e-Pathshala पाठशाला
A Gateway to All Post Graduate Courses

MHRD Govt. of India

An MHRD Project under its National Mission on Education throught ICT (NME-ICT)

Subject: Management

Production of Courseware

e-Content for Post Graduate Courses

Paper:15 , Quantitative Techniques for Management Decisions

Module: 32, Correlation: Karl Pearson's Coefficient of Correlation, Spearman Rank Correlation

Principal Investigator	Prof. S P Bansal-Vice Chancellor Maharaja Agrasen University, Baddi
Co-Principal Investigator	Prof. YoginderVerma-PVC Central University of Himachal Pradesh. Kangra, H.P.
Paper Coordinator	Prof. Pankaj Madan Dean- FMS Gurukul Kangri Vishwavidyalaya , Haridwar
Content Writer	Dr Deependra Sharma Associate Professor, Amity University Gurgaon

Items	**Description of Module**
Subject Name	Management
Paper Name	Quantitative Techniques for Management Decisions
Module Title	Correlation: Karl Pearson's Coefficient of Correlation, Spearman Rank Correlation
Module Id	32
Pre-Requisites	Basic Statistics
Objectives	After studying this paper, you should be able to - 1) Clearly define the meaning of Correlation and its characteristics. 2) Understand different types of correlation and their application in statistics. 3) Define methods of studying correlation with suitable mathematical examples. 4) Comprehend the concept of 'coefficient of determination' and will be able to interpret it.
Keywords	Linear Correlation, Non-linear Correlation, Positive Correlation, Negative Correlation.

Module-32 Correlation: Karl Pearson's Coefficient of Correlation, Spearman Rank Correlation
Topics covered 1. Meaning of correlation 2. Characteristics of Correlation 3. Types of correlation 4. Various techniques of studying correlation 5. Scatter diagram technique of correlation 6. Karl Pearson's coefficient 7. Assumptions and limitations of Karl Pearson's coefficient of correlation 8. Coefficient of determination 9. Interpretation of coefficient of determination 10. Spearman's Rank Correlation 11. Merits and demerits of Rank Correlation 12. Calculation of Spearman's Coefficient of correlation and numerical illustrations. 13. Summary.

Correlation Analysis

Learning Objectives

After reading this module , student will be able to -

1. Clearly define the meaning of Correlation and its characteristics.
2. Understand different types of correlation and their application in statistics.
3. Define methods of studying correlation with suitable mathematical examples.
4. Comprehend the concept of ‘coefficient of determination’ and will be able to interpret it.

Correlation

Correlation refers to connection, in correlation analysis we study the connection or the relation between two or more variables. If two variables vary in such a way that the change in one variable is accompanied by changes in other variable, these variables are said to be correlated. For example, relationship between the height and weight of students in a class, there exists some relation between earning of family and amount spent on luxurious items, relationship between dose of insulin and the blood sugar, etc. There may be ‘n’ number of variables which may affect each other and relationship might exists among all of them for example there are three Variables X, Y and Z, X may be related to Z and Y both and Z may relate to Y, hence we can state that the correlation is a statistic tool to find out the relationship between two or more variables.

Definition:

Tuttle defined correlation as: “An analysis of the co-variation of two or more variables is usually called correlation”. Correlation is the degree of interrelatedness of two or more Variables. It is a process to determine the amount of relationship between variables with the help of tools and techniques provided by statistics. Many authors have given different definitions of correlation in simple terms correlation is the scale of relationship between variables.

Characteristics of Correlation:

1. Although correlation analysis establishes the degree of relationship between variables but it fails to throw light on cause-effect relationship.

2. Existence of correlation between the variables may be due to chance, especially when sample taken is small in number. If we write some data points or some observation of the change in a particular variable there is a possibility of some pattern or relationship in the observation which may not be intentional likewise it is also possible between different data sets or observations of different variables which may not be true in real situation. For example if we collect data sets of number of bikes sold in NCR and amount of rainfall.

Year	Number of Bikes sold in NCR(In Lakhs)	Total Rainfall in NCR(MM)
2001	25	120

2002	35	140
2003	45	160
2004	55	180
2005	65	200
2006	75	220
2007	85	240
2008	95	260

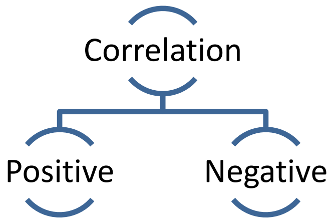
In the example given above we see a perfect positive relationship between Number of bikes sold in NCR and Total rainfall in NCR that is as the number of bikes sold is increasing the total rainfall is also increasing and the ration of change between two variables is the same, However we know that in real world there is no relation in both of the variables. To handle these situation researcher should always apply common sense.

3. It should be noted that correlated variables may be influenced by one or more variables. It is close to impossible that we are able to find out complete cause and effect relationship between two or more variables, in most of the cases one variable is affected or influenced by many other variables in some proportion, even if our data is showing complete perfect relationship between two variables there may be third hidden variable which may be affecting the relationship, for example- if a television brand is trying to understand the buying decision of consumers and they are collecting data of both the husband and wives behavior and they stablish a relationship between the two observed data sets but is also a possibility that the purchase decision of television in the family may be somewhat effected by the number of children in the family for that data has not been collected.

Types of correlation:

Scholars have defined correlation in different types. Basically they all can be defined in three types-

Type-I



In Type I correlation direction of change in the observation is important, it can be understood more once we differentiate between positive and negative correlation.

Positive Correlation:

Let us take two variables X and Y the change direction in both the variables is in such a way that if X decreases Y on an average also decreases and if X increases Y on an average also increases. This correlation is called Positive correlation. For Example:

Example 1

X	Y
10	15
12	20
11	22
18	25
20	37

Example 2

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

In both the cases we see that the change in X and Y is same direction. Examples of positive correlation can be height and weight, water consumption and temperature etc.

Negative Correlation:

Likewise positive correlation in negative correlation the direction of change is important but in opposite direction. If we take same example of X and Y variables and see that if X is increasing and Y on an average is decreasing or if X is decreasing and Y on an average is increasing, variables with this correlation are called negatively correlated, for example-

Situation 1

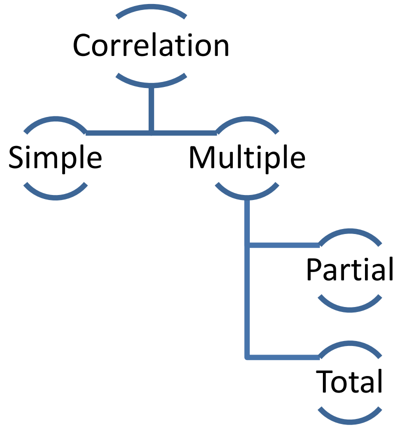
X	Y
20	40
30	30
40	22
60	15
80	16

Situation 2

X	Y
100	10
90	20
60	30
40	40
30	50

In both the case above we see that the change in X and Y is in opposite direction. Law of demand is an example of negative correlation.

Type II



In Type II correlation number of variables studied is important. Variable Numbers decide whether they are simple or multiple, to understand this let us discus one by one-

Simple Correlation:

When in a problem we only study two variables, the problem is said a simple correlation problem. For example if we study the wheat production in an area and the degree of rainfall in that same area, it will be considered as an example of simple correlation.

Multiple Correlations:

When more than two variables are studied then the problem is of multiple correlations. For example if we study the production of rice in an area, the amount of rainfall and the amount of fertilizers used in the same area, it will be treated as a problem of this correlation.

It is of two types as per the influencing nature of Variables-

a) **Partial Correlation**:

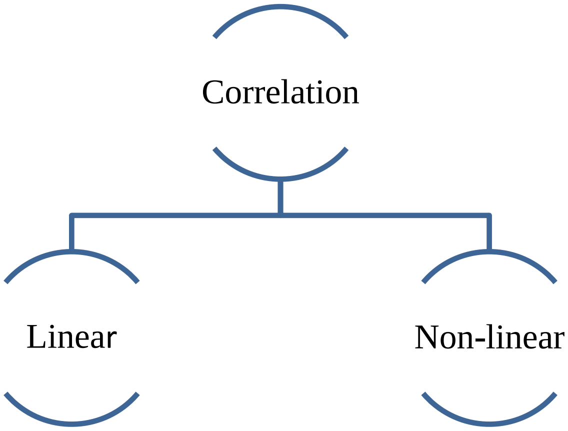
Partial Correlation arrives where in a problem we identify more than two variables but consider only two variables to be influencing each other, assuming other variables as constant.

Eg. In the problem of wheat production, rainfall and temperature if correlation analysis is conducted between wheat production and rainfall and is limited to time periods when a certain average daily temperature existed.

b) Total Correlation:

When we study all the existing variables it is a problem of total correlation. However it is normally impossible or close to impossible.

Type III



In Type III Correlation we differentiate the relationship of variable based on the ratio of change. Ratio of two variables decides whether both the variables are linearly correlated or Non-linearly.

Linear Correlation:

Linear correlation is based on the ratio of change and its consistency between the variables under study. Variables are said to be linearly correlated when the ratio is constant. In other words if the degree of change in one variable and amount of change in the other variable results in the same ratio, then the correlation is said to be linear.

If we draw a graph, of variables having linear relationship will always form a straight line.

For example: below are the examples of linear correlation-

a)X =	10	20	30	40	50	60	70	80
Y =	70	140	210	280	350	420	490	560

Y = 7X

Non-Linear correlation:

When the amount of change in one variable does not hold a constant ratio to the amount of change in the other variables, it is said to be nonlinear correlation.

Eg.- Doubling the rainfall will not result in increasing the production of rice wheat by two times. This correlation is also referred as curvilinear relationship between the variables.

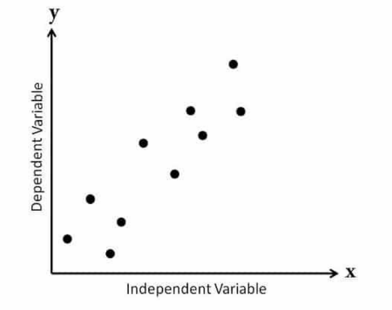
Various techniques of studying correlation:

Following are the techniques used to study correlation:

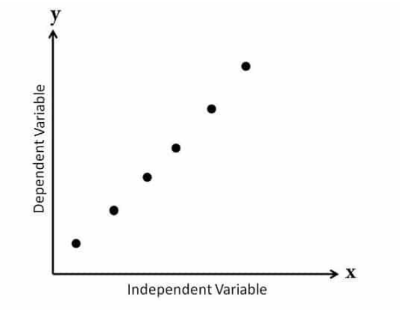
1. Scatter diagram technique of correlation
2. Karl Pearson's Coefficient of correlation
3. Spearman's Rank Correlation

1. Scatter diagram technique:

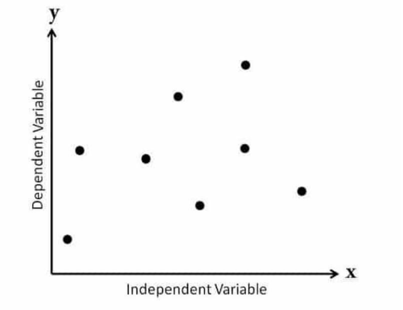
It is a easy and attractive technique of diagrammatic representation. Here, the given data are plotted on a graph sheet in the form of dots. The X variable is plotted on the horizontal axis and y variable on the vertical axis, now we can know the scatter or concentration of the various points. X and Y variables are generally referred as independent and dependent variables respectively.



(A. Low Degree correlation)



(B. High Degree Correlation)



(C. Zero correlation)

2. Karl Pearson's Coefficient of correlation:

While we study two variables a few immediate questions come in mind, first is there any relationship between the behavior of these two variables second if there is any relationship between the variables whether it is statistically significant enough third what is the type of relationship (+ve or – ve) between the variables and the fourth is that can the value of one variable be predicted based on the change in other variables? If these questions can be answered correctly we can measure and predict the behavior of two variables. Correlation analysis answers the first three questions and the last question is answered by regression analysis.

It not only tells about the strength of the relationship among the two variables but also about their direction of change.

It is widely used in practice and is popularly known as Pearson's coefficient of Correlation .It is denoted by 'r'. It is a mathematical method used for measuring the degree of linear relationship between variables. It gives a number that states the strength and the direction of the relationship between the variables.

Karl Pearson’s coefficient of correlation can be measured in two ways:

1. When deviation is taken from Actual Mean, the formula is-

$$r = r(x, y) = \Sigma xy / \sqrt{\Sigma x^2 \Sigma y^2}$$

Illustration 2.1

Find the coefficient between the sales and expenses from the data given below-

Roll No.	Marks in Subject A	Marks in Subject B
1	48	45
2	35	20
3	17	40
4	23	25
5	47	45

Answer:

Let the marks in subject A be denoted by X and that in subject B by Y.

X	(X-34) = x	x^2	Y	(Y-35) = y	y^2	xy
48	14	196	45	10	100	140
35	1	1	20	-15	125	-15
17	-17	289	40	5	25	-85
23	-11	121	25	-10	100	110
47	+13	169	45	10	100	130
$\sum X = 170$	$\sum x = 0$	$\sum x^2=776$	$\sum Y = 175$	$\sum y = 0$	$\sum y^2= 550$	$\sum xy =280$

Mean can be calculated by $\frac{\Sigma X}{5} = 34$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{280}{\sqrt{776 X 550}} = \frac{280}{653.3} = 0.429$$

Since r = 0.429 it means that there is moderate positive correlation between the both the subjects A and B.

2. When deviation is taken from Assumed mean, the formula is-

$$r = r(x,y) = (N\ \Sigma dxdy - \Sigma dx\ \Sigma dy) / (\sqrt{N\ \Sigma dx^2 - (\Sigma dx)^2}\ \sqrt{N\ \Sigma dy^2 - (\Sigma dy)^2})$$

“r” is referred as coefficient of correlation.

Illustration 2.1: The above problem (illustration 2.1) can also be solved by taking assumed mean-

Let the marks in subject A be denoted by X and that in subject B by Y and assumed mean of the marks in subject A and B is 35 and 40 respectively-

X	(X-35) = dx	dx^2	Y	(Y-40) = dy	dy^2	dxdy
48	13	169	45	5	25	65
35	0	0	20	-20	400	0
17	-18	324	40	0	0	0
23	-12	144	25	-15	225	180
47	12	144	45	5	25	60
∑X = 170	∑dx = -5	$\sum dx^2$=781	∑Y = 175	∑dy = -25	$\sum dy^2$=675	∑dxdy =305

Substituting these values if the formula –

$$r= \frac{(5X305 - (-5)(-25))}{\sqrt{(5X\ 781-25)(5X675-625)}} = \frac{(1525-125)}{\sqrt{(3905-25)(3375-625)}} = \frac{1400}{\sqrt{3880X2750}} = \frac{1400}{3266.50} = .429$$

$$r= .429$$

Coefficient of correlation:

Karl Pearson’s Coefficient of correlation has got following observational properties like-

1. The value of r is always in between -1 to +1 or $-1 \leq r \geq +1$.
2. Degree of correlation is expressed by the value of r for example is the value of r is +1 variables are high positively correlated or perfect positive correlation, if it is -1 variables are high negatively correlated or perfect negative correlation and if it is 0 there is no correlation between the variables.
3. Direction of change is indicated by sign (+ve) or (-ve).
4. It is expressed as the geometric mean of two regression coefficient $r= \sqrt{}\ bxy * byx$

Limitations of Karl Pearson’s Coefficient of correlation

There are a few limitations of Karl Pearson’s coefficient of correlation-

a) It only defines the linear relationship between the variables.
b) It is unduly affected by extreme values of two variable values.
c) Computation of Pearson’s coefficient is sometimes cumbersome.

Coefficient of determination:

Coefficient of determination is denoted by ‘r^2’. r^2 measures the proportion of variation in one variable that is explained by the other. Both r and r^2 are symmetric measures of association. In other words, the correlation of X and Y will not different as the correlation of Y and X. it means in both the variables X and Y whichever is taken as dependent variable and as independent variable it makes no difference, what matters is the relationship between both of them.

Degree of determination can be expressed as the ratio of explained variation and total variation-

$$r^2 = \text{Explained Variation} / \text{Total variation}$$

The maximum value of r^2 is 1 because it is possible to explain all of the variation in Y but it is not possible to explain more than all of it.

Interpretation of Coefficient of determination:

Coefficient of determination measures the relationship between two variables in terms of percentage. It is the proportion of explained variation in the value of response variables. For example if the value of r^2 is 0.82 it means 82 percent of the variation in response variable is explainable by explanatory variable and it does not state anything about rest of the 18 percent variation that might be because of any other factor or variable.

a) If $r^2 = 0$ states there is no relationship between two variables.
b) If $r^2 = 1$ is the perfect association between the variables.
c) If $0 \leq r^2 \leq 1$ then the degree of variation in response variable is due to variation in values of explanatory variables. Value of r^2 close to zero shows low proportion of variation in response variable due to variation in values of explanatory variable. While the value of r^2 close to 1 shows that the complete variation in response variable is due to variation in values of explanatory variable.

3. Spearman’s Rank Correlation Coefficient:

In case, it is not possible to measure the variables quantitatively but can be arranged in serial order then Karl Pearson’s coefficient of correlation is not applicable. For these situations, the coefficient of correlation had been given by a British psychologist Charles, Edward Spearman. In this method the individuals in the group are arranged in order thereby obtaining for each individual a number indicating its rank in the group. This type of correlation is often called as non-metric correlation as it is a measure of correlation for two non-metric variables that relies on ranking to compute the correlation.

Merits

- This method is quite simple & is comparatively a easier way to establish a relationship between variables.
- Spearman’s rank correlation coefficient can be calculated easily in case of non-quantitative or qualitative variables.
- Through this method ordinal ranks can be given easily to establish a relationship between two qualitative variables.

Demerits

- It is assumed that both the variables are normally distributed that may not always be true.
- It only describes the linear relationship between the variables, does not interpret anything about nonlinear relationship.
- In case of large data, giving ranks is a bit cumbersome.

- When grouped data is available this method cannot be applied.

Calculation of Spearman’s Rank Correlation:

Rank Coefficient of correlation is calculated under three conditions-

A) When Actual ranks are given the formula is-

$$R = 1- (6 \sum D^2) / N (N^2 – 1) „$$

Where R = Rank correlation coefficient „ D = Difference of rank between paired item in two series. N = Total number of observation.

Calculation Method

Where actual ranks are given computation can be done by following these steps-

- Take the difference of the two ranks (R_1-R_2) this difference is denoted by D.
- Calculate $\sum D^2$ by squaring the differences and take their total.
- Apply the formula-

$$R= 1 - \frac{6\sum D^2}{N^3-N}$$

Illustration 3.1:

The ranks of the ten students in two subjects A and B are given in the table below-

Ranks In Subject A	Ranks in Subject B
3	4
5	6
4	3
8	9
9	10
7	7
1	2
2	1
6	5
10	8

Find out correlation coefficient by using spearman’s rank correlation formula-

Answer:

Step 1: Take the difference of the two ranks and square it-

Ranks In Subject A R_1	Ranks in Subject B R_2	D^2 (R_1-R_2)
3	4	1
5	6	1
4	3	1
8	9	1
9	10	1
7	7	0
1	2	1
2	1	1
6	5	1
10	8	4

Step 2: Take a total of D^2 that is $\sum D^2 = 12$

Step 3: make the computation using spearman's Rank coefficient formula-

$$R = 1 - \frac{6\sum D^2}{N^3 - N}$$

$$R = 1 - \frac{6 \text{ x } 12}{10^3 - 10} = 1 - \frac{72}{990} = 0.927$$

B) When ranks are not given

In case ranks are not given we need to assign the ranks to the available observations. We can start giving ranks by giving first or last to highest or lowest but we must follow the same rule for all other available variables.

Illustration 3.2: Calculate the rank correlation coefficient for the following given marks of two Subjects A and B.

Marks of Subject A	Marks of Subject B
92	86
89	83
87	91
86	77
83	68
77	85
71	52
63	82
53	37
50	57

Answer: Calculation of Rank Correlation Coefficient-

Marks of Subject A	R_1	Marks of Subject B	R_2	$(R_1-R_2)^2 = D^2$
92	10	86	9	1
89	9	83	7	4
87	8	91	10	4
86	7	77	5	4
83	6	68	4	4
77	5	85	8	9
71	4	52	2	4
63	3	82	6	9
53	2	37	1	1
50	1	57	3	4
N= 10				$\sum D^2=44$

$$R= 1 - \frac{6\sum D^2}{N^3-N}$$

$$R= 1 – \frac{6 \text{ x } 44}{10^3-10} = 1- \frac{264}{990} = 0.733$$

C) When Ranks are equal

In some cases two or more data can be equal hence equal ranks are given, in those cases we assign average ranks to each of them for example if two individuals are equal at fourth place then each given the average rank $\frac{4+5}{2}$ = 4.5 similarly if three are ranked equal at fourth place, they are given an average rank of $\frac{4+5+6}{3}$ = 5.

In case of equal ranks formula for calculating Spearman’s coefficient of correlation is –

$$R = 1 - \frac{6\left\{\sum D^2+\frac{1}{12}(m1^3-m1)+\frac{1}{12}(m2^3-m2)\ldots\ldots\ldots\right\}}{N^3-N}$$

Where m is number of items whose ranks are common.

Illustration 3.3:

Marks in the two subjects X and Y of eight applicants are shown below. Calculate rank coefficient of correlation-

Applicants	Subject X	Subject Y
A	15	40
B	20	30
C	28	50
D	12	30
E	40	20
F	60	10

G	20	30
H	80	60

Answer:

Applicants	Marks in Subject X	Rank Assigned	Marks in Subject Y	Rank Assigned	$(R_1\text{-}R_2)^2$ $= D^2$
A	15	2	40	6	16
B	20	3.5	30	4	0.25
C	28	5	50	7	4
D	12	1	30	4	9
E	40	6	20	2	16
F	60	7	10	1	36
G	20	3.5	30	4	0.25
H	80	8	60	8	0
N= 8					$\sum D^2$= 81.5

$$R = 1 - \frac{6\left\{\sum D^2+\frac{1}{12}(m1^3-m1)+\frac{1}{12}(m2^3-m2)\ldots\ldots\ldots\ldots\right\}}{N^3-N}$$

In Subject X 20 is repeated two times hence m1 = 2 and in subject Y 30 is repeated three times hence m2= 3, putting these values in the formula-

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

$$= 1 – \frac{6(81.5+0.5+2)}{504}$$

$$= 1 - \frac{6\ X\ 84}{504}$$

$$= 0$$

Since R= 0 it means there is no correlation between marks obtained in two subjects.

Summary

When two variables change in such a way that at any point the value of one variable is related with the value of another variable both are called correlated variables. Correlation analysis is a statistical method of measuring the degree of relationship between two or more variables. Correlation analysis establishes relationship between the variables but it does not say which variable is a cause and which variable is an effect or in other words Causal relationship between the variables cannot be established by correlation analysis.

Correlation can be of different types-

a) Positive or negative

b) Simple or multiple
c) Linear or nonlinear

There are different methods of studying correlation, stated below-

1) Scatter Diagram Method
2) Karl Pearson’s Coefficient of correlation
3) Spearman’s Rank Correlation

In Correlation analysis a very important term is measured ‘coefficient of correlation’ and is denoted by ‘r’. The value of ‘r’ defines the strength of relationship between the variables. The value of coefficient of correlation can be between 0 and 1. As much r is closer to 1 the variable relationship is that much stronger also the sign of the numerical value of r shows positive or negative direction of the relationship. Coefficient of correlation is also used to determine the value of ‘coefficient of determination’ by just squaring it and is denoted by r^2. Coefficient of determination r^2 states the percentage of variation in one variable explained by another variable.