Correlation

Correlation– Types, Degree, Correlation and Causation

Objective

After going through this lesson you shall be able to understand the following concepts.

- Meaning of Correlation
- Types of Correlation
  - Positive and Negative Correlation
  - Linear and Curvilinear Correlation
  - Simple and Multiple Correlation
- Degree of Correlation
  - Perfect Correlation
  - Zero Correlation
  - Limited Degree of Correlation
- Correlation and Causation

Introduction

Observe carefully around yourself. You will notice that there are many such pairs of variables where one variable is related to the other. Take for example, the amount of rainfall and crop yield. The crop yield is directly related to the amount of rainfall. Similarly, consider the temperature and the demand for ice-creams.

The two are related in the sense that as temperature rises, the demand for ice creams also rises. One can find the existence of a relationship between many such variables like the number of hours put to study and the marks obtained in the examination; the price of a commodity and its supply; number of vehicles and pollution level and so on.

The relationship between two variables is studied with the help of a statistical tool i.e. 'Correlation'. Correlation is a statistical tool that measures the quantitative relationship between different variables. It studies the degree and intensity of the relationship between the two variables.

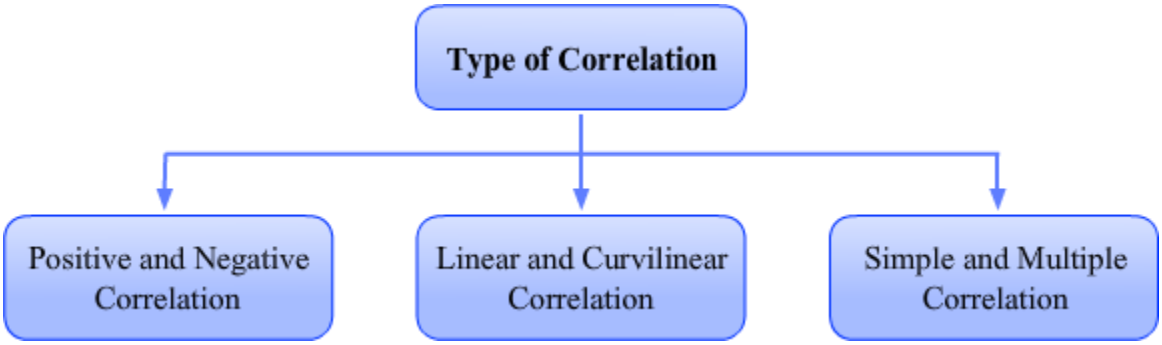
Types of Correlation

Based on the nature of relationship between the two variable, correlation can be broadly classified into the following three categories:

(i) Positive and Negative Correlation

(ii) Linear and Curvilinear Correlation

(iii) Simple and Multiple Correlation



Positive and Negative Correlation

Positive correlation: A positive correlation between two variables exists when both of them ***move in the same direction***. In other words, if with the increase in one variable, the other also increases and with the decrease in one variable the other also decreases then, the two variables are said to be positively correlated. For example, in summers as the temperature rises, the demand for soft drinks rises. Thus, the demand for soft drinks and temperatures are positively correlated. Consider the following two variables *X* and *Y.*

X	10	20	30	40	50	60
Y	15	22	29	36	43	50

In the above example, as the value of *X* increases from 10 to 20 to 30 and so on, the value of *Y* also increases from 15 to 22 to 29 and so on. This suggests a positive correlation between *X* and *Y.* In economics, we can cite various examples, where the two variables are positively correlated. For example, there exists a positive correlation between price of a commodity and its supply. As the price of a commodity increases, its supply also increases and as the price decreases, its supply also decreases. Similarly, income of a consumer and expenditure are positively correlated. As the income increases, the expenditure also increases and vice-versa.

Negative correlation: Two variables are said to be negatively correlated, if the two variables move in the ***opposite direction***. In other words, when one variable increases and the other variable falls, the two variables are said to be negatively correlated. For example, in winters as the temperature falls, the demand for room heaters increases. Thus, the demand for room heaters and temperature is negatively correlated. Consider the following two variables *A* and *B.*

A	10	20	30	40	50	60
B	90	80	70	60	50	40

In the above example, as the value of *A* increases, the value of *B* decreases. One of the common example for negative correlation in economics is the price of a commodity and its quantity demanded. As the price of a commodity rises, its demand falls and vice-versa. Similarly, there exists a negative correlation between the interest rate and the demand for loans. As the interest rate rises, the demand for loans falls and vice-versa.

Linear and Curvilinear Correlation

Linear correlation: If the ratio for change between the two variables is constant or fixed, then the two variable are said to be linearly correlated. For example, consider the following two variables *P* and *Q.*

P	50	55	60	65	70	75
Q	40	42	44	46	48	50

For every increase in the variable *P* by 5 units, the value of variable *Q* increases by 2 units. In this case, there exists a ***linear positive correlation*** between the two variables. ***A linear positive correlation between two variables is depicted by a positively sloped straight line graph***.

On the other hand, consider the variables *A* and *D*.

A	70	60	50	40	30	20
D	55	50	45	40	35	30

Here, for every decrease in the variable *A* by 10 units, the variable *D* decreases by 5 units. In other words, there exists a linear negative correlation between the two variables. ***A linear negative correlation between two variables is depicted by a negatively sloped straight line graph***.

Curvilinear correlation (Non-linear correlation): As against, linear correlation, if the ratio of change between the two variables is not constant, then the two variables are

said to be curvilinearly correlated. For example, consider the following two variables, *K* and *R.*

K	50	55	60	65	70	75
R	10	12	17	30	35	45

Here, although both the variables increase, the ratio of increase is not the constant. In other words, the value of variable *R* does not change in a constant ratio with a change in the value of *K*. Thus, there exists a curvilinear correlation between the two.

Similarly, consider the following two variables *A* and *B.*

A	50	45	40	35	30	25
B	20	18	14	9	7	2

Here as the value of *A* decreases, the value of *B* also decreases. However, the ratio of decrease in the value of *B* is not constant. Thus, here exists a curvilinear relationship between the two variables.

Such type of correlation is ***depicted by quadratic graph, parabola, hyperbola etc.***

Simple and Multiple Correlation

Simple correlation: The study of relationship between only two variables is known as simple correlation. For example, relationship between price and demand.

Multiple correlation: The study of relationship among three or more than three variables simultaneously is called multiple correlation. For example, study of the relationship between price, demand, tastes and income of the consumers

Degrees of Correlation

The degree or the extent of correlation between two variables is described by the value of the correlation coefficients. (*The calculation of the correlation coefficient will be discussed in the next lesson*)

The degrees of correlation are:

(i) Perfect Correlation

(ii) Zero Correlation

(iii) Limited Degree of Correlation

Perfect correlation: Perfect correlation exists when two variables change in exactly equal proportion. It can be further classified into two categories.

(a) Perfect Positive Correlation

(b) Perfect Negative Correlation

*Perfect positive correlation***:** When the proportional change in two variables is in the same direction, then it is called perfect positive correlation. For a perfect positive correlation between two variables the value of correlation coefficient is equal to **+**1.

Perfect negative correlation: There exists a perfect negative correlation between two variables if the proportional change in the two variables is in the opposite direction. The value of correlation coefficient for a perfect negative correlation is **–** 1.

Zero correlation: If there is no relation between two variables, i.e. change in one variable has no effect on the change in the other, then the variable lacks correlation. The value of correlation coefficient for zero correlation equals zero.

Note: A zero correlation between any two variables does not mean that there is no relationship at all between them. In fact, it should be interpreted that the two variables are not linearly related. However, it may be possible that the two variables may be non-linearly related with each other.

Limited degree of correlation: Between the extremes of perfect correlation and zero correlation, there exists limited degree of correlation. This implies that although the two variables are related, an increase (or decrease) in one variable is not accompanied by an equal proportionate increase (or decrease) in the other variable. In this case, the value of correlation coefficient lies between zero and one.

The degrees of correlation are summarised in the following table.

Degrees of Correlation	**Positive**	**Negative**

Perfect Correlation	+1	–1
Very high correlation	Between + 0.75 and + 1	Between – 0.75 and – 1
Moderate degree	Between + 0.25 and + 0.75	Between – 0.25 and – 0.75
Low degree	Between 0 and + 0.25	Between 0 and – 0.25
Zero	0	0

Correlation and Causation

Correlation between any two variables simply suggests that there exists a relationship between the two variables. However, it does not imply that one variable causes the other. In other words, a relation between cause and effect is not a prerequisite for the correlation.

Correlation only measures the degree and intensity of the relationship between the two variables, but surely not the cause and effect relationship between them. It might happen that any two variables are just coincidentally related. For example, if correlation is found between yield of wheat and yield of cotton, then this might be purely by chance. In other words, existence of correlation does not always imply causation.

Methods of Measuring Correlation Scattered Diagram and Karl Pearson's Coefficient of Correlation

Objective

In this lesson, we will try to understand the following methods of measuring correlation.

- Scattered Diagram
- Karl Pearson's Coefficient of Correlation
  - Actual Mean Method
  - Direct Method

- Short-Cut Method/Assumed Mean Method
- Step-Deviation Method

Introduction

In the previous lesson, we studied about the meaning and the various types of correlation. In this lesson, we will understand how correlation can be measured. We will study three methods of measuring correlation namely, Scattered diagram, Karl Pearson's coefficient of correlation and Spearman's rank correlation coefficient. (Spearman's rank correlation coefficient will be dealt in the next lesson.)

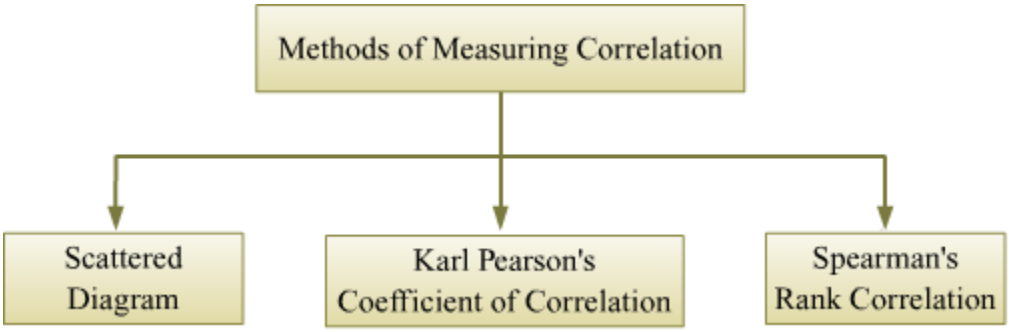
Methods of Correlation

The following are the three main methods of measuring correlation.

(i) Scattered Diagram

(ii) Karl Pearson's Coefficient of Correlation

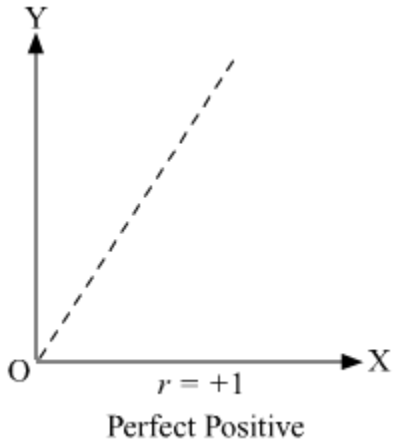
(iii) Spearman's Rank Correlation Coefficient



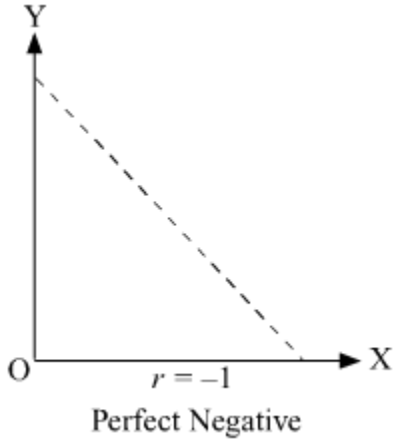
Scattered Diagram

Scattered diagram is a graphical measure of estimating the direction, magnitude and degree of correlation existing between the two variables. Under this method, different values of the two variables are plotted on a graph. The set of points obtained is called as the scatter diagram. An analysis of the scatter diagram can give a fair idea of the degree of correlation between the two variables. The following graphs explain the various degrees of correlation.

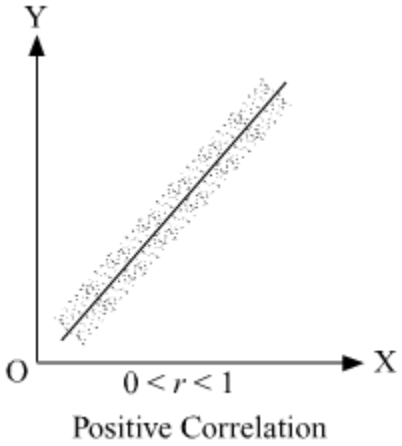
(i) *Perfect positive correlation*: When the two variables are perfectly positively correlated (r = +1), the points obtained on the scatter diagram lie on a positive sloped straight line.



(ii) *Perfect negative correlation*: In case of perfect negative correlation between two variables, the points obtained on a scatter diagram will lie on a downward sloping straight line from the upper left hand corner to the lower right corner.

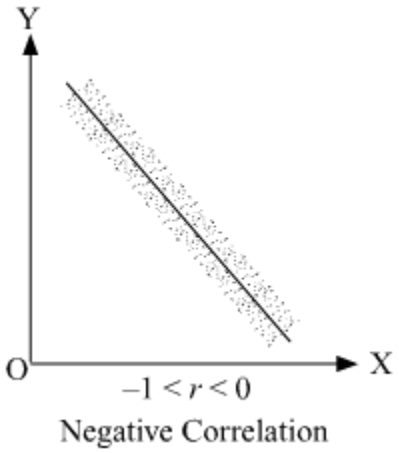


(iii) *High degree of positive correlation*: In case the two variables share a high degree of positive correlation, the points obtained on the scatter diagram lie very close to each other and reflect an upward trend.



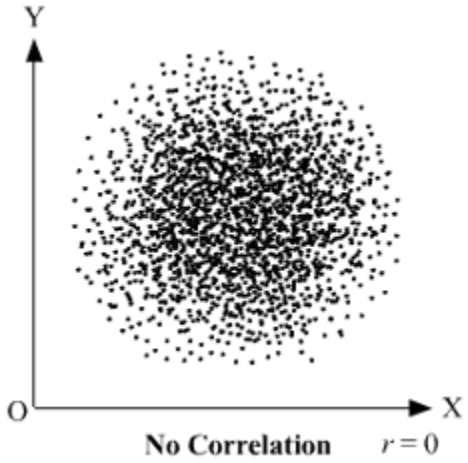
Positive Correlation

(iv) ***High degree of negative correlation***: If there exists high degree of negative correlation between two variables, the points on the scatter diagram will reflect a downward trend.



Negative Correlation

(v) ***No correlation***: If there does not exist any correlation between two variables, no trend can be deduced from the points on the scatter diagram.



The following points reflect some of the merits and demerits of a scatter diagram.

Merits of Scatter Diagram

i. A scatter diagram is very ***easy to draw and understand***.

ii. Construction of a scatter diagram ***does not involve any tedious and difficult calculation*** process like other methods.

iii. It is ***not affected by the presence of the extreme values*** in the series.

iv. A scatter diagram ***reveals the type of correlation between two variables merely at a glance***.

Demerits of Scatter Diagram

i. A scatter diagram ***presents only a rough estimation of correlation*** between the variables. It does not help in ascertaining the exact degree of correlation.

ii. It helps us in knowing only the type of the correlation, i.e. positive or negative. However, it ***fails to reveal anything about the magnitude and degree of the correlation***.

iii. This method of correlation ***fails in case of ascertaining correlation between more than two variables***.

iv. Scatter diagram ***fails to reveal the direction of causation***. In other words, whether, *X* causes *Y* or *Y* causes *X* remains unknown.

Karl Pearson's Coefficient of Correlation

Karl Pearson's coefficient of correlation was given by a British statistician Karl Pearson. As against scatter diagram, Karl Pearson's coefficient of correlation is a mathematical measure of correlation. It is denoted by '*r*' and calculated using the following formula.

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y}$$

where,

$r =$ Coefficient of Correlation

$x = X - \bar{X}$

$y = Y - \bar{Y}$

$\sigma_x =$ Standard deviation of X series

$\sigma_y =$ Standard deviation of Y series

$N =$ Number of Observations

The above formula for the Karl Pearson's coefficient of correlation can be applied using the following four methods.

i. Actual Mean Method

ii. Direct Method

iii. Assumed Mean Method/Short-Cut Method

iv. Step-Deviation Method

(i) *Actual mean method*

The following steps are involved in the calculation of Karl Pearson's coefficient of correlation using the actual mean method.

***Step* 1**: Calculate the arithmetic mean of *x*-series ($\bar{X}$) and arithmetic of *y*-series ($\bar{Y}$).

***Step* 2**: From ($\bar{X}$) obtain the deviation of each item of the *x*-series and denote it as *x.* Similarly, from ($\bar{Y}$) obtain the deviations for *y*-series and denote it by *y.*

Step* 3**: Square the deviations of *x* and *y* series as obtained in ***step 2 and obtain their sum as Σx^2 and Σy^2 respectively.

Step* 4**: Multiply the deviations obtained in ***step 2 and obtain their sum as *Σxy*.

***Step* 5**: The following formula is applied to calculate the Karl Pearson's coefficient of correlation.

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}}$$

where,

$r =$ Coefficient of Correlation

$x = X - \bar{X}$

$y = Y - \bar{Y}$

Example: For the following data calculate Karl Pearson's correlation coefficient.

X	5	7	10	15	19	
Y	2	4	7	12	15	

Solution

X	***Y***	***x*** $(x-\bar{x})$	x^2	***y*** $(y-\bar{y})$	y^2	***xy***
5	2	– 10	100	– 10	100	100
7	4	– 8	64	– 8	64	64
10	7	– 5	25	– 5	25	25
15	12	0	0	0	0	0
19	15	4	16	3	9	12
22	19	7	49	7	49	49
27	25	12	144	13	169	156
©*X* = 105	©*Y* = 84		©x^2 = 398		©y^2 = 416	©*xy* = 406

$$\overline{X}=\frac{\sum X}{N}=\frac{105}{7}=15$$

$$\overline{Y}=\frac{\sum Y}{N}=\frac{84}{7}=12$$

Now, Karl Pearson's Correlation Coefficient (r)

$$r=\frac{\sum xy}{\sqrt{\sum x^2\times\sum y^2}}$$

Substituting the values in the formula we get,

$$r=\frac{406}{\sqrt{398\times416}}$$

$$\text{or, } r=\frac{406}{406.9}=0.99$$

Thus, the value of Karl Pearson's correlation coefficient is 0.99. This suggests that there exists a high degree of positive correlation between values of *x* and values of *y*.

Example: The following data presents the marks obtained by 10 students in English and Science. Calculate the value of correlation coefficient.

Marks in English	15	17	22	26	30	32	35	39	42	52
Marks in Science	10	12	16	22	27	35	40	48	52	58

Solution

Marks in English *X*	**Marks in Science *Y***	***x*** $(x-\overline{x})$	***y*** $(y-\overline{y})$	x^2	y^2	***xy***
15	10	– 16	– 22	256	484	352
17	12	– 14	– 20	196	400	280
22	16	– 9	– 16	81	256	144
26	22	– 5	– 10	25	100	50
30	27	– 1	– 5	1	25	5
32	35	1	3	1	9	3
35	40	4	8	16	64	32
39	48	8	16	64	256	128
42	52	11	20	121	400	220
52	58	21	26	441	676	546

©X = 310	©Y = 320			©x^2= 1,202	©y^2= 2,670	©xy = 1,760

$$\overline{X}=\frac{\sum X}{N}=\frac{310}{10}=31$$

$$\overline{Y}=\frac{\sum Y}{N}=\frac{320}{10}=32$$

Now,

$$\text{Correlation Coefficient}, r=\frac{\sum xy}{\sqrt{\sum x^2\times\sum y^2}}$$

Substituting the values in the formula.

$$r=\frac{1,760}{\sqrt{1,202\times2,670}}=\frac{1,760}{1,791.46}=0.98$$

Thus, correlation coefficient is 0.98. This suggests that there exists a high degree of positive correlation between the marks obtained in English and the marks obtained in Science.

Example: From the following data calculates Karl Pearson's correlation coefficient.

	X **Series**	***Y*** **Series**
Number of observations	10	10
Arithmetic mean	25	27
Standard deviation	2	4
Sum of product of deviation of *X* series and *Y* series is 25		

Solution

We know,

$$r=\frac{\Sigma xy}{N\sigma_x\sigma_y}$$

Given,

Sum of product of deviation of *X* series and *Y* series, (Σxy) = 25

Number of observations, N = 10

Standard deviation of *X* series, (σ_x) = 2

Standard deviation of *Y* series, (σ_y) = 4

Substituting the given values in the formula we get,

$$r = \frac{25}{10 \times 2 \times 4} = \frac{25}{80} = 0.312$$

Hence, the value of correlation coefficient is 0.312.

Example: For the given data, calculate Karl Pearson's correlation coefficient.

	X **Series**	***Y*** **Series**
Number of items	10	10
Variance	12	15
Covariance of *x* and *y* series is 10.2		

Solution

We know:

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y}$$

$$\text{Covariance} = \frac{\Sigma xy}{N} = 10.2$$

Variance of *X* = 12

4 $\text{Standared Deviation of } X, (\sigma_x) = \sqrt{12} = 3.46$

Similarly,

Variance of *Y* = 15

4 $\text{Standared Deviation of } Y (\sigma_y) = \sqrt{15} = 3.87$

Substituting the values in the formula for correlation coefficient we get,

$$r = \frac{10.2}{3.46 \times 3.87} = \frac{10.2}{13.39} = 0.76$$

Hence, correlation coefficient is 0.76.

Example: From the following data calculate the number of items for the *X* and *Y* series.

Correlation coefficient = 0.5

Σxy = 150

Standard Deviation of *y* = 10

Sum of square of deviation of *x* (Σx^2) = 100.

Solution

We know:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

Substituing the given values in the formula

$$0.5 = \frac{150}{\sqrt{100 \times \sum y^2}}$$

Taking square of both sides

$$0.25 = \frac{22500}{100 \times \sum y^2}$$

$$\text{or}, \sum y^2 = \frac{22500}{100 \times 0.25} = 900$$

Now,

$$\sigma_y = \sqrt{\frac{\sum y^2}{N}} = \sqrt{\frac{900}{N}}$$

$$\text{or}, 10 = \sqrt{\frac{900}{N}}$$

$$\text{or}, 100 = \frac{900}{N}$$

$$\text{or}, N = \frac{900}{100} = 9$$

Hence, there are 9 items in *X* and *Y* series.

(ii) *Direct method*

The following steps are involved in the calculation of Karl Pearson's coefficient of correlation using the direct method.

***Step* 1**: Calculate the arithmetic mean of *x*-series ($\overline{X}$) and arithmetic of y-series ($\overline{Y}$).

***Step* 2**: For the *x*-series obtain the square of each value and obtain the sum as ΣX^2. Similarly, obtain the square of the values of the *y*-series and obtain the sum as ΣY^2.

***Step* 3**: Multiply each value of the *x* series with the corresponding value of the *y*-series and obtain the total as ΣXY.

***Step* 4**: The The following formula is applied to calculate the Karl Pearson's coefficient of correlation.

$$r = \frac{\Sigma XY - N\left(\frac{\Sigma X}{N}\right)\times\left(\frac{\Sigma Y}{N}\right)}{\sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2} \times \sqrt{\frac{\Sigma Y^2}{N} - \left(\frac{\Sigma Y}{N}\right)^2}}$$

OR

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2} \times \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

Where,

ΣXY = Sum of multiples of X and Y values

ΣX^2 = Sum of squares of X values

ΣY^2 = Sum of squares of Y values

N = Total number of observations

Example: For the following data, calculate Karl Pearson's coefficient of correlation using the direct method.

X	***Y***	***X*²**	***Y*²**	***XY***
5	18	25	324	90
7	16	49	256	112
8	13	64	169	104
12	16	144	256	192
13	6	169	36	78
15	3	225	09	45
ΣX =60	ΣY =72	ΣX^2 = 676	ΣY^2 =1,050	ΣXY= 621

Solution

$$\overline{X}=\frac{60}{6}=10$$

$$\overline{Y}=\frac{72}{6}=12$$

$$r=\frac{\Sigma XY-N\cdot\overline{X}\cdot\overline{Y}}{\sqrt{\Sigma X^2-N\left(\overline{X}\right)^2}\times\sqrt{\Sigma y^2-N\left(\overline{Y}\right)^2}}$$

$$=\frac{621-6\times10\times12}{\sqrt{676-6\times100}\times\sqrt{1,050-6\times144}}$$

$$=\frac{621-720}{\sqrt{76}\times\sqrt{186}}$$

$$=-\frac{99}{8.71\times13.63}=-0.83$$

Thus, there exists high degree of negative correlation between *X* and *Y*.

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

X	2	5	9	12	17	12	18	20
Y	4	6	10	14	20	15	10	14

Solution

X	***Y***	X^2	Y^2	***XY***
2	4	4	16	8
5	6	25	36	30
9	10	81	100	90
12	14	144	196	168
17	20	289	400	340
12	15	144	225	180
18	10	324	100	180
20	14	400	196	280
17	20	289	400	340
10	17	100	289	170
©*X* = 122	©*Y* = 130	©X^2= 1,800	©Y^2= 1,958	©*XY*= 1,786

$$\text{Correlation Coefficient, } r = \frac{\Sigma XY - N\left(\frac{\Sigma X}{N}\right)\times\left(\frac{\Sigma Y}{N}\right)}{N\sqrt{\frac{\Sigma X^2}{N}-\left(\frac{\Sigma X}{N}\right)^2}\times\sqrt{\frac{\Sigma y^2}{N}-\left(\frac{\Sigma Y}{N}\right)^2}}$$

$$= \frac{1,786 - 10\left(\frac{122}{10}\right)\times\left(\frac{130}{10}\right)}{10\sqrt{\frac{1,800}{10}-\left(\frac{122}{10}\right)^2}\times\sqrt{\frac{1,958}{10}-\left(\frac{130}{10}\right)^2}}$$

$$= \frac{200}{10\times 28.85} = 0.693$$

Hence, the Karl Pearson's correlation coefficient is 0.693

(iii) *Assumed mean method/ Short-cut method*

The calculation procedure described in the actual mean method and the direct method can become quite tedious and lengthy if the arithmetic mean of the series is in fractions rather than in whole numbers. To simplify the calculation procedure in such cases we use the assumed mean method or the short-cut method. In this method, the calculation procedure remains the same with a slight difference that rather than taking deviations from the actual value of arithmetic mean, they are taken from the value of an assumed mean.

The following steps are involved in the calculation of Karl Pearson's coefficient of correlation using the assumed mean method.

***Step* 1**: Decide an assumed mean for the *x* and *y*-series.

***Step* 2**: From the value of the assumed mean for the *x*-series, obtain deviations and denote it by d_x . Obtain their total, Σd_x

***Step* 3**: Similarly, from the value of the assumed mean for the *y*-series, obtain deviations and denote it by d_y. Obtain their total, Σd_y.

***Step* 4**: Square the deviations obtained in ***step* 2** and obtain their sum, Σd_x^2.

***Step* 5**: Square the deviations obtained in ***step* 3** and obtain their sum, Σd_y^2.

***Step* 6**: Multiply the deviations obtained in ***step* 2** and ***step* 3** and obtain the sum, $\Sigma d_x d_y$.

***Step* 7**: The following formula is applied to calculate the Karl Pearson's coefficient of correlation.

$$r = \frac{\sum d_x d_y - \frac{\left(\sum d_x\right) \times \left(\sum d_y\right)}{N}}{\sqrt{\sum d_x^2 - \frac{\left(\sum d_x\right)^2}{N}} \times \sqrt{\sum d_y^2 - \frac{\left(\sum d_y\right)^2}{N}}}$$

where,

d_x = Deviation of X series from assumed mean

d_y = Deviation of Y series from assumed mean

$\sum d_x d_y$ = Sum of multiples of d_x and d_y

$\sum d_x^2$ = Sum of squares of d_x

$\sum d_y^2$ = Sum of squares of d_y

$\sum d_x$ = Sum of deviations of X series

$\sum d_y$ = Sum of deviations of Y series

N = Total numbers of observations

Example: For the following series calculate the Karl Pearson's coefficient of correlation using the assumed mean method.

X	***Y***
31	55
34	57
35	59
37	64
39	66
43	69

Solution

X	***Y***	(d_x) ***X – A***	$(d_x)^2$	(d_y) ***Y – A***	$(d_y)^2$	$d_x d_y$
31	55	–4	16	–4	16	16
34	57	–1	1	–2	04	2
A = 35	*A* = 59	0	0	0	0	0

37 39 43	64 66 69	2 4 8	4 16 64	5 7 10	25 49 100	10 28 80
		$\Sigma d_x = 9$	$\Sigma d_{x^2} = 101$	$\Sigma d_y = 16$	$\Sigma d_{y^2} = 194$	$\Sigma d_x d_y = 136$ 2

$$r = \frac{\Sigma d_x dy - \frac{(\Sigma d_x) \times (\Sigma dy)}{N}}{\sqrt{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}} \times \sqrt{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}}$$

$$= \frac{136 - \frac{9 \times 16}{6}}{\sqrt{101 - \frac{(9)^2}{6}} \times \sqrt{194 - \frac{(16)^2}{6}}}$$

$$= \frac{112}{\sqrt{87.5} \times \sqrt{151.34}} = 0.97$$

Thus, there exists very high degree of positive correlation.

(iv) *Step-Deviation method*

The step-deviation method further simplifies the calculation procedure of the assumed mean method. In this method, the value of the deviations is further reduced by dividing them by a common factor. The following is the formula for the calculation of Karl Pearson's correlation coefficient using the step-deviation method.

$$r = \frac{\sum d'_x d'_y - \frac{\left(\sum d'_x\right) \times \left(\sum d'_y\right)}{N}}{\sqrt{\sum d'^2_x - \frac{\left(\sum d'_x\right)^2}{N}} \times \sqrt{\sum d'^2_y - \frac{\left(\sum d'_y\right)^2}{N}}}$$

where,

$d'_x = \frac{d_x}{h}$ and $d'_y = \frac{d_y}{i}$

$h =$ common factor for X series

$i =$ common factor for Y series

d_x = Deviation of X series from assumed mean

d_y = Deviation of Y series from assumed mean

$\sum d'_x d'_y =$ Sum of multiples of d'_x and d'_y

$\sum d'^2_x =$ Sum of squares of d'_x

$\sum d'^2_y =$ Sum of squares of d'_y

$\sum d_x =$ Sum of deviations of X series

$\sum d_y =$ Sum of deviations of Y series

$N =$ Total numbers of observations

Example: For the series given below calculate the value of Karl Pearson's correlation coefficient.

X	*Y*
150	340
160	350
170	360
180	370
190	380
200	390
220	400
220	410

Solution

X	Y	(dx) $X - A$	d'_x $\left(\frac{X-A}{10}\right)$	d'^2_x	(d_y) $Y - A$	d'_y $\left(\frac{Y-A}{10}\right)$	d'^2_y	d'
150	340	–30	–3	9	–30	–3	9	
160	350	–20	–2	4	–20	–2	4	
170	360	–10	–1	1	–10	–1	1	
A = 180	A = 370	0	0	0	0	0	0	
190	380	10	1	1	10	1	1	
200	390	20	2	4	20	2	4	
220	400	30	3	9	30	3	9	
220	410	40	4	16	40	4	16	1
			$\Sigma d'_x = 4$	$\Sigma d'^2_x =$ 44		$\Sigma d'_y = 4$	$\Sigma d'^2_y = 44$	$\Sigma d_x d$

$$r = \frac{\Sigma d'_x d'_y - \frac{(\Sigma d'_x) \times (\Sigma d'_y)}{N}}{\sqrt{\Sigma d'^2_x - \frac{(\Sigma d'_x)^2}{N}} \times \sqrt{\Sigma d'^2_y - \frac{(\Sigma d'_y)^2}{N}}}$$

$$= \frac{44 - \frac{(4) \times (4)}{8}}{\sqrt{44 - \frac{16}{8}} \times \sqrt{44 - \frac{16}{8}}}$$

$$= \frac{42}{42} = 1$$

Thus, there exists perfect positive correlation

Properties of Correlation Coefficient

The following are some of the important properties of correlation coefficient.

(i) *Pure number*: Correlation coefficient is free from any units. That is, it is a pure number.

(ii) *Indication of relationship*: A negative correlation coefficient indicates inverse relation between the two variables, while a positive correlation coefficient indicates a positive relation.

(iii) *Can Indicate absence of relation*: If correlation coefficient is zero, then it implies that there is no linear relation between the two variables.

(iv) *Value*: The value of correlation coefficient lies between –1 and +1. If r = +1 then it indicates perfect positive correlation while if r = –1 then it indicates perfect negative correlation.

(v) *Independent of origin and scale*: The value of correlation coefficient is independent of its origin and any change in the scale of the graph.

Merits of Karl Pearson's Correlation Coefficient

The following are some of the merits.

i. It is the ***most common and an ideal method*** of calculating correlation.

ii. The value of the correlation coefficient ***helps in assessing the type and magnitude of the linear relationship*** between the two variables.

iii. It ***helps in measuring the exact correlation*** between the two variables.

Demerits of Karl Pearson's Correlation Coefficient

The following are some of the demerits.

i. It is ***affected by the presence of extreme values***.

ii. It ***involves a tedious and time consuming calculation*** process.

iii. It only studies the linear relationship between the two variables and ***fails to study non-linear relationship*** such as quadratic relations etc.

Strengthen this topic TAKE A TOPIC TEST

SCROLL DOWN FOR THE NEXT TOPIC

Spearman's Rank Correlation Coefficient

Objective

After going through this lesson, you shall be able to understand the computation of Spearman's rank correlation coefficient.

Introduction

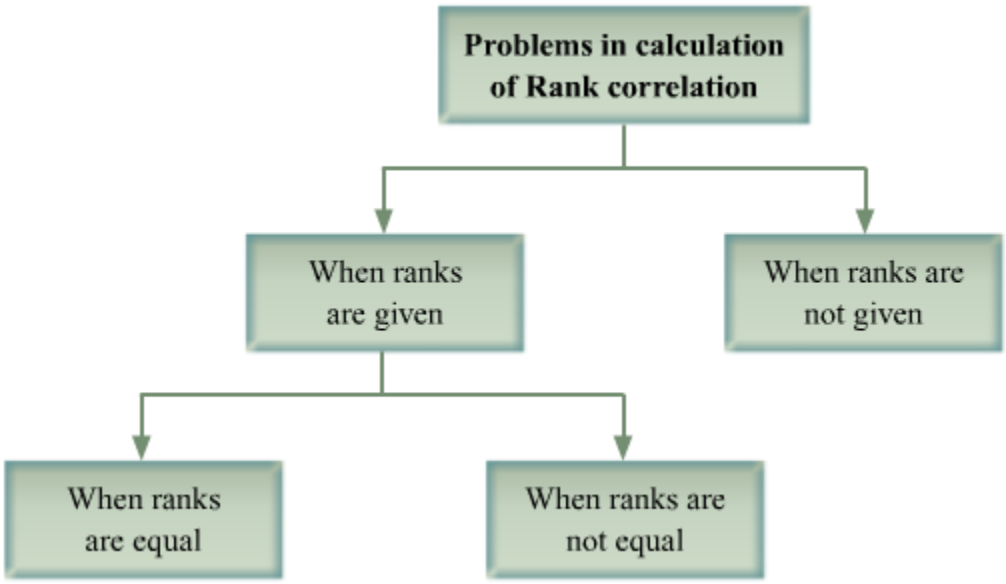
In the previous lesson, we studied the scatter diagram and the Karl Pearson's coefficient of correlation. The two methods of estimating the correlation are useful only where the variables are quantifiable (or measureable). However, a correlation may also exists in two or more qualitative variables. For the estimation of correlation coefficient in case of qualitative variables, the Spearman's rank correlation is used. This method of estimating the correlation coefficient was developed by a British psychologist 'Charles Edward Spearman' in the year 1904. In this method, rather than an absolute value, the variables are assigned ranks and correlation is estimated between the rankings of the two variables.

Calculation of Spearman's Rank Correlation Coefficient

The calculation of Spearman's rank correlation coefficient can be studied under the following two broad categories.

i. When ranks are given

ii. When ranks are not given



Calculation of Rank Correlation when the ranks are given

Steps to calculate Spearman's rank correlation coefficient when ranks are given and are not equal

***Step* 1**: Ascertain the difference (*D*) between the ranks assigned to the two series, i.e. ($R_1 - R_2$)
***Step* 2**: Square the differences i.e. D^2
***Step* 3**: Sum-up all the square of differences, i.e. $\sum D^2$

$$\text{Spearman's Rank Correlation Coefficient } (r_k) = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

where,
r_k = Coefficient of Rank Correlation
D = Rank Differences i.e. $(R_1 - R_2)$
D^2 = Squares of Rank Differences i.e. $(R_1 - R_2)^2$
N = Number of Observations

Example: In a debate competition, the two judges assigned the following ranks to the contestants. Calculate the rank correlation coefficient.

Rank Assigned by Judge 1 R_1	**Rank Assigned by Judge 2** R_2
4	9
6	5
3	6
9	7
8	4
2	1
1	3

Solution

R_1	R_2	D $(R_1 - R_2)$	D^2
4	9	–5	25
6	5	1	1
3	6	–3	9
9	7	2	4
8	4	4	16
2	1	1	1
1	3	–2	4

			$?D^2 = 60$

$$\text{Spearman's Rank Correlation Coefficient } (r_k) = 1 - \frac{6\sum D^2}{N^3 - N}$$
$$= 1 - \frac{6 \times 60}{7^3 - 7}$$
$$= 1 - \frac{360}{336}$$
$$= -0.07$$

There exists a negative correlation. This suggests that to some extent the two judges do not agree with each other in terms of assigning ranks.

Calculation of Rank Correlation when the ranks are not given

Here, the calculation procedure can be divided into the following two broad categories.

i. When no two ranks are tied

ii. When ranks are tied

i. ***When no two ranks are tied.***

In this case, following steps are followed to calculate Spearman's Rank Correlation Coefficient.
***Step* 1**: Allot ranks to the given items of raw data by assigning the first rank to the smallest item followed by the second rank to the subsequent item and so on. In this way, the last rank will be assigned to the largest item.
***Step* 2**: Ascertain the difference (D) between the ranks assigned to the two series i.e. ($R_1 - R_2$).
***Step* 3**: Square the differences i.e. D^2
***Step* 4:** Sum-up all the square of differences, i.e. $\sum D^2$
Now, apply the following formula to ascertain Spearman's Rank Correlation Coefficient.

$$\text{Spearman's Rank Correlation Coefficient } (r_k) = 1 - \frac{6\sum D^2}{N^3 - N}$$

where,

r_k = Coefficient of Rank Correlation
D = Rank differences i.e. ($R_1 - R_2$)

D^2 = Squares of rank differences i.e. $(R_1 - R_2)^2$
N = Number of observations

Example: The following data presents the marks in two subjects statistics and accountancy for 7 students. Calculate the rank correlation coefficient between the marks of the two subjects.

X	Y
25	31
37	28
71	55
29	67
48	92
64	45
56	38

Solution

X	Y	R_1	R_2	(D) $R_1 - R_2$	D^2
25	31	7	6	1	1
37	28	5	7	–2	4
71	55	1	3	–2	4
29	67	6	2	4	16
48	92	4	1	3	9
64	45	2	4	–2	4
56	38	3	5	–2	4
					$\sum D^2 = 42$

$$\text{Spearman's Rank Correlation Coefficient } (r_k) = 1 - \frac{6\sum D^2}{N^3 - N}$$
$$= 1 - \frac{6 \times 42}{7^3 - 7}$$
$$= 1 - \frac{252}{336}$$
$$= 0.25$$

Thus, there exists a low degree of positive correlation in the marks of the two subjects.

ii. ***When the ranks are tied or repeated.***
In this case, following steps are followed to calculate Spearman's rank correlation coefficient.

Step 1*:* Allot ranks to the given items of raw data by assigning first rank to the smallest item followed by the second rank to the subsequent item and so on. In this way, the last rank will be assigned to the largest item. In case, any two items have the same ranks, then the average rank is assigned to both the items.
***Step 2*:** Ascertain the difference (*D*) between the ranks assigned to the two series i.e. ($R_1 - R_2$).
***Step 3*:** Square the differences i.e. D^2
***Step 4*:** Sum-up all the square of differences, i.e. $\sum D^2$
Step 5: Apply the following formulae to ascertain Spearman's Rank Correlation Coefficient.

$$r_k = 1 - \frac{6\left[\sum D^2 + \frac{1}{12}\left(M_1^3 - M_1\right) + \frac{1}{12}\left(M_2^3 - M_2\right) + \ldots\ldots\ldots\ldots\right]}{N^3 - N}$$

where,
$M =$ Number of items of equal ranks
$D = R_1 - R_2$
$D^2 =$ Sum of squares of difference of Rank1 and Rank2
$N =$ Number of observations

Example: Marks by two judges to the participants in a singing competition are given. Calculate Spearman's rank correlation coefficient.

X	*Y*
12	14

17	16
15	12
14	17
19	18
16	12
14	15

Solution

X	Y	R_1	R_2	D
12	14	1	3	–2
17	16	6	5	1
15	12	4	1.5	2.5
14	17	2.5	6	–3.5
19	18	7	7	0
16	12	5	1.5	–3.5
14	15	2.5	4	–1.5

$$r_k = 1 - \frac{6\left[\sum D^2 + \frac{1}{12}\left(M_1^3 - M_1\right) + \frac{1}{12}\left(M_2^3 - M_2\right) + \ldots\ldots\ldots\ldots\right]}{N^3 - N}$$

$$= 1 - \frac{6\left[38 + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(2^3 - 2\right)\right]}{7^3 - 7}$$

$$= 0.31$$

Thus, value of the rank correlation coefficient is 0.31

Note: $M_1 = 2$ (As the Judge X has given 14 marks to two students)
$M_2 = 2$ (As the Judge Y has given 12 marks to two students)

Example: Calculate the value of rank correlation coefficient for the following data.

X	Y
10	9
12	13
15	17
10	13
16	18
14	13

Solution

X	Y	R_1	R_2	D
10	9	1.5	1	0.5
12	13	3	3	0
15	17	5	5	0
10	13	1.5	3	–1.5
16	18	6	6	0
14	13	4	3	1

$$r_k = 1 - \frac{6\left[\sum D^2 + \frac{1}{12}\left(M_1^3 - M_1\right) + \frac{1}{12}\left(M_2^3 - M_2\right) + \ldots\ldots\ldots\ldots\right]}{N^3 - N}$$

$$= 1 - \frac{6\left[3.5 + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(3^3 - 3\right)\right]}{6^3 - 6}$$

$$= 0.83$$

Hence, the value of rank correlation coefficient is 0.83

Note: $M_1 = 2$ (As the Judge X has given 10 marks to two students)
$M_2 = 3$ (As the Judge Y has given 13 marks to three students)

Example: Calculate the value of rank correlation coefficient for the following data.

X	Y
10	9
12	9
13	17
12	13
16	9
14	13

Solution

X	Y	R_1	R_2	D
10	9	1	2	–1
12	9	2.5	2	0.5
13	17	4	6	–2
12	13	2.5	4.5	–2
16	9	6	2	4

14	13	5	4.5	0.5

$$r_k = 1 - \frac{6\left[\sum D^2 + \frac{1}{12}\left(M_1^3 - M_1\right) + \frac{1}{12}\left(M_2^3 - M_2\right) + \frac{1}{12}\left(M_3^3 - M_3\right) \ldots\ldots\right]}{N^3 - N}$$

$$= 1 - \frac{6\left[25.5 + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(3^3 - 3\right)\right]}{6^3 - 6}$$

$$= 0.19$$

Hence, the value of rank correlation coefficient is 0.19

Note: $M_1 = 2$ (As the Judge X has given 10 marks to two students)
$M_2 = 2$ (As the Judge Y has given 13 marks to two students)
$M_3 = 3$ (As the Judge Y has given 9 marks to three students)

Example: The coefficient of rank correlation of marks obtained by 10 contestants in a dance competition and debate competition was found to be 0.7. However, later it was discovered that the difference in the ranks in the two competitions, for one of the students, was wrongly taken to be 4 instead of 6. Calculate the correct rank correlation coefficient.

Solution

Given, $r_k = 0.7$
$N = 10$
Wrong $D = 4$
Correct $D = 6$.

Now, $r_k = 1 - \frac{6\sum D^2}{N^3 - N}$

Substituting the given values in the formula we get,

$$0.7 = 1 - \frac{6\sum D^2}{10^3 - 10}$$

$$\text{or, } (1 - 0.7) = \frac{6\sum D^2}{990}$$

$$\text{or, } 0.3 \times 990 = 6\sum D^2$$

$$\text{or, } \sum D^2 = 49.5$$

Therefore, wrong ΣD^2 is 49.5
Correct ΣD^2 = 49.5 – Wrong D^2 + Correct D^2
= 49.5 – 4^2 + 6^2
= 69.5

Substituting the value of correct ΣD^2 in the formula for correlation coefficient:

Correct Rank Correlation Coefficient $= 1 - \frac{6 \times 69.5}{990}$

$= 1 - \frac{417}{990}$

$= 1 - 0.421$

$= 0.579$

Hence, the correct rank correlation coefficient is 0.579.

Merits of Spearman's Rank Correlation Coefficient

The following are some of the merits.

i. ***Easy to calculate***: The calculation of Spearman's rank correlation is easier than the Pearson's correlation.

ii. ***Used for qualitative variables***: It can be used for qualitative variables such as honestly, beauty, intelligence etc.

iii. ***Helps in assessing the type and magnitude*:** The value of the correlation coefficient helps us in assessing the type and magnitude of the linear relationship between the two variables.

iv. ***Actual values are not necessary*:** This method can also be used even when only ranks are given and not the actual values of observations.

Demerits of Spearman's Rank Correlation Coefficient

The following are some of the demerits.

i. ***Cannot be used in continuous series***: This method cannot be used in case of open-ended series or continuos series. That is, this method works out for discrete and individual series only.

ii. ***Cannot be used for large number of observations***: It cannot be used for large number of observations. It works only if the number of observations is less than 30.

iii. ***Lacks precision***: Compared to Karl Pearson's method, the rank correlation method lacks precision. This is because, it does not use all the information, (actual observations) rather uses only ranks.

Similarities and Dissimilarities Between Karl Pearson's Correlation Coefficient (r_{xy}) and Spearman's Rank Correlation Coefficient (r_k)

The following points can be observed with regard to the similarities and dissimilarities in the two correlation coefficients.

i. Generally, all the properties of Karl Pearson's coefficient of correlation are similar to that of the rank correlation coefficient.

ii. Rank correlation coefficient is generally lower or equal to Karl Pearson's coefficient.

iii. Rank correlation coefficient is usually preferred to measure the correlation between the two qualitative variables.

iv. The difference between the two coefficients is because the rank correlation coefficient uses ranks whereas the Karl Pearson's coefficient uses full set of observations.

v. If the precisely measured data are available, then both the coefficients will be identical.

vi. If extreme values are present in the data, then the rank correlation coefficient is more precise and reliable and consequently its value differs from that of the Karl Pearson's coefficient.

vii. When the values are not repeated, the value of Karl Pearson's correlation coefficient will be same as that of the Spearman's Rank correlation coefficient.