For commenting on skewness, we calculate γ_1 (as here μ_3 is negative).

$$\gamma_1 = \sqrt{\beta_1} = \sqrt{\frac{{\mu_3}^2}{{\mu_2}^3}} = \frac{\mu_3}{\sigma^3} = \frac{-36}{(4)^3} = \frac{-36}{64} = -0.5625$$

The distribution is negatively skewed.

For kurtosis, we calculate β_2

$$\beta_2 = \frac{\mu_4}{{\mu_2}^2} = \frac{120}{(16)^2} = 0.469$$

Since $\beta_2 < 3$, the distribution is platykurtic.

Questions

1. What are moments ?
2. Mention the two types of moments.
3. What are raw moments ?
4. What are Central moments ?
5. What is skewness ?
6. How does skewness differ from dispersion ?
7. Draw the various skewed curves and indicate the rough position of mean, median and mode.
8. What is the relationship between mean and mode for a positively skewed distribution ?
9. Mention two properties of symmetrical distribution.
10. Can the values of, $\overline{x}$, M, Z be the same? If yes, state the situation.
11. Define the term Kurtosis.
12. Name the distribution when $\beta_2 > 3$.
13. Draw a neat diagram to show different peaks for symmetrical distribution and name each of them.

Exrcise Problems

1. Calculate Karl Pearson's coefficient of skewness from the data given below.

x	10	20	30	40	50	60	70
f	1	5	12	22	17	9	4

Ans. : +0.23

2. Compute Karl Pearson's coefficient of skewness for the following distribution.

Marks	Above 0	Above 10	Above 20	Above 30	Above 40	Above 50	Above 60	Above 70	Above 80
No. of students	150	140	100	80	80	70	30	14	0

Ans. : - 0.75

3. Consider the following data of two distributions.

	Distribution A	Distribution B
Mean	100	90
Median	90	80
Standard Deviation	10	10

Find out whether both the distributions have the same degree of skewness.

4. Calculate Karl-Pearson' s coefficient of skewness from the following data.

C.I.	70-80	60-70	50-60	40-50	30-40	20-30	10-20	0-10
f	11	12	30	35	21	11	6	5

Ans : 0.046

5. For a set of 10 observations, $\Sigma x=452$, $\Sigma x^2 = 24270$ and mode = 43.7. Find Pearson' s coefficient of skewness. Anx.: 0.08

6. Calculate Bowley' s coefficient of skewness from the data given below.

C.I.	30-40	40-50	50-60	60-70	70-80	80-90	90-100
f	1	3	11	21	43	32	9

Ans. : - 0.035

7. Compute the coefficient of skewness based on quartiles.

C.I.	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89
f	5	9	14	20	25	15	8	4

Ans. : - 0.103

8. The Karl-Pearson' s coefficient of skewness of a distribution is 0.32. Its standard deviation is 6.5 and the mean is 29.6. Find the mode.

Ans : 27.52

9. In a distribution, $\overline{x}$ = 65, Z = 80 and coefficient of Skewness= -0.6, Find median and coefficient of variation. Ans. : 70, 38.46%

3. Consider the following data of two distributions.

	Distribution A	Distribution B
Mean	100	90
Median	90	80
Standard Deviation	10	10

Find out whether both the distributions have the same degree of skewness.

4. Calculate Karl-Pearson' s coefficient of skewness from the following data.

C.I.	70-80	60-70	50-60	40-50	30-40	20-30	10-20	0-10
f	11	12	30	35	21	11	6	5

Ans : 0.046

5. For a set of 10 observations, $\Sigma x=452$, $\Sigma x^2 = 24270$ and mode = 43.7. Find Pearson' s coefficient of skewness. Anx.: 0.08

6. Calculate Bowley' s coefficient of skewness from the data given below.

C.I.	30-40	40-50	50-60	60-70	70-80	80-90	90-100
f	1	3	11	21	43	32	9

Ans. : - 0.035

7. Compute the coefficient of skewness based on quartiles.

C.I.	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89
f	5	9	14	20	25	15	8	4

Ans. : - 0.103

8. The Karl-Pearson' s coefficient of skewness of a distribution is 0.32. Its standard deviation is 6.5 and the mean is 29.6. Find the mode.

Ans : 27.52

9. In a distribution, $\overline{x}$ = 65, Z = 80 and coefficient of Skewness= -0.6, Find median and coefficient of variation. Ans. : 70, 38.46%

10. For a moderately skewed distribution, arithmetic mean = 160, mode = 157 and S.D. = 50, find coefficient of skewness. Ans. : 0.06

11. In a frequency distribution, the coefficient of skewness based on quartiles is 0.6. If the sum of the upper and lower quartiles is 100 and the median is 38, find the value of the upper quartile. Ans. : 70

12. Given Bowley's coefficient of skewness = - 0.8, Q_1 = 40 and Q_3 = 60, find Q_2. Ans : 58

13. The first four central moments are 0, 40, 100 and 200. Comment on the skewness and Kurtosis.

*****

Unit-VI

ANALYSIS OF BIVARIATE DATA

Introduction :

So far we have confined ourselves to univariate distributions. That is , the distributions involving only one variable. When data regarding two or more variables are available , we may have to study the relative variation of these variables.

For example , there exists some relationship between the height of son and height of father , price of a commodity and amount of goods demanded , examination scores of a student and the number of hours of study etc..

If variables show related variations they are said to be correlated.

Correlation :

'Correlation is a statistical device which helps in analyzing the covariation of two or more variables.' It is a study of inter dependence between the variables.

Correlation analysis determines the degree of relationship between variables. It refers to the techniques used in measuring the closeness of relationship between the variables .

Correlation may be

i) Simple Correlation.

ii) Multiple Correlation.

iii) Partial Correlation.

Simple correlation concerns with related variation between two related variables. **Multiple Correlation** and **Partial Correlation** concern with the related variations among three or more variables. In Multiple Correlation three or more variables are studied simultaneously. In Partial Correlation, we recognize more than two variables but consider only two variables to be influencing each other and the effect of other influencing variables being kept constant.

Two variables are said to be correlated when they tend to vary either in the same direction or in the opposite directions. If the variation in one variable is accompanied by a definite variation in the other variable then, the two variables are said to be **correlated**.

For example, there exists some relationship between the age of the husband and the age of his wife, increase in the rainfall up to a point and production of paddy etc. So we say that they are correlated.

When the variables x and y are correlated, there may be three types of relationship.

i) x is the cause and y is the effect.

ii) x is the effect and y is the cause.

iii) x and y are effects of some other causes.

Examples :

i) Rainfall (cause) and yield of paddy (effect).

ii) Price (effect) and demand (cause).

iii) Yield per acre of rice and tea are both effects of rainfall.

A cause and effect relationship between two variables is called **Causation.**

Examples :

i) Production and price of electronic goods show causation. Here, an increase in production causes decrease in price.

ii) Increased demand and price of commodity. Here, increased demand of a commodity due to growth of population causes increase in price.

Even when there is absence of causation, variables may show correlation. For instance, population of India and population of China in different years show correlated variation. But, they do not show cause and effect relationship. Sales of pigs and sales of pig–iron show correlated variation. But, there will be no causation. Thus correlation in the absence of causation is called **non-sense correlation** or **spurious correlation**. So correlation doesn't necessarily imply causation. But, the existence of causation always implies correlation.

Significance (utility) of the study of correlation :

The study of correlation is of immense use in day to day life because of the following reasons:

1) With the help of correlation analysis we can measure the degree of relationship that exists between the two variables.
2) Once we know that the two variables are related, we can estimate the value of one variable for a given value of the other variable.
3) Progressive development in the methods of science and philosophy has been characterized by the rich knowledge of relationship. In nature also one finds multiplicity of inter related forces.
4) The purpose of the study of correlation is to reduce the range of uncertainty in any field by studying the relationship between the variables. The predictions based on correlation analysis are likely to be more reliable and realistic.
5) Correlation analysis is useful in economic field and business.

Types of correlation :

1) Positive Correlation : If the variables vary in the same direction (if both the variables increase or decrease together) then, correlation is said to be **Positive** or **Direct.**

For example,

i) The height and weight of a group of persons.

ii) Income and expenditure of different families.

2) Negative Correlation : If the variables vary in the opposite directions (if one variable increases the other variable decreases) then, correlation is said to be **Negative** or **Inverse.**

For example,

i) The sales of woolen garments and the atmospheric temperature.

ii) Production and price of vegetables.

3) Non correlation : If the two variables do not show the associated variation, they are said to be **non correlated**.

For example,

i) Purchasing power of money and atmospheric temperature.

ii) Sales of umbrella and the sales of student notebook.

4) Perfect Correlation : If the variables vary in the same proportion (the variables show exact linear relationship) then, the correlation is said to be **perfect.** Perfect correlation may be positive or negative. Correlation is perfect positive when the variations are in the same proportion and in the same direction.

For example,

i) 10% increase in x leads to 10% increase in y.

ii) Consider the following table :

x	1	2	3	4	5
y	5	10	15	20	25

Correlation is perfect negative when the variations are in the same proportion but in the opposite directions.

For Example,

i) 5% increase in x leads to 5% decrease in y.

ii) Consider the following table :

x	1	2	3	4	5
y	20	18	16	14	12

Measurement of Correlation :

The important methods of ascertaining whether two variables are correlated or not are,

i) Scatter diagram method.

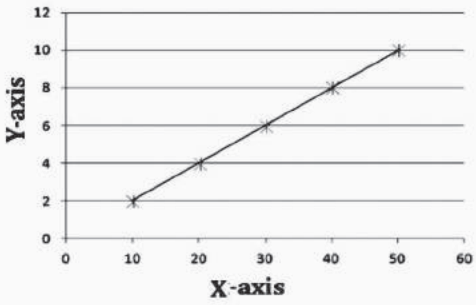
ii) Karl Pearson's coefficient of correlation.

iii) Spearman's coefficient of rank correlation.

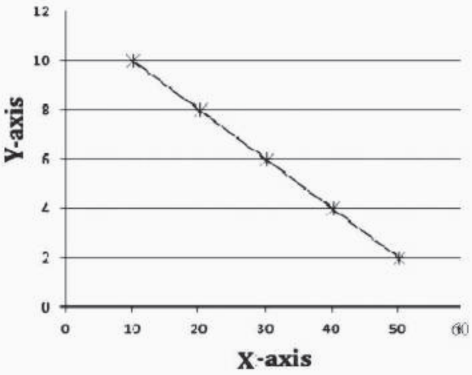
Scatter diagram method :

This is a non mathematical method of measuring correlation. When the two variables are plotted on a graph by taking the two variables along the two axes, a set of scattered points will be obtained on the graph. It is called scatter diagram. Thus, scatter diagram is a graph, which shows the degree of relationship between two variables. Usually for related variables the points on a scatter diagram will have a direction. The direction can be seen by drawing a line close to the plotted points on a scatter diagram. Depending on the direction of the line the correlation will be studied. If all the points lie on a straight line rising from the lower left-hand corner to the upper right-hand corner, correlation is said to be perfect positive.

On the other hand, if all the points are lying on a straight line falling from upper left hand corner to the lower right hand corner of the diagram then, the correlation is said to be perfect negative.

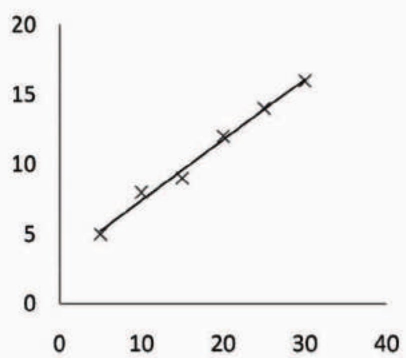


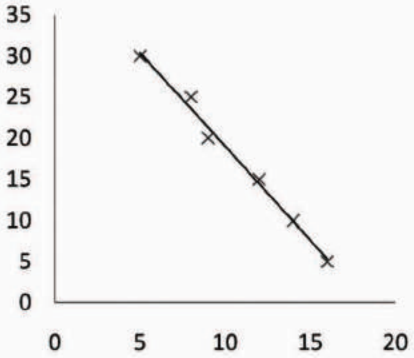
Perfect positive correlation (r=+1)



Perfect negative correlation (r=-1)

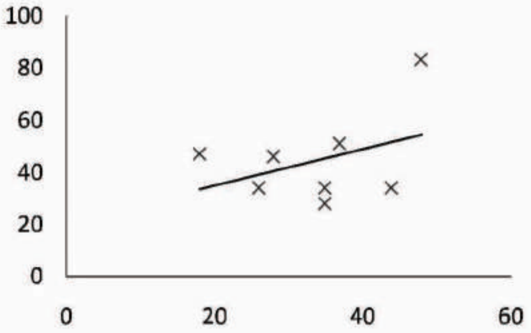
If the plotted points are close to the drawn line then there exists higher degree positive or higher degree negative correlation depending on the direction of the line.

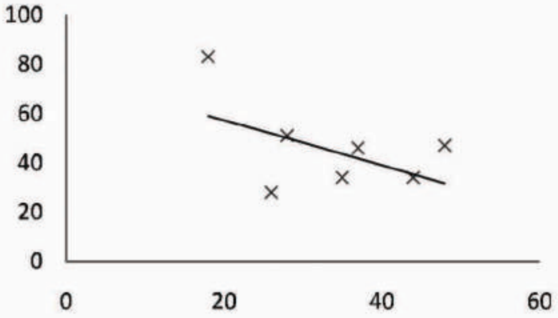




Higher positive correlation. Higher negative correlation.

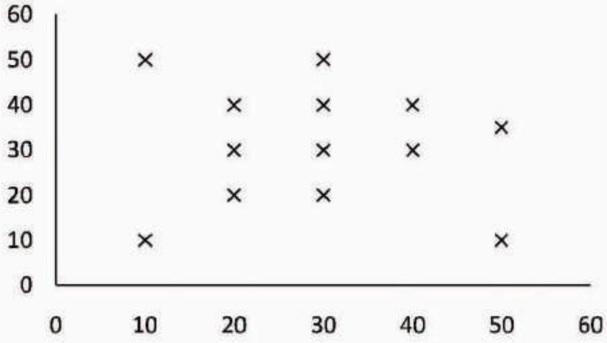
If the path formed by the dots is very wide then the correlation is of lower degree positive or negative depending on the directions of the points.



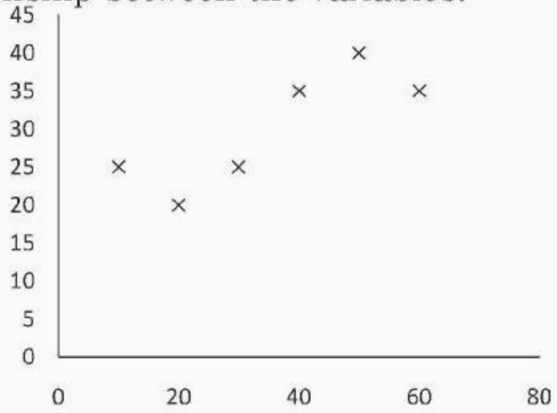


Lower degree positive correlation. Lower degree negative correlation.

If the plotted points are in a haphazard manner, it shows the absence of any relationship between the variables. So the variables are non correlated (Independent).



Any other curve form of arrangement of plotted points indicates curvilinear relationship between the variables.



Merits :

i) It is a simple and non mathematical method of studying correlation between the variables.

ii) It is not influenced by the size of extreme items whereas most of the mathematical methods of finding correlation are influenced by extreme items.

iii) Making a scatter diagram usually is the first step in investigating the relationship between two variables.

Demerits :

By applying this method we can get an idea about the direction of correlation and also whether it is high or low. But we cannot establish the exact degree of correlation between the variables.

Karl Pearson's (Product moment) coefficient of correlation :

Of the several mathematical methods of measuring correlation, the Karl Pearson's method, popularly known as Pearson's coefficient of correlation, is most widely used in practice. This coefficient of correlation measures in one figure the exact degree of correlation existing between two variables.

The product moment correlation coefficient is defined as

$$\Upsilon_{xy} = \frac{\textbf{Covariance (x, y)}}{\sqrt{\textbf{variance(x)} \times \textbf{Variance(y)}}}$$

$$\Upsilon_{xy} \quad = \frac{\textbf{Cov (x, y)}}{\sigma_x \times \sigma_y}$$

$$\Upsilon_{xy} \quad = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{[\Sigma(x-\bar{x})^2][\Sigma(y-\bar{y})^{\ 2}]}}$$

γ_{xy} is a relative measure of correlation. Here, $\bar{x}$ and σ_x are the mean and standard deviation of variable x respectively. $\bar{y}$ and σ_y are the mean and standard deviation of variable y respectively. For the purpose of calculations the above formula is used in a simplified form as follows.

For ungrouped data,

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

Here, 'n' is number of pairs of observations.

For grouped (tabulated) data,

$$\Upsilon_{xy} = \frac{N\Sigma\Sigma xyf_{xy} - (\Sigma xf_x)(\Sigma yf_y)}{\sqrt{[N\Sigma x^2f_x - (\Sigma xf_x)^2]\left[N\Sigma y^2f_y - (\Sigma yf_y)^2\right]}}$$

Where,

f_{xy}- bivariate frequencies.

f_x- marginal frequencies of x.

f_y- marginal frequencies of y.

N- total frequency of the given bivariate frequency distribution.

We know that,

$$\gamma_{xy} = \frac{\mathrm{Cov}(x, y)}{\sigma_x \times \sigma_y}$$

If x and y are positively correlated, higher values of x will be associated with higher values of y and lower values of x will be associated with lower values of y. So if $(x-\bar{x})$ is positive, $(y-\bar{y})$ will also be positive and if $(x-\bar{x})$ is negative, $(y-\bar{y})$ will also be negative. There may be exceptions to this, the exceptions will be more if correlation is low. So the product $(x-\bar{x})(y-\bar{y})$ will be positive. Therefore, cov(x, y) will be positive and so coefficient of correlation will be positive. So the γ_{xy} will be higher if correlation is high and it will be lower if correlation is low.

If x and y are negatively correlated, higher values of x will be associated with lower values of y and lower values of x will be associated with higher values of y. So if $(x-\bar{x})$ is positive, $(y-\bar{y})$ will be negative and if $(x-\bar{x})$ is negative, $(y-\bar{y})$ will be positive. There may be exceptions to this, the exceptions will be more if correlation is low. So the product $(x-\bar{x})(y-\bar{y})$ will be negative. Therefore, cov(x, y) will be negative and so coefficient of correlation will be negative.

If x and y are non correlated, some of the products will be positive and other products will be negative. The positive and negative products add up to zero (or a very low value) and so, coefficient of correlation will be zero.

Interpretation of coefficient of correlation :

i. A positive value of γ indicates positive correlation.

ii. A negative value of γ indicates negative correlation.

iii. If $\gamma=+1$, then the correlation is perfect positive.

iv. If $\gamma=-1$, then the correlation is perfect negative.

v. $\gamma=0$, then the variables are non correlated.

vi. If $|\gamma| \geq 0.5$, then the correlation will be of higher degree.

Properties of correlation coefficient :

i. The coefficient of correlation is a pure number independent of units of measurement of the variables.

ii. The coefficient of correlation is not affected by the changes in origin and scale.

iii. The coefficient of correlation cannot be numerically greater than 1. i.e., $-1 \leq \gamma_{xy} \leq 1$, $\left(\left|\gamma_{xy}\right| \leq 1\right)$.

iv. The coefficient of correlation is not affected by the order,

i.e, $\gamma_{xy} = \gamma_{yx}$

Merit : The correlation coefficient summarizes in one figure not only the degree of correlation but also the direction. That is, whether correlation is positive or negative.

Limitations :

i. The correlation coefficient always assumes linear relationship regardless of the fact whether that assumption is correct or not.

ii. Great care must be exercised in interpreting the value of this coefficient as very often the coefficient is misinterpreted.

iii. The value of the coefficient is unduly affected by the extreme items.

iv. As compared with other methods this method takes more time to compute the value of correlation coefficient.

Spearman's coefficient of rank correlation :

The Karl Pearson's method is based on the assumption that the population under study is distributed normally. When it is known that the population is not normal or the shape of the distribution is not known, there is a need for a measure of correlation which involves no assumption about the parameters of the population. Also the Karl Pearson's coefficient of correlation can be calculated only if the characteristics under study are quantitative. It cannot be used in cases where the characteristics under study are qualitative. For example, honesty, efficiency, intelligence, etc. In such cases, Spearman's coefficient of rank correlation can be calculated. In this method we can avoid making any assumptions about the populations under study, by ranking of the observations according to size and calculations based on the ranks rather than the original observations. If it is possible to assign ranks to the units in the two characteristics, coefficient of rank correlation is the product moment coefficient of correlation between the ranks. The formula for computing coefficient of rank correlation denoted by ρ (Rho) is,

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

Where, d= $R_1 - R_2$, d-refers to the difference of ranks between paired items in the two series. R_1 and R_2 are the ranks given to paired items in two series respectively. 'n' stands for the number of pairs.

Since 'ρ' is the product moment coefficient of correlation between the ranks, its value also lies between -1 and +1. The value of this coefficient, interpreted in the same way as Karl Pearson's correlation coefficient. When 'ρ' is +1, there is complete agreement in the order of the ranks and the ranks are in the same direction. When 'ρ' is -1, there is complete agreement in order of the ranks and they are in the opposite directions. Spearman's coefficient of rank correlation can be easily calculated than Karl Pearson's coefficient of correlation.

In some cases it may be found necessary to rank two or more individuals or entries as equal. In such case it is customary to give each individual an average rank. Thus if two individuals are ranked equal at 5^{th} place, they are given the rank $\frac{5+6}{2}$, that is, 5.5 each. If three are ranked equal at 5^{th} place, they are given the rank $\frac{5+6+7}{3}=6$ each. In other words, where two or more items are to be ranked equal, the ranks are assigned for the purpose of calculating coefficient of correlation is the average of the ranks which these individuals would have got.

When equal ranks are assigned to some entries, an adjustment in the formula for calculating the rank correlation coefficient is made. The adjustment consists of adding $\frac{1}{12}\left(m^3 - m\right)$ to the value of Σd^2, where 'm' stands for the number of items whose ranks are common. If there is more than one such group of items with common ranks, then this value is added as many times as the number of such groups. The formula can thus be written as :

$$\rho \ = 1 - \left(\frac{6[\Sigma d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \ldots\ldots]}{n^3 - n}\right)$$

Merits :

i. This method is simple to understand and easy to apply when compared with the Karl Pearson's method. In both the methods we get the same answer for a given set of bivariate data, when there will be no repetitions of the values in a series.

ii. When the data are of a qualitative nature, this is the only method to be used. For example, the workers of two factories can be ranked in the order of efficiency, ranks given by the two judges in beauty contest, the degree of correlation is established by applying this method only.

iii. Even when actual data are given, rank method can be applied for ascertaining correlation.

Limitations :

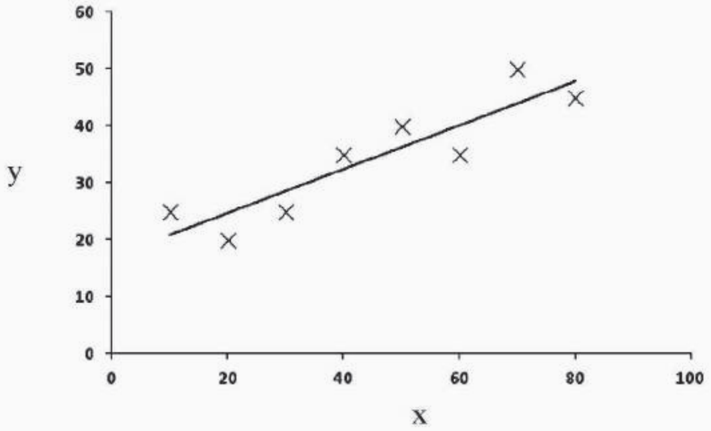
i. This method cannot be used for finding out correlation in the case of a grouped frequency distribution.

ii. When the number of items exceeds 30 the calculations become quite tedious and time consuming. Therefore, this method should not be applied where the number of items exceeds 30, unless, we are given the ranks and not the actual values of the variable.

Examples on scatter diagram :

Example1. Draw scatter diagram and conclude about correlation.

X	10	20	30	40	50	60	70	80
Y	25	20	25	35	40	35	50	45

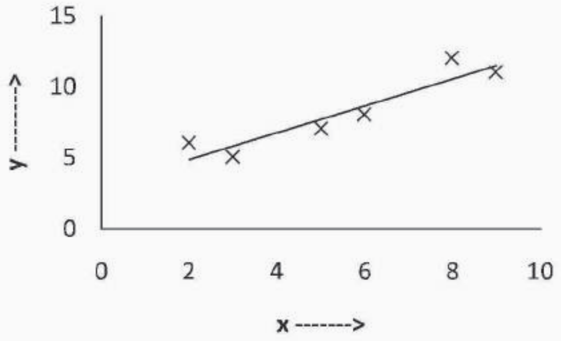


From the diagram we find that there exists a higher degree positive correlation between x and y.

Example2. Draw scatter diagram and conclude about correlation.

x	2	3	5	6	8	9
y	6	5	7	8	12	11

Solution :

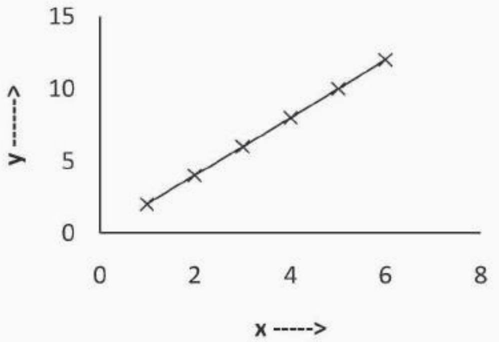


From the diagram we find that there exists a higher degree positive correlation between x and y.

Example3. Draw scatter diagram and interpret.

x	1	2	3	4	5	6
y	2	4	6	8	10	12

Solution :

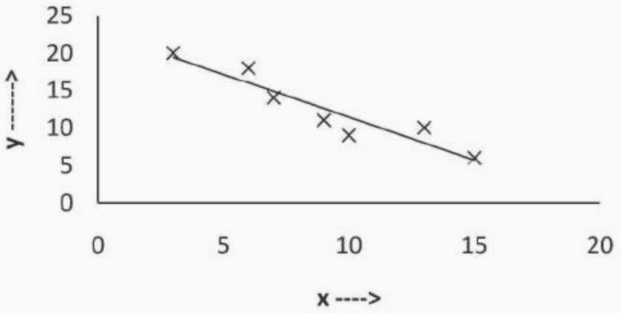


From the diagram we find that there exists a perfect positive correlation between x and y

Example4. Draw scatter diagram and comment about correlation.

x	3	6	7	9	10	13	15
y	20	18	14	11	9	10	6

Solution :



From the diagram we find that there exists a higher degree negative correlation between x and y.

Examples on Karl Pearson's coefficient of correlation :

Example : 5. Calculate the product moment coefficient of correlation between the following marks (out of 10) in statistics and mathematics of six students.

Student	1	2	3	4	5	6
Statistics(x)	7	4	6	9	3	8
Mathematics(y)	8	5	4	8	3	6

Solution:

x	y	x^2	y^2	xy
7	8	49	64	56
4	5	16	25	20
6	4	36	16	24
9	8	81	64	72
3	3	9	9	9
8	6	64	36	48
37	**34**	**255**	**214**	**229**

Product moment coefficient of correlation is,

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

$$= \frac{6(229) - (37)(34)}{\sqrt{[6(255) - 37^2][6(214) - 34^2]}}$$

$$= \frac{116}{\sqrt{(161)(128)}}$$

$$\gamma_{xy} = \mathbf{0.8081}$$

There exists a positive correlation of higher degree between x and y.

Example : 6 Calculate Karl Pearson's coefficient of correlation for the following data.

x	70	71	72	73	74	75	76	77
y	30	29	25	25	22	20	18	16

.

Solution:

x	y	u = x-74	v = y-25	u^2	v^2	uv
70	30	-4	5	16	25	-20
71	29	-3	4	9	16	-12
72	25	-2	0	4	0	0
73	25	-1	0	1	0	0
74	22	0	-3	0	9	0
75	20	1	-5	1	25	-5
76	18	2	-7	4	49	-14
77	16	3	-9	9	81	-27
Total		**-4**	**-15**	**44**	**205**	**-78**

Karl Pearson's coefficient of correlation is,

$$\Upsilon_{xy} = \Upsilon_{uv} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{[n\Sigma u^2 - (\Sigma u)^2][n\Sigma v^2 - (\Sigma v)^2]}}$$

$$= \frac{8(-78) - (-4)(-15)}{\sqrt{[8(44) - (-4)^2][8(205) - (-15)^2]}}$$

$$= \frac{-624 - 60}{\sqrt{[352 - 16][1640 - 225]}}$$

$$= \frac{-684}{\sqrt{(336)(1415)}}$$

$$\mathbf{= -\,0.9912}$$

There exists a negative correlation of higher degree between x and y.

Example : 7.

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

x	300	350	400	450	500	550	600	650	700
y	800	900	1000	1100	1200	1300	1400	1500	1600

Solution:

x	y	$u = \frac{(x-500)}{50}$	$v = \frac{(y-1200)}{100}$	u^2	v^2	uv
300	800	-4	-4	16	16	16
350	900	-3	-3	9	9	9
400	1000	-2	-2	4	4	4
450	1100	-1	-1	1	1	1
500	1200	0	0	0	0	0
550	1300	1	1	1	1	1
600	1400	2	2	4	4	4
650	1500	3	3	9	9	9
700	1600	4	4	16	16	16
Total		**0**	**0**	**60**	**60**	**60**

Karl Pearson's coefficient of correlation is,

$$\gamma_{xy} = \gamma_{uv} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{[n\Sigma u^2 - (\Sigma u)^2][n\Sigma v^2 - (\Sigma v)^2]}}$$

$$= \frac{9(60) - (0)(0)}{\sqrt{[9(60) - (0)^2\,][9(60) - (0)^2]}}$$

$$= \frac{540}{\sqrt{[540][540]}}$$

$$= \frac{540}{540}$$

$$\gamma_{xy} = \mathbf{1}$$

There exists a perfect positive correlation between x and y

Example : 8.

Calculate Karl Pearson's coefficient of correlation from the following data and comment on the result.

x	12	14	16	18	20	22	24	26	28	30
y	36	33	30	27	24	21	18	15	12	9

Solution:

x	y	u= (x-20)/2	v= (y-21)/3	u^2	v^2	uv
12	36	-4	5	16	25	-20
14	33	-3	4	9	16	-12
16	30	-2	3	4	9	-6
18	27	-1	2	1	4	-2
20	24	0	1	0	1	0
22	21	1	0	1	0	0
24	18	2	-1	4	1	-2
26	15	3	-2	9	4	-6
28	12	4	-3	16	9	-12
30	9	5	-4	25	16	-20
Total		**5**	**5**	**85**	**85**	**-80**

Karl Pearson's coefficient of correlation is,

$$\Upsilon_{xy} = \Upsilon_{uv} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{[n\Sigma u^2 - (\Sigma u)^2][n\Sigma v^2 - (\Sigma v)^2]}}$$

$$= \frac{10(-80) - (5)(5)}{\sqrt{[10(85) - (5)^2][10(85) - (5)^2]}}$$

$$= \frac{-800 - 25}{\sqrt{[850 - 25][850 - 25]}}$$

$$= \frac{-825}{825}$$

$$\Upsilon_{xy} = \mathbf{-1}$$

There exists perfect negative correlation between x and y.

Example : 9.

Calculate the coefficient of correlation between the number of male children and the number of female children from the following data.

No. of male	No. of female children				
children	0	1	2	3	4
0	3	4	2	-	-
1	4	8	8	2	-
2	-	7	12	8	4
3	-	3	8	8	5
4	-	-	3	5	6

Solution:

	y					f_x	uf_x	u^2f_x	uv f_{xy}
x	0	1	2	3	4				
0	$\boxed{0}$ 3	$\boxed{0}$ 4	$\boxed{0}$ 2	_	_	9	0	0	0
1	$\boxed{0}$ 4	$\boxed{8}$ 8	$\boxed{16}$ 8	$\boxed{6}$ 2	_	22	22	22	30
2	-	$\boxed{14}$ 7	$\boxed{48}$ 12	$\boxed{48}$ 8	$\boxed{32}$ 4	31	61	124	142
3	-	$\boxed{9}$ 3	$\boxed{48}$ 8	$\boxed{72}$ 8	$\boxed{60}$ 5	24	72	216	189
4	-	-	$\boxed{24}$ 3	$\boxed{60}$ 5	$\boxed{96}$ 6	14	56	224	180
f_y	7	22	33	23	15	**N=100**	**212**	**586**	**541**
yf_y	0	22	66	69	60	**217**			
y^2f_y	0	22	132	207	240	**601**			
xyf_{xy}	0	31	136	186	188	**541**			

Karl Pearson's coefficient of correlation is,

$$\Upsilon_{xy} = \frac{N\Sigma\Sigma xyf_{xy} - (\Sigma xf_x)\left(\Sigma yf_y\right)}{\sqrt{[N\Sigma x^2f_x - (\Sigma xf_x)^2]\left[N\Sigma y^2f_y - \left(\Sigma yf_y\right)^2\right]}}$$

$$= \frac{(100)(541) - (212)(217)}{\sqrt{[(100)(586) - (212)^2][(100)(601) - (217)^2]}}$$

$$= \frac{54100 - 46004}{\sqrt{[58600 - 44944][60100 - 47089]}}$$

$$= \frac{8096}{\sqrt{(13656)(13\,011)}}$$

$$= \frac{8096}{13329.598}$$

Υ_{xy} **= 0.6074**

Example : 10. The following table gives the frequency according to age groups and marks obtained by 67 students in an intelligence test. Calculate Karl Pearson's coefficient of correlation between age and intelligence.

Test marks	Age in years			
	18	19	20	21
200-250	4	4	2	1
250-300	3	5	4	2
300-350	2	6	8	5
350-400	1	4	6	10

Solution:

x	y				f_x	u	uf_x	u^2f_x	uvf_{xy}
	18	**19**	20	21					
225	$\boxed{4}$ 4	$\boxed{0}$ 4	$\boxed{-2}$ 2	$\boxed{-2}$ 1	11	-1	-11	11	0
275	$\boxed{0}$ 3	$\boxed{0}$ 5	$\boxed{0}$ 4	$\boxed{0}$ 2	14	0	0	0	0
325	$\boxed{-2}$ 2	$\boxed{0}$ 6	$\boxed{8}$ 8	$\boxed{10}$ 5	21	1	21	21	16
375	$\boxed{-2}$ 1	$\boxed{0}$ 4	$\boxed{12}$ 6	$\boxed{40}$ 10	21	2	42	84	50
f_y	10	19	20	18	**N = 67**		**52**	**116**	**66**
v	-1	0	1	2					
vf_y	-10	0	20	36	**46**				
v^2f_y	10	0	20	72	**102**				
uvf_{xy}	0	0	18	48	**66**				

Karl Pearson’s coefficient of correlation is,

$$\Upsilon_{xy} = \Upsilon_{uv} = \frac{N\Sigma\Sigma uvf_{xy} - (\Sigma uf_x)(\Sigma vf_y)}{\sqrt{[N\Sigma u^2 f_x - (\Sigma uf_x)^2]\left[N\Sigma v^2 f_y - (\Sigma vf_y)^2\right]}}$$

$$= \frac{67x66 - (52)(46)}{\sqrt{[(67)(116) - (52)^2][(67)(102) - (46)^2]}}$$

$$= \frac{4422 - 2392}{\sqrt{[7772 - 2704][6834 - 2116]}}$$

$$= \frac{2030}{\sqrt{(5068)(4718)}}$$

$$= \frac{2030}{4889.8694}$$

Υ_{xy} = **0.4151**

Example : 11.

Calculate Karl Pearson’s coefficient of correlation from the following data and comment on the result.

	y			
x	80 - 90	90 - 100	100 - 110	110 - 120
40 - 44	2	2	4	7
44 - 48	-	2	3	4
48 - 52	-	4	5	1
52 - 56	2	6	4	-
56 - 60	3	4	1	-

Solution :

x	y				f_x	u	uf_x	u^2f_x	uvf_{xy}
	85	**95**	105	115					
42	4 2	0 2	-8 4	-28 7	15	-2	-30	60	-32
46	- -	0 2	-3 3	-8 4	9	-1	-9	9	-11
50	- -	0 4	0 5	0 1	10	0	0	0	0
54	-2 2	0 6	4 4	- -	12	1	12	12	2
58	-6 3	0 4	2 1	- -	8	2	16	32	-4
f_y	7	18	17	12	**N = 54**		**-11**	**113**	**-45**
v	-1	0	1	2					
vf_y	-7	0	17	24	**34**				
$v2f_y$	7	0	17	48	**72**				
uvf_{xy}	-4	0	-5	-36	**-45**				

Karl Pearson's coefficient of correlation is,

$$\Upsilon_{xy} = \Upsilon_{uv} = \frac{N\Sigma\Sigma uvf_{xy} - (\Sigma uf_x)(\Sigma vf_y)}{\sqrt{[N\Sigma u^2 f_x - (\Sigma uf_x)^2]\left[N\Sigma v^2 f_y - (\Sigma vf_y)^2\right]}}$$

$$= \frac{(54)(-45) - (-11)(34)}{\sqrt{[(54)(113) - (-11)^2][(54)(72) - (34)^2]}}$$

$$= \frac{-2430 + 374}{\sqrt{[6102 - 121][3888 - 1156]}}$$

$$= \frac{-2056}{\sqrt{(5981)(2732)}}$$

$$= \frac{-2056}{4042.2879}$$

Υ_{xy} **= -0.5068**

Comment: Higher negative correlation exists between the variables x and y.

Example : 12. Coefficient of correlation between two variables x and y is 0.8. Their covariance is 20. The variance of x is 16. Find the S.D. of y.

Solution:

Given that $\Upsilon_{xy} = 0.8$, Cov(x,y) = 20 and $\sigma_x{}^2 = 16$

We know that,

$$\Upsilon_{xy} = \frac{\text{Cov}(x, y)}{(\sigma_x)(\sigma_y)}$$

$$\therefore 0.8 = \frac{20}{(4)(\boldsymbol{\sigma}_y)}$$

$$\sigma_y = \frac{20}{4 \times 0.8} = \frac{20}{3.2} = 6.25$$

Example : 13. In a bivariate data, V(x) = 40, V(y) = 35 and Cov(x,y)= -34 , find the coefficient of correlation.

Solution :

Given that $\sigma_x = \sqrt{40}$, $\sigma_y = \sqrt{35}$ and Cov(x,y)= -34

We know that

$$\Upsilon_{xy} = \frac{\text{Cov}(x, y)}{(\sigma_x)(\sigma_y)}$$

$$= \frac{-34}{(\sqrt{40})(\sqrt{35})}$$

$$= \frac{-34}{37.4166}$$

$$\boldsymbol{\Upsilon_{xy} = -0.9087}$$

Example : 14. In a bivariate data on x and y, if standard deviations of x and y are 7 and 9 respectively. If Cov(x,y) = 10, find the coefficient of correlation.

Solution :

Given that: $\sigma_x = 7$, $\sigma_y = 9$ and Cov(x,y) = 10

We know that, $\boldsymbol{\Upsilon_{xy}} = \frac{\text{Cov}(x, y)}{(\sigma_x)(\sigma_y)} = \frac{10}{(7)(9)} = \mathbf{0.1587}$

Example : 15.

If $\Sigma(x-\bar{x})^2 = 160$, $\Sigma(x-\bar{y})^2 = 438$ and $\Sigma(x-\bar{x})(y-\bar{y}) = 240$, find Υ_{xy}

Solution :

We know that,

$$\Upsilon = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{[\Sigma(x-\bar{x})^2][\Sigma(y-\bar{y})^2]}} = \frac{240}{\sqrt{[160][438]}} = \frac{240}{264.7263} = \mathbf{0.9066}$$

Example : 16.

If n=9, $\Sigma x = 2$, $\Sigma y = 7$, $\Sigma x^2 = 134$, $\Sigma y^2 = 165$ and $\Sigma xy = -129$, find γ

Solution :

We know that, $\Upsilon_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$

$$= \frac{9(-129) - 2(7)}{\sqrt{[9x134 - (2)^2][9x165 - (7)^2]}}$$

$$= \frac{-1161 - 14}{\sqrt{[1206 - 4][1485 - 49]}}$$

$$= \frac{-1175}{\sqrt{[1202][1436]}}$$

$$= \frac{-1175}{1313.8}$$

Υ_{xy} = **-0.8944**

Examples on Spearman's Rank correlation coefficient :

Example : 17.

The ranks of the same 16 students in tests in Mathematics and Statistics were as follows.

Ranks in Mathematics	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Ranks in Statistics	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13

Calculate the rank correlation coefficient for the above data.

Solution:

R_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
R_2	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13	
$d = R_1 - R_2$	0	-8	0	0	0	-1	5	2	1	-1	-4	3	-1	2	-1	3	Total
d^2	0	64	0	0	0	1	25	4	1	1	16	9	1	4	1	9	136

Rank correlation coefficient is,

$$\rho = 1 - \left[\frac{6\sum d^2}{n^3 - n}\right]$$

$$= 1 - \left[\frac{6 \times 136}{16^3 - 16}\right]$$

$$= 1 - \left[\frac{816}{4080}\right] = 1 - 0.2$$

ρ **= 0.8**

Example : 18

Calculate the Spearman's rank correlation coefficient for the following data.

x	35	37	38	42	44	46	51	54	55	56
y	40	32	39	42	41	31	50	52	46	55

Solution:

x	35	37	38	42	44	46	51	54	55	56	
y	40	32	39	42	41	31	50	52	46	55	
R_1	10	9	8	7	6	5	4	3	2	1	
R_2	7	9	8	5	6	10	3	2	4	1	
$d = R_1 - R_2$	3	0	0	2	0	-5	1	1	-2	0	
d^2	9	0	0	4	0	25	1	1	4	0	$\sum d^2 = 44$

Rank correlation coefficient is

$$\rho = 1 - \left[\frac{6\sum d^2}{n^3 - n}\right] = 1 - \left[\frac{6 \times 44}{10^3 - 10}\right] = 1 - \left[\frac{264}{990}\right] = 1 - 0.2667 = \mathbf{0.7333}$$

Example : 19.

Quotations of index numbers of equity share prices of a certain joint stock company and the prices of preference shares are given below.

Years	2003	2004	2005	2006	2007	2008	2009
Equity Shares	97.5	99.4	98.6	96.2	95.1	98.4	97.1
Preference Shares	75.1	75.9	77.1	78.2	79	74.6	76.2

Using the method of rank correlation, determine the relationship between equity shares and preference shares prices.

Solution:

Equity Shares	Preference Shares	R_1	R_2	$d = R_1 - R_2$	d^2
97.5	75.1	4	6	-2	4
99.4	75.9	1	5	-4	16
98.6	77.1	2	3	-1	1
96.2	78.2	6	2	4	16
95.1	79	7	1	6	36
98.4	74.6	3	7	-4	16
97.1	76.2	5	4	1	1
					$\sum d^2 = 90$

Rank correlation coefficient is,

$$\rho = 1 - \left[\frac{6\sum d^2}{n^3 - n}\right]$$

$$= 1 - \left[\frac{6 \times 90}{7^3 - 7}\right]$$

$$= 1 - \left[\frac{540}{336}\right] = 1 - 1.6071$$

$$\rho = -\mathbf{0.6071}$$

Example : 20. Calculate the coefficient of rank correlation.

x	18	28	35	44	35	26	37	48
y	83	51	34	34	34	28	46	47

Solution:

x	y	R_1	R_2	$d = R_1 - R_2$	d^2
18	83	8	1	7	49
28	51	6	2	4	16
35	34	4.5	6	-1.5	2.25
44	34	2	6	-4	16
35	34	4.5	6	-1.5	2.25
26	28	7	8	-1	1
37	46	3	4	-1	1
48	47	1	3	-2	4
					$\sum d^2 = 91.5$

In x series, 35 is repeted 2 times $\therefore m_1 = 2$. The ranks of 35 = $\frac{4+5}{2} = 4.5$.

Similarly, in y series, 34 is repeted 3 times $\therefore m_2 = 3$. The ranks of 34 = $\frac{5+6+7}{3} = 6$

Rank correlation coefficient is

$$\rho = 1 - \frac{6\left[\sum d^2 + \frac{1}{12}(m_1{}^3 - m_1) + \frac{1}{12}(m_2{}^3 - m_2)\right]}{n^3 - n}$$

$$= 1 - \frac{6\left[91.5 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3)\right]}{8^3 - 8}$$

$$= 1 - \frac{6\,[91.5 + 0.5 + 2]}{504}$$

$$= 1 - \frac{6 \times 94}{504}$$

$$= 1 - \frac{564}{504}$$

$$= 1 - 1.119$$

$$\rho = -\mathbf{0.119}$$

Example : 21.

Calculate Spearman's correlation coefficient for the following data.

Marks in English	25	32	25	30	28	30	34	36	24
Marks in Kannada	40	42	44	46	48	32	36	38	34

Solution:

x	25	32	25	30	28	30	34	36	24	Total
y	40	42	44	46	48	32	36	38	34	
R_1	7.5	3	7.5	4.5	6	4.5	2	1	9	
R_2	5	4	3	2	1	9	7	6	8	
$d = R_1 - R_2$	2.5	-1	4.5	2.5	5	-4.5	-5	-5	1	
d^2	6.25	1	20.3	6.25	25	20.3	25	25	1	130

Rank correlation coefficient is,

$$\rho = 1 - \frac{6\left[\sum d^2 + \frac{1}{12}(m_1{}^3 - m_1) + \frac{1}{12}(m_2{}^3 - m_2)\right]}{n^3 - n}$$

$$= 1 - \frac{6\left[130 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)\right]}{9^3 - 9}$$

$$= 1 - \frac{6\,[130 + 0.5 + 0.5]}{720}$$

$$= 1 - \frac{6 \times 131}{720}$$

$$= 1 - 1.0917$$

$$= \mathbf{-0.0917}$$

Example : 22.

Calculate Spearman's rank correlation coefficient from the following data.

x	18	16	20	22	12	24	15	20	17	20	23
y	15	21	18	23	20	24	16	17	19	25	22

Solution:

x	18	16	20	22	12	24	15	20	17	20	23	
y	15	21	18	23	20	24	16	17	19	25	22	
R_1	7	9	5	3	11	1	10	5	8	5	2	Total
R_2	11	5	8	3	6	2	10	9	7	1	4	
$d = R_1 - R_2$	-4	4	-3	0	5	-1	0	-4	1	4	-2	
d^2	16	16	9	0	25	1	0	16	1	16	4	104

$$\rho = 1 - \frac{6\left[\sum d^2 + \frac{1}{12}({m_1}^3 - m_1)\right]}{n^3 - n}$$

$$= 1 - \frac{6\left[104 + \frac{1}{12}(3^3 - 3)\right]}{11^3 - 11}$$

$$= 1 - \frac{6\,[104 + 2]}{1320}$$

$$= 1 - \frac{6 \times 106}{1320} = 1 - 0.4818$$

$\rho =$ **0.5182**

Example : 23.

If $\Sigma d^2 = 95$ and n = 12, find rank correlation coefficient.

Solution :

$$\rho = 1 - \left[\frac{6\sum d^2}{n^3 - n}\right] = 1 - \left[\frac{6 \times 95}{12^3 - 12}\right] = 1 - \left[\frac{570}{1716}\right] = 1 - 0.3322 \ \mathbf{= 0.6678}$$

Example : 24.

The coefficient of rank correlation between marks in Accountancy and marks in Statistics for a certain group of students is 0.75. If the sum of the squares of the difference in ranks is given to be 30, find the number of students in the group.

Solution : Given that, $\rho = 0.75$ and $\Sigma d^2 = 30$

We know that, $\rho = 1 - \left[\frac{6\sum d^2}{n^3 - n}\right]$

$$0.75 = 1 - \left[\frac{6 \times 30}{n^3 - n}\right]$$

$$\text{i.e.,} \frac{180}{n^3 - n} = 0.25$$

$$\text{i.e.,} n(n^2 - 1) = \frac{180}{0.25} = 720$$

$$\text{i.e.,} n(n^2 - 1) = 9 \times 80 = 9(9^2 - 1)$$

$\therefore$ **n = 9**

Questions

1. What is correlation ?
2. Name two variables which are correlated.
3. Mention types of correlation.
4. Give an example for negative correlation between two variables.
5. Give an example for positive correlation between two variables.
6. Give an example where correlation does not exist.
7. What is causation ?
8. What do you mean by spurious correlation ?
9. What is 'perfect' correlation ?
10. What is the nature of correlation between the variables 'expenditure on advertisement' and 'amount of sales' ?
11. What is the nature of correlation between the variables 'number of employees' and 'expenditure on salary' ?
12. What is the nature of correlation between the variables 'amount of investment' and 'amount of sales' ?
13. Indicate the sign of correlation when the variables are varying in the same direction.
14. Mention which type of correlation is associated with 'value of rupee' and 'atmospheric temperature'.
15. Give an example of spurious correlation.
16. Mention various methods of computing correlation.
17. What is scatter diagram ?

18. Draw a scatter diagram to show positive correlation between two variables.
19. Draw a scatter diagram to show that there exists perfect negative correlation between two variables.
20. Mention one demerit of scatter diagram.
21. What do you understand by the term 'coefficient of correlation' ?
22. Define Karl Pearson's coefficient of correlation.
23. What is the value of 'γ' when two variables are uncorrelated ?
24. What is the nature of correlation when $\gamma = -1$?
25. If $\gamma = 1$, what is your conclusion ?
26. What is the range for Karl Pearson's coefficient of correlation ?
27. Mention a property of Karl Pearson's coefficient of correlation.
28. Mention a merit of Karl Pearson's coefficient of correlation.
29. What is the value of 'γ' when two variables are independent ?
30. Which method is used to calculate correlation coefficient when the data is qualitative in nature ?
31. Mention the limits of Spearman's coefficient of rank correlation.
32. When do you calculate Spearman's correlation coefficient ?
33. If $\sum d^2 = 0$, what is the value of spearman's rank correlation coefficient ?
34. Mention one limitation of Spearman's coefficient of rank correlation.
35. Define the term 'correlation'. Give an example.
36. What is 'positive correlation' ? Give an example.
37. Mention which type of correlation is associated with

    a) Production and price of vegetable.

    b) Production of pigs and the production of the pig-iron.
38. Mention two utility of the study of correlation.
39. Indicate scatter diagrams for $\gamma = +1$ and $\gamma = -1$.
40. Mention two merits of scatter diagrams.

41. Mention two properties ofγ .

42. In a bivariate data, the variances and covariance are equal. Find the coefficient of correlation. Ans: 1

43. Mention two demerits of Karl Pearson's coefficient of correlation.

44. Write the formula for Spearman's coefficient of rank correlation when one rank repeats 'm' times.

45. Mention two merits of Spearman's coefficient of rank correlation.

46. Explain 'scatter diagram' with necessary diagrams.

47. Explain types of correlation giving an example for each.

Exercise Problems

1. Draw a scatter diagram for the data given below and interpret.

x	15	18	20	19	14	12	22	11
y	14	16	13	15	18	18	11	20

2. Draw a scatter diagram for the data given below and interpret.

x	50	100	150	200	250	300	350
y	12	14	13	14	16	18	19

3. In a bivariate data on x and y , Var (x) = 9, Var (y) = 49 and Cov (x,y) = 20. Find γ_{xy}. Ans: 0.9525

4. Given, Cov (X, Y) = -100, V(x) = 400 and V(y) = 25. Find γ_{xy}.
Ans. -1

5. In a bivariate data covariance is 20, variances are 25 and 36 respectively. Findγ_{xy}. Ans: 0.6667

6. If Cov(x, y) = 1125, σ_x = 47.5 and σ_y = 39.6, Find γ_{xy}. Ans. 0.5981

7. If Cov(x, y) = 129.65,σ_x = 12.83 and σ_y = 15.79 , Find γ_{xy}.
Ans: 0.64

8. If $\sum(x-\bar{x})^2 = 6000$, $\sum(y-\bar{y})^2 = 920$ and $\sum(x-\bar{x})(y-\bar{y}) = 240$, Find γ_{xy}.
Ans: 0.1022

9. If $\sum(x-\bar{x})^2 = 88$, $\sum(y-\bar{y})^2 = 120$ and $\sum(x-\bar{x})(y-\bar{y}) = 93$, Find γ_{xy}.
Ans: 0.905

10. If $n = 7$, $\Sigma xy = 676$, $\Sigma x = 70$, $\Sigma y = 63$, $\Sigma x^2 = 728$, $\Sigma y^2 = 651$, find γ. Ans: 0.9485

11. If $n = 10$, $\Sigma xy = 3377$, $\Sigma x = 155$, $\Sigma y = 239$, $\Sigma x^2 = 2485$, $\Sigma y^2 = 5925$, find γ. Ans: -0.962

12. Coefficient of correlation between two variables 'x' and 'y' is 0.32. Their covariance is 10.56. The variance of x is 9. Find Standard deviation of 'y'. Ans: 11

13. If n=9 and $\Sigma d^2 = 24$, find the coefficient of rank correlation. Ans: 0.8

14. If n=10 and $\Sigma d^2 = 182$, find the coefficient of rank correlation. Ans: -0.1030

15. If the rank correlation coefficient is -0.6 and $\Sigma d^2 = 56$, find n. Ans: 6

16. Find Karl Pearson's coefficient of correlation from the following data.

x	100	101	102	102	100	99	97	98	96	95
y	98	99	99	97	95	92	95	94	90	91

Ans: 0.847

17. Calculate Pearson's coefficient of correlation from the following data.

x	12	9	8	10	11	13	7
y	14	8	6	9	11	12	3

Ans: 0.9485

18. Calculate Pearson's coefficient of correlation from the following data.

x	104	111	104	114	118	117	105	108	106	100	104	105
y	57	55	47	45	45	50	64	63	66	62	69	61

Ans: 0.6742

19. Calculate Pearson's coefficient of correlation from the following data.

x	40	42	46	48	50	56
y	10	12	15	23	27	30

Ans: 0.9562

20. Calculate Karl Pearson's coefficient of correlation from the following data.

x	36	41	46	59	46	65	31	68	41	70
y	48	60	53	36	50	42	66	44	58	65

Ans:-0.4022

21. Calculate the coefficient of correlation by Karl Pearson's method from the following data relating to overhead expenses and cost of production.

Overheads('000Rs)	80	90	100	110	120	130	140	150	160
Cost('000Rs)	15	15	16	19	17	18	16	18	19

Ans: -0.6928

22. Calculate Karl Pearson's coefficient of correlation for the following data.

y x	80 – 90	90 – 100	100 – 110	110 – 120	120 – 130
52.5	1	3	7	5	2
57.5	2	4	10	7	4
62.5	1	5	12	10	7
67.5	-	3	8	6	3

Ans : 0.0945

23. Calculate Karl Pearson's coefficient of correlation from the data given below.

Marks	Age in years				
	18	19	20	21	22
20 – 25	3	2	-	-	-
15 – 20	-	5	4	-	-
10 – 15	-	-	7	10	-
5 – 10	-	-	-	3	2
0 – 5	-	-	-	3	1

Ans: -0.837

24. Calculate Karl Pearson's coefficient of correlation.

x y	20 – 29	30 – 39	40 – 49	50 – 59
10 – 14	10	10	-	-
14 – 18	-	20	8	-
18 – 22	-	10	25	6
22 – 26	-	-	7	4

Ans: 0.7401

25. Calculate Spearman's rank correlation coefficient.

Advertisement cost (in '000 Rs)	39	65	62	90	82	75	25	98	36	78
Sales (in lakh Rs)	47	54	58	86	62	68	60	91	51	84

Ans : 0.8182

26. Following are the marks of 8 students in Statistics and Mathematics. Find coefficient of rank correlation.

Marks in Statistics	25	43	27	35	54	61	37	45
Marks in Mathematics	35	47	20	37	63	54	28	40

Ans: 0.8333

27. Following are the sales statistics of 6 sales representatives in two different weeks. Find the Spearman's coefficient of rank correlation.

Representatives	1	2	3	4	5	6
I week sales	60	110	65	40	70	20
II week sales	90	100	80	30	70	20

Ans: 0.7714

28. Calculate the coefficient of rank correlation from the following data.

x	80	78	75	75	68	67	60	59
y	12	13	14	14	14	16	15	17

Ans: -0.9286

29. The following data relate to the age of 10 employees and the number of days on which they reported sick in a month.

Age	28	32	38	42	46	52	54	57	58	59
Sick in a month	0	1	3	4	2	5	4	6	7	8

Calculate Spearman's coefficient of correlation and interpret its value. Ans: 0.935

Regression Analysis

After having established the fact that two variables are closely related, it may be interesting to estimate (predict) the value of one variable for a given value of another variable. For example, if we know that the expenditure on advertisement and sales are correlated, we may have to find expected amount of sales for a given advertising expenditure or the expected amount of expenditure for attaining a given amount of sales. Similarly, if we know that the yield of rice and rainfall are closely related we may have to find out the amount of rain required to achieve a certain production figure. Regression analysis reveals average relationship between two variables and this makes estimation or prediction possible .

The term **"Regression"** literally means "return to the origin" or "stepping back towards the average". A line describing this tendency to regress or step back is called a '**Regression line**'. Since we can estimate with fair amount of accuracy the position of this line, it is also called an 'estimating line'.

The term 'regression' was first used by a British biometrician, **Sir Francis Galton** (1822-1911), in connection with the inheritance of stature. Galton found that the offspring of abnormally tall or short parents tend to "regress" or "step back" to the average population height. But the term "Regression" as now used in statistics is only a convenient term without having any reference to biometry.

If two variables are correlated, the unknown value of one of the variables can be estimated by using the known value of the other variable.

The property of the tendency of the actual value to lie close to the estimated value is called 'Regression'. In a wider usage**, regression is the theory of estimation of unknown value of a variable with the help of known values of the other variables.**

Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data. It is a statistical device with the help of which we are in a position to estimate (or predict) the unknown values of one variable from a known values of another variable.

In regression analysis, there are two types of variables. The variable whose value is influenced or is to be predicted is called Dependent variable and the variable which influences the values or is used for prediction, is called Independent variable. In regression analysis independent variable is also known as **Regressor** or **Predictor** or **Explanator**. While the dependent variable is also known as **Regressed** or **Explained** variable.

If the variables in a bivariate data are correlated, we may find that the points in the scatter diagram will cluster in the form of a curve or in the form of a straight line. In the former case it is known as the **"Curve of Regression"** and the latter case it is known as **"line of regression".** Two variables are said to have linear relationship when the change in the value of the independent variable by one unit leads to a constant absolute change in the value of the dependent variable. When the two variables have linear relationship the regression equations can be used to find out the values of dependent variables.

The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable. Thus the line of regression is the 'line of best fit' and is obtained by the principles of least squares.

For a bivariate data on 'x' and 'y', the regression equation obtained with the assumption that 'x' is dependent on 'y' is called regression equation of x on y. The regression equation of x on y is : $(\mathbf{x}-\bar{\mathbf{x}}) = \mathbf{b}_{\mathbf{xy}} (\mathbf{y}-\bar{\mathbf{y}})$.

The regression equation obtained with the assumption that ' y ' is dependent on 'x' is called regression equation of y on x. The regression equation of y on x is: $(\mathbf{y}-\bar{\mathbf{y}}) = \mathbf{b}_{\mathbf{yx}} (\mathbf{x}-\bar{\mathbf{x}})$.

Here, the constants 'b_{xy}' and 'b_{yx}' are the regression coefficients.

The regression equation of x on y is used for the estimation of x values and the regression equation of y on x is used for the estimation of y values.

Graphical representation of the regression equations are called regression lines.

For ungrouped data,

$$\mathbf{b_{xy}} = \frac{\gamma\sigma_x}{\sigma_y} = \frac{n\Sigma xy-(\Sigma x)(\Sigma y)}{n\Sigma y^2-(\Sigma y)^2}$$

$$\mathbf{b_{yx}} = \frac{\gamma\sigma_y}{\sigma_x} = \frac{n\Sigma xy-(\Sigma x)(\Sigma y)}{n\Sigma x^2-(\Sigma x)^2}$$

For grouped data,

$$\mathbf{b_{xy}} = \frac{N\sum\sum xyf_{xy}-(\sum xf_x)(\sum yf_y)}{N\sum y^2f_y-(\sum yf_y)^2}$$

$$\mathbf{b_{yx}} = \frac{N\sum\sum xyf_{xy}-(\sum xf_x)(\sum yf_y)}{N\sum x^2f_x-(\sum xf_x)^2}$$

Where, f_{xy} bivariate frequencies, f_x marginal frequencies of x, f_y - marginal frequencies of y and N- Total frequency of given bivariate frequency distribution.

Properties of regression coefficients :

Regression coefficients are the coefficients of the independent variables in the regression equations.

i. The regression coefficient b_{xy} is the change occurring in x for a unit change in y. The regression coefficient b_{yx} is the change occurring in y for a unit change in x.

ii. Regression coefficients are independent of the change of origin but not of scale.

iii. The geometric mean of the regression coefficients is equal to the coefficient of correlation (numerically). That is

$$\gamma = \pm\sqrt{\mathbf{b_{xy}}.\mathbf{b_{yx}}}$$

we know that $\mathbf{b_{xy}} = \gamma\frac{\sigma_x}{\sigma_y}$ and $\mathbf{b_{yx}} = \gamma\frac{\sigma_y}{\sigma_x}$

$$b_{xy}.\,b_{yx} = (\gamma\frac{\sigma_x}{\sigma_y})\,(\gamma\frac{\sigma_y}{\sigma_x})$$

$$\therefore \quad b_{xy} . b_{yx} = \gamma^2$$

$$\therefore \quad \gamma = \pm\sqrt{b_{xy} . b_{yx}}$$

iv. Since, the coefficient of correlation numerically cannot be greater than one, the product of regression coefficients cannot be greater than one. That is, $\mathbf{b_{xy}.b_{yx} \leq 1}$

So, if one of the regression coefficients is greater than one, the other must be less than one.

v. The two regression coefficients cannot be of opposite signs. That is, both are positive together or both are negative together or zero together.

Note : 1. If 'γ' is positive then, both the regression coefficients will be positive.

2. If 'γ' is negative then, both the regression coefficients will be negative.

3. If 'γ' is zero then, both the regression coefficients will be zero.

Properties of Regression lines (Equations) :

i. The two regression lines intersect at $(\overline{x}, \overline{y})$ (mean values).

ii. If the variables are positively correlated then the regression lines will have positive slope. When the slope is negative, the variables are negatively correlated.

iii. If there is perfect correlation, the regression lines coincide (there will be only one regression line).

iv. When correlation does not exist (γ=0), the two regression lines are perpendicular to each other.

Uses of regression analysis :

i. Regression analysis helps in establishing a functional relationship between two or more variables. So, regression analysis provides an estimated value of the dependent variable for a given value of the independent variable. This can be used for predictions or estimation of the future production, prices, sales, investment, income, profits etc. . It is of immense use in business forecasting .

ii. With the help of regression coefficients, we can calculate the correlation coefficient which measures the degree of correlation that exists between the variables.

iii. Regression analysis is widely used in statistical estimations of demand curves, supply curves, production functions, cost functions, consumption functions, etc.

Difference between correlation and regression analysis :

Correlation	Regression
1. Correlation analysis deals with the association between two or more variables.	1. Regression analysis is concerned with the derivation of an appropriate functional relationship between variables.
2. Correlation need not imply cause and effect relationship between the variables under study.	2. Regression analysis clearly indicates the cause and effect relationship.
3. There may be nonsense correlation between two variables which is purely due to chance and has no practical relevance.	3. There is nothing like nonsense (spurious) regression.
4. r_{xy} and r_{yx} are symmetric ($r_{xy} = r_{yx}$). That is, it is immaterial which of x and y is dependent variable and which is independent variable.	4. b_{xy} and b_{yx} are not symmetric (b_{xy}? b_{yx}). Hence it definitely makes a difference as to which variable is dependent and which is independent.
5. Correlation coefficient is independent of change of scale and origin.	5. Regression coefficients are independent of change of origin but not of scale.

Problems on Regression equations :

Example : 1.

The following data relates to the age of husbands and wives.

Age of husband (Years)	25	28	30	32	35	36	38	39	42	45
Age of wife (Years)	20	26	29	30	25	18	26	35	35	46

Obtain the two regression equations and determine the most likely age of the husband when wife's age is 25 years. Also, determine the most likely age of the wife when husband's age is 30 years.

Solution : Let x denotes the age of husband and y denotes the age of wife.

x	y	u = x - 35	v = y - 30	u^2	v^2	uv
25	20	-10	-10	100	100	100
28	26	-7	-4	49	16	28
30	29	-5	-1	25	1	5
32	30	-3	0	9	0	0
35	25	0	-5	0	25	0
36	18	1	-12	1	144	-12
38	26	3	-4	9	16	-12
39	35	4	5	16	25	20
42	35	7	5	49	25	35
45	46	10	16	100	256	160
350	290	0	-10	358	608	324

$$\bar{x} = \frac{\sum x}{n} = \frac{350}{10} = 35$$

$$\bar{y} = \frac{\sum y}{n} = \frac{290}{10} = 29$$

$$b_{xy} = \frac{n\sum uv - \sum u \sum v}{n\sum v^2 - (\sum v)^2} = \frac{10 \times 324 - 0 \times (-10)}{10 \times 608 - (-10)^2} = 0.5418$$

$$b_{yx} = \frac{n\sum uv - \sum u \sum v}{n\sum u^2 - (\sum u)^2} = \frac{10 \times 324 - 0 \times (-10)}{10 \times 328 - (0)^2} = 0.905$$

Regression equation of x on y is,

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

i.e., $(x - 35) = 0.5418\ (y - 29)$

i.e., $x = 0.5418y - 15.7122 + 35$

i.e., $x = 0.5418y + 19.2878$

When $y = 25$, $x = 0.5418 \times 25 + 19.2878 =$ **32.83**

Thus the most likely age of husband is 33 years.

Regression equation of y on x is,

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

i.e., $(y - 29) = 0.905(x - 35)$

i.e., $y = 0.905x - 31.675 + 29$

i.e., $y = 0.905x - 2.675$

When $x = 30$, $y = 0.905 \times 30 - 2.675 =$ **24.48**

Thus the most likely age of wife is 24 years.

Example : 2.

Compute the regression equation of y on x from the following data.

x	2	4	5	6	8	11
y	18	12	10	8	7	5

Solution:

x	y	x^2	xy
2	18	4	36
4	12	16	48
5	10	25	50
6	8	36	48
8	7	64	56
11	5	121	55
36	60	266	293

$$\bar{x} = \frac{\sum x}{n} = \frac{36}{6} = 6$$

$$\bar{y} = \frac{\sum y}{n} = \frac{60}{6} = 10$$

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{6 \times 293 - 36 \times 60}{6 \times 266 - (36)^2} = \frac{-402}{300} = -1.3333$$

Regression equation of y on x is,

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

i.e., $(y - 10) = -1.3333(x - 6)$

i.e., $y = -1.3333x + 7.9998 + 10$

i.e., $y = -1.3333x + 17.9998$

Example : 3 Find the regression equation of x on y and predict the value of x when y is 9.

x	3	6	5	4	4	6	7	5
y	3	2	3	5	3	6	6	4

Solution:

x	3	6	5	4	4	6	7	5	$\sum x =$ 40
y	3	2	3	5	3	6	6	4	$\sum y =$ 32
y^2	9	4	9	25	9	36	36	16	$\sum y^2 = 144$
xy	9	12	15	20	12	36	42	20	$\sum xy = 166$

$$\bar{x} = \frac{\sum x}{n} = \frac{40}{8} = 5$$

$$\bar{y} = \frac{\sum y}{n} = \frac{32}{8} = 4$$

$$b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2} = \frac{8 \times 166 - 40 \times 32}{8 \times 144 - (32)^2} = 0.375$$

Regression equation of x on y is,

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

i.e., $(x - 5) = 0.375\ (y - 4)$

i.e., $x = 0.375y - 1.5 + 5$

i.e., $x = 0.375y + 3.5$

When $y = 9$, $x = 0.375 \times 9 + 3.5 = \mathbf{6.875 \simeq 7}$

Example : 4

Find the two regression lines from the following data.

x	55	57	58	59	59	60	61	62	64
y	74	77	78	75	78	82	82	79	81

Solution:

x	y	x^2	y^2	xy
55	74	3025	5476	4070
57	77	3249	5929	4389
58	78	3364	6084	4524
59	75	3481	5625	4425
59	78	3481	6084	4602
60	82	3600	6724	4920
61	82	3721	6724	5002
62	79	3844	6241	4898
64	81	4096	6561	5184
535	706	31861	55448	42014

$$\bar{x} = \frac{\sum x}{n} = \frac{535}{9} = 59.4444$$

$$\bar{y} = \frac{\sum y}{n} = \frac{706}{9} = 78.4444$$

$$b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2} = \frac{9 \times 42014 - 535 \times 706}{9 \times 55448 - (706)^2} = 0.698$$

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{9 \times 42014 - 535 \times 706}{9 \times 31861 - (535)^2} = 0.7939$$

Regression equation of x on y is,

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

i.e., $(x - 59.4444) = 0.698\ (y - 78.4444)$

i.e., $x = 0.698y - 54.7542 + 59.4444$

i.e., x = 0.698y + 4.6902

Regression equation of y on x is,

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

i.e., $(y - 79.4444) = 0.7939(x - 59.4444)$

i.e., $y = 0.7939x - 47.1929 + 78.4444$

i.e., $y = 0.7939x + 31.2515$

Example : 5.

Following is the distribution of students according to their height (inches) and weight (lbs).

Height	Weight			
	90 - 100	100 - 110	110 - 120	120 - 130
50 - 55	4	7	5	2
55 - 60	6	10	7	4
60 - 65	6	12	10	7
65 - 70	3	8	6	3

Solution:

	y								
x	95	**105**	115	125	f_x	u	uf_x	u^2f_x	uvf_{xy}
52.5	4 4	0 7	-5 5	-4 2	18	-1	-18	18	-5
57.5	0 6	0 10	0 7	0 4	27	0	0	0	0
62.5	-6 6	0 12	10 10	14 7	35	1	35	35	18
67.5	-6 3	0 8	12 6	12 3	20	2	40	80	18
f_y	19	37	28	16	**N = 100**		**57**	**133**	**31**
v	-1	0	1	2					
vf_y	-19	0	28	32	**41**				
v^2f_y	19	0	28	64	**111**				
uvf_{xy}	-8	0	17	22	**31**				

$$\bar{x} = a + c\bar{u} = 57.5 + 10 \times \frac{57}{100} = \mathbf{60.35}$$

$$\bar{y} = b + d\bar{v} = 105 + 10 \times \frac{41}{100} = \mathbf{109.1}$$

$$b_{xy} = \frac{c}{d}\left[\frac{N\sum\sum uvf_{xy} - \sum uf_x \sum vf_y}{N\sum v^2 f_y - \left(\sum vf_y\right)^2}\right]$$

$$= 0.5 \left[\frac{100 \times 31 - 57 \times 41}{100 \times 111 - 41^2}\right]$$

$$= 0.5 \left[\frac{763}{9419}\right]$$

$$= 0.5 \times 0.081$$

$$= \mathbf{0.0405}$$

$$b_{yx} = \frac{d}{c}\left[\frac{N\sum\sum uvf_{xy} - \sum uf_x \sum vf_y}{N\sum u^2 f_x - (\sum uf_x)^2}\right]$$

$$= 2 \left[\frac{100 \times 31 - 57 \times 41}{100 \times 133 - 57^2}\right]$$

$$= 2 \left[\frac{763}{10051}\right]$$

$$= 2 \times 0.0759$$

$$= \mathbf{0.1518}$$

Regression equation of x on y is,

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

i.e., $(x - 60.35) = 0.0405\,(y - 109.1)$

i.e., $x = 0.0405y - 4.4186 + 60.35$

i.e., $x = 0.0405y + 55.9314$

Regression equation of y on x is,

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

i.e., $(y - 109.1) = 0.1518(x - 60.35)$

i.e., $y = 0.1518x - 9.1611 + 109.1$

i.e., $y = 0.1518x + 99.9389$

Example : 6. Obtain the regression equation of y on x and hence estimate y, when x = 26.

x	y				
	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
0 – 10	4	6	-	-	-
10 – 20	-	5	9	1	-
20 – 30	-	-	4	3	3
30 – 40	-	-	-	6	3
40 – 50	-	-	-	4	2

Solution:

x	y					fx	u	uf_x	$u^2 f_x$	$uv\ f_{xy}$
	5	15	**25**	35	45					
5	$\boxed{16}$ 4	$\boxed{12}$ 6	–	–	–	10	-2	-20	40	28
15	–	$\boxed{5}$ 5	$\boxed{0}$ 9	-1 1	–	15	-1	-15	15	4
25	–	–	$\boxed{0}$ 4	$\boxed{0}$ 3	$\boxed{0}$ 3	10	0	0	0	0
35	–	–	–	$\boxed{6}$ 6	$\boxed{6}$ 3	9	1	9	9	12
45	–	–	–	8 4	8 2	6	2	12	24	16
fy	4	11	13	14	8	**N = 50**		**-14**	**88**	**60**
v	-2	-1	0	1	2					
vf_y	-8	-11	0	14	16	**11**				
v^2f_y	16	11	0	14	32	**73**				
uvf_{xy}	16	17	0	13	14	**60**				

$$\bar{x} = a + c\bar{u} = 25 + 10 \times \frac{-14}{50} = \mathbf{22.2}$$

$$\bar{y} = b + d\bar{v} = 25 + 10 \times \frac{11}{50} = \mathbf{27.2}$$

$$b_{yx} = \frac{d}{c}\left[\frac{N\sum\sum uvf_{xy} - \sum uf_x \sum vf_y}{N\sum u^2 f_x - (\sum uf_x)^2}\right]$$

$$= \frac{10}{10}\left[\frac{50 \times 60 - (-14)11}{50 \times 88 - (-14)^2}\right]$$

$= \mathbf{0.7502}$

Regression equation of y on x is,

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

i.e., $(y - 27.2) = 0.7502(x - 22.2)$

i.e., $y = 0.7502x - 16.6544 + 27.2$

i.e., $y = 0.7502x + 10.5456$

When x = 26, $y = 0.7502 \times 26 + 10.5456 = 30.0508 \simeq \mathbf{30}$

Example : 7 Obtain the regression equation of x on y and hence estimate x, when y = 20.

	y				
x	15	16	17	18	19
50 – 52	4	1	-	-	-
52 – 54	3	5	2	1	-
54 – 56	-	3	4	2	3
56 – 58	-	1	3	3	2
58 – 60	-	-	1	1	1

Solution:

	y									
x	15	16	**17**	18	19	f_x	u	uf_x	u^2f_x	uvf_{xy}
51	$\boxed{16}$ 4	$\boxed{2}$ 1		–	–	5	-2	-10	20	18
53	$\boxed{6}$ 3	$\boxed{5}$ 5	$\boxed{0}$ 2	$\boxed{-1}$ 1	–	11	-1	-11	11	10
55	–	$\boxed{0}$ 3	$\boxed{0}$ 4	$\boxed{0}$ 2	$\boxed{0}$ 3	12	0	0	0	0
57	–	$\boxed{-1}$ 1	$\boxed{0}$ 3	$\boxed{3}$ 3	$\boxed{4}$ 2	9	1	9	9	6
59	–	–	0 1	2 1	4 1	3	2	6	12	6
f_y	7	10	10	7	6	**N = 40**		**-6**	**52**	**40**
v	-2	-1	0	1	2					
vf_y	-14	-10	0	7	12	**-5**				
v^2f_y	28	10	0	7	24	**69**				
uvf_{xy}	22	6	0	4	8	**40**				

$$\bar{x} = a + c\bar{u} = 55 + 2 \times \frac{-6}{40} = \mathbf{54.7}$$

$$\bar{y} = b + d\bar{v} = 17 + 1 \times \frac{-5}{40} = \mathbf{16.875}$$

$$b_{xy} = \frac{c}{d}\left[\frac{N\sum\sum uvf_{xy} - \sum uf_x \sum vf_y}{N\sum v^2 f_y - \left(\sum vf_y\right)^2}\right]$$

$$= \frac{2}{1}\left[\frac{40 \times 40 - (-6)(-5)}{40 \times 69 - (-5)^2}\right]$$

$$= 2\left[\frac{1570}{2735}\right]$$

$$\mathbf{= 1.1481}$$

Regression equation of x on y is,

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

i.e., (x – 54.7) = 1.1481 (y – 16.875)

i.e., x = 1.1481y – 19.3742 + 54.7

i.e., x = 1.1481y + 35.3258

When y = 20,

x = 1.1481× 20 + 35.3258 = 58.2878 ≃ **58**

Example : 8.

In a correlation analysis between production and price of a commodity, the following data are obtained,

	Production index (x)	Price index (y)
Arithmetic Mean	110	98
Standard Deviation	12	5

Coefficient of correlation between production and price is -0.4. Write down the regression equation of price on production and estimate the price index when the production index is 116.

Solution :

Given that $\bar{x} = 110, \bar{y} = 98, \sigma_x = 12, \sigma_y = 5$ and $\gamma = -0.4$

We need the regression equation of y on x,

$$b_{yx} = \frac{\gamma\sigma_y}{\sigma_x} = \frac{-0.4 \times 5}{12} = -\mathbf{0.1667}$$

Regression equation of y on x is,

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

i.e., $(y - 98) = -0.1667(x - 110)$

i.e., $y = -0.1667x + 18.337 + 98$

i.e., $y = -0.1667x + 116.337$

When x = 116,

$y = -0.1667 \times 116 + 116.337 = 96.9998 \simeq \mathbf{97}$

So the price index is **97** when the production index is 116.

Example : 9.

From the following data regarding the amount of rainfall (x) and the production of rice(y). Find the most likely production corresponding to the rainfall of 40cm.

	Rainfall(cm)	Production(Quintals)
Mean	35	50
Variance	25	64

Solution:

Given that, $\bar{x} = 35, \bar{y} = 50, \sigma_x = \sqrt{25} = 5, \sigma_y = \sqrt{64} = 8$ and $\gamma = 0.8$

We need the regression equation of y on x,

$$b_{yx} = \frac{\gamma\sigma_y}{\sigma_x} = \frac{0.8 \times 8}{5} = \mathbf{1.28}$$

Regression equation of y on x is,

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

i.e., $(y - 50) = 1.28(x - 35)$

i.e., $y = 1.28x - 44.8 + 50$

i.e., $y = 1.28x + 5.2$

When $x = 40$, $y = 1.28 \times 40 + 5.2 =$ **56.4**

Thus, the most likely production is 56.4 quintals when the rainfall is 40cm.

Example : 10. If $b_{xy} = \frac{-3}{4}$ and $b_{yx} = \frac{-1}{3}$, find γ_{xy}.

Solution :

We know that, $|\gamma| = \sqrt{b_{xy} . b_{yx}}$

$$= \sqrt{\frac{-3}{4} \times \frac{-1}{3}} = 0.5$$

$$\therefore \gamma_{xy} = -0.5$$

Example : 11.

Estimate the value of x when y = 20, using the following data.

	x	y
Mean	25	30
S.D.	5	4

and $\gamma = 0.8$.

Solution :

Given that $\bar{x} = 25, \bar{y} = 30, \sigma_x = 5, \sigma_y = 4$ and $\gamma = 0.8$

We need the regression equation of x on y.

$$b_{xy} = \frac{\gamma\sigma_x}{\sigma_y} = \frac{0.8 \times 5}{4} = 1$$

Regression equation of x on y is,

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

i.e., $(x - 25) = 1\ (y - 30)$

i.e., $x = y - 30 + 25$

i.e., $x = y - 5$

When y = 20,

x= 20 – 5 = **15**

Example : 12. If γ_{xy}= 0.8, $\sigma_x = 3$ and $b_{xy} = 0.5$ find σ_y

Solution :

We know that, $b_{xy} = \gamma \dfrac{\sigma_x}{\sigma_y}$

$$\text{i.e.}, 0.5 = \frac{0.8 \times 3}{\sigma_y}$$

$$\therefore \sigma_y = \frac{2.4}{0.5} = \mathbf{4.8}$$

Example : 13. If γ_{xy}= 0.8 and b_{yx}= 0.75, find b_{xy}.

Solution:

We know that, $|\gamma| = \sqrt{b_{xy}.b_{yx}}$

$$\therefore \gamma^2 = b_{xy}.b_{yx}$$

$$\text{i.e.}, b_{xy} = \frac{\gamma^2}{b_{yx}} = \frac{(0.5)^2}{0.75} = \mathbf{0.3333}$$

Example : 14.

If the two regression equations are 2x – y + 3 = 0 and x – 3y + 6 =0. Find $\bar{x}$, $\bar{y}$ and γ_{xy}.

Solution:

Given that,

2x – y + 3 = 0------------------>(1)

x – 3y + 6 =0------------------->(2)

By solving equations (1) and (2) we get, x = -0.6 and y = 1.8.

$\therefore$ $\mathbf{\bar{x} = -0.6}$ **and** $\mathbf{\bar{y} = 1.8}$

Consider 2x – y + 3 = 0

$\text{i.e.}, x = \dfrac{1}{2}(y - 3)$

$\therefore\ \mathbf{b_{xy} = \frac{1}{2}}$

Consider x – 3y + 6 =0

$\text{i.e.,}\ y = \frac{1}{3}(x + 6)$

$\therefore\ \mathbf{b_{yx} = \frac{1}{3}}$

So, $|\gamma| = \sqrt{b_{xy}.b_{yx}}$

$= \sqrt{\frac{1}{2} \times \frac{1}{3}}$

$= \mathbf{0.4082}$

Example : 15.

If the two regression equations are 3x + 5y = 3 and 4x + 3y = 4, find the mean values of x and y and also the coefficient of correlation between x and y.

Solution : Given that,

3x + 5y = 3----------------------->(1)

4x + 3y = 4----------------------->(2)

By solving equations (1) and (2) we get, x = 1 and y = 0.

$\therefore\ \mathbf{\bar{x} = 1\ \text{and}\ \bar{y} = 0}$

Consider 3x + 5y = 3

$\text{i.e.,}\ y = \frac{-3}{5}\ (x - 1)$

$\therefore\ \mathbf{b_{yx} = \frac{-3}{5}}$

Consider 4x + 3y = 4

$\text{i.e.,}\ x = \frac{-3}{4}y + 1$

$$\therefore\ \mathbf{b_{xy}} = \frac{\mathbf{-3}}{\mathbf{4}}$$

So, $|\gamma| = \sqrt{b_{xy}.b_{yx}} = \sqrt{\frac{-3}{4} \times \frac{-3}{5}} = \mathbf{0.6708}$

i.e., $\gamma = -\ \mathbf{0.6708}$

Note : If the two regression equations are taken in the other way then the value of $|\gamma|$ becomes greater than 1. This is not possible.

Example:16.

In a bivariate data, $\sum x = 30$, $\sum y = 400$, $\sum xy = 850$, $\sum x^2 = 196$, $\sum y^2 = 46500$ and n = 10. Estimate the value of y corresponding to the value of x = 5.

Solution :

We need the regression equation of y on x.

$$\bar{x} = \frac{\sum x}{n} = \frac{30}{10} = 3$$

$$\bar{y} = \frac{\sum y}{n} = \frac{400}{10} = 40$$

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{10 \times 850 - 30 \times 400}{10 \times 196 - (30)^2} = -3.3$$

Regression equation of y on x is,

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

i.e., $(y - 40) = -3.3(x - 10)$

i.e., $y = -3.3x + 33 + 40$

i.e., $y = -3.3x + 73$

When x = 5, $y = -3.3 \times 5 + 73 = \mathbf{56.5}$

Questions

1. What is meant by regression ?
2. Write the regression equation of x on y.
3. Write the regression equation of y on x.
4. If y = 0.26x + 7.94 is the regression equation of y an x, then find y when x= 6. Ans: y = 9.5
5. Write the relationship between correlation coefficient and regression coefficients.
6. What is your conclusion when the regression lines are perpendicular ?
7. Write the coordinates of the point of intersection of the two regression equations.
8. What is your conclusion when the regression lines coincide ?
9. Mention two properties of regression coefficients.
10. Mention two properties of regression lines.
11. Prove that $\gamma = \pm\sqrt{b_{xy}\, b_{yx}}$.
12. Mention two uses of regression analysis.
13. Mention two differences between correlation and regression analysis.
14. Mention the properties of regression coefficients.
15. Mention the properties of the regression lines.
16. Distinguish between correlation and regression.

Exercise Problems

1. The following figures relate to years of service and income in thousands of rupee of the employees of an organization. Find the initial start for a person applying for a job after having served in another factory for a period of 12 years in a similar capacity.

Length of service (years)	11	7	9	5	8	6	10
Income (thousands of rupee)	7	5	3	2	6	4	8

Ans. y=0.75x-1, Rs. 8000

2. Find the coefficient of correlation between x and y from the following data.

x	3	6	5	4	4	6	7	5
y	3	2	3	5	3	6	6	4

Also obtaine the regression equation of y on x and predict the average value of y when x is 9. Ans. $\gamma = 0.433$, $y = 0.5x+1.5$, $y=6$

3. From the following data regarding the age of husband and the age of wife, form the two regression lines and estimate the age of husband when the age of wife is 16 years.

Husband's age (Years)	36	23	27	28	28	29	30	31	33	35
Wife's age (Years)	29	18	20	22	27	21	29	27	29	28

Ans. $x = 0.75y+11.25$, $y = 0.8913x-1.739$, $x = 23.25$

4. The height (x cm) and weight (y kg) of 6 students are as follows. Obtain the two regression equations. Also, find the expected height of a student whose weight is 60 kg.

x	153	157	168	160	170	163
y	48	50	50	49	54	53

Ans. $x = 2.04y+58.46$, $x=180.86$cm, $y = 0.264x+7.9469$

5. You are given with the following information about the expenditure on advertisement and sales :

	Advertisement Expenditure (Crore Rs.)	Sales (Crore Rs.)
Mean	20	120
S.D.	5	2

Correlation coefficient = 0.8

i. Obtain the two regression equations.

ii. Find the likely sales when the expenditure on advertisement is Rs. 25 crores.

iii. What should be the budget on advertisement if the company wants to attain a sales target of Rs. 150 Crores ?

Ans. x=2y-220, y=0.32x+113, y=121 crores, x=80 crores.

6. Following are the details of the marks scored by students in Malayalam and English examination.

	Malayalam	English
Mean	40	50
S.D.	10	16

and coeffient of correlation is 0.3.

Estimate the scores in English when the score in Malayalam is 50 and also, the scores in Malayalam when the score in English is 30.

Ans. x-40=0.1875(y-50), x=36.25, y-50=0.48(x-40),y=54.8

7. The regression equations of a bivariate distribution are :

Regression equation of y on x is 4y = 9x + 15

Regression equation of x on y is 25x = 6y + 7.

Find $\bar{x}$, $\bar{y}$ and γ. Ans: $\bar{x} = 2.57$, $\bar{y} = 9.52$, $\gamma = 0.735$

8. In a laboratory experiment on correlation research study, the equation to the two regression lines were found to be 2x - y + 1 =0 and 3x - 2y + 7 = 0. Find the means of x and y. Also, workout the values of regression coefficients and the coefficient of correlation between the two variables x and y.

$$\text{Ans}: \bar{x}=5,\ \bar{y}=11,\ b_{xy}=\frac{1}{2},\ b_{yx}=\frac{3}{2} \text{ and } \gamma=0.866$$

9. The two regression lines obtained from a certain data were y=x+5 and 16x=9y-94. Find the variance of 'x if, the variance of 'y' is 16. Also, find covariance between 'x' and 'y'.

$$\text{Ans}: b_{xy}=\frac{9}{16},\ b_{yx}=1,\ \gamma=\frac{3}{4},\ V(x)=9 \text{ and Covariance} = 9$$

10. The following data relates to the regression analysis over a paired sample of six items. $\Sigma x=18$, $\Sigma y=216$, $\Sigma xy=738$, $\Sigma x^2=370$ and $\Sigma y^2=8332$. Obtain the two regression equations. Also, estimate y when x=4.

Ans. x=0.1619y-35.1456, y=0.2848x-2.8273, y=-1.6881

11. From the following data obtain the two regression equations.

$n=10,\ \Sigma x=550,\ \Sigma y=680,\ \Sigma xy=45900,\ \Sigma x^2=38500,\ \Sigma y^2=56000.$

Ans. x-55=0.8709(y-68), y-68=1.0303(x-55)

12. From the following data regarding the marks obtained by 130 students in two tests, obtain the two regression equations.

Test-1(x)
Test-2(y)
	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80

20 – 30	2	5	3	-	-
30 – 40	1	8	12	6	-
40 – 50	-	5	22	14	1
50 – 60	-	2	16	9	2
60 – 70	-	1	8	6	1
70 – 80	-	-	2	4	2

Ans. y=0.7x+8.14, x=0.29y+42.93

13. Calculate the two regression equations from the following bivariate table and determine the value of γ_{xy}.

Y X	0 – 10	10 – 20	20 – 30	30 – 40
10 – 20	5	4	3	-
20 – 30	7	6	7	6
30 – 40	-	5	-	7

Ans. y=0.667x+3.125, γ=0.413, x=0.256y+19.931

14. Obtain the two regression equations from the following data regarding the marks in physics(x) and chemistry(y).

y x	5 – 15	15 – 25	25 – 35	35 – 45
0 – 10	1	1	-	-
10 – 20	3	6	5	1
20 – 30	1	8	9	2
30 – 40	-	3	9	3
40 – 50	-	-	4	4

Ans. y=0.67x+8.91, x=0.4452y+14.98

15. Obtain the lines of regression for the following bivariate frequency distribution.

Sales revenue (in '000 Rs)	Advertisement expenditure (in '000 Rs)			
	5 – 15	15 – 25	25 – 35	35 – 45
75 – 125	4	1	-	-
125 – 175	7	6	2	1
175 – 225	1	3	4	2
225 – 275	1	1	3	4

Ans. x=2.6579y+118.9472, y=0.1337x-1.39

16. If b_{xy} = -7.3 and b_{yx}= -0.11, find γ_{xy}. Ans. -0.8961

17. If $b_{xy} = \frac{-4}{3}$ and $b_{yx} = \frac{-2}{3}$, find γ_{xy}. Ans. -0.9428

18. If b_{xy} = 0.4 and b_{yx}= 1.6, find γ. Ans. 0.8

19. If b_{xy} = 0.25 and b_{yx}= 0.5, find γ. Ans. 0.3536

20. If y = 0.45x + 2 is regression equation of y on x and x = 0.5y – 4 is regression equation of x on y, find the values of b_{xy} and b_{yx}.

Ans. 0.5, 0.45

21. If γ = 0.6, σ_x = 10 , σ_y = 15, find the values of b_{xy} and b_{yx}.

Ans. 0.4, 0.9

22. If γ = 0.4, σ_x = 12 , σ_y = 15, find the values of b_{xy} and b_{yx}.

Ans. 0.32 , 0.5

*****

Unit-VII

ASSOCIATION OF ATTRIBUTES

As we know that correlation coefficient measures the degree of relationship between the variables; such as height and weight of persons, ages of husbands and wives, demand and supply of items etc. On the other hand the method of association of attributes measures the degree of relationship between the attributes; such as sex and literacy, literacy and employment, smoking and tea drinking, intelligence and employment etc.

Notations and Terminology :

In the discussion of attributes the data can be classified as presence and absence of a particular characteristic. Capital letters A and B are used to represent the presence of attributes and (A), (B) are used to represent the number of units which possess A and B. Greek letters as 'α' (alpha) and 'β' (beta) are used to represent the absence of attributes A and B respectively.

The observations in different classes are called 'class frequencies'. The class frequencies are denoted by (A), (α), (AB), $(A\beta)$ etc.

For example, if (A) represents number of men, (α) represents number of women and if (B) represents number of smokers, (β) represents number of non-smokers. The combinations of attributes are denoted by (AB), $(\alpha\,\beta)$, (αB) and $(A\beta)$. Thus (AB) represents men smokers, $(A\beta)$ represents men non-smokers, (αB) represents women smokers and $(\alpha\,\beta)$ represents women non-smokers.

The above information can be presented in the form a 2 × 2 contingency table as :

	A	α	Total
B	(AB)	(αB)	(B)
β	$(A\beta)$	$(\alpha\beta)$	(β)
Total	(A)	(α)	N

Note : (1) The data classified according to attributes and represented by a table called contingency table.

(2) In the above table excluding the marginal totals there are two rows and two columns, therefore the table known as 2×2 contingency table. The table has nine cells. Hence, it is also called as nine square table.

Number of class frequencies : In the study of 'n' attributes the total number of class frequencies is given by $3^{n.}$ For one attribute, total frequencies are $3^1 = 3$. They are in the order 1+2=3.

For two attribute, total frequencies are $3^2 = 9$. They are in the order 1+4+4=9.

Order of the Classes : The order of the class depends upon the number of attributes specified. A class having one attribute is known as the class of the first order, class of the combination two attributes as class of the second order and so on. The total number of observations is denoted by N.

Here, N is frequency of zero order, because it has no attribute to indicate.

(A), (B), (α), (β) are called frequencies of the first order, and

(AB), (αB), $(A\beta)$, $(\alpha\beta)$ are called frequencies of the second order.

Methods of studying Association : Important methods of measuring the association of two attributes are

i. Comparison of Observed and Expected frequencies method.

ii. Proportion method.

iii. Yule's Coefficient of Association.

iv. Coefficient of Colligation.

v. Coefficient of Contingency.

Here, our discussion is confined only to Yule's Coefficient of Association.

Yule's Coefficient of Association : The degree or the extent of the two attributes is associated can be measured using Yule's coefficient of association. That is, whether the attributes are negatively associated, positively associated or independent and extent of association. The Yule's Coefficient of association is denoted by symbol 'Q' and is obtained by applying the formula :

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

Interpretation :

On the basis of the Yule's Coefficient of Association (Q) we can interpret the association between the two attributes as below :

The value of this coefficient lies between ±1 (i.e., $-1 \leq Q \leq 1$). When the value of Q is +1 there is perfect positive association between the attributes. When Q is -1 there is perfect negative association (perfect dissociation) between the attributes and when the value Q is zero the two attributes are independent.

Example : 1.

In case of two attributes, if N=250, (AB)=30, (A)=100 and (B)=50. Find the remaining classes and their frequencies.

Solution :

Here, the missing frequencies are calculated, by putting the known frequencies in a nine square table :

	A	α	Total
B	(AB)= 30	(αB)= 20*	(B)= 50
β	(Aβ)= 70*	(αβ)=130*	(β)=200*
Total	(A)= 100	(α)=150*	N=250

Note : In the above table figures with '*' indicates, calculated frequencies.

Example : 2.

Prepare a 2×2 contingency table from the following information. Calculate the Yule's Coefficient of Association and interpret the result.

N=760, (α) = 558, (B) = 180, (AB) = 18.

Solution:

By putting the known values in a 2×2 contingency table find the unknown values to determine the Yule's Coefficient of Association.

	A	α	Total
B	(AB) 18	(αB)**162**	(B)180
β	(Aβ)**184**	($\alpha\beta$)**396**	(β)**580**
Total	(A)**202**	(α)558	N=760

$$\text{Yule's Coefficient of Association: } Q = \frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$$

$$= \frac{(18)(396)-(184)(162)}{(18)(396)+(184)(162)}$$

$$= \frac{7128-29808}{7128+29808}$$

$$= \frac{-22680}{36936}$$

$$\therefore \mathbf{Q = -\ 6140}$$

There exists a negative association between the attributes.

Example : 3.

Prepare a nine square with the following information. Calculate the Yule's Coefficient of Association and interpret the result.

(A) = 450, (B) = 600, (Aβ) = 100, N=1000.

Solution :

By putting the known values in the nine square tables find the unknown values to determine the Yule's Coefficient of Association.

	A	α	Total
B	(AB)**350**	(αB)**250**	(B)600
β	(Aβ)100	($\alpha\beta$)**300**	(β)**400**
Total	(A)450	(α)**550**	N=1000

$$\text{Yule's Coefficient of Association: } Q = \frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$$

$$= \frac{(350)(300)-(100)(250)}{(350)(300)+(100)(250)}$$

$$= \frac{105000-25000}{105000+25000}$$

$$= \frac{80000}{130000}$$

$$\therefore \mathbf{Q = 0.6154}$$

There exists a positive association between the attributes.

Example : 4.

In a survey regarding the effect of vaccination in prevention of smallpox, the following information was obtained;

"Out of 2000 persons in a village exposed to smallpox, 450 were attacked, 365 were vaccinated and of these only 50 were attacked."

Can vaccination be regarded as a preventive measure for smallpox ?

Smallpox	A(vaccinated)	α (not vaccinated)	Total
B(attacked)	(AB)50	(αB) **400**	(B) 450
β (not attacked)	(Aβ)**315**	(αβ)**1235**	(β)**1550**
Total	(A)365	(α)**1635**	N=2000

$$\text{Yule's Coefficient of Association: } Q = \frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$$

$$= \frac{(50)(1235)-(315)(400)}{(50)(1235)+(315)(400)}$$

$$= \frac{61750-126000}{61750+126000}$$

$$= \frac{-64250}{187750}$$

$$\therefore \mathbf{Q = -\ 0.3422}$$

There exists a low degree of negative association between vaccination (A) and attack of smallpox (B).

Example : 5.

Following are the survey results of a literate persons and the employment at a village. Find Yule's Coefficient of Association and interpret.

Total Adults	= 5000
Literates	= 645
Employed	= 695
Literate employed	= 410

Solution :

Let A and B denotes the literates and employed, α and β denotes the illiterates and unemployed. Now, identify the class frequencies and they are written in their notations as :

N=5000, (A)=645, (B)=695, (AB)=410. Putting the known values in the nine square table and find the missing values.

	A	α	Total
B	(AB)410	(αB) **285**	(B)695
β	(Aβ)**235**	(αβ)**4070**	(β)**4305**
Total	(A)645	(α)**4355**	N=5000

Yule's Coefficient of Association: $Q = \dfrac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$

$$= \frac{(410)(4070)-(235)(285)}{(410)(4070)+(235)(285)}$$

$$= \frac{1668700-66975}{668700+66975}$$

$$= \frac{1601725}{1735675}$$

$\therefore$ **Q = 0.9228**

There exists a high degree of positive association between literacy and employment.

Example : 6.

In a co-educational institution, out of 200 students 150 were boys. In an examination 160 students were passed, 10 girls had failed. Is there any association between sex and success in the examination ?

Solution :

Let A and B denotes boys and passed the examination, α and β denotes girls and failed.

Therefore, the class frequencies are :

N=200, (A)=150, (B)=160, $(\alpha\beta)$=10. Putting the known values in the nine square table and the missing values are calculated.

	A	α	Total
B	(AB)**120**	(αB)**40**	(B)160
β	(Aβ) **30**	(αβ)10	(β) **40**
Total	(A)150	(α)**50**	N=200

Yule's Coefficient of Association: $Q = \dfrac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$

$$= \frac{(120)(10)-(30)(40)}{(120)(10)+(30)(40)}$$

$$= \frac{1200-1200}{1200+1200}$$

$$= \frac{0}{2400}$$

$\therefore$ **Q = 0**

Hence, there is no association between sex and success in the examination.

Questions

1. What is meant by Association of Attribute ? How does it differ from correlation?

2. What is the difference between coefficient of correlation and association of attributes?

3. Write the formula of Yule's coefficient of Association.

4. Calculate Yule's coefficient of Association between marriage and failure of students from the following data pertaining to 525 students.

	Passed	Failed
Married	90	65
Unmarried	260	110

Ans: - 0.2612

5. Eighty eight residents of a city who were interviewed during a sample survey are classified below, according to their smoking and tea drinking habits. Calculate Yule's coefficient of Association and comment on its value.

	Smokers	Non-smokers
Tea drinkers	40	33
drinkers	3	12

Ans: 0.6580

6. Compute Yule's coefficient of Association from the following data.

   (AB)=150, N=1000, (A)=200, (B)= 300. Ans : 0 .8571

7. Compute Yule's co-efficient of Association from the following data .

   N=250, (Aβ)=70, (A)=100, (B)=50. Ans : 0 .4717

8. Given N=500, (αβ)=280, (A)=160 and (B)=200. Calculate Yule's coefficient of Association. Ans : 0 .9406

9. Given N=2500, (AB)=400, (α)=2100 and (β)=900

   Calculate Yule's coefficient of Association. Ans : 1

10. Find the association between intelligence of fathers and the intelligence of sons from the following data:

    Intelligent fathers with intelligent sons : 50

    Dull fathers with intelligent sons : 100

    Dull fathers with dull sons : 300

    Intelligent fathers with dull sons : 200

    Ans: -0.142

11. 2000 candidates appeared for a competitive examination. 400 came out successful. 350 had attended a coaching class and of these 200 had come out successful. Estimate the utility of coaching classes, using Yule's coefficient of Association.

    Ans : 0 .8125

12. 200 candidates appeared for II PUC Examination in a college and 60 of them succeeded. 35 received a special coaching in tutorial class and out of them 20 candidates succeeded. Using Yule's co-efficient, discuss whether the special coaching is effective or not.

    Ans : 0 .6129

*****

Unit-VIII

INTERPOLATION AND EXTRAPOLATION

If there are two variables which are interdependent, that is there is a **cause** and **effect** relationship between them, then one variable is known as dependent variable and the other as independent variable.

Suppose, we have two variables x and y, where x is independent and y is dependent, then a functional relationship can be written as follows :

$$y = f\,(x) \text{ or } y_{x.}$$

Where f(x) is a function of x. it can be represented in a tabular form as follows.

x	y
x_0	y_0
x_1	y_1
x_2	y_2
-	-
-	-
-	-
x_n	y_n

The procedure of estimating the missing value of y for a given value of x, where x is within the limits x_0 and x_n is called **Interpolation.** That is **'Interpolation is a procedure of estimating the unknown value of dependent variable for a given value of independent variable which is within the limits or the range of the independent variable'.**

But if the value of y is to be estimated for a value of x which is outside the limits $x_{0,}$ and $x_{n,}$ then procedure is known as the technique of **Extrapolation.**

Extrapolation is a procedure of estimating the unknown value of dependent variable for a given value of independent variable which is outside the limits or the range of the independent variable'.

Assumptions of Interpolation and Extrapolation :

In making use of the techniques of interpolation and extrapolation the following assumptions are made

1. There are no sudden jumps in the values of dependent variable from one period to another.
2. The rate of change of figures of the dependent variable from one period to another must be uniform.

Methods of Interpolation:

The methods of interpolation or extrapolation may be broadly classified as follows :

(i) Graphic Method.

(ii) Algebraic Method .

According to the syllabus of I PUC Statistics only Algebraic method will be discussed.

Algebraic Methods :

A number of algebraic methods, based on different assumptions have been developed for interpolation or extrapolation. Some of the commonly used methods are :

(i) The Binomial Expansion Method.

(ii) Method of finite differences.

(iii) Lagrange's method.

Our discussion is confined only to binomial expansion method.

Before we consider this method we shall study some mathematical preliminaries.

(a) Polynomial.

The mathematical expression $a_0 + a_1x + a_2x^2 + a_3x^3 + ...+ a_nx^n$ is called polynomial of order n.

The equation, $y = a_0 + a_1x + a_2x^2 + ... + a_nx^n$ is called the nth degree equation.

(b) Binomial Expansion.

The mathematical forms, a^2, $5^{1/2}$ $(p + q)^3$ etc., are called exponential forms.

In $(2a - 3b)^3$, $2a - 3b$ is called the base and 3 is called the exponent or index or power.

The expression x + y, a-b, 2x+3 etc. Which have two terms are known as binomial terms.

We known that $(x+y)^2 = x^2 + 2xy + y^2$ The right hand side of this equation is called the binomial expansion of the left hand side. Similarly,

$$(x+y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$$

and $(x-y)^3 = x^3 - 3x^2y + 3xy^2 - y^3$

In general, $(x+y)^n = x^n + \frac{nx^{n-1}y}{1!} + \frac{n(n-1)x^{n-2}y^2}{2!} + ... + y^n$

For example,

$$(y-1)^5 = y^5 - \frac{5}{1!}y^4 + \frac{5(5-1)}{2!}y^3 - \frac{5(5-1)(5-2)}{3!}y^2 + \frac{5(5-1)(5-2)(5-3)}{4!}y^1 - y^0$$

$$= y^5 - 5y^4 + 10y^3 - 10y^2 + 5y - 1$$

Method of Binomial Expansion :

As the name suggests, this method is based on the binomial expansion. This method is applicable under the following conditions.

(i) The arguments (independent variable) are equidistant like, 25, 30, 35, 40 etc. That is the difference between any two successive values of x is a constant mathematically, the values of x should be in arithmetic progression.

(ii) The value of x for which the value of y is to be estimated must be one of the values of x.

For example,

x	15	20	25	30	35	40
y	32	?	47	54	61	65

Here the values of x are equidistant. The value of y which is to be estimated corresponds to a value of x. But if we have to estimate the value of y for x = 24 or x = 28 this method cannot be used for interpolation.

The Binomial expansion method involves the following steps:

(1) Find the number of known values of y. Let us say in an example there are n known values of y.

(2) Take $(y-1)^n = 0$.

(3) Expand the left hand side of the equation using the binomial expansion, but take the exponents of y as the subscripts. For example, take

$(y-1)^2$ as $y_2 - 2y_1 + y_0$.

4) Substitute the given values in the equation.

(5) Solve the equation for the unknown.

In the following table, the equations corresponding to few values of the 'n' are given.

n	Expansion of	The equation that should be taken
2	$(y-1)^2$	$y_2 - 2y_1 + y_0 = 0$
3	$(y-1)^3$	$y_3 - 3y_2 + 3y_1 - y_0 = 0$
4	$(y-1)^4$	$y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0$
5	$(y-1)^5$	$y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0$
6	$(y-1)^6$	$y_6 - 6y_5 + 15y_4 - 20y_3 + 15y_2 - 6y_1 + y_0 = 0$
7	$(y-1)^7$	$y_7 - 7y_6 + 21y_5 - 35y_4 + 35y_3 - 21y_2 + 7y_1 - y_0 = 0$
8	$(y-1)^8$	$y_8 - 8y_7 + 28y_6 - 56y_5 + 70y_4 - 56y_3 + 28y_2 - 8y_1 + y_0 = 0$

The expansion of the binomial formula can also be obtained by the following simple procedure:

(a) The first subscript of y will be the number equivalent of which we have to find the binomial expansion. Thus if $(y-1)^5= 0$ is to be expanded, the first y will be y_5. After that each y's subscript will be reduced by 1 till it reaches y_0. That is y_5, y_4, y_3, y_2, y_1, y_0.

(b) The plus and minus signs are to be placed alternatively, starting from the first, which will be plus. That is $+ y_5$, $- y_4$, $+ y_3$, $- y_2$, $+ y_1$, $- y_0$.

(c) The numerical coefficients are found by referring to 'Pascal Triangle' given below :

Pascal Triangle

n																		Sum
1								1		1								2
2							1		2		1							4
3						1		3		3		1						8
4					1		4		6		4		1					16
5				1		5		10		10		5		1				32
6			1		6		15		20		15		6		1			64
7		1		7		21		35		35		21		7		1		128
8	1		8		28		56		70		56		28		8		1	256

Example : 1.

In the following table the values of X represent the degrees of freedom and the Y values represent the chi-square values at 5% level if significance. Find the missing value.

x	2	3	4	5	6	7
y	5.99	7.81	9.49	11.07	?	14.07

Solution :

x	y
2 (x_0)	5.99 (y_0)
3 (x_1)	7.81 (y_1)
4 (x_2)	9.49 (y_2)
5 (x_3)	11.07 (y_3)
6 (x_4)	-
7 (x_5)	14.07 (y_5)

The values of x increase by 1 and the unknown term is $y_{4,}$ which corresponds to $x_{4.}$ Hence we can apply the method of binomial expansion

In the given example, the known values of y are 5. Hence we have to consider $(y-1)^5 = 0$. This leads to the equation.

$$y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0$$

Substituting the values, the equation takes the form.

$$14.07 - 5y_4 + 10(11.07) - 10(9.49) + 5(7.81) - 5.99 = 0$$

i.e., $14.07 - 5y_4 + 110.70 - 94.90 + 39.05 - 5.99 = 0$

i.e., $5y_4 = 62.93$

$\Rightarrow$ $\mathbf{y_4 = 12.586}$

Example : 2.

Use the binomial expansion method to ascertain the most likely index number for the year 1983.

Year	1981	1982	1983	1984	1985
Index No.	100	107	?	157	212

Solution :

x	1981 (x_0)	1982 (x_1)	1983 (x_2)	1984 (x_3)	1985 (x_4)
y	100 (y_0)	107 (y_1)	? (y_2)	157 (y_3)	212 (y_4)

We have to estimate $y_{2.}$ Since there are 4 known values of y, the formula of estimation is based on the expansion of $(y-1)^4=0$.

We get the equation,

$y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0.$

Substituting the known values in the equation, we obtain,

i.e., $212 - 4(157) + 6y_2 - 4y_2(107) + 100 = 0$

$$212 - 628 + 6y_2 - 428 + 100 = 0$$

$$\text{i.e., } 6y_2 = 744$$

$$\Rightarrow \mathbf{y_2 = 124}$$

Hence the probable value of the index number of **1983 is 124.**

Example : 3.

Using suitable method of interpolation, estimate the sales (in lakhs) in 2004 from the following data :

Year	2001	2002	2003	2004	2005	2006
Sales in lakhs of Rs.	150	235	365	?	525	780

Solution :

x	y
2001 (x_0)	150 (y_0)
2002 (x_1)	235 (y_1)
2003 (x_2)	365 (y_2)
2004 (x_3)	- (y_3)
2005 (x_4)	525 (y_4)
2006 (x_5)	780 (y_5)

We have estimate $y_{3.}$ Since there are 5 known values of y, the formula of estimation is based on the expansion of $(y-1)^5 = 0$.

We get the equation,

$Y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0.$

Substituting the known values in the equation, we obtain,

$780-5(525)+10y_3-10(365)+5(235)-150 = 0$

$780-2625+10y_3-10y_3365+1175-150 = 0$

$10y_3 - 4470 = 0$

$10y_3 = 4470$

$\Rightarrow \mathbf{y_3 = 447}$

Hence the probable sales (in lakhs) in **2004 is 447.**

Example : 4.

Use the binomial expansion method to estimate the index number for 2004.

Year	2000	2001	2002	2003	2004
Index number	100	107	124	157	?

Solution :

x	2000(x_0)	2001 (x_1)	2002 (x_2)	2003 (x_3)	2004 (x_4)
y	100 (y_0)	107 (y_1)	124 (y_2)	157 (y_3)	- (y_4)

We have estimate $y_{4.}$ Since there are 4 known values of y, the formula of estimation is based on the expansion of $(y-1)^4 = 0$.

We get the equation,

$y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0$

Substituting the known values in the equation, we obtain,

$y_4 - 4(157) + 6(124) - 4(107) + 100 = 0$

$y_4 - 1056 + 844 = 0$

$y_4 - 212 = 0$

$\therefore \mathbf{y_4 = 212}$

Questions

1. What is Interpolation ?
2. Mention the situations where the technique of interpolation is used.
3. What are the assumptions made in interpolation ?
4. What is Extrapolation ?
5. Distinguish between interpolation and extrapolation.
6. Interpolate the export of handlooms during 2008 from the following data.

Year	2000	2002	2004	2006	2008	2010	2012
Export of handlooms (Rs. In crores)	10	13	15	23	26	?	32

7. Interpolate the missing figure.

Year	2001	2002	2003	2004	2005	2006	2007
Sales ('000 Rs.)	100	120	150	180	210	-	320

8. Interpolate the index for 2008 from the following data.

Year	2006	2007	2008	2009	2010
Index No.	278	281	-	313	322

9. From the following data interpolate the production of cement in 2007.

Year	2005	2006	2007	2008	2009	2010
Production (lakh tons)	44	90	-	160	270	390

10. Use the binomial expansion method to estimate the index number for 2005.

Year	2001	2002	2003	2004	2005
Value	74	78	84	92	-

*****

Unit-IX

PROBABILITY THEORY

Introduction :

During 17th century the concept of probability originated as " Theory of games and chances". The theory of probability is developed by Galileo, James Bernoulli, P. S. Laplace, A. N. Kolmogorov and others.

The word chance or probable is used very often. For example : " Probably he may get rank in the examination." Such chance statement becomes precise if we express this in terms of proportion or percentage as " There is a 90% chance that he may get rank in the examination". Thus, **Probability is a numerical measure of chance of occurrence of an event.**

Probability theory provides a base for statistical inference. On the basis of this theory only Insurance Companies fix the premiums. Therefore, the concept of probability is very much needed to develop various techniques in Statistics, Business and even in Economics.

We need the following terms to understand the concept of probability.

Experiment :

Any process which yields well defined results or statistical data is called an **experiment.** A result of an experiment is called an **outcome.**

There are two types of experiments : Deterministic and Random.

An experiment which has a unique (same) outcome, when repeated under identical (same) conditions is a **deterministic experiment.**

Ex: 1. Noting the water temperature (in boiling process) at different time points.

Ex: 2. Measuring the area of a circle of a specified (with fixed) radius.

An experiment which does not have a unique outcome, when repeated under identical conditions is a random experiment. Thus, an experiment whose outcome is not always unique is a **random experiment (trial).**

Ex : 1. Noting the body temperature of a patient at different time points.

Ex : 2. Rolling a die and observing the outcomes.

Ex : 3. Number of heads obtained when two coins are tossed.

Probability theory is concerned only with analysis of random experiment.

Sample space :

The set of all possible outcomes (sample points) of a random experiment is a **sample space**. It is denoted by **'S'**.

Ex : When a die is rolled its sample space is S = {1, 2, 3, 4, 5, 6}.

A sample space containing only a finite number of outcomes is a **finite sample space.**

Ex : When a coin is tossed then S = {H, T}.

Ex : If two coins are tossed then S = {HH, HT, TH, TT}.

A sample space with uncountable number of outcomes is an **infinite sample space.**

Ex : Body temperature may be between 98°F and 104°F. Here the points in the interval cannot be counted.

In a random experiment of rolling two dice, the exhaustive coutcomes are as shown below: Hence, its sample space is

6	(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)
5	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
4	(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
3	(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
2	(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
1	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
	1	2	3	4	5	6

S = { (1,1), (1,2), (1,3), (1,4), (1,5), (1,6),
(2,1), (2,2), (2,3), (2,4), (2,5), (2,6),
(3,1), (3,2), (3,3), (3,4), (3,5), (3,6),
(4,1), (4,2), (4,3), (4,4), (4,5), (4,6),
(5,1), (5,2), (5,3), (5,4), (5,5), (5,6),
(6,1), (6,2), (6,3), (6,4), (6,5), (6,6) }

As we have seen a deck (pack) of playing cards consists of 52 cards, which are divided into 4 suits of 13 cards each. Hearts ($\heartsuit$), diamonds ($\diamondsuit$), spades ($\spadesuit$) and clubs ($\clubsuit$). Hearts and diamonds are of red colour, while spades and clubs are of black colour. The 13 cards in each suit are : Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen and King. Jacks, queens and kings are called face cards, face cards with aces become honour cards. Excluding face cards, the remaining cards in a suit are called number cards.

The sample space of drawing one card from a pack of 52 playing cards is -

$$S = \{ \boxed{\heartsuit A}, \boxed{\heartsuit 2}, \boxed{\heartsuit 3}, \boxed{\heartsuit 4}, \boxed{\heartsuit 5}, \boxed{\heartsuit 6}, \boxed{\heartsuit 7}, \boxed{\heartsuit 8}, \boxed{\heartsuit 9}, \boxed{\heartsuit 10}, \boxed{\heartsuit J}, \boxed{\heartsuit Q}, \boxed{\heartsuit K},$$

$$\boxed{\diamondsuit A}, \boxed{\diamondsuit 2}, \boxed{\diamondsuit 3}, \boxed{\diamondsuit 4}, \boxed{\diamondsuit 5}, \boxed{\diamondsuit 6}, \boxed{\diamondsuit 7}, \boxed{\diamondsuit 8}, \boxed{\diamondsuit 9}, \boxed{\diamondsuit 10}, \boxed{\diamondsuit J}, \boxed{\diamondsuit Q}, \boxed{\diamondsuit K},$$

$$\boxed{\spadesuit A}, \boxed{\spadesuit 2}, \boxed{\spadesuit 3}, \boxed{\spadesuit 4}, \boxed{\spadesuit 5}, \boxed{\spadesuit 6}, \boxed{\spadesuit 7}, \boxed{\spadesuit 8}, \boxed{\spadesuit 9}, \boxed{\spadesuit 10}, \boxed{\spadesuit J}, \boxed{\spadesuit Q}, \boxed{\spadesuit K} \}$$

$$\boxed{\clubsuit A}, \boxed{\clubsuit 2}, \boxed{\clubsuit 3}, \boxed{\clubsuit 4}, \boxed{\clubsuit 5}, \boxed{\clubsuit 6}, \boxed{\clubsuit 7}, \boxed{\clubsuit 8}, \boxed{\clubsuit 9}, \boxed{\clubsuit 10}, \boxed{\clubsuit J}, \boxed{\clubsuit Q}, \boxed{\clubsuit K} \}$$

Events :

An event is a set of outcomes of a random experiment. It is a subset of sample space. Events are denoted by capital letters **A, B, C** etc and their outcomes are denoted by small letters **a, b, c** etc.

An event which does not contain any outcome is **null** (impossible) **event.** It is denoted byϕ. An event which has only one outcome is a **simple** (elementary) **event .** An event which has more than one outcome is a **compound event.** An event which contains all the outcomes of a random experiment is a **sure** (certain) **event.** It is nothing but sample space.

Ex : 1. While throwing a die, A is an event of getting a multiple of 3, B is an of event of getting a multiple of 4 and C is an event of getting a multiple of 7. That is, A= {3, 6}, B = {4}, C = Φ. Here, A is a compound event, B is a simple event and C is a null event.

Ex : 2. In tossing two coins, A is an event that the toss results in one head, B is an event that the toss results in two heads and C is an event that the toss results in three heads. i.e., A = {HT, TH}, B = {HH}, C = Φ. Here, A is a compound event, B is a simple event and C is a null event.

The outcomes which entail the occurrence of an event are said to be **favourable outcomes** of that event.

Ex : In throwing a die, let A be an event of getting an even number. That is, A= {2, 4, 6}. Here, favourable outcomes are 2, 4 and 6.Whenever the throw results in 2 or 4 or 6 then the event A is said to occur.

Exhaustive Events (Cases) :

Exhaustive events (cases) are the total number of possible outcomes of any random experiment. Thus, a set of events is exhaustive if one or the other of the events in the set occurs whenever the random experiment is conducted. The union of exhaustive events is sample space.

In tossing 'n' coins, exhaustive cases are 2^n

In throwing 'n' dice, exhaustive cases are 6^n

Ex : 1. In throwing a die, the events A = { 1, 3, 5 } and B = { 2, 4, 6} } together are exhaustive.

Ex : 2. In throwing a die, the events A = {1, 4} and B = {2, 3, 5} together are not exhaustive

Equally likely events (Equiprobable events) :

Two or more events are said to be equally likely if they have an equal chance of occurrence.

Ex : 1. While tossing a fair coin, the events 'Head' and 'Tail' are equally likely events.

Ex : 2. In throwing a die, the events A = {1, 3, 5} and B = {2, 4, 6} are equally likely.

Union of events :

Union of two or more events is the event of occurrence of at least one of those events.

Thus, union of two events A and B is the event of occurrence of at least one of them and it is denoted by (A∪B) or (A+B) or (A or B)

Ex: 1. While tossing two coins simultaneously, if A is the event of occurrence of two heads and B is the event of occurrence of one head then (A∪B) is the event of occurrence of at least one head.

Here, A = {HH}, B = {HT, TH} then (A∪B) = {HH, HT, TH}

Ex : 2. While throwing a die, if A is th e event of getting a multiple of 2 and B is the event of getting a multiple of 3 then, (A∪B) is the event of getting a multiple of 2 or 3.

Here, A = {2, 4, 6}, B = {3, 6} then (A∪B) = {2, 3, 4, 6}

Intersection of events :

Intersection of two or more events is the event of simultaneous occurrence of those events.

Thus, intersection of two events A and B is event of occurrence of both of them and it is denoted by (A∩B) or (AB) or (A and B).

Ex :1. While throwing a die, if A is the event of getting a multiple of 2 and B is the event of getting a multiple of 3 then, (A∩B) is the event of getting a multiple of 2 and 3.

Here, A={ 2, 4, 6 }, B = { 3, 6 } then (A∩B) = { 6 }

Ex :2. While tossing two coins simultaneously, if A is the event of occurrence of two heads and B is the event of occurrence of one head then, (A∩B) is the event of occurrence of both.

Here, A= {HH}, B = {HT, TH} then (A∩B) = ϕ (null event)

Mutually exclusive events (Disjoint events) :

Two or more events are said to be mutually exclusive, if occurrence of one event prevents the occurrence of all other events in a single trial.

Two or more events are mutually exclusive if only one of them can occur at a time. That is, if intersection of events is a null event then events are said to be mutually exclusive.

Ex : 1. While tossing a coin, the events 'Head ' and ' Tail ' are mutually exclusive because when coin is tossed once, the result cannot be Head as well as Tail.

Ex : 2. In throwing a die, the events A = {1, 3, 5}, B = {2, 4, 6} are mutually exclusive $\therefore (A \cap B) = \Phi$

Complement of an event :

For any event A, the **complement of A is the event of non-occurrence of A.** It is the event constituted by all sample points which are not in A. The complement of A is denoted by A^C or $\overline{A}$ (A-bar) or A^I (A-dash).

Ex : While throwing a die, if A = { 3, 6 } its complement is A^I = { 1, 2, 4, 5 }. Here, A is an event of getting a multiple of 3 then A^I is an event of getting a non-multiple of 3.

Note : Events A and A^I are always mutually exclusive and exhaustive. That is, $(A \cap A^I) = \Phi$ and $(A \cup A^I) = S$

Methods of assigning probabilities :

Important methods of assigning the probabilities to events are Classical method and Statistical method.

Classical method :

If a fair coin is tossed the chance of occurrence of head and tail are same, which is 50% or ½. When a fair die with 6 faces is rolled the probability of a particular face turning up is 1/6. This method of assigning the probabilities to events is called **classical or mathematical method.** It is based on the count of all possible outcomes of a random experiment.

Statistical method :

An alternative procedure, known as the **statistical method** can be used to assign the probabilities to events if the random experiment is **repeated underidentical conditions.** If a coin is tossed 'n' times and 'm' of these tosses result in heads, then the proportion m/n approaches a value p as n becomes large. The fact that the limiting value exists cannot be established mathematically. This method of assigning the probabilities to events is **statistical or empirical method.** It is based on law of large numbers.

To assign probabilities there are some basic rules called axioms which are specified in axiomatic definition of probability.

Classical (Mathematical or a priori) definition :

In a random experiment let S be the sample space and A be an event. Then, $P(A) = \frac{n(A)}{n(S)}$

OR

Let a random experiment have n possible outcomes which are equally likely, mutually exclusive and exhaustive. Let m of these outcomes be favourable to an event A. Then, probability of A is

$$P(A) = \frac{m}{n} = \frac{\text{Number of favourable outcomes}}{\text{Number of exhaustive outcomes}}$$

Limitations of classical definition :

This definition is applicable only when : i) The outcomes are equally likely, mutually exclusive and exhaustive. ii) The possible number of outcomes n is finite.

Statistical (Empirical, Posteriori) definition :

(Probability as relative frequency)

Let a random experiment be repeated n times essentially under identical conditions. Let m of these repetitions results in the

occurrence of an event A. Then, the probability of event A is the limiting value of the ratio m/n as n increases indefinitely.

That is, $P(A) = \lim\limits_{n \to \infty} \frac{m}{n}$

Here, it is assumed that a unique limit exists.

Note : In practice, it is not possible to repeat an experiment infinite number of times. Therefore, the probability is obtained from a sufficiently high number of repetitions.

Axiomatic definition :

Let A and B be the events of a sample space S. Let P(A) and P(B) are the real numbers (probabilities) assigned to these events. Then, **P(A) is the probability of A** if, the following axioms are satisfied.

Axiom (i) : $P(A) \geq 0$ (non-negativity condition)

Axiom (ii) : P(S) = 1. 'S' being the sure event.

Axiom (iii) : For any two disjoint events A and B,

$$P(A+B) = P(A)+P(B)$$

Note : The third axiom can be generalized for any number of mutually exclusive events.

Results: (1) $P(\Phi) = 0$, (2) $P(S) = 1$, (3) $0 \leq P(A) \leq 1$, (4) $P(A) + P(A^{I}) = 1$

These results are proved as follows:

1) $P(\Phi) = 0$. That is, probability of null event is zero.

Proof : By definition, $P(\Phi) = \frac{n(\Phi)}{n(S)} = \frac{0}{n} = 1$

2) $P(S) = 1$. That is, probability of sure event is one.

Proof : By definition, $P(S) = \frac{n(S)}{n(S)} = \frac{n}{n} = 1$

3) $0 \leq P(A) \leq 1$. That is, P(A) is the value between 0 and 1 (The limits of probability are 0 and 1).

Proof : Here, by definition of probability of an event A is $P(A)=\frac{m}{n}$

The least and highest possible values of m are zero and n.

That is, $0 \leq m \leq n$; dividing by n

$$\frac{0}{n} \leq \frac{m}{n} \leq \frac{n}{n}$$

i.e., $0 \leq P(A) \leq 1$

4) $P(A) + P(A^I) = 1$. That is, the sum of probabilities of complementary events is 1.

Proof : We know that,

$$(A \cup A^I) = S$$

$$\therefore P(A \cup A^I) = P(S)$$

$$P(A) + P(A^I) = 1 \ (\because \text{ A and } A^I \text{ are mutually exclusive events})$$

Alternative proof

Out of ‘n’ outcomes, if ‘m’ outcomes are favourable to event A, then the remaining (n – m) outcomes are favourable to event A^I .

By definition, $P(A) = \frac{m}{n}$ and $P(A^I) = \frac{n-m}{n}$

$$P(A) + P(A^I) = \frac{m}{n} - \frac{n-m}{n} = \frac{m+n-m}{n} = \frac{n}{n} = 1$$

Examples on mathematical definition:

Example : 1.

When a coin is tossed, what is the probability of getting head ?

Solution :

Let A be an event of getting head.

Here, S = { H, T } and A= { H }

$$\therefore P(A) = \frac{\text{Number of outcomes of A}}{\text{Number of outcomes of S}} = \frac{n(A)}{n(S)} = \frac{1}{2} = \mathbf{0.5}$$

Example : 2.

When two different coins are tossed, find the probability of getting:

a) 2 heads, b) 1 head.

Solution :

Let A be an event of getting 2 heads and B be an event of getting 1 head.

Here, S = { HH, HT, TH, TT }, A ={ HH } and B = { HT, TH }

a) $P(A) = \frac{n(A)}{n(S)} = \frac{1}{4} = \mathbf{0.25}$

b) $P(B) = \frac{n(B)}{n(S)} = \frac{2}{4} = \mathbf{0.5}$

Example : 3.

A die is thrown once, what is the probability of getting :

a) an odd number, b) an even number, c) a multiple of 3 ?

Solution :

Let ' O ' be an event of getting an odd number, ' E ' be an event of getting an even number and ' M ' be an event of getting a multiple of 3.

Here, S = { 1, 2, 3, 4, 5, 6 }, O = { 1, 3, 5 }, E = { 2, 4, 6} and M = { 3, 6 }

a) $P(O) = \frac{n(O)}{n(S)} = \frac{3}{6} = \frac{1}{2} = \mathbf{0.5}$

b) $P(E) = \frac{n(E)}{n(S)} = \frac{3}{6} = \frac{1}{2} = \mathbf{0.5}$

c) $P(M) = \frac{n(M)}{n(S)} = \frac{2}{6} = \frac{1}{3} = \mathbf{0.3333}$

Example : 4.

A bag contains 3 red and 2 white balls. A ball is drawn from this bag, what is the probability that it is : a) Red, b) White ?

Solution :

Here, number of red balls is 3, number of white balls is 2 and of total number of balls are 5.

a) $P(R) = \frac{{}^{3}C_{1}}{{}^{5}C_{1}} = \frac{3}{5} = \mathbf{0.6}$ and

b) $P(W) = \frac{{}^{2}C_{1}}{{}^{5}C_{1}} = \frac{2}{5} = \mathbf{0.4}$

Example : 5.

A card is drawn from a pack of 52 playing cards. What is the probability that it is a : a) Club ? b) King ?

Solution: Let C be an event of getting a Club (♣)and K be an event of getting a king.

Here, $C = \{ \boxed{\clubsuit A}, \boxed{\clubsuit 2}, \boxed{\clubsuit 3}, \boxed{\clubsuit 4}, \boxed{\clubsuit 5}, \boxed{\clubsuit 6}, \boxed{\clubsuit 7}, \boxed{\clubsuit 8}, \boxed{\clubsuit 9}, \boxed{\clubsuit 10}, \boxed{\clubsuit J}, \boxed{\clubsuit Q}, \boxed{\clubsuit K} \}$

$K = \{ \boxed{\heartsuit K}, \boxed{\diamondsuit K}, \boxed{\spadesuit K}, \boxed{\clubsuit K} \}$ and S is the Sample space.

a) $P(C) = \frac{n(C)}{n(S)} = \frac{13}{52} = \mathbf{0.25}$

b) $P(K) = \frac{n(K)}{n(S)} = \frac{4}{52} = \mathbf{0.0769}$

Example : 6.

A bag contains 6 black and 4 white balls. Two balls are drawn from this bag. What is the probability that they are : a) Black b) White c) one Black and one White ?

Solution :

Here, total number of balls is 10.

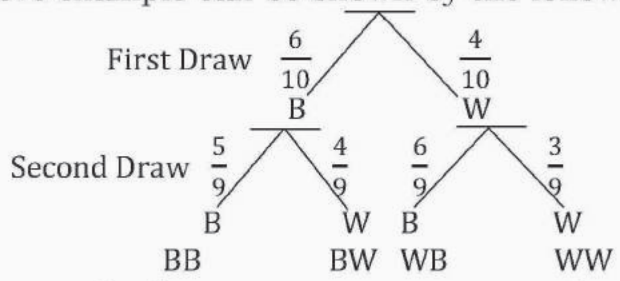
a) $P(B) = \frac{{}^{6}C_{2}}{{}^{10}C_{2}} = \frac{15}{45} = \mathbf{0.3333}$

b) $P(W) = \frac{{}^{4}C_{2}}{{}^{10}C_{2}} = \frac{6}{45} = \mathbf{0.1333}$

c) $P(1B \text{ and } 1\,W) = \frac{{}^{6}C_{1} \times {}^{4}C_{1}}{{}^{10}C_{2}} = \frac{6 \times 4}{45} = \mathbf{0.5333}$

Alternative method

The above example can be shown by the following probability tree



i.e., $P(BB) = \frac{6}{10} \times \frac{5}{9} = \mathbf{0.3333}$ and $P(WW) = \frac{4}{10} \times \frac{3}{9} = \mathbf{0.1333}$

$$P(1B \text{ and } 1W) = P(BW) + P(WB) = \frac{6}{10} \times \frac{4}{9} + \frac{4}{10} \times \frac{6}{9} = \mathbf{0.5333}$$

Example : 7.

Two cards are drawn from a pack of 52 playing cards. What is the probability that they are: a) Kings b) Hearts ?

Solution :

a) $P(K) = \frac{4_{C_2}}{52_{C_2}} = \frac{6}{1326} = \mathbf{0.0045}$

b) $P(H) = \frac{13_{C_2}}{52_{C_2}} = \frac{78}{1326} = \mathbf{0.0588}$

Example :8.

A card is drawn from a well-shuffled deck of 52 playing cards. Calculate the probability that the card will be (i) an ace, (ii) not an ace.

Solution :

If A is an event of an ace then, $A^{\prime}$ is an event of not an ace.

Here , $P(A) = \frac{4}{52} = \mathbf{0.0769} \Rightarrow P(A') = 1 - P(A) = 1 - \frac{4}{52} = \frac{48}{52} = \mathbf{0.9231}$

Example : 9.

The chance of player A winning a match is 3/5 . Find the chance that he does not win.

Solution :

Let A be an event of winning the match, then A^I is the event of not winning the match. A and A^I are complementary events.

We know that, $P(A) + P(A') = 1$

$\therefore$ $P(A^I) = 1 - P(A) = 1 - \frac{3}{5} = \frac{2}{5} =$ **0.4**

It is simplified using calculator as:
1 $\boxed{M+}$ 3 ÷ 5 $\boxed{M-}$ $\boxed{MR}$ $\boxed{MC}$

Example : 10.

In a lot of 20 bulbs, 4 are defective. One bulb is selected at random from this lot. What is the probability that selected bulb is non defective?

Solution :

Let, D be an event of selecting a defective bulb. Then, D^I is the event of selecting a non-defective bulb.

$\therefore$ D and D^I are complementary events

We know that, $P(D) + P(D^I) = 1$

$\therefore$ $P(D^I) = 1 - P(D) = 1 - \frac{4}{20} = \frac{16}{20} =$ **0.8**

It is simplified using calculator as:
1 $\boxed{M+}$ 4 ÷ 20 $\boxed{M-}$ $\boxed{MR}$ $\boxed{MC}$

Example : 11.

A die is thrown once, what is the probability of getting a :

a) Multiple of 4 ? b) Non-multiple of 4 ?

Solution :

If A is an event of getting a multiple of 4. Then, A^I is an event of getting a non-multiple of 4.

Here, S = { 1, 2, 3, 4, 5, 6 }, A ={ 4 } and A' = { 1, 2, 3, 5, 6 }

a) $P(A) = \frac{n(A)}{n(S)} = \frac{1}{6} =$ **0.1667**

b) $P(A') = \frac{n(A')}{n(S)} = \frac{5}{6} =$ **0.8333**

OR

$$P(A') = 1 - P(A) = 1 - \frac{1}{6} = \frac{5}{6} = \mathbf{0.8333}$$

Example : 12:

If $P(A) = \frac{1}{4}$ then , find $P(A')$.

Solution :

We know that, $P(A) + P(A') = 1$. i.e., $P(A') = 1 - P(A) = 1 - \frac{1}{4} = \frac{3}{4} = \mathbf{0.75}$

Example : 13.

If $P(A) = 0.6$ then, find $P(A^I)$.

Solution :

We know that, $P(A) + P(A^I) = 1$. i.e., $P(A^I) = 1 - P(A) = 1 - 0.6 = \mathbf{0.4}$

Addition theorem of probability for two non-mutually exclusive events: (Theorem of total probability)

Statement :

Let A and B be two events with respective probabilities P(A) and P(B). Then, the probability of occurrence of at least one of these two events is –

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad \text{or} \quad P(A+B) = P(A) + P(B) - P(AB)$$

Proof :

Let S be the sample space, A and B be two events.

By definition,

$$P(A) = \frac{n(A)}{n(S)}, \; P(B) = \frac{n(B)}{n(S)}, \; P(A \cap B) = \frac{n(A \cap B)}{n(S)} \text{ and } P(A \cup B) = \frac{n(A \cup B)}{n(S)}$$

$$\text{Consider } P(A \cup B) = \frac{n(A \cup B)}{n(S)}$$

$$= \frac{n(A) + n(B) - n(A \cap B)}{n(S)}$$

By the result of number of elements in a set we know that, $n(A \cup B) = n(A) + n(B) - n(A \cap B)$

$$= \frac{n(A)}{n(S)} + \frac{n(B)}{n(S)} - \frac{n(A \cap B)}{n(S)}$$

$$= P(A) + P(B) - P(A \cap B)$$

Alternative Proof

Out of 'n' exhaustive outcomes of a random experiment, if ' a ' outcomes are favourable to event A, 'b' outcomes are favourable to event B and ' c ' outcomes are common to both A and B.

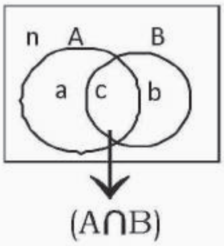
Then, $P(A) = \frac{a}{n}$, $P(B) = \frac{b}{n}$ and $P(A \cap B) = \frac{c}{n}$

Here, favourable outcomes to event (A or B) are a + b – c

$$\therefore P(A \cup B) = \frac{a+b-c}{n}$$

$$= \frac{a}{n} + \frac{b}{n} - \frac{c}{n}$$

$$= P(A) + P(B) - P(A \cap B)$$



Note: If A, B and C are three events, then

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(C \cap A) + P(A \cap B \cap C)$$

Example : 14.

A die is thrown once. What is the probability of getting a multiple of 2 or 3 ?

Solution :

Let A be an event of getting a multiple of 2, B be an event of getting a multiple of 3, (A∩B) is an event of getting a multiple of 2 and 3 (i.e., both) and (A∪B) is an event of getting a multiple of 2 or 3 (i.e., at least one).

Here, S = {1, 2, 3, 4, 5, 6}, A = {2, 4, 6}, B = {3, 6} and A∩B = { 6 }

$\therefore P(A) = \frac{n(A)}{n(S)} = \frac{3}{6}$, $P(B) = \frac{n(B)}{n(S)} = \frac{2}{6}$