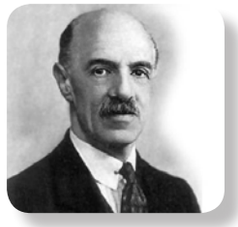
CHAPTER

4 CORRELATION ANALYSIS

Karl Pearson

Karl Pearson (1857-1936) was a English Mathematician and Biostatistician. He founded the world's first university statistics department at University College, London in 1911. The linear correlation coefficient is also called Pearson product moment correlation coefficent. It was developed by Karl Pearson with a related idea by Francis Galton (see Regression analysis - for Galton's contribution). It is the first of the correlation measures developed and commonly used.



Charles Spearman

Charles Edward Spearman (1863-1945) was an English psychologist and ,after serving 15 years in Army he joined to study PhD in Experimental Psychology and obtained his degree in 1906. Spearman was strongly influenced by the work of Galton and developed rank correlation in 1904.He also pioneered factor analysis in statistics.

"When the relationship is of a quantitative nature, the appropriate statistical tool for discovering the existence of relation and measuring the intensity of relationship is known as correlation"

—CROXTON AND COWDEN

LEARNING OBJECTIVES

The student will be able to

- learn the meaning, definition and the uses of correlation.
- identify the types of correlation.
- understand correlation coefficient for different types of measurement scales.
- differentiate different types of correlation using scatter diagram.
- calculate Karl Pearson's coefficient of correlation, Spearman's rank correlation coefficient and Yule's coefficient of association.
- interpret the given data with the help of coefficient of correlation.

Introduction

"Figure as far as you can, then add judgment"

The statistical techniques discussed so far are for ***only one variable***. In many research situations one has to consider two variables simultaneously to know whether these ***two variables*** are related linearly. If so, what type of relationship that exists between them. This leads to bivariate (two variables) data analysis namely correlation analysis. If two quantities vary in such a way that movements (upward or downward) in one are accompanied by the movements(upward or downward) in the other, these quantities are said to be co-related or correlated.

The correlation concept will help to answer the following types of questions.

- Whether study time in hours is related with marks scored in the examination?
- Is it worth spending on advertisement for the promotion of sales?
- Whether a woman's age and her systolic blood pressure are related?
- Is age of husband and age of wife related?
- Whether price of a commodity and demand related?
- Is there any relationship between rainfall and production of rice?

4.1 DEFINITION OF CORRELATION

Correlation is a statistical measure which helps in analyzing the interdependence of two or more variables. In this chapter the dependence between only two variables are considered.

1. **A.M. Tuttle** defines correlation as:

   "An analysis of the co-variation of two or more variables is usually called correlation"

2. **Ya-kun-chou** defines correlation as:

   "The attempts to determine the degree of relationship between variables".

Correlation analysis is the process of studying the strength of the relationship between two related variables. High correlation means that variables have a strong linear relationship with each other while a low correlation means that the variables are hardly related. The type and intensity of correlation is measured through the correlation analysis. The measure of correlation is the correlation coefficient or correlation index. It is an absolute measure.

Uses of correlation

- Investigates the type and strength of the relationship that exists between the two variables.
- Progressive development in the methods of science and philosophy has been characterized by the rich knowledge of relationship.

4.2 TYPES OF CORRELATION

1. ***Simple (Linear) correlation*** (2 variables only) : The correlation between the given two variables. It is denoted by r_{xy}
2. ***Partial correlation (more than 2 variables):*** The correlation between any two variables while removing the effect of other variables. It is denoted by $r_{xy.z\ldots}$

3. ***Multiple correlation (more than 2 variables) :*** The correlation between a group of variables and a variable which is not included in that group. It is denoted by $R_{y.(xz...)}$

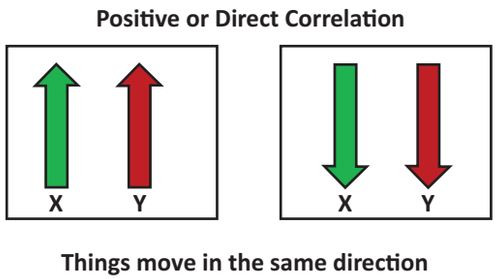
In this chapter, we study simple correlation only, multiple correlation and partial correlation involving three or more variables will be studied in higher classes .

4.2.1 Simple correlation or Linear correlation

Here, we are dealing with data involving two related variables and generally we assign a symbol '*x* ' to scores of one variable and symbol '*y* ' to scores of the other variable. There are five types in simple correlation. They are

1. Positive correlation (Direct correlation)
2. Negative correlation (Inverse correlation)
3. Uncorrelated
4. Perfect positive correlation
5. Perfect negative correlation

1) Positive correlation: (Direct correlation)

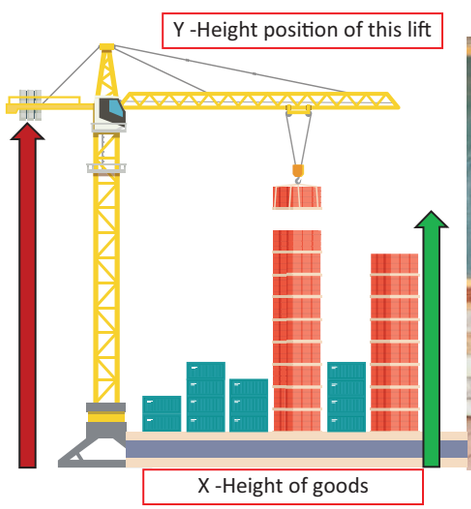


The variables are said to be positively correlated if larger values of *x* are associated with larger values of *y* and smaller values of *x* are associated with smaller values of *y*. In other words, if both the variables are varying in the *same direction* then the correlation is said to be positive. In other words, if one variable increases, the other variable (on an average) also increases or if one variable decreases, the other (on an average)variable also decreases.

For example,

i) Income and savings

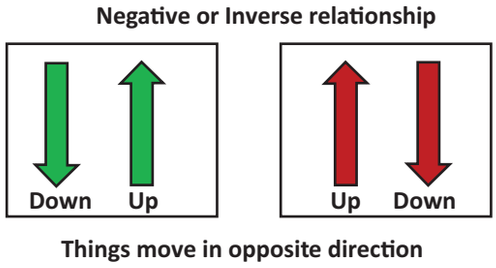
ii) Marks in Mathematics and Marks in Statistics. (*i.e.*,Direct relationship pattern exists).



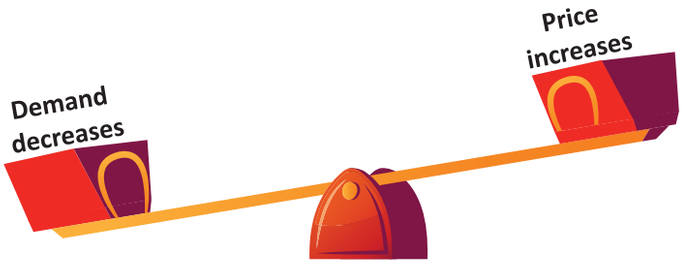
Height of the Lift increases / decreases according to the Height of goods increases / decreases.

The starting position of writing depends on the height of the writer.

2) Negative correlation: (Inverse correlation)



The variables are said to be negatively correlated if smaller values of x are associated with larger values of y or larger values x are associated with smaller values of y. That is the variables varying in the ***opposite directions*** is said to be negatively correlated. In other words, if one variable increases the other variable decreases and vice versa.



For example,

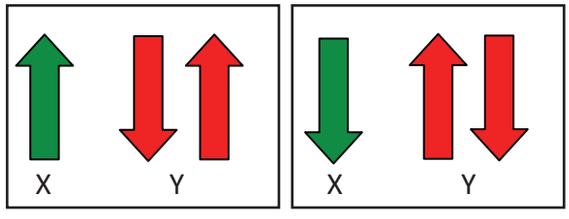
i) Price and demand

ii) Unemployment and purchasing power

3) Uncorrelated:

The variables are said to be uncorrelated if smaller values of x are associated with smaller or larger values of y and larger values of x are associated with larger or smaller values of y. If the two variables do not associate linearly, they are said to be uncorrelated. Here $r = 0$.

Important note: Uncorrelated does not imply independence. This means "do not interpret as the two variables are independent instead interpret as there is no specific linear pattern exists but there may be non linear relationship".



4) Perfect Positive Correlation

If the values of x and y increase or decrease ***proportionately*** then they are said to have perfect positive correlation.

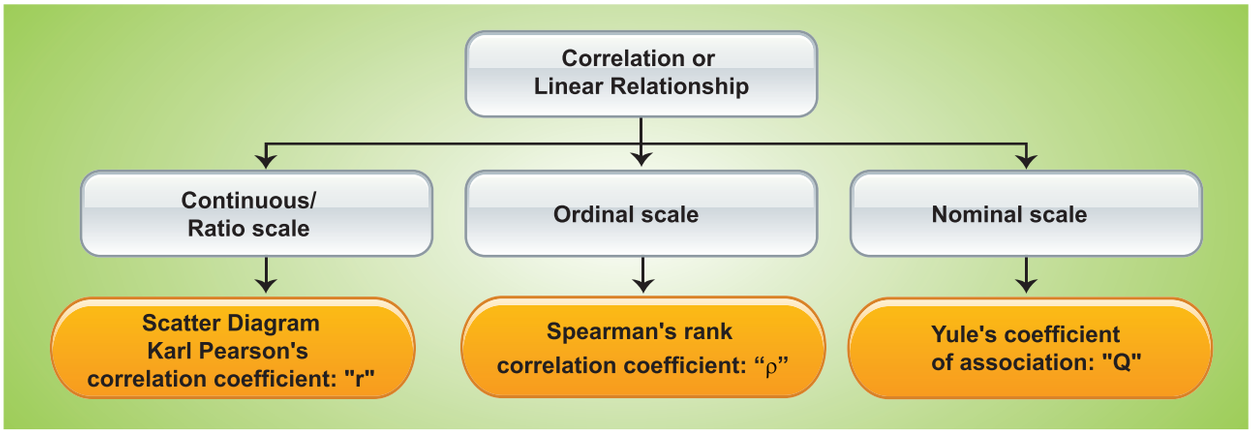
5) Perfect Negative Correlation

If x increases and y decreases ***proportionately*** or if x decreases and y increases ***proportionately***, then they are said to have perfect negative correlation.

Correlation Analysis

The purpose of correlation analysis is to find the existence of linear relationship between the variables. However, the method of calculating correlation coefficient depends on the types of measurement scale, namely, ratio scale or ordinal scale or nominal scale.

Statistical tool selection



Methods to find correlation

1. Scatter diagram
2. Karl Pearson's product moment correlation coefficient : 'r'
3. Spearman's Rank correlation coefficient: 'ρ'
4. Yule's coefficient of Association: 'Q'

NOTE

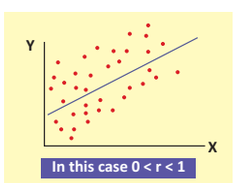
For higher order dimension of nominal or categorical variables in a contingency table, use chi-square test for independence of attributes. (Refer Chapter 2)

4.3 SCATTER DIAGRAM

A scatter diagram is the simplest way of the diagrammatic representation of bivariate data. One variable is represented along the X-axis and the other variable is represented along the Y-axis. The pair of points are plotted on the two dimensional graph. The diagram of points so obtained is known as scatter diagram. The direction of flow of points shows the type of correlation that exists between the two given variables.

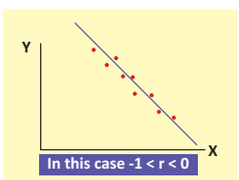
1) Positive correlation

If the plotted points in the plane form a band and they show the rising trend from the lower left hand corner to the upper right hand corner, the two variables are positively correlated.



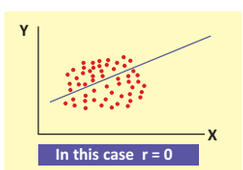
2) Negative correlation

If the plotted points in the plane form a band and they show the falling trend from the upper left hand corner to the lower right hand corner, the two variables are negatively correlated.



3) Uncorrelated

If the plotted points spread over in the plane then the two variables are uncorrelated.

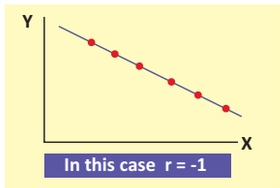


4) Perfect positive correlation

If all the plotted points lie on a straight line from lower left hand corner to the upper right hand corner then the two variables have perfect positive correlation.

5) Perfect Negative correlation

If all the plotted points lie on a straight line falling from upper left hand corner to lower right hand corner, the two variables have perfect negative correlation.



4.3.1 Merits and Demerits of scatter diagram

Merits

- It is a simple and non-mathematical method of studying correlation between the variables.
- It is not influenced by the extreme items
- It is the first step in investigating the relationship between two variables.
- It gives a rough idea at a glance whether there is a positive correlation, negative correlation or uncorrelated.

Demerits

- We get an idea about the direction of correlation but we cannot establish the exact strength of correlation between the variables.
- No mathematical formula is involved.

4.4 KARL PEARSON'S CORRELATION COEFFICIENT

When there exists some relationship between two measurable variables, we compute the degree of relationship using the correlation coefficient.

Co-variance

Let (X,Y) be a bivariable normal random variable where $V(X)$ and $V(Y)$ exists. Then, covariance between X and Y is defined as

$$\text{cov}(X,Y) = E[(X-E(X))(Y-E(Y))] = E(XY) - E(X)E(Y)$$

If (x_i,y_i), i=1,2, ..., n is a set of n realisations of (X,Y), then the sample covariance between X and Y can be calculated from

$$\text{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{n}\sum_{i=1}^{n}x_iy_i - \bar{x}\,\bar{y}$$

4.4.1 Karl Pearson's coefficient of correlation

When X and Y are linearly related and (X,Y) has a bivariate normal distribution, the co-efficient of correlation between X and Y is defined as

$$r(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{V(X)V(Y)}}$$

This is also called as product moment correlation co-efficient which was defined by Karl Pearson.

Based on a given set of n paired observations (x_i,y_i), i=1,2, ... n the sample correlation co-efficient between X and Y can be calculated from

$$r(X,Y) = \frac{\frac{1}{n}\sum_{i=1}^{n}x_iy_i - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2 - \bar{y}^2}}$$

or, equivalently

$$r(X,Y)=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2}}$$

4.4.2 Properties

1. The correlation coefficient between X and Y is same as the correlation coefficient between Y and X (*i.e*, $r_{xy} = r_{yx}$).
2. The correlation coefficient is free from the units of measurements of X and Y
3. The correlation coefficient is unaffected by change of scale and origin.

   Thus, if $u_i=\frac{x_i-A}{c}$ and $v_i=\frac{y_i-B}{d}$ with $c \neq 0$ and $d \neq 0$ $\quad i=1,2,\ldots,n$

$$r=\frac{n\sum_{i=1}^{n}u_iv_i-\sum_{i=1}^{n}u_i\sum_{i=1}^{n}v_i}{\sqrt{n\sum_{i=1}^{n}u_i^2-\left(\sum_{i=1}^{n}u_i\right)^2}\sqrt{n\sum_{i=1}^{n}v_i^2-\left(\sum_{i=1}^{n}v_i\right)^2}}$$

where A and B are arbitrary values.

Remark 1: If the widths between the values of the variabls are not equal then take $c = 1$ and $d = 1$.

Interpretation

The correlation coefficient lies between -1 and +1. *i.e.* $-1 \leq r \leq 1$

- A positive value of 'r' indicates positive correlation.
- A negative value of 'r' indicates negative correlation
- If $r = +1$, then the correlation is perfect positive
- If $r = -1$, then the correlation is perfect negative.
- If $r = 0$, then the variables are uncorrelated.
- If $|r| \geq 0.7$ then the correlation will be of higher degree. In interpretation we use the adjective 'highly'
- If X and Y are independent, then $r_{xy} = 0$. However the converse need not be true.

Example 4.1

The following data gives the heights(in inches) of father and his eldest son. Compute the correlation coefficient between the heights of fathers and sons using Karl Pearson's method.

Height of father	65	66	67	67	68	69	70	72
Height of son	67	68	65	68	72	72	69	71

Solution:

Let x denote height of father and y denote height of son. The data is on the ratio scale. We use Karl Pearson's method.

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

Calculation

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
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
544	552	37028	38132	37560

$$r = \frac{8\times 37560 - 544\times 552}{\sqrt{8\times 37028 - (544)^2}\sqrt{8\times 38132 - (552)^2}} = 0.603$$

Heights of father and son are positively correlated. It means that on the average , if fathers are tall then sons will probably tall and if fathers are short, probably sons may be short.

Short-cut method

Let $A = 68$, $B = 69$, $c = 1$ and $d = 1$

x_i	y_i	$u_i = (xi - A)/c$ $= xi - 68$	$v_i = (y_i - B)/d$ $= y_i - 69$	u_i^2	v_i^2	$u_i v_i$
65	67	-3	-2	9	4	6
66	68	-2	-1	4	1	2
67	65	-1	-4	1	16	4
67	68	-1	-1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
Total		0	0	36	44	24

$$r = \frac{n\sum_{i=1}^{n} u_i v_i - \sum_{i=1}^{n} u_i \sum_{i=1}^{n} v_i}{\sqrt{n\sum_{i=1}^{n} u_i^2 - \left(\sum_{i=1}^{n} u_i\right)^2}\sqrt{n\sum_{i=1}^{n} v_i^2 - \left(\sum_{i=1}^{n} v_i\right)^2}}$$

$$r = \frac{8\times 24 - 0\times 0}{\sqrt{8\times 36 - (0)^2}\ \sqrt{8\times 44 - (0)^2}}$$

$$r = \frac{8\times 24}{\sqrt{8\times 36}\ \sqrt{8\times 44}}$$

$$= 0.603$$

Note: The correlation coefficient computed by using direct method and short-cut method is the same.

Example 4.2

The following are the marks scored by 7 students in two tests in a subject. Calculate coefficient of correlation from the following data and interpret.

Marks in test-1	12	9	8	10	11	13	7
Marks in test-2	14	8	6	9	11	12	3

Solution:

Let x denote marks in test-1 and y denote marks in test-2.

	x_i	y_i	x_i^2	y_i^2	x_iy_i
	12	14	144	196	168
	9	8	81	64	72
	8	6	64	36	48
	10	9	100	81	90
	11	11	121	121	121
	1	12	169	144	156
	7	3	49	9	21
Total	70	63	728	651	676

$$r = \frac{n\sum_{i=1}^{n} x_iy_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\ \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$\sum_{i=1}^{n} x_i = 70 \quad \sum_{i=1}^{n} x_i^2 = 728 \quad \sum_{i=1}^{n} x_iy_i = 676$$

$$\sum_{i=1}^{n} y_i = 63 \quad \sum_{i=1}^{n} y_i^2 = 651 \quad n = 7$$

$$r = \frac{7\times 676 - 70\times 63}{\sqrt{\left[7\times 728 - 70^2\right]} \times \sqrt{\left[7\times 651 - 63^2\right]}}$$

$$= \frac{4732 - 4410}{\sqrt{[5096 - 4900]} \times \sqrt{[7\times 651 - 3969]}}$$

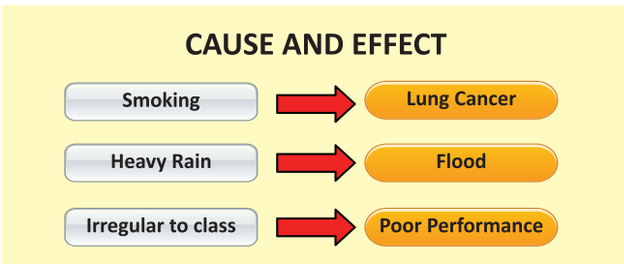
$$= \frac{322}{\sqrt{196} \times \sqrt{588}} = \frac{322}{14\times 24.25} = \frac{322}{339.5} = 0.95$$

There is a high positive correlation between test-1 and test-2. That is those who perform well in test-1 will also perform well in test-2 and those who perform poor in test-1 will perform poor in test- 2.

The students can also verify the results by using shortcut method.

4.4.3 Limitations of Correlation

Although correlation is a powerful tool, there are some limitations in using it:

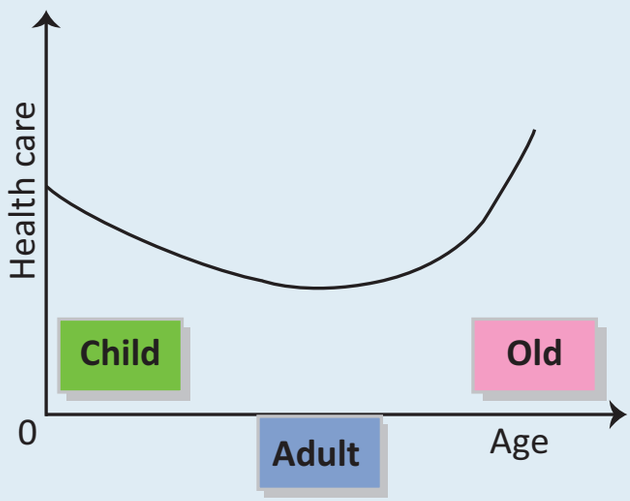


1. Outliers (extreme observations) strongly influence the correlation coefficient. If we see outliers in our data, we should be careful about the conclusions we draw from the value of r. The outliers may be dropped before the calculation for meaningful conclusion.
2. Correlation does not imply causal relationship. That a change in one variable causes a change in another.

NOTE

1. **Uncorrelated** : Uncorrelated ($r = 0$) implies no 'linear relationship'. But there may exist non-linear relationship (curvilinear relationship).

   Example: Age and health care are related. Children and elderly people need much more health care than middle aged persons as seen from the following graph.



However, if we compute the linear correlation r for such data, it may be zero implying age and health care are uncorrelated, but non-linear correlation is present.

2. **Spurious Correlation** : The word '**spurious**' from Latin means **'false'** or 'illegitimate'. *Spurious correlation means an association extracted from correlation coefficient that may not exist in reality.*

4.5 SPEARMAN'S RANK CORRELATION COEFFICIENT

If the data are in ordinal scale then Spearman's rank correlation coefficient is used. It is denoted by the Greek letter ρ (**rho**).

Spearman's correlation can be calculated for the subjectivity data also, like competition scores. The data can be ranked from low to high or high to low by assigning ranks.

Spearman's rank correlation coefficient is given by the formula

$$\rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2 - 1\right)}$$

where $D_i = R_{1i} - R_{2i}$

R_{1i} = rank of i in the first set of data

R_{2i} = rank of i in the second set of data and

n = number of pairs of observations

Interpretation

Spearman's rank correlation coefficient is a statistical measure of the strength of a monotonic (increasing/decreasing) relationship between paired data. Its interpretation is similar to that of Pearson's. That is, the closer to the ±1 means the stronger the monotonic relationship.

Positive Range	Negative Range
0.01 to 0.19: "Very Weak Agreement"	(-0.01) to (-0.19): "Very Weak Disagreement"
0.20 to 0.39:"Weak Agreement"	(-0.20) to (-0.39): "Weak Disagreement"
0.40 to 0.59: "Moderate Agreement"	(-0.40) to (-0.59): "Moderate Disagreement"
0.60 to 0.79: "Strong Agreement"	(-0.60) to (-0.79): "Strong Disagreement"
0.80 to 1.0: "Very Strong Agreement"	(-0.80) to (-1.0): "Very Strong Disagreement"

Example 4.3

Two referees in a flower beauty competition rank the 10 types of flowers as follows:

Referee A	1	6	5	10	3	2	4	9	7	8
Referee B	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient and find out what degree of agreement is between the referees.

Solution:

Rank by 1st referee R_{1i}	Rank by 2nd referee R_{2i}	$D_i = R_{1i} - R_{2i}$	D_i^2
1	6	-5	25
6	4	2	4
5	9	-4	16
10	8	2	4
3	1	2	4
2	2	0	0
4	3	1	1
9	10	-1	1
7	5	2	4
8	7	1	1
			$\sum_{i=1}^{n} D_i^2 = 60$

Here n = 10 and $\sum_{i=1}^{n} D_i^2 = 60$

$$\rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2 - 1\right)}$$

$$= 1 - \frac{6 \times 60}{10\left(10^2 - 1\right)} = 1 - \frac{360}{10(99)} = 1 - \frac{360}{990} = 0.636$$

Interpretation: Degree of agreement between the referees 'A' and 'B' is 0.636 and they have "strong agreement" in evaluating the competitors.

Example 4.4

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

Candidates	1	2	3	4	5
Marks in Tamil	75	40	52	65	60
Marks in English	25	42	35	29	33

Solution:

Tamil		English		$D_i = R_{1i} - R_{2i}$	D_i^2
Marks	Rank (R_{1i})	Marks	Rank (R_{2i})		
75	1	25	5	-4	16
40	5	42	1	4	16
52	4	35	2	2	4
65	2	20	4	-2	4
60	3	33	3	0	0
					40

$$\sum_{i=1}^{n} D_i^2 = 40 \text{ and } n = 5$$

$$\rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2 - 1\right)}$$

$$= 1 - \frac{6 \times 40}{5\left(5^2 - 1\right)} = 1 - \frac{240}{5(24)} = -1$$

Interpretation: This perfect negative rank correlation (-1) indicates that scorings in the subjects, totally disagree. Student who is best in Tamil is weakest in English subject and vice-versa.

Example 4.5

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

Years	2013	2014	2015	2016	2017	2008	2009
Equity shares	97.5	99.4	98.6	96.2	95.1	98.4	97.1
Reference 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 share	R_{1i}	R_{2i}	$D_i = R_{1i} - R_{2i}$	D_i^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	79.0	7	1	6	36
98.4	74.6	3	7	-4	16
97.1	76.2	5	4	1	1
					$\sum_{i=1}^{n} D_i^2 = 90$

$\sum_{i=1}^{n} D_i^2 = 90$ and $n = 7$.

Rank correlation coefficient is

$$\rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2 - 1\right)}$$

$$= 1 - \frac{6 \times 90}{7\left(7^2 - 1\right)} = 1 - \frac{540}{7 \times 48} = 1 - \frac{540}{336} = 1 - 1.6071 = -0.6071$$

Interpretation: There is a negative correlation between equity shares and preference share prices. There is a strong disagreement between equity shares and preference share prices.

4.5.1 Repeated ranks

When two or more items have equal values (i.e., a tie) it is difficult to give ranks to them. In such cases the items are given the average of the ranks they would have received. For example, if two individuals are placed in the 8th place, they are given the rank $\frac{8+9}{2} = 8.5$ each, which is common rank to be assigned and the next will be 10; and if three ranked equal at the 8th place, they are given the rank $\frac{8+9+10}{3} = 9$ which is the common rank to be assigned to each; and the next rank will be 11.

In this case, a different formula is used when there is more than one item having the same value.

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

where m_i is the number of repetitions of i^{th} rank

Example 4.6

Compute the rank correlation coefficient for the following data of the marks obtained by 8 students in the Commerce and Mathematics.

Marks in Commerce	15	20	28	12	40	60	20	80
Marks in Mathematics	40	30	50	30	20	10	30	60

Solution:

Marks in Commerce (X)	Rank (R_{1i})	Marks in Mathematics (Y)	Rank (R_{2i})	$D_i = R_{1i} - R_{2i}$	D_i^2
15	2	40	6	-4	16
20	3.5	30	4	-0.5	0.25
28	5	50	7	-2	4
12	1	30	4	-3	9
40	6	20	2	4	16
60	7	10	1	6	36
20	3.5	30	4	-0.5	0.25
80	8	60	8	0	0
				Total	$\sum D^2 = 81.5$

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

Repetitions of ranks

In Commerce (X), 20 is repeated two times corresponding to ranks 3 and 4. Therefore, 3.5 is assigned for rank 2 and 3 with m_1=2.

In Mathematics (Y), 30 is repeated three times corresponding to ranks 3, 4 and 5. Therefore, 4 is assigned for ranks 3,4 and 5 with m_2=3.

Therefore,

$$\rho = 1 - 6\left[\frac{81.5 + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(3^3 - 3\right)}{8\left(8^2 - 1\right)}\right]$$

$$= 1 - 6\frac{[81.5 + 0.5 + 2]}{504} = 1 - \frac{504}{504} = 0$$

Interpretation: Marks in Commerce and Mathematics are uncorrelated

4.6 YULE'S COEFFICIENT OF ASSOCIATION

Udny yule

This measure is used to know the existence of relationship between the two attributes A and B (binary complementary variables). Examples of attributes are drinking, smoking, blindness, honesty, etc.

Udny Yule (1871 – 1951), was a British statistician. He was educated at Winchester College and at University College London. After a year dong research in experimental physics, he returned to University College in 1893 to work as a demonstrator for Karl Pearson. Pearson was beginning to work in statistics and Yule followed him into this new field. Yule was a prolific writer, and was active in Royal Statistical Society and received its Guy Medal in Gold in 1911, and served as its President in 1924–26.The concept of Association is due to him.

Coefficient of Association

Yule's Coefficient of Association measures the strength and direction of association. "Association" means that the attributes have some degree of agreement.

2×2 Contingency Table

Attribute A ⇩	Attribute B		Total
	Yes B	No β	
Yes A	(AB)	$(A\beta)$	(A)
No α	(αB)	$(\alpha\beta)$	(α)
Total	(B)	(β)	N

Yule's coefficient: $Q = \dfrac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$

Note 1: The usage of the symbol α is not to be confused with level of significance.

Note 2: (AB): Number with attributes AB etc.

This coefficient ranges from –1 to +1. The values between –1 and 0 indicate inverse relationship (association) between the attributes. The values between 0 and +1 indicate direct relationship (association) between the attributes.

Example 4.7

Out of 1800 candidates appeared for a competitive examination 625 were successful; 300 had attended a coaching class and of these 180 came out successful. Test for the association of attributes attending the coaching class and success in the examination.

Solution:

$N = 1800$

A: Success in examination

B: Attended the coaching class

α: No success in examination

β: Not attended the coaching class

$(A) = 625,\ (B) = 300,\ (AB) = 180$

	B	β	Total
A	180	445	625
α	120	1055	1175
Total	300	1500	$N = 1800$

Yule's coefficient: $Q = \dfrac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$

$$= \frac{180\times1055-445\times120}{180\times1055+445\times120}$$

$$= \frac{189900-53400}{189900+53400}$$

$$= \frac{136500}{243300}$$

$$= 0.561 > 0$$

Interpretation: There is a positive association between success in examination and attending coaching classes. Coaching class is useful for success in examination.

Remark: Consistency in the data using contingency table may be found as under.

Construct a 2 × 2 contingency table for the given information. If at least one of the cell frequencies is negative then there is inconsistency in the given data.

Example 4.8

Verify whether the given data: N = 100, (A) = **75,** (B) = 60 and (AB) = 15 is consistent.

Solution:

The given information is presented in the following contingency table.

	B	β	Total
A	15	60	75
α	45	(-20)	25
Total	60	40	N = 100

Notice that $(\alpha\beta)$ = –20

Interpretation: Since one of the cell frequencies is negative, the given data is "Inconsistent".

POINTS TO REMEMBER

- ❖ Correlation study is about finding the linear relationship between two variables. Correlation is not causation. Sometimes the correlation may be spurious.
- ❖ Correlation coefficient lies between –1 and +1.
- ❖ Pearson's correlation coefficient provides the type of relationship and intensity of relationship, for the data in ratio scale measure.
- ❖ Spearman's correlation measures the relationship between the two ordinal variables.
- ❖ Yule's coefficient of Association measures the association between two dichotomous attributes.

EXERCISE 4

I. Choose the best answer.

1. The statistical device which helps in analyzing the co-variation of two or more variables is

   (a) variance (b) probability

   (c) correlation coefficient (d) coefficient of skewness

2. "The attempts to determine the degree of relationship between variables is correlation" is the definition given by

   (a) A.M. Tuttle (b) Ya-Kun-Chou

   (c) A.L. Bowley (d) Croxton and Cowden

3. If the two variables do not have linear relationship between them then they are said to have

   (a) positive correlation (b) negative correlation

   (c) uncorrelated (d) spurious correlation

4. If all the plotted points lie on a straight line falling from upper left hand corner to lower right hand corner then it is called

   (a) perfect positive correlation (b) perfect negative correlation

   (c) positive correlation (d) negative correlation

5. If $r = +1$, then the correlation is called

   (a) perfect positive correlation (b) perfect negative correlation

   (c) positive correlation (d) negative correlation

6. The correlation coefficient lies in the interval

   (a) $-1 \leq r \leq 0$ (b) $-1 < r < 1$ (c) $0 \leq r \leq 1$ (d) $-1 \leq r \leq 1$

7. Rank correlation coefficient is given by

   (a) $1+\frac{6\sum_{i=1}^{n} D_i^2}{n^3-n}$ (b) $1-\frac{6\sum_{i=1}^{n} D_i^2}{n^3-n}$ (c) $1-\frac{6\sum_{i=1}^{n} D_i^2}{n^3+n}$ (d) $1-\frac{6\sum_{i=1}^{n} D_i^3}{n\left(n^2-1\right)}$

8. If $\sum D^2 = 0$, rank correlation is

   (a) 0 (b) 1 (c) 0.5 (d) –1

9. Rank correlation was developed by

   (a) Pearson (b) Spearman (c) Yule (d) Fisher

10. Product moment coefficient of correlation is

    (a) $r = \frac{\sigma_x \ \sigma_y}{\text{cov}\left(x, y\right)}$ (b) $r = \sqrt{\sigma_x \ \sigma_y}$ (c) $r = \frac{\text{cov}\left(x, y\right)}{\sigma_x \ \sigma_y}$ (d) $r = \frac{\text{cov}\left(x, y\right)}{\sigma_{xy}}$

11. The purpose of the study of _____ is to identify the factors of influence and try to control them for better performance.

(a) mean (b) correlation (c) standard deviation (d) skewness

12. The height and weight of a group of persons will have _____ correlation.

(a) positive (b) negative

(c) zero (d) both positive and negative

13. _____ correlation studies the association of two variables with ordinal scale.

(a) A.M. Tuttle rank (b) Croxton and Cowdon rank

(c) Karl Pearson's rank (d) Spearman's rank.

14. _____ presents a graphic description of quantitative relation between two series of facts.

(a) scatter diagram (b) bar diagram (c) pareto diagram (d) pie diagram

15. _____ measures the degree of relationship between two variables.

(a) standard deviation (b) correlation coefficient

(c) moment (d) median

16. The correlation coefficient of x and y is symmetric. Hence

(a) $r_{xy} = r_{yx}$ (b) $r_{xy} > r_{yx}$ (c) $r_{xy} < r_{yx}$ (d) $r_{xy} \neq r_{yx}$

17. If cov $(x, y) = 0$ then its interpretation is

(a) x and y are positively correlated (b) x and y are negatively correlated

(c) x and y are uncorrelated (d) x and y are independent

18. Rank correlation is useful to study data in _____ scale.

(a) ratio (b) ordinal (c) nominal (d) ratio and nominal

19. If $r = 0$ then cov(x, y) is

(a) 0 (b) +1 (c) -1 (d) α

20. If cov$(x, y) = \sigma_x, \sigma_y$ then

(a) $r = 0$ (b) $r = -1$ (c) $r = +1$ (d) $r = \alpha$

II. Give very short answer to the following questions.

21. What is correlation?

22. Write the definition of correlation by A.M. Tuttle.

23. What are the different types of correlation?

24. What are the types of simple correlation?

25. What do you mean by uncorrelated?

26. What you understand by spurious correlation?

27. What is scatter diagram?

28. Define co-variance.

29. Define rank correlation.

30. If $\sum D^2 = 0$ what is your conclusion regarding Spearman's rank correlation coefficient?

31. Give an example for (i) positive correlation

(ii) negative correlation (iii) no correlation

32. What is the value of 'r' when two variables are uncorrelated?

33. When the correlation coefficient is +1, state your interpretation.

III. Give short answer to the following questions.

34. Write any three uses of correlation.

35. Define Karl Pearson's coefficient of correlation.

36. How do you interpret the coefficient of correlation which lies between 0 and +1?

37. Write down any 3 properties of correlation?

38. If rank correlation coefficient $r = 0.8$, $\sum D^2 = 3$ then find n?

39. Write any three merits of scatter diagram.

40. Given that cov$(x, y) = 18.6$, variance of $x = 20.2$, variance of $y = 23.7$. Find r.

41. Test the consistency of the following data with the symbols having their usual meaning. $N = 1000$, $(A) = 600$, $(B) = 500$, $(AB) = 50$.

IV. Give detailed answer to the following questions.

42. Explain different types of correlation.

43. Explain scatter diagram.

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

x	9	8	7	6	5	4	3	2	1
y	15	16	14	13	11	12	10	8	9

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

Wages	100	101	102	102	100	99	97	98	96	95
Cost of living	98	99	99	97	95	92	95	94	90	91

How are the wages and cost of living correlated?

46. Calculate the Karl Pearson's correlation coefficient between the marks (out of 10) in statistics and mathematics of 6 students.

Student	1	2	3	4	5	6
Statistics	7	4	6	9	3	8
Mathematics	8	5	4	8	3	6

47. In a marketing survey the prices of tea and prices of coffee in a town based on quality was found as shown below. Find the rank correlation between prices of tea and prices of coffee.

Price of tea	88	90	95	70	60	75	50
Price of coffee	120	134	150	115	110	140	100

48. Calculate the Spearman's rank correlation coefficient between price and supply from the following data.

Price	4	6	8	10	12	14	16	18
Supply	10	15	20	25	30	35	40	45

49. A random sample of 5 college students is selected and their marks in Tamil and English are found to be:

Tamil	85	60	73	40	90
English	93	75	65	50	80

Calculate Spearman's rank correlation coefficient.

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

x	53	98	95	81	75	71	59	55
y	47	25	32	37	30	40	39	45

51. Calculate the coefficient of correlation for the following data using ranks.

Mark in Tamil	29	24	25	27	30	31
Mark in English	29	19	30	33	37	36

52. From the following data calculate the rank correlation coefficient.

x	49	34	41	10	17	17	66	25	17	58
y	14	14	25	7	16	5	21	10	7	20

Yule's coefficient

53. Can vaccination be regarded as a preventive measure of Hepatitis B from the data given below. Of 1500 person in a locality, 400 were attacked by Hepatitis B. 750 has been vaccinated. Among them only 75 were attacked.

ANSWERS

I 1. (c) 2. (b) 3. (c) 4. (b) 5. (a)
6. (d) 7. (b) 8. (b) 9. (b) 10. (b)
11. (b) 12. (a) 13. (d) 14. (a) 15. (b)
16. (a) 17. (c) 18. (b) 19. (a) 20. (c)

II 30. $r = 1$

III 38. $n = 10$

40. $r = 0.85$

41. $(\alpha\beta) = -50$, The given data is inconsistent

IV 44. $r = 0.95$ it is highly positively correlated

45. $r = 0.847$ wages and cost of living are highly positively correlated.

46. $r = 0.8081$. Statistics and mathematics marks are highly positively correlated.

47. $\rho = 0.8929$ price of tea and coffee are highly positively correlated.

48. $\rho = 1$ (perfect positive correlation)

49. $\rho = 0.8$

50. $\rho = -0.905$ x and y are highly negatively

51. $\rho = -0.78$ marks in Tamil and English are negatively correlated.

52. $\rho = +0.733$

53. There is a negative association between attacked and vaccinated. There is a positive association between not attacked and not vaccinated. Hence vaccination can be regarded as a preventive measure of Hepatitis B.