UNIT-5

LINEAR CORRELATION & REGRESSION

Linear Correlation: Types of Correlation, Methods of studying Correlation: Scatter diagram method, Karl Pearson's coefficient of correlation and Rank Correlation.

Linear Regression Analysis: Linear and Non-Linear regression, lines of regression, coefficients of regression.

Course Outcome: Estimate the level of correlation, the linear relationship using the regression lines for the given bivariate data.

Chapter Outlines:

- Understand the types of correlation
- Studying correlation through Scatter Diagram
- Calculation of Karl Pearson's coefficient of correlation for a bivariate (X, Y) data.
- Use simple linear regression for building empirical models
- Understand Regression coefficients and its properties
- Calculation of Spearman's Rank correlation coefficient.

INTRODUCTION

Correlation and Regression analysis are statistical techniques that are broadly used in physical geography to examine causal relationships between variables. Regression and correlation measure the degree of relationship between two or more variables in two different but related ways.

Correlation, as the name suggests is a word formed by combining 'co' and 'relation'. It refers to the analysis of the relationship that is established between two variables in each dataset. It helps in understanding (or measuring) the linear relationship between two variables.

Correlation is a statistical method used to determine the extent to which two variables are related. Correlation analysis measures the degree of association between two or more variables.

Definition: (Correlation) Two variables X and Y are said to be correlated when a change in the value of one variable results in a corresponding change in the value of the other variable. This could be a direct or an indirect change in the value of variables. This indicates a relationship between both the variables.

Correlation is a statistical measure that deals with the strength of the relation between the two variables in question.

Correlation can be a positive or negative value.

Types of correlation

- **Positive correlation**

A positive correlation is a relationship between two variables where if one variable increases, the other one also increases. A positive correlation also exists in one decrease and the other also decreases.

Examples:

- The more time you spend running on a treadmill, the more calories you will burn.
- As the temperature goes up, ice cream sales also go up.
- The less time I spend marketing my business, the fewer new customers I will have.

➢ **Negative correlation**

A negative correlation means that there is an inverse relationship between two variables - when one variable decreases, the other increases. The vice versa is a negative correlation too, in which one variable increase and the other decreases.

Examples:

- A student who has many absences has a decrease in grades.
- If the sun shines more, a house with solar panels requires less use of other electricity.
- Price and demand of a commodity.
-

METHOD OF SCATTER DIAGRAM

What is a Scatter Diagram?

A simple and attractive method of measuring correlation by diagrammatically representing bivariate distribution for determination of the nature of the correlation between the variables is known as the Scatter Diagram Method. This method gives the investigator/analyst a visual idea of the nature of the association between the two variables. It is the simplest method of studying the relationship between two variables as there is no need to calculate any numerical value.

How to draw a Scatter Diagram?

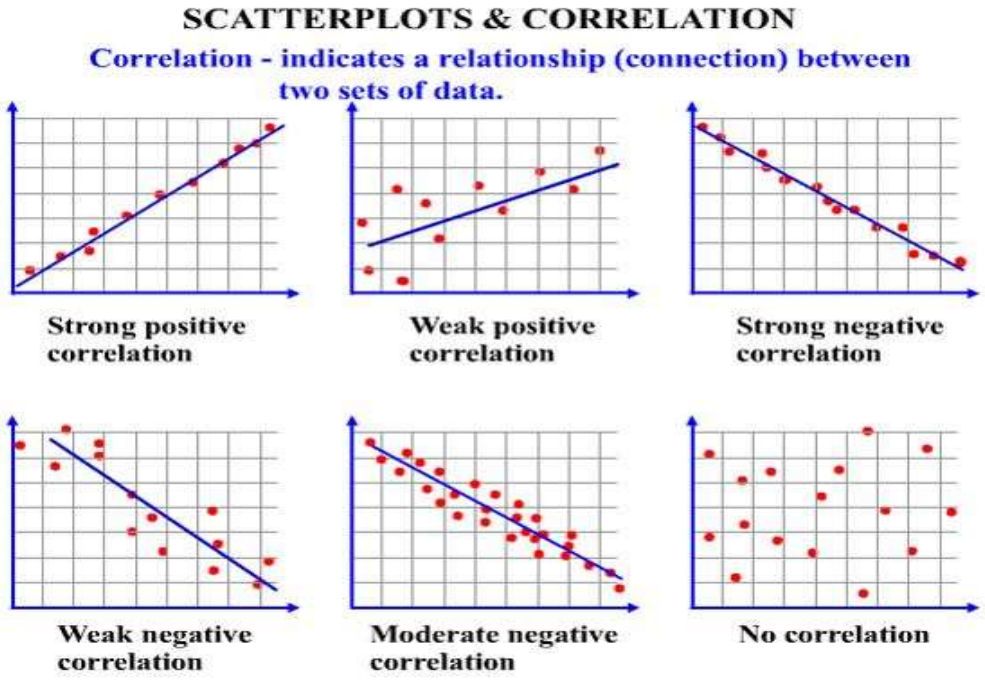
The two steps required to draw a Scatter Diagram or Dot Diagram are as follows:

1. Plot the values of the given variables (say X and Y) along the X-axis and Y-axis, respectively.
2. Show these plotted values on the graph by dots. Each of these dots represents a pair of values.

Interpretation of Scatter Diagram

After observing the pattern of dots, one can know the presence or absence of correlation and its type. Besides, it also gives an idea of the nature and intensity of the relationship between the two variables.

The scatter diagram can be interpreted in the following ways:



UNIT-5

Karl Pearson's Correlation Coefficient (*The limits are -1 and +1)*

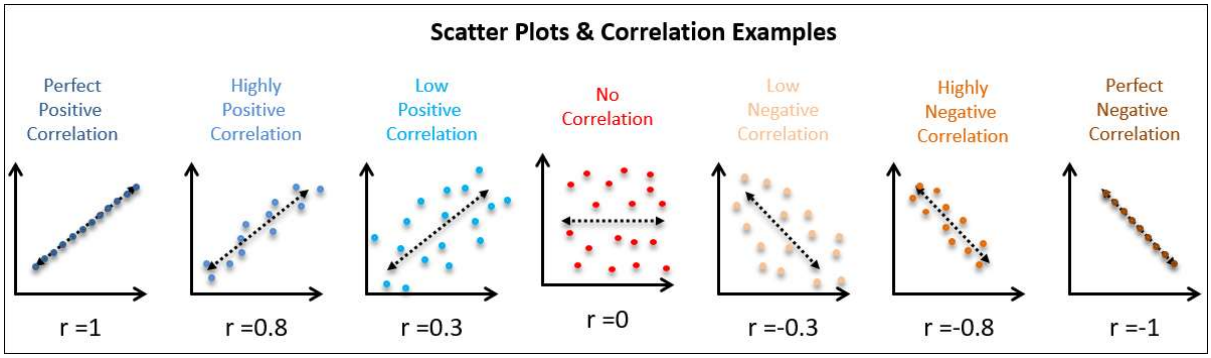
$$r(X,Y) = r_{xy} = \frac{COV(X,Y)}{\sigma_X \cdot \sigma_Y} == \frac{\left(\frac{\sum XY}{n}\right) - \left(\overline{X}.\overline{Y}\right)}{\sqrt{\left(\frac{\sum X^2}{n}\right) - \left(\overline{X}\right)^2}\ \sqrt{\left(\frac{\sum Y^2}{n}\right) - \left(\overline{Y}\right)^2}}$$

Where, cov(x,y) is the covariation between the two variables. And, σ_X *and* σ_Y are the standard deviations of x and y variables, respectively.

Properties of 'r':

- It lies between -1 and +1, both included. In both the extreme cases, there is either perfect negative or perfect positive correlation, respectively.
- A high value of 'r' indicates strong linear relationship, and vice versa.
- A positive value indicates positive correlation.
- The value of 'r' is unaffected by a change of origin or change of scale.
- If r=+1 the variables X and Y are perfectly positively correlated.
- If r= -1 the variables X and Y are perfectly negatively correlated.
- If r=0 the variables X and Y are uncorrelated i.e., no correlation.

The Pearson correlation coefficient (r) is an indicator of the strength of a *linear* relationship between two variables,



Regression Analysis

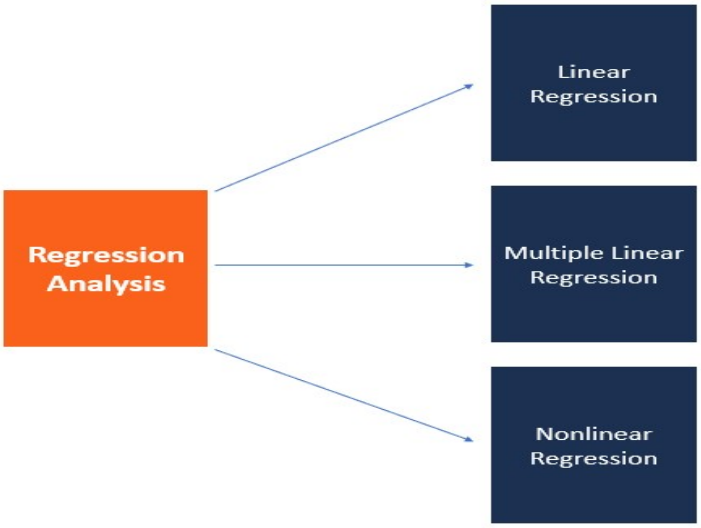
What is Regression Analysis?

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modelling the future relationship between them.

UNIT-5

Regression analysis includes several variations, such as linear, multiple linear, and nonlinear. The most common models are simple linear and multiple linear. Nonlinear regression analysis is commonly used for more complicated data sets in which the dependent and independent variables show a nonlinear relationship.

Regression analysis offers numerous applications in various disciplines, including finance.



WHAT IS LINEAR REGRESSION?

Definition: Linear regression is one of the most straight forward yet powerful tools in management studies. It models the relationship between a dependent variable y and one or more independent variables x. The simplest form is the equation of a straight line:

$$y = a + bx$$

Here, y is the dependent variable you're trying to predict, x is the independent variable you're using for prediction, 'b' is the slope of the line, and a is the intercept.

Example: Let's dive into an example. Suppose you're a data analyst at a retail company, and you want to predict monthly sales based on advertising spend. Here's a small dataset:

Advertising Spend (x)	Monthly Sales (y)
10	40
20	50
30	60
40	70
50	80

Using linear regression, you find the equation of the line that best fits this data: **Y = 1.2x + 28**

This means that for every additional unit of advertising spend, monthly sales increase by 1.2 units, and if there were no advertising spend, you'd expect sales to be 28 units.

WHAT IS NONLINEAR REGRESSION?

Definition: Nonlinear regression is used when the relationship between the dependent and independent variables is not linear. Unlike linear regression, which fits a straight line, nonlinear regression fits a curve to the data.

Types: There are various types of nonlinear regression models, each suited for different kinds of data patterns:

- **Polynomial Regression:** Fits a polynomial equation to the data, such as $y = ax^2 + bx + c$.
- **Exponential Regression:** Models the data with an exponential function, such as y=ae^{bx}.
- **Logarithmic Regression:** Uses a logarithmic function, such as $y = a \log(x) + b$.

UNIT-5

A ***Regression line*** is a fundamental concept in statistics and data analysis used to understand the relationship between two variables. It represents the best-fit line that predicts the dependent variable based on the independent variable.

Regression lines

1. The regression line of Y on X is given by $Y = a + bX$

The corresponding regression equation is $Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}\left(X - \bar{X}\right)$

2. The regression line of X on Y is given by $X = a + bY$

The corresponding regression equation is $X - \bar{X} = r\frac{\sigma_X}{\sigma_Y}\left(Y - \bar{Y}\right)$

Example 10·1. *Calculate the correlation coefficient for the following heights (in inches) of fathers (X) and their sons (Y):*

X :	65	66	67	67	68.	69	70	72
Y :	67	68	65	68	72	72	69	71

Solution.

CALCULATIONS FOR CORRELATION COEFFICIENT

X	Y	X^2	Y^2	XY
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830
72	71	5184	5041	5112
Total 544	552	37028	38132	37560

From the calculations table we have $n = 8$

$\sum X = 544,\ \sum Y = 552,\ \sum X^2 = 37028,\ \sum Y^2 = 38132,\ \sum XY = 37560$

Hence $Mean\ of\ X\ \ \bar{X} = \frac{\sum X}{n} = \frac{544}{8} = 68$

$Mean\ of\ Y\ \ \bar{Y} = \frac{\sum Y}{n} = \frac{552}{8} = 69$

UNIT-5

Corelation $r = \frac{COV(X,Y)}{\sigma_X \cdot \sigma_Y} == \frac{\left(\frac{\sum XY}{n}\right) - \left(\overline{X}.\overline{Y}\right)}{\sqrt{\left(\frac{\sum X^2}{n}\right) - \left(\overline{X}\right)^2}\ \sqrt{\left(\frac{\sum Y^2}{n}\right) - \left(\overline{Y}\right)^2}}$

$$= \frac{\left[\frac{37560}{8}\right] - (68*69)}{\sqrt{\left(\frac{37028}{8}\right) - (68)^2}\ \sqrt{\left(\frac{38132}{8}\right) - (69)^2}} = \frac{4695-4692}{\sqrt{4628.5-4624}\ \sqrt{4766.5-4761}}$$

$$= \frac{3}{\sqrt{4.5}\ \sqrt{5.5}} = \frac{3}{(2.12)*(2.345)} = 0.6$$

Regression lines

1. The regression line of Y on X is given by Y=a+bX

The corresponding regression equation is

$$Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x})$$

2. The regression line of X on Y is given by X=a+bY

The corresponding regression equation is

$$X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$

From the Example 10.1 of Correlation we can derive the lines of regression as follows

$$\bar{X} = 68;\quad \bar{Y} = 69;\quad \sigma_x = 2.12;\quad \sigma_y = 2.345;\quad r = 0.6$$

Equation of line of regression of Y on X is

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

$$i.e.,\ Y = 69 + 0{\cdot}6 \times \frac{2{\cdot}35}{2{\cdot}12}(X - 68) \Rightarrow Y = 0{\cdot}665\,X + 23{\cdot}78$$

Equation of line of regression of X on Y is

$$X - \bar{X} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y})$$

$$\Rightarrow \quad X = 68 + 0{\cdot}6 \times \frac{2{\cdot}12}{2{\cdot}35}(Y - 69)\ \ i.e.,\ X = 0{\cdot}54Y + 30{\cdot}74$$

Estimation of Y value when Y=70

To estimate X for given Y, we use the line of regression of X on Y. If $Y = 70$, estimated value of X is given by

$$\hat{X} = 0{\cdot}54 \times 70 + 30{\cdot}74 = 68{\cdot}54,$$

UNIT-5

Example2: The following data is obtained from 10 observations.

$\sum x = 250, \sum x^2 = 6500, \sum y = 300, \sum y^2 = 10000$ and $\sum xy = 7900$.

Determine i) Coefficient of correlation ii) regression line y on x iii) regression line x on y.

Solution: It is given n=10 $\overline{X} = \frac{\sum X}{n} = \frac{250}{10} = 25$ *and* $\overline{Y} = \frac{\sum Y}{n} = \frac{300}{10} = 30$

i) We have the **Karl Pearson's coefficient of correlation**

$$r(X,Y) = r_{xy} = \frac{\left(\frac{\sum XY}{n}\right) - (\overline{X}.\overline{Y})}{\sqrt{\left(\frac{\sum X^2}{n}\right) - (\overline{X})^2}\sqrt{\left(\frac{\sum Y^2}{n}\right) - (\overline{Y})^2}} = \frac{\left(\frac{7900}{10}\right) - (25 \times 30)}{\sqrt{\left(\frac{6500}{10}\right) - (25)^2}\sqrt{\left(\frac{10000}{10}\right) - (30)^2}}$$

$$= \frac{790 - 750}{\sqrt{650 - 625}\sqrt{1000 - 900}} = \frac{40}{5 \times 10} = \frac{4}{5} = 0.8$$

From these calculations we can have

$$\overline{X} = 25; \overline{Y} = 30; \ \sigma_X = \sqrt{\left(\frac{1}{n}\sum X^2\right) - (\overline{X})^2} = 5; \quad \sigma_Y = \sqrt{\left(\frac{1}{n}\sum Y^2\right) - (\overline{Y})^2} = 10$$

and Correlation Coefficient $r = 0.8$

ii) The regression line of Y on X $\quad Y - \overline{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \overline{X})$

$$Y - \overline{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \overline{X}) \Rightarrow Y - 30 = 0.8\left(\frac{10}{5}\right)(X - 25)$$

$$\Rightarrow Y - 30 = 1.6(X - 25)$$

$$\Rightarrow Y = 1.6X - 40 + 30$$

$$\Rightarrow Y = 1.6X - 10$$

ii) The regression line of X on Y $\quad X - \overline{X} = r\frac{\sigma_X}{\sigma_Y}(Y - \overline{Y})$

$$X - \overline{X} = r\frac{\sigma_X}{\sigma_Y}(Y - \overline{Y}) \Rightarrow X - 25 = 0.8\left(\frac{5}{10}\right)(Y - 30)$$

$$\Rightarrow X - 25 = 0.4(Y - 30)$$

$$\Rightarrow X = 0.4Y - 12 + 25$$

$$\Rightarrow X = 0.4Y + 13$$

Similar Questions for Practice

1. Determine the coefficient of correlation and two lines of regression between expenses and sales from the following data and estimate (i) Sales when expenses are 12 lakhs (ii) expenses when target sales are 20 lakhs.

Expenses (lakhs)	7	10	9	4	11	5	3
sales (lakhs)	12	14	13	5	15	7	4

2. The coefficient of correlation between the marks in two subjects A & B was found to be 0.8. The average Marks in subject A was 24 and that of subject B was 21. The standard deviations of marks were 4 and 5 respectively. Find with the help of regression equations that (i) The expected Marks of a student in subject A when Subject B marks is 20. (ii) The expected marks in subject B when subject A marks is 25.
3. (a)Obtain the two regression lines from the following data n=20, $\sum X=80$, $\sum Y=40$, $\sum X^2=1680$, $\sum Y^2=320$ and $\sum XY=480$.
   (b) Develop the two regression lines based on following data,
   $\sum x = 50$, $\sum y = 60$, $\bar{x} = 5$, $\bar{y} = 6$ $\sum xy = 350$, also given that Variance of x=4 and variance of y=9.

Regression Coefficients:

From the regression equations of Y on X $Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}\left(X - \bar{X}\right)$, the constant 'b', the slope of the line of regression equation of Y on X is also called the coefficient of regression of Y on X. It represents the increment in the value of dependent variable Y corresponding to a unit change in the value of independent variable X.

More precisely, we write

$$b_{YX} = \text{Regression coefficient of } Y \text{ on } X = \frac{\mu_{11}}{\sigma_X^2} = r\frac{\sigma_Y}{\sigma_X}$$

Similarly, from the regression equations of X on Y $X - \bar{X} = r\frac{\sigma_X}{\sigma_Y}\left(Y - \bar{Y}\right)$, the coefficient of regression of X on Y indicates the change in the value of variable X corresponding to a unit change in the value of variable Y and is given by

$$b_{XY} = \text{Regression coefficient of } X \text{ on } Y = \frac{\mu_{11}}{\sigma_Y^2} = r\frac{\sigma_X}{\sigma_Y}$$

Properties of Regression Coefficients.

(a) Correlation coefficient is the geometric mean between the regression coefficients.

$$b_{XY} \times b_{YX} = r\frac{\sigma_X}{\sigma_Y} \times r\frac{\sigma_Y}{\sigma_X} = r^2$$

$$r = \pm\sqrt{b_{XY} \times b_{YX}}$$

(b) The sign of correlation coefficient is the same as that of regression coefficients, since the sign of each depends upon the co-variance term.

(c) If the regression coefficients are positive, 'r' is positive and if the regression coefficients are negative 'r' is negative.

(d) If one of the regression coefficients is greater than unity. The other must be less than unity.

(e) Arithmetic mean of the regression coefficients is greater than the correlation coefficient r, provided r > O.

$$\frac{1}{2}(b_{YX} + b_{XY}) \geq r$$

(f) Regression coefficients are independent of the change of origin but not of scale.

UNIT-5

Example 3: You are given the following data:

	X	Y
Arithmetic Mean	36	85
Standard Deviation	11	8

If the Correlation coefficient between X and Y is 0.66, then find (i) the two regression coefficients, (ii) the most likely value of Y when X=40

Solution: Given that $\bar{X} = 36;\ \bar{Y} = 85;\ \sigma_x = 11;\ \sigma_y = 8;\ r = 0.66$

(i) Regression coefficient of Y on X

$$b_{YX} = r\frac{\sigma_y}{\sigma_x} = 0.66 * \frac{8}{11} = 0.48$$

Regression coefficient of Y on X

$$b_{XY} = r\frac{\sigma_x}{\sigma_y} = 0.66 * \frac{11}{8} = 0.9075$$

ii) The most likely value of Y when X=40

we use the Regression line of Y on X

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X}) \Rightarrow Y - 85 = 0.66\left(\frac{8}{11}\right)(40 - 36)$$

$$\Rightarrow Y - 85 = 0.48\ (4)$$

$$\Rightarrow Y = 1.92 + 85$$

$$\Rightarrow Y = 86.92$$

Example 4: The equations of two lines of regression obtained in a correlation analysis are as follows 2X+3Y-8=0 and X+2Y-5=0 obtain (i) the mean values of X and Y (ii) the regression coefficients (iii) the correlation coefficient.

Solution: (i) Mean values of X and Y

Since both the lines passes through its mean values $(\bar{X}, \bar{Y})$ we can write the given equations as $2\bar{X} + 3\bar{Y} - 8 = 0$ *and* $\bar{X} + 2\bar{Y} - 5 = 0$

After solving these two equations we get $\bar{X} = 1$ *and* $\bar{Y} = 2$

(ii) Regression coefficients

Consider equation-1 { 2X+3Y-8=0} is the regression line of X on Y (i.e., of the form X=a+bY) write the equation-1 as

$$2X = 8 - 3Y \Rightarrow X = \frac{8}{2} - \frac{3}{2}Y$$

$$\text{then we have } b_{XY} = -\frac{3}{2}$$

Now consider equation-2 { X+2Y-5=0 } is the regression line of Y on X (i.e., form of Y=a+bX) Writing the equation-2 as

$$2Y = 5 - X \Rightarrow Y = \frac{5}{2} - \frac{1}{2}X$$

$$\text{then we have } b_{YX} = -\frac{1}{2}$$

UNIT-5

(iii) The correlation coefficient

As per the properties of regression coefficients we have

Correlation coefficient $r = -\sqrt{b_{XY} \times b_{YX}}$ $(\because \textit{both the coefficients are negative})$

Hence $$r = -\sqrt{b_{XY} \times b_{YX}} = -\sqrt{\frac{3}{2} \times \frac{1}{2}} = -\sqrt{\frac{3}{4}} = -0.866$$

Note: In this example suppose if we consider the equation-1 is the line of regression of Y on X then we have $Y = \frac{8}{3} - \frac{2}{3}X$ *hence* $b_{YX} = -\frac{2}{3}$

And equation-2 is the line of X on Y the we have $X = 5 - 2Y$ *hence* $b_{XY} = -2$

With these coefficients, correlation $r = -\sqrt{b_{XY} \times b_{YX}} = -\sqrt{\frac{2}{3} \times 2} = -\sqrt{\frac{4}{3}} = -1.1547$

This is not correct value of correlation, since correlation should be in between -1 and +1

So, *we must consider the equation-(1) as the line X on Y and equation-(2) is the line of Y on X.*

Similar Questions for Practice

1. Find the most likely price in Mumbai corresponding to the price of Rs.70 at Kolkata, and the most likely price in Kolkata corresponding to the price of Rs.68 in Mumbai from the following data.

	Kolkata	Mumbai
Average	65	67
Standard deviation	2.5	3.5

Correlation Coefficient between the prices of commodities in two cities r = 0.8

2. **(a)** For the given lines of regression 3X–2Y=5and X–4Y=7. Find (i) Means of X and Y (ii) Regression coefficients (ii) Coefficient of correlation.

(b) The two regression lines were found to be 4X–5Y+33=0 and 20X–9Y–107=0 . Find the mean values and coefficient of correlation between X and Y.

Spearman's Rank Correlation Coefficient

Coefficient of correlation between the ranks of x_i's and y_i's is called the rank correlation coefficient between A and B for that group of individuals.

$$\rho = 1 - \left[\frac{6 * (\sum d_i^2)}{n(n^2 - 1)}\right]$$

This is the Spearman's formula for the rank correlation coefficient.

The limits of Spearman's Rank Correlation are -1 and +1

UNIT-5

Example1: The ranks of some 16 students in Mathematics and Physics are as follows. Two numbers within brackets denote the ranks of the students in Mathematics and Physics.

(1, 1) (2, 10) (3, 3) (4, 4) (5, 5) (6, 7) (7, 2) (8, 6) (9, 8)
(10, 11) (11, 15) (12, 9) (13, 14) (14, 12) (15, 16) (16, 13).

Calculate the **rank correlation coefficient** for proficiencies of this group in Mathematics and Physics.

Solution.

Ranks in Maths. (X)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
Ranks in Physics(Y)	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13	
$d = X - Y$	0	−8	0	0	0	−1	5	2	1	−1	−4	3	−1	2	−1	3	0
d^2	0	64	0	0	0	1	25	4	1	1	16	9	1	4	1	9	136

Rank correlation coefficient is given by

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 136}{16 \times 255} = 1 - \frac{1}{5} = \frac{4}{5} = 0.8$$

Example2: From the following information relating to the stock exchange quotations for two shares A and B, determine **Spearman's Rank correlation coefficient** between shares A and B.

Price of Share A (Rs.)	160	164	172	182	166	170	178
Price of Share B (Rs.)	292	280	260	234	266	254	230

Solution: Since the data is not given the ranks, we have to assign the ranks for both the series values, so that we can calculate the rank correlation

Price of A (X)	Price of B (Y)	ranks of A (Rx)	ranks of B (Ry)	$d_i^2 = (R_x - R_y)^2$
160	292	7	1	36
164	280	6	2	16
172	260	3	4	1
182	234	1	6	25
166	266	5	3	4
170	254	4	5	1
178	230	2	7	25
				108

Spearman's Rank correlation coefficient

$$\rho = 1 - \left[\frac{6\sum d_i^2}{n(n^2-1)}\right] = 1 - \left[\frac{6 \times 108}{7(49-1)}\right] = 1 - \left[\frac{648}{336}\right] = 1 - 1.9285 = -0.9285$$

UNIT-5

Repeated Ranks:

$$\rho = 1 - \left[\frac{6\left[\sum d_i^2 + T_X + T_Y\right]}{n\left(n^2 - 1\right)} \right] \quad where \quad T_X = T_Y = \sum \frac{m\left(m^2 - 1\right)}{12}$$

Example1: **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

Solution.

CALCULATIONS FOR RANK CORRELATION

X	Y	Rank X (x)	Rank Y (y)	d = x − y	d^2
68	62	4	5	−1	1
64	58	6	7	−1	1
75	68	2·5	3·5	−1	1
50	45	9	10	−1	1
64	81	6	1	5	25
80	60	1	6	−5	25
75	68	2·5	3·5	−1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
				$\Sigma d = 0$	$\Sigma d^2 = 72$

In the X-series we see that the value 75 occurs 2 times. The common rank given to these values is 2·5 which is the average of 2 and 3, the ranks which these values would have taken if they were different. The next value 68, then gets the next rank which' is 4. Again we see that value 64 occurs thrice. The common rank given to it is 6 which is the average of 5, 6 and 7.

Similarly in the Y-series, the value 68 occurs twice and its common rank is 3·5 which is the average of 3 and 4. As a result of these common rankings, the formula for 'p' has to be corrected.

In the X -series the correction is to be applied twice, once for the value 75 which occurs twice (m = 2) an the value 64 which occurs thrice (m = 3). $T_X = \frac{5}{2}$ *and* $T_Y = \frac{1}{2}$

The total correction for X-series is

$$\frac{2(4-1)}{12} + \frac{3(9-1)}{12} = \frac{5}{2}$$

Similarly, this correction for the Y-series is $\frac{2(4-1)}{12} = \frac{1}{2}$, as the value 68 occurs twice.

Thus $\rho = 1 - \frac{6\left[\Sigma d^2 + \frac{5}{2} + \frac{1}{2}\right]}{n(n^2-1)} = 1 - \frac{6(72+3)}{10 \times 99} = 0.545$

UNIT-5

Similar Questions for Practice

1. Calculate the rank correlation coefficient between the rankings of 10 students in accordance with their performance in two subjects Mathematics and Computer science given in the following table.

Rank in Mathematics	6	5	3	10	2	4	9	7	8	1
Rank in Computer science	3	8	4	9	1	6	10	7	5	2

2. From the following information relating to the stock exchange quotations for two shares A and B, determine Spearman's Rank correlation coefficient between shares A and B.

Price of Share A(Rs.)	160	164	172	182	166	170	178
Price of Share B(Rs.)	292	280	260	234	266	254	230

3. The table below shows the I.Q . scores of 10 fathers and their eldest sons:

Father's IQ	98	97	102	103	103	105	110	114	116	102
Son's IQ	102	94	105	115	113	111	105	112	120	105

Calculate rank correlation coefficient between the I.Q. of father 's and their eldest son's.

UNIT-5

SUMMARY

1. Correlation, Positive Correlation, Negative Correlation.

2. Karl Pearson's Coefficient of Correlation

$$r(X,Y)=r_{xy}=\frac{COV(X,Y)}{\sigma_X\ .\ \sigma_Y}=\frac{\left(\frac{1}{n}\sum XY\right)-\left(\overline{X}.\overline{Y}\right)}{\sqrt{\left(\frac{1}{n}\sum X^2\right)-\left(\overline{X}\right)^2}\sqrt{\left(\frac{1}{n}\sum Y^2\right)-\left(\overline{Y}\right)^2}}$$

[

3. Limits of Karl Pearson's coefficient of correlation $\mathbf{-1 \le\ r_{xy}\ \le +1}$

4. Regression line of (i) Y on X Y=a+bX (ii) X on Y is X=a+bY

i) The regression equation of Y on X → $Y-\overline{Y}=r\frac{\sigma_Y}{\sigma_X}\left(X-\overline{X}\right)$

ii) The regression equation of X on Y → $X-\overline{X}=r\frac{\sigma_X}{\sigma_Y}\left(Y-\overline{Y}\right)$

5. Regression Coefficients $b_{yx}=r\frac{\sigma_y}{\sigma_x};\ \ b_{yx}=r\frac{\sigma_y}{\sigma_x}$

6. **Properties of Regression Coefficients.**

7. Spearman's Rank Correlation Coefficient $\rho=1-\left[\frac{6\sum d_i^2}{n\left(n^2-1\right)}\right]$

8. Spearman's Rank Correlation Coefficient for equal (Repeated Ranks)

$$\rho=1-\left[\frac{6\left[\sum d_i^2+T_X+T_Y\right]}{n\left(n^2-1\right)}\right]\quad where\ \ T_X=T_Y=\sum\frac{m\left(m^2-1\right)}{12}$$

$$where\ \ m\ is\quad 'Number\ \ of\ \ times\ \ a\ \ Value\ \ is\ \ repeated'$$