

Coefficient of Correlation

Correlation Analysis

Correlation analysis is applied in quantifying the association between two continuous variables, for example, a dependent and independent variables or among two independent variables.

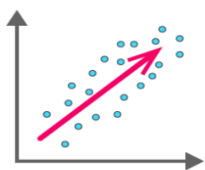
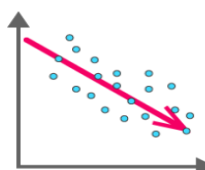
The sample of a correlation coefficient is estimated in the correlation analysis. It ranges between -1 and +1, denoted by r and quantifies the strength and direction of the linear association among two variables. The sign of the coefficient of correlation shows the direction of the association. The magnitude of the coefficient shows the strength of the association.

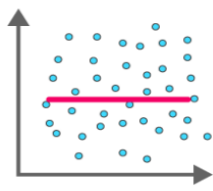
The correlation coefficient is measured on a scale that varies from + 1 through 0 to - 1. Complete correlation between two variables is expressed by either + 1 or -1. When one variable increases as the other increases the correlation is positive; when one decreases as the other increases it is negative. Complete absence of correlation is represented by 0.

For example, a correlation of $r = 0.8$ indicates a positive and strong association among two variables, while a correlation of $r = -0.3$ shows a negative and weak association. A correlation near to zero shows the non-existence of linear association among two continuous variables.

There are mainly two types of correlations:

- **Positive Correlation**
- **Negative Correlation**

Positive Correlation	The value of one variable increases linearly with increase in another variable. This indicates a similar relation between both the variables. So its correlation coefficient would be positive or 1 in this case.	 Positive correlation © Byjus.com
Negative Correlation	When there is a decrease in values of one variable with increase in values of other variable. In that case, correlation coefficient would be negative.	 Negative correlation © Byjus.com

Zero Correlation or No Correlation	There is one more situation when there is no specific relation between two variables.	 <p>No correlation</p> <p>© Byjus.com</p>
---	---	---

Correlation Coefficient Formula

Let X and Y are the two random variables.

The **population correlation coefficient** for X and Y is given by the formula:

$$\rho_{xy} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y}$$

Where,

ρ_{xy} = Population correlation coefficient between X and Y

μ_x = Mean of the variable X

μ_y = Mean of the variable Y

σ_x = Standard deviation of X

σ_y = Standard deviation of Y

E= Expected value operator

Cov= Covariance

The above formulas can also be written as:

$$\rho_{x,y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E(X)^2} \cdot \sqrt{E(Y^2) - E(Y)^2}}$$

Formally, the **sample correlation coefficient** is defined by the following formula, where

s_x and s_y are the sample standard deviations, and s_{xy} is the sample covariance.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Similarly, the **population correlation coefficient** is defined as follow, where σ_x and σ_y are the population standard deviations, and σ_{xy} is the population covariance.

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

The sample correlation coefficient formula or called Pearson Correlation Coefficient is

$$r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n \sum x_i^2 - (\sum x_i)^2][n \sum y_i^2 - (\sum y_i)^2]}}$$

or

$$r = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sqrt{[\sum X^2 - (\sum X)^2/n][\sum Y^2 - (\sum Y)^2/n]}}$$

or

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

The above are many formulas used to find the value of correlation coefficient for the given data all of them give the same value but different formula.

Based on the value obtained through these formulas, we can determine how much strong the association is between given two variables.

Here,

n = Number of values or elements

$\sum x_i$ = Sum of x values

$\sum y_i$ = Sum of y values

$\sum x_i y_i$ = Sum of the product x and y values

$\sum x_i^2$ = Sum of squares of x values

$\sum y_i^2$ = Sum of squares of y values

Therefore, Correlation illustrates the relationship between two or more variables.

It is expressed in the form of a number that is known as correlation coefficient.

Correlation Coefficient Meaning and Examples

The exact meaning of the correlation coefficient is explained here on the basis of the value of r varying between -1 and +1. Let us discuss here:

- If $r = 1$, that means for each positive increase in one variable, there is a positive increase in another variable too, in a fixed proportion. For example, the size of the cloth increases in correlation with the height of the person.
- If $r = -1$, that means for each positive increase in one variable, there is a negative decrease in another variable too, in a fixed proportion. For example, the distance decreases in correlation with the increasing speed vehicle.
- If $r = 0$, there is no increase or decrease of another variable with respect to first. They are not related to each other here.

Example on Correlation Coefficient

Example Calculate the Correlation coefficient of given data

x	50	51	52	53	54
y	3.1	3.2	3.3	3.4	3.5

Solution:

Here n = 5

x	50	51	52	53	54
y	3.1	3.2	3.3	3.4	3.5
xy	155	163.2	171.6	180.2	189
x²	2500	2601	2704	2809	2916
y²	9.61	10.24	10.89	11.56	12.25

$$\sum x = 260$$

$$\sum y = 16.5$$

$$\sum xy = 859$$

$$\sum x^2 = 13530$$

$$\sum y^2 = 54.55$$

We apply these values for the above table by substituting all the values in formula to obtain the correlation coefficient of x and y

$$r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n \sum x_i^2 - (\sum x_i)^2][n \sum y_i^2 - (\sum y_i)^2]}}$$

we got $r = 1$. This shows a positive correlation coefficient.

Usually, in statistics, there are three types of correlations: Pearson correlation, Kendall rank correlation and Spearman correlation.

Spearman correlation

Spearman's rank correlation coefficient allows us to identify easily the strength of correlation within a data set of two variables, and whether the correlation is positive or negative. The Spearman coefficient is denoted with the Greek letter rho (ρ).

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

For example to calculate the Spearman correlation is,

Spearman correlation from the data				
Child number	Rank height	Rank dead space	d	d²
1	1	3	2	4
2	2	1	-1	1
3	3	2	-1	1
4	4	4	0	0
5	5	5.5	0.5	0.25
6	6	11	5	25
7	7	7	0	0
8	8	5.5	-2.5	6.25
9	9	8	-1	1
10	10	13	3	9
11	11	10	-1	1
12	12	9	-3	3
13	13	12	-1	1
14	14	15	1	1
15	15	14	-1	1
Total				60.5

$$\rho_s = 1 - \frac{6 \times 60.5}{15 \times (225 - 1)} = 0.8920$$

Thus we get that the value is very close to that of the Pearson correlation coefficient. Also, there are two kinds of correlation coefficient

1- Partial correlation coefficient

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

or
$$r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}}$$

or
$$r_{23.1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{(1 - r_{21}^2)(1 - r_{31}^2)}}$$

Example

Assume that X1 is the income family for every month, X2 represent spending family for every month also and X3 the no. of person in the family and no. of random sample is 20 family, the following of corr. Coef. are

$$r_{12}=0.91, r_{13}=0.39 \text{ and } r_{23}=0.62, \text{ and } r_{12}=r_{21}, r_{13}=r_{31} \text{ and } r_{23}=r_{32}$$

Calculate partial correlation coefficient between income and spending by no. of person in family.

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} = \frac{0.91 - 0.39 \times 0.62}{\sqrt{(1 - 0.39^2)(1 - 0.62^2)}} =$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} = \frac{0.39 - 0.91 \times 0.62}{\sqrt{(1 - 0.91^2)(1 - 0.62^2)}} =$$

$$r_{23.1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{(1 - r_{21}^2)(1 - r_{31}^2)}} = \frac{0.62 - 0.91 \times 0.39}{\sqrt{(1 - 0.91^2)(1 - 0.39^2)}} =$$

Theorem If $r_{12}=r_{13}=r_{32}=r$ prove that $r_{12.3}=r_{13.2}=r_{23.1}=\frac{r}{1+r}$

Sol:-

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

$$\because r_{12}=r_{13}=r_{32}=r = \frac{r - r \times r}{\sqrt{(1 - r^2)(1 - r^2)}} = \frac{r - r^2}{\sqrt{(1 - r^2)^2}} = \frac{r(1 - r)}{(1 - r)(1 + r)} = \frac{r}{1 + r}$$

The same procedure you do for $r_{13.2}$ and $r_{23.1}$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} = ? \quad \text{Complete it}$$

$$r_{23.1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{(1 - r_{21}^2)(1 - r_{31}^2)}} = ? \quad \text{Complete it}$$

Multiple correlation coefficient

If X_1, X_2, X_3 have three random variables then multiple corr. Coef. between X_1 and both X_2, X_3 than calculate according to the following formula.

$$r_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

Where r_{12} is the corr. between x_1, x_2

Where r_{13} is the corr. between x_1, x_3

Where r_{23} is the corr. between x_2, x_3

Example

Assume that X_1 represent of producing wheat, X_2 represent the type of seeds and X_3 represent the type used of manufactures, calculated corr. coef. between these variables with the following results

Calculate the multiple corr. coef. between X_1 and both X_2 and X_3 .

$r_{12}=0.70$, $r_{13}=0.80$ and $r_{23}=0.55$

$$\begin{aligned} r_{1.23} &= \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \\ &= \sqrt{\frac{0.70^2 + 0.80^2 - 2(0.70)(0.80)(0.55)}{1 - (0.55)^2}} = \end{aligned}$$

Theorem

If $r_{12}=r_{13}=r_{23}=r$ prove that $r_{1.23}=r_{2.13}=r_{3.12}=\sqrt{\frac{2r^2}{1+r}}$

Sol:-

$$\begin{aligned} r_{1.23} &= \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \\ \because r_{12}=r_{13}=r_{32}=r &= \sqrt{\frac{r^2 + r^2 - 2.r.r.r}{1 - r^2}} = \sqrt{\frac{2r^2 - 2.r^3}{1 - r^2}} = \sqrt{\frac{2r^2(1-r)}{(1-r)(1+r)}} = \sqrt{\frac{2r^2}{1+r}} \end{aligned}$$

The same procedure you do for $r_{2.13}$ and $r_{3.12}$.