NARSIMHA REDDY ENGINEERING COLLEGE
UGC AUTONOMOUS INSTITUTION
Maisammaguda (V), Kompally - 500100, Secunderabad, Telangana State, India
UGC - Autonomous Institute
Accredited by NBA & NAAC with 'A' Grade
Approved by AICTE
Permanently affiliated to JNTUH

UNIT-V

Applied Statistics

CORRELATION :

Introduction

In a bivariate distribution and multivariate distribution, we may be interested to find if there is any relationship between the two variables under study. Correlation refers to the relationship between two or more variables. The correlation expresses the relationship or interdependence of two sets of variables upon each other.

Definition Correlation is a statistical tool which studies the relationship b/w 2 variables & correlation analysis involves various methods & techniques used for studying & measuring the extent of the relationship b/w them.

Two variables are said to be correlated if the change in one variable results in a corresponding change in the other.

The Types of Correlation

1) **Positive and Negative Correlation**: If the values of the 2 variables deviate in the same direction

i.e., if the increase in the values of one variable results in a corresponding increase in the values of other variable (or) if the decrease in the values of one variable results in a corresponding decrease in the values of other variable is called Positive Correlation.

e.g. Heights & weights of the individuals If the increase (decrease) in the values of one variable results in a corresponding decrease (increase) in the values of other variable is called Negative Correlation.

e.g, Price and demand of a commodity.

2) **Linear and Non-linear Correlation :**The correlation between two variables is said to be Linear if the corresponding to a unit change in one variable there is a constant change in the other variable over the entire range of the values (or) two variables , are said to be linearly related if there exists a relationship of the form y = a + bx.

e.g when the amount of output in a factory is doubled by doubling the number of workers.

Two variables are said to be Non-linear or curvilinear if corresponding to a unit change in one variable the other variable does not change at a constant rate but at fluctuating rate. i.e Correlation is said to be non-linear if the ratio of change is not constant. In other words, when all the points on the scatter diagram tend to lie near a smooth curve, the correlation is said to be non-linear (curvilinear).

3) **Partial and Total correlation:** The study of two variables excluding some other variables is called Partial Correlation.

e.g. We study price and demand eliminating the supply.

In Total correlation all the facts are taken into account.

e.g Price, demand &supply ,all are taken into account.

4) **Simple and Multiple correlation :**When we study only two variables, the relationship is described as Simple correlation.

E.g quantity of money and price level, demand and price

Karl Pearson's Coefficient of Correlation (Pearson's Correlation Coefficient)

Definition

Karl Pearson's Coefficient of Correlation is a statistical measure that indicates the **strength and direction of the linear relationship** between two variables $X$and Y.

It is denoted by r.

The value of $r$always lies between **−1 and +1**.

$$\boxed{-1 \leq r \leq +1}$$

Interpretation of r

Value of r	Interpretation
$+1$	Perfect positive correlation
0	No correlation
-1	Perfect negative correlation
$0 < r < 1$	Positive correlation
$-1 < r < 0$	Negative correlation

Formula

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

Where

- n= Number of observations
- $\sum x$= Sum of X values
- $\sum y$= Sum of Y values
- $\sum xy$= Sum of the products of X and Y
- $\sum x^2$= Sum of squares of X
- $\sum y^2$= Sum of squares of Y

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

X	**1**	**2**	**3**	**4**	**5**
Y	**2**	**4**	**5**	**4**	**5**

Sol :

Step 1: Prepare the Table

X	Y	XY	X^2	Y^2
1	2	2	1	4
2	4	8	4	16
3	5	15	9	25
4	4	16	16	16
5	5	25	25	25

Step 2: Calculate Totals

$n = 5 \quad \sum X = 15 \quad \sum Y = 20 \quad \sum XY = 66 \quad \sum X^2 = 55 \quad \sum Y^2 = 86$

Step 3: Apply the Formula

$$r = \frac{5(66) - 15(20)}{\sqrt{(5 \times 55 - 15^2)(5 \times 86 - 20^2)}}$$

Step 4: Simplify

Numerator:

$$330 - 300 = 30$$

Denominator:

$$\sqrt{(275 - 225)(430 - 400)}$$
$$= \sqrt{50 \times 30}$$
$$= \sqrt{1500} = 38.73$$

Step 5: Final Answer

$$r = \frac{30}{38.73} = 0.775$$

$$\boxed{r = 0.775}$$

Interpretation

Since $r = 0.775$, there is a **strong positive correlation** between the variables $X$and Y.

Problem 2

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

X	28	41	40	38	35	33	40	32	36	33
Y	23	34	33	34	30	26	28	31	36	38

What inference would you draw from the estimate.

Sol: To calculate **Karl Pearson's Coefficient of Correlation**, use the formula

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

Given Data

X	**28**	**41**	**40**	**38**	**35**	**33**	**40**	**32**	**36**	**33**
Y	23	34	33	34	30	26	28	31	36	38

Step 1: Prepare the Calculation Table

X	**Y**	**XY**	**X^2**	**Y^2**
28	23	644	784	529
41	34	1394	1681	1156
40	33	1320	1600	1089

X	Y	XY	X²	Y²
38	34	1292	1444	1156
35	30	1050	1225	900
33	26	858	1089	676
40	28	1120	1600	784
32	31	992	1024	961
36	36	1296	1296	1296
33	38	1254	1089	1444

Step 2: Find the Totals

$n = 10 \quad \sum X = 35 \sum Y = 313 \quad \sum XY = 11220 \quad \sum X^2 = 12832 \quad \sum Y^2 = 9991$

Step 3: Substitute into the Formula

$$r = \frac{10(11220) - 356(313)}{\sqrt{[10(12832) - 356^2][10(9991) - 313^2]}}$$

Step 4: Simplify

Numerator:

$$10(11220) - 356(313) = 112200 - 111428 = 772$$

Denominator:

$$\sqrt{(128320 - 126736)(99910 - 97969)}$$
$$= \sqrt{1584 \times 1941} = \sqrt{3074544} \approx 1753.44$$

Therefore,

$$r = \frac{772}{1753.44} \approx 0.44$$

Final Answer

$$\boxed{r \approx 0.44}$$

Inference

Since

$$\boxed{r = 0.44}$$

- The correlation is **positive**, meaning that as **X increases, Y tends to increase**.
- The value **0.44** indicates a **moderate positive correlation** (not a strong relationship, but a noticeable positive linear association).

Rank Correlation Coefficient :

A **Rank Correlation Coefficient** is a statistical measure that determines the **degree of association between the ranks of two variables**. It is used when data are in the form of ranks rather than actual numerical values.

The two most common rank correlation coefficients are:

1. **Spearman's Rank Correlation Coefficient (ρ or r_s)**
   - Measures the strength and direction of the relationship between two ranked variables.
   - Formula (when there are no tied ranks):

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Where:

- d= Difference between the ranks of corresponding observations.
- n= Number of observations.

2. **Kendall's Rank Correlation Coefficient (τ)**
   - Measures association based on the number of concordant and discordant pairs.
   - Preferred when the sample size is small or there are many tied ranks.

Interpretation of Spearman's Coefficient

- $r_s = +1$: Perfect positive rank correlation.
- $r_s = -1$: Perfect negative rank correlation.
- $r_s = 0$: No rank correlation.

Uses

- Measuring correlation when data are ordinal.
- Analyzing non-linear but monotonic relationships.
- Comparing rankings given by different judges.

Example: If two judges rank 10 students in a competition, Spearman's rank correlation coefficient can be used to determine how similar their rankings are.

Problem 1:

Following are the rank obtained by 10 students in two subjects' statistics and Mathematics. To what extent the knowledge of the students in two subjects is related.

statistics	1	2	3	4	5	6	7	8	9	10
Mathematics	2	4	1	5	3	9	7	10	6	8

Sol: 1. Spearman's Rank Correlation Coefficient

Given the ranks of 10 students in Statistics and Mathematics.

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

$$\sum d^2 = 1 + 4 + 4 + 1 + 4 + 9 + 0 + 4 + 9 + 4 = 40$$

Number of students,

$$n = 10$$

Using Spearman's rank correlation formula,

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

$$= 1 - \frac{6(40)}{10(10^2 - 1)}$$

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

$$= 1 - 0.2424$$

$$\boxed{r_s = 0.758}$$

Interpretation
There is a high positive correlation between the students' performance in Statistics and Mathematics. Students who perform well in one subject tend to perform well in the other.

Problem 2:
Obtain the rank correlation coefficient for the following data

x	68	64	75	50	64	80	75	40	55	64
y	62	58	68	45	81	60	68	48	50	70

Sol:

Rank Correlation Coefficient (with Tied Ranks)
Step 1: Assign Ranks
Ranks of X

X	Rank
80	1
75	2.5
75	2.5
68	4
64	6
64	6
64	6
55	8
50	9
40	10

Ranks of Y

Step 2: Compute d and d^2

X	Y	Rank X	Rank Y	d	d^2
68	62	4	5	-1	1
64	58	6	7	-1	1
75	68	2.5	3.5	-1	1
50	45	9	10	-1	1
64	81	6	1	5	25
80	60	1	6	-5	25
75	68	2.5	3.5	-1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16

Y	Rank
81	1
70	2
68	3.5
68	3.5
62	5
60	6
58	7
50	8
48	9
45	10

$$\sum d^2 = 72$$

Step 3: Tie Correction

For X:

- 64 occurs 3 times

$$\frac{3^3 - 3}{12} = \frac{24}{12} = 2$$

- 75 occurs 2 times

$$\frac{2^3 - 2}{12} = \frac{6}{12} = 0.5$$

Total correction for X

$$T_x = 2.5$$

For Y:

- 68 occurs twice

$$T_y = \frac{2^3 - 2}{12} = 0.5$$

Total tie correction

$$T = T_x + T_y = 3$$

Corrected value

$$\sum d^2 = 72 + 3 = 75$$

Step 4: Spearman's Rank Correlation

$$r_s = 1 - \frac{6(75)}{10(10^2 - 1)} = 1 - \frac{450}{990} = 1 - 0.4545$$

$$\boxed{r_s = 0.545}$$

Equal or Repeated Ranks:

Definition: Equal or Repeated Ranks (Tied Ranks) occur when **two or more observations have the same value**. Since these observations cannot be assigned different ranks, they are given the **average of the ranks** they would have occupied.

This situation commonly arises while calculating **Spearman's Rank Correlation Coefficient**.

Spearman's Rank Correlation with Tied Ranks

Tie Correction Formula

When tied ranks are present, Spearman's rank correlation requires a tie correction.

For a group of m tied observations,

$$T = \frac{m^3 - m}{12}$$

Where:

- m= Number of equal observations in one group.

The adjusted formula is

$$r_s = 1 - \frac{6\left(\sum d^2 + T_x + T_y\right)}{n(n^2 - 1)}$$

where:

- d= Difference between corresponding ranks
- T_x= Tie correction for variable X
- T_y= Tie correction for variable Y
- n= Number of observations .

Problems:

Example 1: One Pair of Equal Ranks Data

Student	**A**	**B**	**C**	**D**	**E**	**F**
Marks	85	70	85	60	95	75

Step 1: Arrange in Descending Order

Marks Student Rank

95	E	1
85	A	2.5
85	C	2.5
75	F	4
70	B	5
60	D	6

Explanation:

The two students scoring **85** occupy ranks **2 and 3**.

Average rank

$$\frac{2+3}{2} = 2.5$$

Hence, both receive **Rank = 2.5**.

Tie correction

$$T = \frac{2^3 - 2}{12} = \frac{8-2}{12} = \frac{6}{12} = 0.5$$

Example 2: Three Equal Values Data

X	45	60	60	80	60	90

Step 1: Arrange in Ascending Order

Value Rank

45 1

60 3

60 3

60 3

80 5

90 6

Explanation

The three values **60** occupy ranks **2, 3, and 4**.

Average rank

$$\frac{2+3+4}{3} = 3$$

Hence, each 60 receives **Rank = 3**.

Tie correction

$$T = \frac{3^3 - 3}{12} = \frac{27 - 3}{12} = \frac{24}{12} = 2$$

Example 3: Four Equal Values Data

X	25	35	35	35	35	50	60

Ascending Order

Value Rank

25	1
35	3.5
35	3.5
35	3.5
35	3.5

50	6
60	7

The four observations **35** occupy ranks **2, 3, 4, and 5**.

Averagerank $\frac{2+3+4+5}{4} = 3.5$

Each observation receives **Rank = 3.5**.

Tie correction

$$T = \frac{4^3 - 4}{12} = \frac{64 - 4}{12} = \frac{60}{12} = 5$$

Example 4: Two Different Groups of Tied Ranks

Data

X 20 30 30 40 50 50 50 70

Ascending Order

Value Rank

20	1
30	2.5
30	2.5
40	4
50	6
50	6
50	6
70	8

Explanation

30 appears twice

Ranks occupied: 2 and 3

Average rank $\frac{2+3}{2} = 2.5$

50 appears three times

Ranks occupied: 5, 6, and 7

Average rank $\frac{5+6+7}{3} = 6$

Tie Corrections

For 30 $T_1 = \frac{2^3-2}{12} = 0.5$

For 50 $T_2 = \frac{3^3-3}{12} = 2$

Total tie correction $T = 0.5 + 2 = 2.5$

Example 5: Tied Ranks in Spearman's Correlation

Given

X	**70**	**80**	**80**	**60**	**90**
Y	65	85	75	55	95

Assign Ranks (Descending)

X	Rank X	Y	Rank Y
90	1	95	1
80	2.5	85	2
80	2.5	75	3
70	4	65	4
60	5	55	5

Compute Differences

Rank X	Rank Y	d	d^2
4	4	0	0
2.5	2	0.5	0.25
2.5	3	-0.5	0.25
5	5	0	0
1	1	0	0

$\sum d^2 = 0.5$

Tie correction for X $T = \frac{2^3-2}{12} = 0.5$

Since Y has no ties $T_Y = 0$

Using Spearman's formula $r_s = 1 - \frac{6(\sum d^2 + T_X + T_Y)}{n(n^2-1)}$

$$= 1 - \frac{6(0.5 + 0.5)}{5(25-1)}$$

$$= 1 - \frac{6}{120} = 0.95$$

Prob : The ranks assigned by two judges to ten competitors are given below.

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

Calculate Spearman's Rank Correlation Coefficient and interpret the result.

Sol : Spearman's Rank Correlation Coefficient is

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

where

- n= Number of competitors
- d= Difference between the two ranks
- $\sum d^2$= Sum of squared differences

Step 2: Prepare the Calculation Table

Competitor	Judge X (R_1)	Judge Y (R_2)	$d = R_1 - R_2$	d^2
A	1	2	-1	1
B	2	1	1	1
C	3	4	-1	1
D	4	3	1	1
E	5	5	0	0
F	6	7	-1	1
G	7	6	1	1
H	8	10	-2	4
I	9	8	1	1
J	10	9	1	1

Competitor	Judge X (R_1)	Judge Y (R_2)	$d = R_1 - R_2$	d^2
Total				12

Thus,

$$\sum d^2 = 12$$

Number of competitors,

$$n = 10$$

Step 3: Substitute in the Formula

$$r_s = 1 - \frac{6(12)}{10(10^2 - 1)}$$

$$= 1 - \frac{72}{10(99)}$$

$$= 1 - \frac{72}{990}$$

$$= 1 - 0.0727$$

$$r_s = 0.9273$$

Prob : The ranks assigned by two judges to ten competitors are given below. Some competitors have obtained the same rank.

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

Calculate Spearman's Rank Correlation Coefficient after applying the correction for tied ranks.

Sol: **Judge X**

Ranks are 1, 2, 3, 3, 5, 6, 7, 8, 9, 10

Competitors **C** and **D** are tied for ranks **3 and 4**.

Average rank

$$\frac{3+4}{2} = 3.5$$

Hence,

Competitor	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Corrected Rank (R_1)	1	2	3.5	3.5	5	6	7	8	9	10

Judge Y

Ranks are

2,1,4,4,5,7,6,9,8,10

Competitors **C** and **D** are tied for ranks **4 and 5**.

Average rank

$$\frac{4+5}{2} = 4.5$$

Hence,

Competitor	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Corrected Rank (R_2)	2	1	4.5	4.5	5	7	6	9	8	10

Step 2: Compute $d$and d^2

Competitor	R_1	R_2	$d = R_1 - R_2$	d^2
A	1	2	-1	1
B	2	1	1	1
C	3.5	4.5	-1	1
D	3.5	4.5	-1	1
E	5	5	0	0
F	6	7	-1	1
G	7	6	1	1
H	8	9	-1	1
I	9	8	1	1
J	10	10	0	0

$$\sum d^2 = 8$$

Step 3: Calculate Tie Corrections

The correction term for each group of tied ranks is

$$\frac{m^3 - m}{12}$$

where $m$is the number of tied observations.

Judge X

One group of **2 tied ranks**

$$\frac{2^3-2}{12}=\frac{8-2}{12}=\frac{6}{12}=0.5$$

Judge Y

One group of **2 tied ranks**

$$\frac{2^3-2}{12}=0.5$$

Total correction

$$0.5+0.5=1$$

Step 4: Apply Spearman's Rank Correlation Formula

The corrected formula is

$$r_s=1-\frac{6(\sum d^2+\text{Correction})}{n(n^2-1)}$$

Substitute the values:

- $n=10$
- $\sum d^2=8$
- Total correction $=1$

$$r_s=1-\frac{6(8+1)}{10(10^2-1)}$$

$$=1-\frac{54}{10(99)}$$

$$=1-\frac{54}{990}$$

$$=1-0.054545$$

$$\boxed{r_s=0.9455}$$