U N I T – 3 :C O R R E L A T I O N

After reading this material, the learners are expected to :

- Understand the meaning of correlation and regression
- Learn the different types of correlation
- Understand various measures of correlation., and,
- Explain the regression line and equation

3.1. Introduction

By now we have a clear idea about the behavior of single variables using different measures of Central tendency and dispersion. Here the data concerned with one variable is called 'univariate data' and this type of analysis is called 'univariate analysis'. But, in nature, some variables are related. For example, there exists some relationships between height of father and height of son, price of a commodity and amount demanded, the yield of a plant and manure added, cost of living and wages etc. This is a a case of 'bivariate data' and such analysis is called as 'bivariate data analysis. Correlation is one type of bivariate statistics. Correlation is the relationship between two variables in which the changes in the values of one variable are followed by changes in the values of the other variable.

3.2. Some Definitions

1. "When the relationship is of a quantitative nature, the approximate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation"—(Craxton and Cowden)
2. 'correlation is an analysis of the co-variation between two or more variables"—(A.M Tuttle)
3. "Correlation analysis attempts to determine the degree of relationship between variables"—(Ya Lun Chou)
4. "Correlation analysis deals with the association between two or more variables"—(Simpson and Kofka)

Thus, the association of any two variates is kown as correlation. Correlation is the numerical measurement showing the degree of correlation between two variables. One variable may be called “independent” and the other “dependent” variable.

3.3. Utility of Correlation

Correlation coefficient is a useful tool for many statistical purposes. Let us what these are:

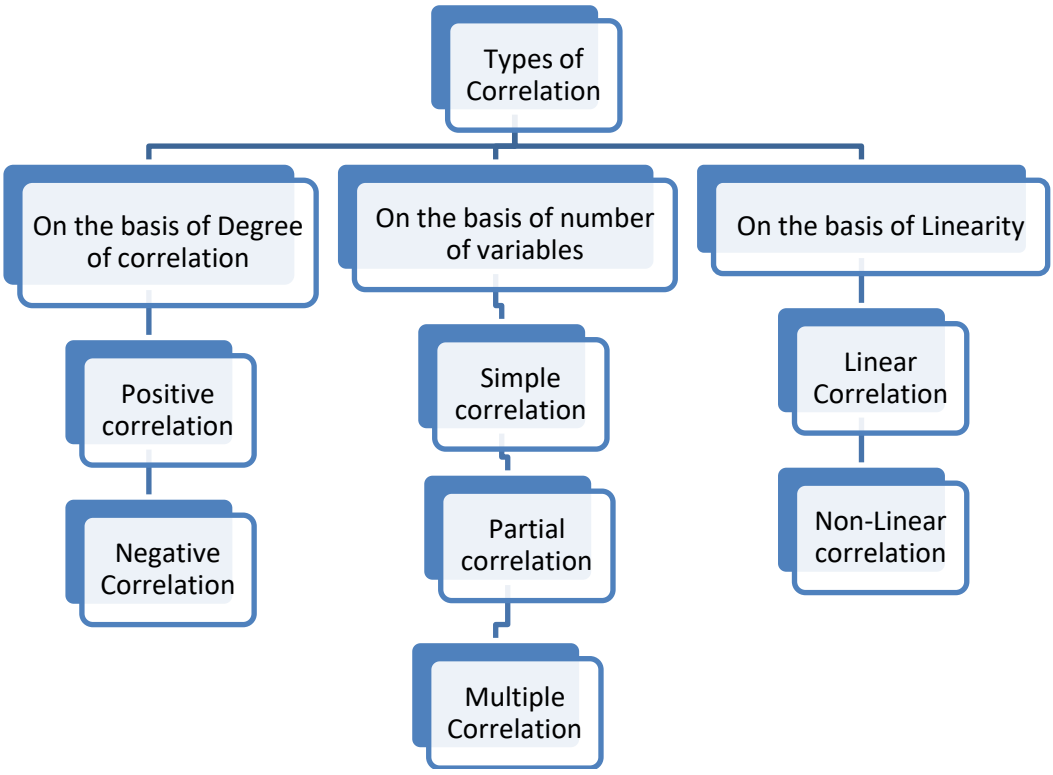
- Correlation is very useful for Economists to study relationships between variables.
- It helps in measuring the degree of relationship between the variables
- We can also test the significance of the relationship
- Sampling error can also be calculated., and
- Correlation is the basis for the study of regression

3.4. Correlation and Causation

The measure of correlation is only a measure of co-variation. It is a numerical measurement of the extent to which correlation can be found between two or more than two variables. It does not prove causation. Correlation may happen because of several reasons like: (i) due to pure chance, (ii) both the correlated variables may be influenced by one or more other variables., and , (iii) both the variables may be mutually influencing each other so that neither can be designated as the cause and the other the effect. Thus, whether the changes in variables indicate causation or not, must be decided on other evidence than on the degree of correlation.

Correlation is best used, therefore, as a suggestive and descriptive piece of analysis, rather than a technique which gives definitive answers. It is often a preparatory piece of analysis, which gives some clues to what the data might yield, to be followed more sophisticated techniques such as regression.

3.5. Types of Correlation



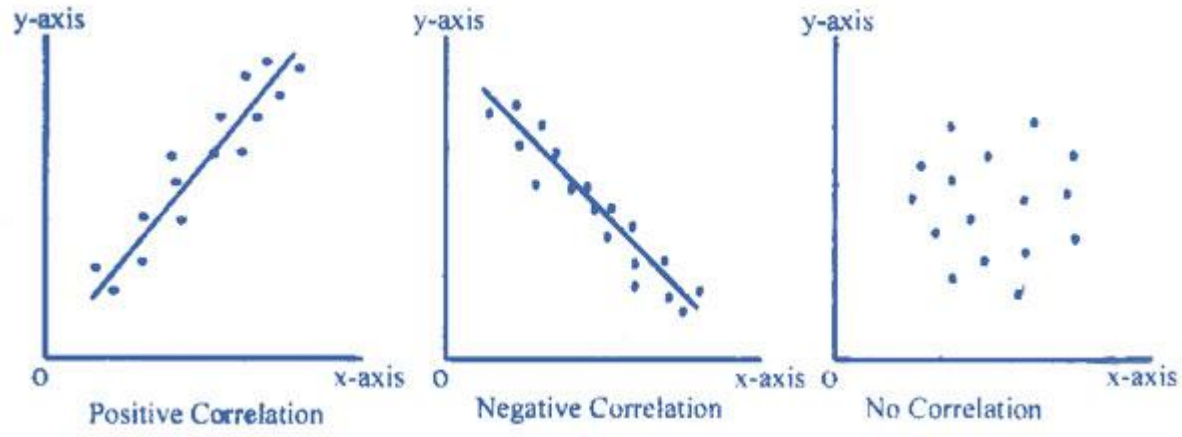
3.5.1. Positive and negative correlation

Whether correlation is positive or negative would depend up on the direction of change of the variables. Correlation is said to be positive when the values of the two variables move in the same direction so that an increase in the value of one variable is followed by an increase in the value of the other variable. Or a decrease in the value of one variable is followed by a decrease in the value of the other variable. Examples of positive correlation are:

(a) Heights and weights.., (b) amount of rainfall and yield of crop., (c) price and supply of a commodity

Correlation is said to be negative when the values of the two variables move in the opposite direction so that an increase in the values of one variable is followed by a decrease in the value of the other variable. Examples of negative correlation are:

(a) price and demand of a commodity. (b) Sales of woolen garments and temperature.. (c) Vaccinations and illness: The more that people are vaccinated for a specific illness, the less that illness occurs.

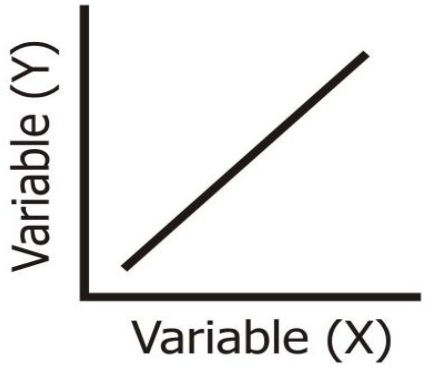


3.5.2. Linear and non-linear correlation

If the amount of change in one variable tends to bear a constant ratio to the amount of change in other variable, then the correlation is said to be linear. Observe the following two variables X and Y.

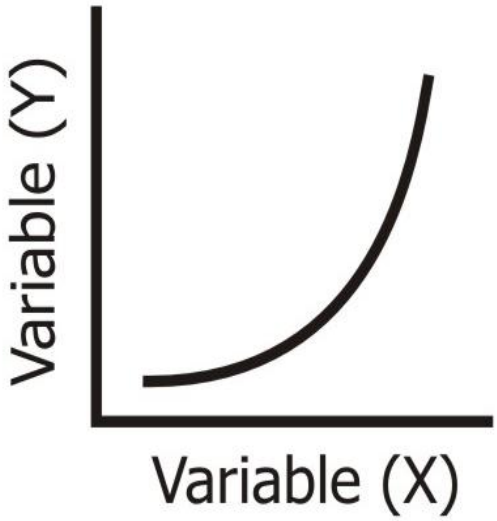
- If we plot these coordinates on a graph, we'll get a straight line.

X	:	Y
1	:	10
2	:	20
3	:	30



Correlation is non linear, if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable. If we plot these coordinates on a graph, we'll get a curve.

X	:	Y
1	:	10
2	:	15
3	:	40



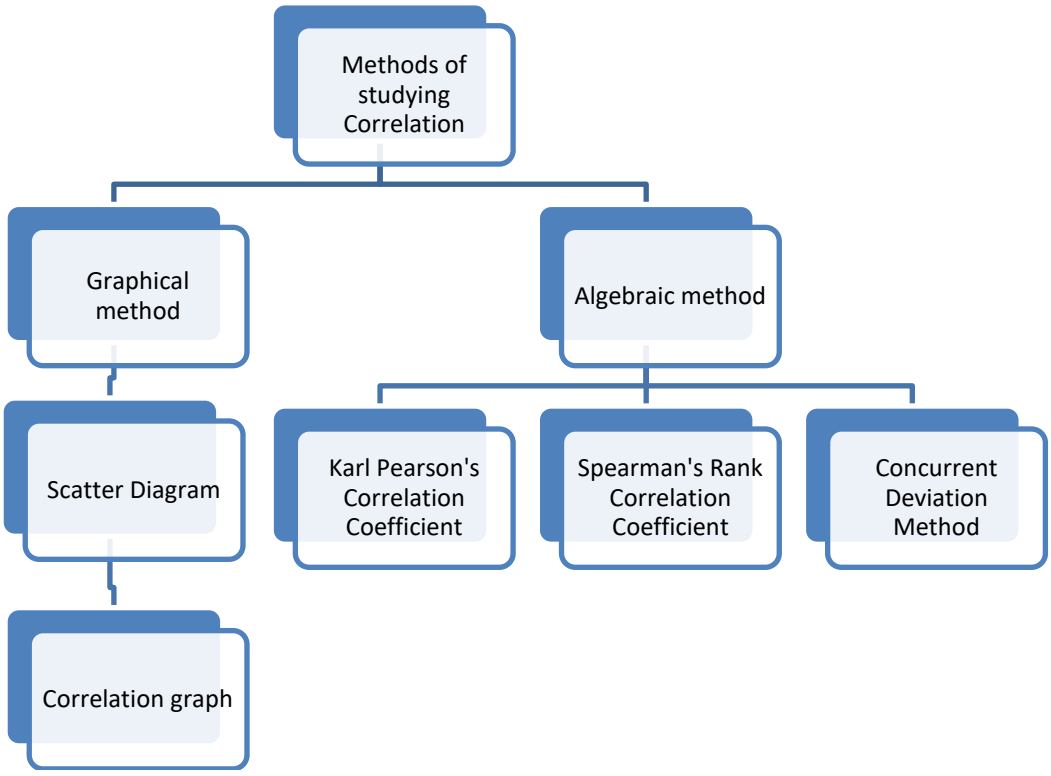
3.5.3. Simple, Partial & Multiple Correlation

Simple Correlation – When we consider only two variables and check the correlation between them it is said to be Simple Correlation. For example, radius and circumference of a circle., Price and quantity demanded

Multiple Correlation – When we consider three or more variables for correlation simultaneously, it is termed as Multiple Correlation. For example, Price of Cola Drink, Temperature, Income and Demand for Cola. Another example, when we study the relationship between the yield of rice per acre and both the amount of rainfall and the amount of fertilizers used, it is a problem of multiple correlation.

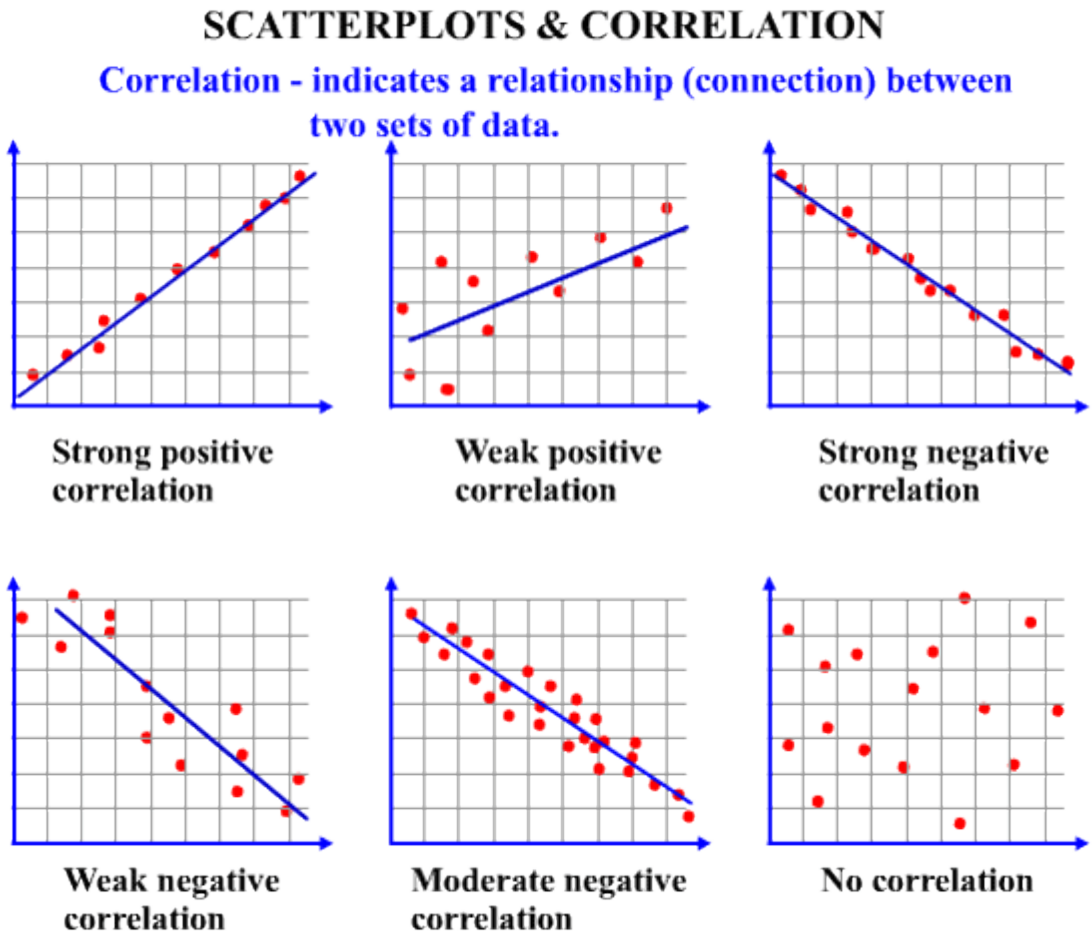
Partial Correlation – When one or more variables are kept constant and the relationship is studied between others, it is termed as Partial Correlation. For example, If we keep Price of Cola constant and check the correlation between Temperature and Demand for Cola, it is termed as Partial Correlation.

3.6.Methods of studying Correlation



3.6.1. Scatter Diagram

This is the simplest method of studying correlation between two variables. The two variables x and y are taken on the X and Y axes of a graph paper. Each pair of x and y value we mark a dot and we get as many points as the number of pairs of observation. By looking through the scatter of points, we can form an idea as whether the variables are related or not. If all the plotted points lie on a straight line rising from the lower left hand corner to the upper right hand corner, correlation is said to be perfectly positive. If all the plotted points lie on a straight line falling from the upper left hand corner to the lower right hand corner of the diagram, correlation is said to be perfectly negative. If all the plotted points fall in a narrow line and the points are rising from the lower left hand corner to the upper right hand corner of the diagram, there is degree of positive correlation between variables. If the plotted points fall in a narrow bank and the points are lying from the upper left hand corner to the right hand corner, there high degree of negative correlation. If the plotted points lie scattered all over the diagram, there is no correlation between the two variables.



Merits and limitations of Scatter diagram

Merits

It is simple and non-mathematical method of studying correlation between variables.

Making a scatter diagram usually is the first step in understanding the relationship between two variables.

Limitations

In this method we cannot measure the exact degree of correlation between the variables.

3.6.2. KARL PEARSON'S CO-EFFICIENT OF CORRELATION

The Karl Pearson's product-moment correlation coefficient (or simply, the Pearson's correlation coefficient) is a measure of the strength of a linear association between two variables and is denoted by *r* or r_{xy}(x and y being the two variables involved).This method of correlation attempts to draw a line of best fit through the data of two variables, and the value of the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit.It

is a mathematical method for measuring correlation between two variables and was suggested by Karl Pearson, a British Statistician. It is the most widely used method for measuring correlation. It is defined as:

$$r = \frac{Covariance\ (x,y)}{S.D.(x)S.D.(y)}$$

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

Interpretation of 'r'

The value of the coefficient of correlation will always lie between -1 and +1., i.e., $-1 \leq r \leq 1$. When r = +1, it means, there is perfect positive correlation between the variables. When r = -1, there is perfect negative correlation between the variables. When r = 0, there is no relationship between the two variables. The coefficient correlation describes not only the magnitude of correlation but also its direction. Thus, +0.8 indicates that correlation is positive because the sign of r is plus and the degree of correlation is high because the numerical value of r(0.8) is close to 1. If r = -0.4, it indicates that there is low degree of negative correlation because the sigh of r is negative e and the numerical value of r is less than 0.5.

Assumptions

While calculating the Pearson's Correlation Coefficient, we make the following assumptions –

- There is a linear relationship (or any linear component of the relationship) between the two variables
- We keep Outliers either to a minimum or remove them entirely

Properties of the Pearson's Correlation Coefficient

1. r lies between -1 and +1, or $-1 \leq r \leq 1$, or the numerical value of r cannot exceed one (unity)
2. The correlation coefficient is independent of the change of origin and scale.
3. Two independent variables are uncorrelated but the converse is not true.

Example 1: calculate correlation coefficient for the following data:

X	2	4	5	6	8	11
Y	18	12	10	8	7	5

Solution:

X	Y	X^2	Y^2	XY
2	18	4	324	36
4	12	16	144	48
5	10	25	100	50
6	8	36	64	48
8	7	64	49	56
11	5	121	25	55
$\sum X = 36$	$\sum Y = 60$	$\sum X^2 = 266$	$\sum Y^2 = 706$	$\sum(XY) = 293$

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

Where:

- N = number of pairs of scores
- Σxy = sum of the products of paired scores
- Σx = sum of x scores
- Σy = sum of y scores
- Σx^2 = sum of squared x scores
- Σy^2 = sum of squared y scores

Substituting the values in the above formula, we have:

$$r = \frac{6 \times 293 - 36 \times 60}{\sqrt{6 \times 266 - 36^2}\ \sqrt{6 \times 706 - 60^2}}$$

$$= \frac{1758 - 2160}{\sqrt{1590-1296}\ \sqrt{4236 - 3600}}$$

$$= \frac{-402}{17.32 \times 25.22}$$

$$= \frac{-402}{436.81}$$

$$= -0.920$$

(Note: there is high degree of negative correlation)

Example 2: Calculate correlation between X and Y

(X)	2	3	4	5	6	7	8
(Y)	4	5	6	12	9	5	4

Solution:

X	Y	X^2	Y^2	XY
2	4	4	16	8
3	5	9	25	15
4	6	16	36	24
5	12	25	144	60
6	9	36	81	54
7	5	49	25	35
8	4	64	16	32
$\sum X = 35$	$\sum Y = 45$	$\sum X^2 = 203$	$\sum Y^2 = 343$	$\sum(XY) = 228$

Correlation Coefficient Formula

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

$$r = \frac{7 \times 228 - 35 \times 45}{\sqrt{7 \times 203 - 35^2}\ \sqrt{7 \times 343 - 45^2}}$$

$$= \frac{1596 - 1575}{\sqrt{1421-1225}\ \sqrt{2401 - 2025}}$$

$$= \frac{21}{14 \times 19.39}$$

$$= \frac{-402}{436.81}$$

$$= 0.077$$

Probable error of coefficient of Correlation 'r'

Probable error of the coefficient of correlation is a statistical measure which measures reliability and dependability of the value of coefficient of correlation. If probable error is added to or subtracted from the coefficient of correlation it would give two such limits within which we can reasonably expect the value of coefficient of correlation to vary. Usually, the coefficient of correlation is calculated from samples. For different samples drawn from the same population, the coefficient of correlation may vary. But the numerical value of such variations is expected to be less than the probable error. The formula for calculating probable error is:

$$\text{Probable error of 'r'} = \frac{0.6745(1 - r^2)}{\sqrt{n}}$$

Where, 0.6745 is a constant number., 'r' stands for correlation coefficient and 'n' number of pairs of observation.

The limits for the population correlation coefficient are:

$r = \pm P.E(r)$

Example: if r = 0.6 and n = 64, find probable error and standard error.

$$\text{Probable error of 'r'} = \frac{0.6745(1 - r^2)}{\sqrt{n}}$$

$$= \frac{0.6745\ (1 - 0.36)}{\sqrt{64}}$$

$$= \frac{0.6745 \times 0.64}{8} = 0.054$$

$$\text{Standard error} = \frac{(1 - r^2)}{\sqrt{n}}$$

$$= \frac{(1 - 0.36)}{\sqrt{64}}$$

$$= 0.08$$

Interpretation of coefficient of correlation on the basis of probable error.

- If the coefficient of correlation is less than its probable error, it is not at all significant
- If the coefficient of correlation is more than six times its probable error, it is significant
- If the probable error is not much and if the coefficient of correlation is 0.5 or more it is generally considered to be significant.

Coefficient of Determination

The nature and extent of relationship between two variables are indicated by the coefficient of correlation. An effective way of interpreting 'r' is by way of coefficient of determination. The coefficient of determination is defined as the ratio of the explained variance to the total variance and is denoted by r^2 .

$$\text{Coefficient of Determination} = r^2 = \frac{\text{Explained Variance}}{\text{Total variance}}$$

r^2 states what percentage of variations in the dependent variable is explained by the independent variable.

Coefficient of non-determination is the ratio of the unexplained variation to the total variation. It is denoted by K^2.

$$K^2 = \frac{\text{unexplained variance}}{\text{Total variance}}$$

$$= 1 - \frac{\text{Explained Variance}}{\text{Total variance}}$$

$$= 1 - r^2$$

Another concept related in the context is known as coefficient of alienation and it is defined as: $\sqrt{1 - r^2}$

Check you progress 1:

1. Given the following pairs of value of the variables X and Y: (a) make a scatter diagram., (b) Do you think that there is any correlation between the variables x and y? (c) Is it positive or negative? , (d) is it high or low? and, (e) by graphic inspection draw an estimated line.

X	2	3	5	6	8	9
Y	6	5	7	8	12	11

2. The lengths and weights of a sample of six articles manufactured by a factor are given here. Find the Pearson's correlation coefficient. (Ans: r = 0.97)

Length (X)	3	5	6	7	10	11
Weight (Y)	8	12	11	14	16	17

3. Find Karl Pearson's coefficient of correlation between the values of X and Y given here under; Also find probable error . (r=-0.86., PE =.062)

X	46	68	72	75	80	70	93	100
Y	64	50	39	48	12	52	46	30

3.6.3. SPEARMAN'S RANK CORRELATION COEFFICIENT

In 1904, C. Spearman introduced a new method of measuring the correlation between two variables. Instead of taking the values of the variables he considered the ranks (or order) of the observations and calculated Pearson's coefficient of correlation for the ranks. The correlation coefficient so obtained is called rank correlation coefficient. This measure is useful in dealing with qualitative characteristics such as intelligence, beauty, morality, honesty etc. The formula for spearman's rank correlation coefficient is:

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

Where, r_s =spearman rank correlation coefficient
D = differences in ranks between paired items ($R_1 - R_2$)
N = number of pairs of observations

Two types of situations may happen here. One is where we are given ranks and the other is where we are not given any ranks.

(a) When ranks are given:

When actual ranks are given, we can follow the steps as: (i) compute the difference between two ranks (R_1 and R_2) and denote it as 'd'., (ii) square the 'd' and obtain $\sum d^2$, and (iii) substitute the values in the formula.

Example: From the following data, calculate Spearman's rank correlation

Rank in Economics	1	2	3	4	5	6	7	8	9	10
Rank in Statistics	4	8	2	3	5	7	6	9	10	1

Solution:

R_1	R_2	d	d^2	**Steps for solution**
1	4	-3	9	**$r = 1 - \frac{6\sum d^2}{n(n^2-1)}$**
2	8	-6	36	**$= 1 - \frac{6(132)}{10(100-1)}$**
3	2	1	1	**$= 1 - 0.8$**
4	3	1	1	**$= 0.2$**
5	5	0	0	
6	7	-1	1	
7	6	1	1	
8	9	-1	1	
9	10	-1	1	
10	1	9	81	**Interpretation: The result indicates that there is low positive correlation**
Total	--		$\sum d^2 = 132$	

(b)When ranks are not given:

In case we are given actual data, we must give them rank. We can assign ranks by taking the largest value as one or the lowest value as one, next to it give as two and the like.

Example : Find the rank correlation coefficient from the following data:

X	17	13	15	16	6	11	14	9	7	12
Y	36	46	35	24	12	18	27	22	2	8

Solution:

X	Y	Rank X (R_1)	RankY (R_2)	d ($R_1 - R_2$)	d^2	**Solving steps**
17	36	1	2	-1	1	**$r = 1 - \frac{6\sum d^2}{n(n^2-1)}$**
13	46	5	1	4	16	**$= 1 - \frac{6(44)}{10(100-1)}$**
15	35	3	3	0	0	**$= 1 - 0.267$**
16	24	2	5	-3	9	**$= 0.733$**
6	12	10	8	2	4	
11	18	7	7	0	0	
14	27	4	4	0	0	
9	22	8	6	2	4	
7	2	9	10	-1	1	
12	8	6	9	-3	9	**Note: correlation is highly positive**
					44	

Calculation of Spearman's Rank correlation when equal or repeated ranks occur

While assigning rank, if two or more items have equal values (i.e., if there occur a tie), they may be given mid rank. Thus, if two items are on the fifth rank, each may ranked as 5 + 6 /2 = 5.5 and the next item in the order of size would be ranked seventh. When two or more ranks are equal, the following formula is used for computing rank correlation.

$$\text{Rank correlation coefficient} = 1 - 6\left(\sum \mathbf{d^2} + \frac{\mathbf{m3 - m)}}{12}\right) \Big/ \mathbf{n(n^2-1)}$$

Where, m stands for the number of equal ranks. The term, $\frac{\mathbf{(m3 - m)}}{12}$ is to be added in the numerator for each group of equal rank both in x and y series.

Example: Calculate 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:

X	Y	Rank X (R_1)	RankY (R_2)	d	$\mathbf{d^2}$
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
					72

In the above table, one can see that:

75 occurs 2 times, i.e., m = 2 $\quad m^3 - m = 2^3 - 2 = 6$

64 occurs 3 times, i.e., m = 3 $\quad m^3 - m = 3^3 - 3 = 24$

75 occurs 2 times, i.e., m = 2 $\quad m^3 - m = 2^3 - 2 = \underline{6}$

Total $m^3 - m = \sum(m^3 - m) = 36$

Rank correlation coefficient $= 1 - 6 \left(\sum \mathbf{d^2} + \frac{\mathbf{m3 - m)}}{12}\right)$

$$\mathbf{n(n^2-1)}$$

$$= 1 - 6 \left(72 + \frac{\mathbf{36)}}{12}\right)$$

$$\mathbf{10(100\text{-}1)}$$

$$= 1 - \frac{6 \times 75}{990}$$

$$= 0.545$$

Note: How ranks are assigned when there is repetition? In such cases, rank is the average of ranks due for all of them if they are different. For example, in the above problem, highest value is the X series is 80. So, it is given rank 1. Next two values are the same 75. They are given the average of 2 and 3., i.e., 2.5. Then 4th rank is given to the next highest value 68. Then 64 occur 3 times. So each 64 is given the average of ranks 5,6,7., i.e, 6 and so on.

Merits and demerits of Rank Correlation

Merits

1. It is easy to compute and understand
2. It is highly useful when the data are of a qualitative nature like intelligence, beauty etc.
3. When the ranks of different item-values are given, this is the only method for finding the degree of correlation.

Demerits

1. This method cannot be employed for finding out correlation in a grouped frequency distribution.
2. It is difficult to calculate rank correlation, if we have more than 30 items of observation as ranking them requires much labour.
3. Compared to Pearson method, rank correlation is not precise.

Check your progress 2:

1. What is rank correlation coefficient? Find the rank correlation coefficient for the following data:

X	35	36	40	38	37	39	41	40	36	38
Y	65	72	78	77	76	77	80	79	76	75

2. Two judges in a beauty competition rank the 12 entries as follows. What degree of agreement is there between the judges?

Judge 1	1	2	3	4	5	6	7	8	9	10	11	12
Judge 2	12	9	6	10	3	5	4	7	8	2	11	1

3. Below are given the heights of fathers (X), and those of their sons (Y) in centimeters. Calculate Spearman's rank Correlation coefficient.

X	180	155	170	174	160	172	166	170	170
Y	170	165	180	180	164	169	172	170	174

3.6.4. CONCURRENT DEVIATION METHOD

The calculation of correlation coefficient by this method is based on the direction of change or variation in the two paired variables. This is denoted by r_c and varies between +1 and -1. It is calculated by the following formula.

- r_c=Coefficient of Concurrent deviation
- C= no of positive signs after multiplying the change direction of change of X-series and Y-Series
- n=no. of pairs of observations computed

$$r_c = \pm\sqrt{\pm\frac{2c-n}{n}}$$

54

Statistics: Correlation

Steps in calculating correlation by concurrent deviation method

(i) find out the direction of change of X variable, i.e., as compared with the first value, whether the second value is increasing or decreasing or is constant. If it is increasing put (+) sign; if it is decreasing put (-) sign (minus) and if it is constant put zero. Similarly, as compared to second value find out whether the third value is increasing, decreasing or constant. Repeat the same process for other values. Denote this column by D_x.
(ii) In the same manner as discussed above find out the direction of change of Y variable and denote this column by D_y.
(iii) Multiply D_x with D_y, and determine the value of c, i.e., the number of positive signs.
(iv) Apply the above formula, i.e., $r_c = \pm\sqrt{\pm}\ (2C-n)/n$

Note. The significance of $\pm$ signs, both (inside the under root and outside the under root) is that we cannot take the under root of minus sign. Therefore, if 2C – n/n is negative, this negative value of multiplied with the minus sign inside would make it positive and we can take the under root. But the ultimate result would be negative. If $\frac{2C-n}{n}$ is positive, then, of course, we get a positive value of the coefficient of correlation.

Example: the following are the marks obtained a group of 10 students in Economics and Statistics. Calculate correlation by the method of Concurrent Deviation.

ECO	8	36	98	25	75	82	90	62	65	39
STAT	84	51	91	60	68	62	86	58	53	47

Solution:

Marks in Economics (X)	Marks in Statistics (Y)	D_x	D_y	$D_x . D_y$
8	84			
36	51	+	-	-
98	91	+	+	+
25	60	-	-	+
75	68	+	+	+
82	62	+	-	-
90	86	+	+	+
62	58	-	-	+
65	53	+	-	-
39	47	-	-	+
		N=9	N =9	C=6

$$\mathbf{r_c = \pm\sqrt{\pm}\,(2C-n)/n \quad = r_c = \pm\sqrt{\pm}\,(2x6-9)/9 \quad = \pm\sqrt{12-9/9} \quad = \pm\sqrt{3/9} \quad = 0.58}$$

Merits and Demerits of Concurrent Deviation Method

Merits:

1. It is simple to compute
2. It is easy to understand

Demerits:

1. It is not useful if long-term changes are to be considered.
2. The method does not differentiate between small and big variations.
3. It indicates the direction of change only

Check your Progress 3:

1. Calculate the coefficient of concurrent deviation from the following data.(Hints: r_c =-0.5)

X	15	18	23	20	21	28	30	29	50
Y	5	1	15	9	25	20	18	29	24

2. What is the meaning of coefficient of determination?----------
---

---

3. Explain the terms a) Probable error, and b)standard error----------
---

