**Correlation** refers to a process for establishing the relationships exist between two variables. You learned a way to get a general idea about whether or not two variables are related, is to plot them on a “[scatter plot](https://byjus.com/maths/scatter-plot/)”. While there are many measures of association for variables which are measured at the ordinal or higher level of measurement, correlation is the most commonly used approach.

Correlation in Statistics

This section shows how to calculate and interpret correlation coefficients for ordinal and interval level scales. Methods of correlation summarize the relationship between two variables in a single number called the correlation coefficient. The correlation coefficient is usually represented using the symbol r, and it ranges from -1 to +1.

A correlation coefficient quite close to 0, but either positive or negative, implies little or no relationship between the two variables. A correlation coefficient close to plus 1 means a positive relationship between the two variables, with increases in one of the variables being associated with increases in the other variable.

A correlation coefficient close to -1 indicates a negative relationship between two variables, with an increase in one of the variables being associated with a decrease in the other variable. A correlation coefficient can be produced for ordinal, interval or ratio level variables, but has little meaning for variables which are measured on a scale which is no more than nominal.

For ordinal scales, the correlation coefficient can be calculated by uisng Spearman’s rho. For interval or ratio level scales, the most commonly used correlation coefficient is Pearson’s r, ordinarily referred to as simply the correlation coefficient.

## Correlation Coefficient

The correlation coefficient, r, is a summary measure that describes the extent of the statistical relationship between two interval or ratio level variables. The correlation coefficient is scaled so that it is always between -1 and +1. When r is close to 0 this means that there is little relationship between the variables and the farther away from 0 r is, in either the positive or negative direction, the greater the relationship between the two variables.

The two variables are often given the symbols X and Y. In order to illustrate how the two variables are related, the values of X and Y are pictured by drawing the scatter diagram, graphing combinations of the two variables. The scatter diagram is given first, and then the method of determining Pearson’s r is presented. From the following examples, relatively small sample sizes are given. Later, data from larger samples are given.

## Scatter Diagram

A scatter diagram is a diagram that shows the values of two variables X and Y, along with the way in which these two variables relate to each other. The values of variable X are given along the horizontal axis, with the values of the variable Y given on the vertical axis.

Later, when the regression model is used, one of the variables is defined as an independent variable, and the other is defined as a dependent variable. In regression, the independent variable X is considered to have some effect or influence on the dependent variable Y. Correlation methods are symmetric with respect to the two variables, with no indication of causation or direction of influence being part of the statistical consideration. A scatter diagram is given in the following example. The same example is later used to determine the correlation coefficient.

## Types of Correlation

The scatter plot explains the correlation between the two attributes or variables. It represents how closely the two variables are connected. There can be three such situations to see the relation between the two variables –

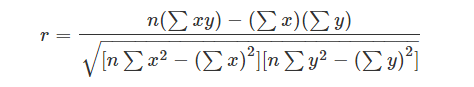
* Positive Correlation – when the value of one variable increases with respect to another.
* Negative Correlation – when the value of one variable decreases with respect to another.
* No Correlation – when there is no linear dependence or no relation between the two variables.

## Correlation Formula

Correlation shows the relation between two variables. Correlation coefficient shows the measure of correlation. To compare two datasets, we use the correlation formulas.

### **Pearson Correlation Coefficient Formula**

The most common formula is the Pearson Correlation coefficient used for linear dependency between the data set. The value of the coefficient lies between -1 to +1. When the coefficient comes down to zero, then the data is considered as not related. While, if we get the value of +1, then the data are positively correlated, and -1 has a negative correlation.



Where n = Quantity of Information

Σx = Total of the First Variable Value

Σy = Total of the Second Variable Value

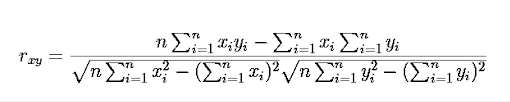
Σxy = Sum of the Product of first & Second Value

Σx2 = Sum of the Squares of the First Value

Σy2 = Sum of the Squares of the Second Value

### **Linear Correlation Coefficient Formula**

The formula for the linear correlation coefficient is given by;



### **Sample Correlation Coefficient Formula**

The formula is given by:

**rxy = Sxy/SxSy**

Where Sx and Sy are the sample standard deviations, and Sxy is the sample covariance.

### **Population Correlation Coefficient Formula**

The population correlation coefficient uses σx and σy as the population standard deviations and σxy as the population covariance.

**rxy = σxy/σxσy**

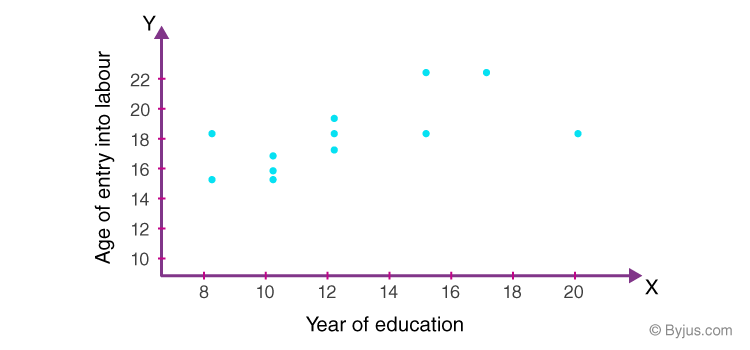
## Correlation Example

Years of Education and Age of Entry to Labour Force Table.1 gives the number of years of formal education (X) and the age of entry into the labour force (Y ), for 12 males from the Regina Labour Force Survey. Both variables are measured in years, a ratio level of measurement and the highest level of measurement. All of the males are aged close to 30, so that most of these males are likely to have completed their formal education.

|  |  |  |
| --- | --- | --- |
| **Respondent Number** | **Years of Education, X** | **Age of Entry into Labour Force, Y** |
| **1** | **10** | **16** |
| **2** | **12** | **17** |
| **3** | **15** | **18** |
| **4** | **8** | **15** |
| **5** | **20** | **18** |
| **6** | **17** | **22** |
| **7** | **12** | **19** |
| **8** | **15** | **22** |
| **9** | **12** | **18** |
| **10** | **10** | **15** |
| **11** | **8** | **18** |
| **12** | **10** | **16** |

Table 1. **Years of Education and Age of Entry into Labour Force for 12 Regina Males**

Since most males enter the labour force soon after they leave formal schooling, a close relationship between these two variables is expected. By looking through the table, it can be seen that those respondents who obtained more years of schooling generally entered the labour force at an older age. The mean years of schooling are X¯ = 12.4 years and the mean age of entry into the labour force is Y¯= 17.8, a difference of 5.4 years.

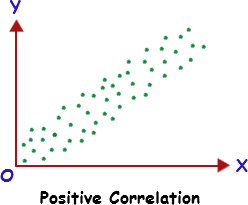


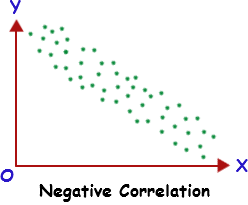
This difference roughly reflects the age of entry into formal schooling, that is, age five or six. It can be seen through that the relationship between years of schooling and age of entry into the labour force is not perfect. Respondent 11, for example, has only 8 years of schooling but did not enter the labour force until the age of 18. In contrast, respondent 5 has 20 years of schooling but entered the labour force at the age of 18. The scatter diagram provides a quick way of examining the relationship between X and Y.

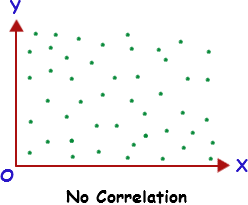
# **Pearson Correlation Formula**

 The name correlation suggests the relationship between two variables as their Co-relation. The correlation coefficient is the measurement of correlation. To see how the two sets of data are connected, we make use of this formula. The linear dependency between the data set is done by the Pearson Correlation coefficient. It is also known as the Pearson product-moment correlation coefficient. The value of the Pearson correlation coefficient product is between -1 to +1.  When the correlation coefficient comes down to zero, then the data is said to be not related. While, if we are getting the value of +1, then the data are positively correlated and -1 has a negative correlation.

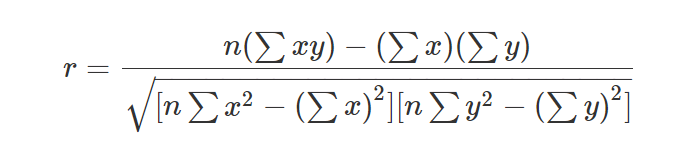
The graphical representation of positive, negative and no correlation is shown below:







The Pearson correlation coefficient is denoted by the letter “r”. The formula for Pearson correlation coefficient r is given by:



Where,  
r = Pearson correlation coefficient  
x = Values in the first set of data  
y = Values in the second set of data  
n = Total number of values.

### **Solved Example**

**Question:**Marks obtained by 5 students in algebra and trigonometry as given below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algebra | 15 | 16 | 12 | 10 | 8 |
| Trigonometry | 18 | 11 | 10 | 20 | 17 |

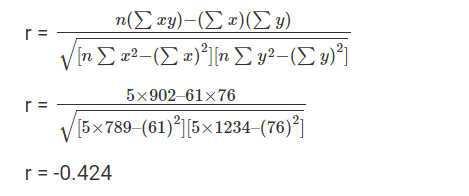
Calculate the Pearson correlation coefficient.

**Solution:**

Construct the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| x | y | x2 | y2 | xy |
| 15 | 18 | 225 | 324 | 270 |
| 16 | 11 | 256 | 121 | 176 |
| 12 | 10 | 144 | 100 | 120 |
| 10 | 20 | 100 | 400 | 200 |
| 8 | 17 | 64 | 289 | 136 |
| ∑x = 61 | ∑y = 76 | ∑x2 = 789 | ∑y2 = 1234 | ∑xy = 902 |

Formula for Pearson correlation coefficient is given by:



# **Correlation Coefficient F**

# Correlation coefficient formula is given and explained here for all of its types. There are various formulas to calculate the correlation coefficient and the ones covered here include Pearson’s Correlation Coefficient Formula, Linear Correlation Coefficient Formula, Sample Correlation Coefficient Formula, and Population Correlation Coefficient Formula. Before going to the formulas, it is important to understand what correlation and correlation coefficient is. A brief introduction is given below and to learn about them in detail, click the linked article.

## Correlation Coefficient FormulaAbout Correlation Coefficient

The correlation coefficient is a measure of the association between two variables. It is used to find the relationship is between data and a measure to check how strong it is. The formulas return a value between -1 and 1, where -1 shows negative correlation and +1 shows a positive correlation.

The correlation coefficient value is positive when it shows that there is a correlation between the two values and the negative value shows the amount of diversity among the two values.

# Correlation Coefficient**or**

## Types of Correlation Coefficient Formula

There are several types of correlation coefficient formulas. But, one of the most commonly used formulas in statistics is Pearson’s Correlation Coefficient Formula. The formulas for all the correlation coefficient are discussed below.

### **Pearson’s Correlation Coefficient Formula**

Also known as bivariate correlation, the [Pearson’s correlation coefficient formula](https://byjus.com/pearson-correlation-formula/) is the most widely used correlation method among all the sciences. The correlation coefficient is denoted by “r”.

To find r, let us suppose the two variables as x & y, then the correlation coefficient r is calculated as:

# **mula**

### **Practice Questions from Coefficient of Correlation Formula**

* **Question 1:**Find the linear correlation coefficient for the following data. X = 4, 8 ,12, 16 and Y = 5, 10, 15, 20.
* **Question 2:**Calculate correlation coefficient for x = 100, 106, 112, 98, 87, 77, 67, 66, 49 and y = 28, 33, 26, 27, 24, 24, 21, 26, 22.
* **Question 3:**What will be the correlation coefficient for X and Y values for the given values: X= (1,2,3,4,5) and Y= {11,22,34,43,56}

# **Linear Correlation Coefficient For**

To find out the relation between two variables in a population, linear correlation formula is used. To see how the variables are connected we will use the linear correlation. Also known as “Pearson’s Correlation”, a linear correlation is denoted by r” and the value will be between -1 and 1.

The elements denote a strong relationship if the product is 1. Similarly, if the coefficient comes close to -1, it has a negative relation. If the Linear coefficient is zero means there is no relation between the data given.

# Where “n” is the number of observations, “xi” and “yi “are the variables.**m**

### **Solved Examples**

**Question 1:**Calculate the linear correlation coefficient for the following data. X = 4, 8 ,12, 16 and Y = 5, 10, 15, 20.

**Solution:**

Given variables are,

X = 4, 8 ,12, 16 and Y = 5, 10, 15, 20

For finding the linear coefficient of these data, we need to first construct a table for the required values.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| x | y | x2 | y2 | XY |
| 4 | 5 | 16 | 25 | 20 |
| 8 | 10 | 64 | 100 | 80 |
| 12 | 15 | 144 | 225 | 180 |
| 16 | 20 | 256 | 400 | 320 |
| Σ x = 40 | Σ y =50 | 480 | 750 | 600 |

# **ul**

The regression equation

Correlation describes the strength of an association between two variables, and is completely symmetrical, the correlation between A and B is the same as the correlation between B and A. However, if the two variables are related it means that when one changes by a certain amount the other changes on an average by a certain amount. For instance, in the children described earlier greater height is associated, on average, with greater anatomical dead Space. If y represents the dependent variable and x the independent variable, this relationship is described as the regression of y on x.

The relationship can be represented by a simple equation called the regression equation. In this context “regression” (the term is a historical anomaly) simply means that the average value of y is a “function” of x, that is, it changes with x.

# The regression equation representing how much y changes with any given change of x can be used to construct a regression line on a scatter diagram, and in the simplest case this is assumed to be a straight line. The direction in which the line slopes depends on whether the correlation is positive or negative. When the two sets of observations increase or decrease together (positive) the line slopes upwards from left to right; when one set decreases as the other increases the line slopes downwards from left to right. As the line must be straight, it will probably pass through few, if any, of the dots. Given that the association is well described by a straight line we have to define two features of the line if we are to place it correctly on the diagram. The first of these is its distance above the baseline; the second is its slope. They are expressed in the following regression equation :

With this equation we can find a series of values of



 the variable, that correspond to each of a series of values of x, the independent variable. The parameters α and β have to be estimated from the data. The parameter signifies the distance above the baseline at which the regression line cuts the vertical (y) axis; that is, when y = 0. The parameter β (the *regression coefficient*) signifies the amount by which change in x must be multiplied to give the corresponding average change in y, or the amount y changes for a unit increase in x. In this way it represents the degree to which the line slopes upwards or downwards.

**a**The regression equation is often more useful than the correlation coefficient. It enables us to predict y from x and gives us a better summary of the relationship between the two variables. If, for a particular value of x, x i, the regression equation predicts a value of y fit , the prediction error is



. It can easily be shown that any straight line passing through the mean values x and y will give a total prediction error



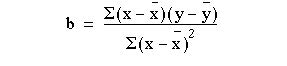
 of zero because the positive and negative terms exactly cancel. To remove the negative signs we square the differences and the regression equation chosen to minimise the sum of squares of the prediction errors,



We denote the sample estimates of Alpha and Beta by a and b. It can be shown that the one straight line that minimises



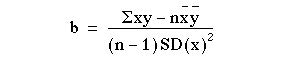
# , the least squares estimate, is given by



And



it can be shown that



which is of use because we have calculated all the components of equation (11.2) in the calculation of the correlation coefficient.

The calculation of the correlation coefficient on the data in table 11.2 gave the following:



Applying these figures to the formulae for the regression coefficients, we have:





Therefore, in this case, the equation for the regression of y on x becomes



This means that, on average, for every increase in height of 1 cm the increase in anatomical dead space is 1.033 ml *over the range of measurements made*.

The line representing the equation is shown superimposed on the scatter diagram of the data in figure 11.2. The way to draw the line is to take three values of x, one on the left side of the scatter diagram, one in the middle and one on the right, and substitute these in the equation, as follows:

If x = 110, y = (1.033 x 110) – 82.4 = 31.2

If x = 140, y = (1.033 x 140) – 82.4 = 62.2

If x = 170, y = (1.033 x 170) – 82.4 = 93.2

Although two points are enough to define the line, three are better as a check. Having put them on a scatter diagram, we simply draw the line through them.

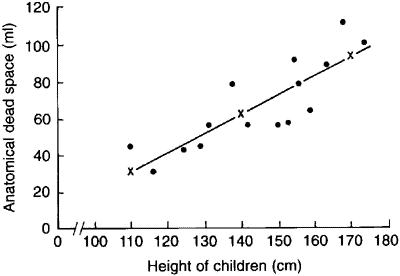
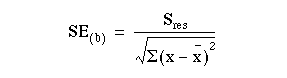


Figure 11.3 Regression line drawn on scatter diagram relating height and pulmonaiy anatomical dead space in 15 children

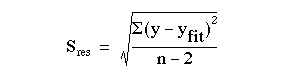
The standard error of the slope SE(b) is given by:



where



 is the residual standard deviation, given by:



This can be shown to be algebraically equal to

This can be shown to be algebraically equal to



We already have to hand all of the terms in this expression. Thus



 is the square root of



. The denominator of (11.3) is 72.4680. Thus SE(b) = 13.08445/72.4680 = 0.18055.

We can test whether the slope is significantly different from zero by:

t = b/SE(b) = 1.033/0.18055 = 5.72.

Again, this has n – 2 = 15 – 2 = 13 degrees of freedom. The assumptions governing this test are:

1. That the prediction errors are approximately Normally distributed. Note this does not mean that the x or y variables have to be Normally distributed.
2. That the relationship between the two variables is linear.
3. That the scatter of points about the line is approximately constant – we would not wish the variability of the dependent variable to be growing as the independent variable increases. If this is the case try taking logarithms of both the x and y variables.

Note that the test of significance for the slope gives exactly the same value of P as the test of significance for the correlation coefficient. Although the two tests are derived differently, they are algebraically equivalent, which makes intuitive sense.

We can obtain a 95% confidence interval for b from



where the tstatistic from has 13 degrees of freedom, and is equal to 2.160.

Thus the 95% confidence interval is

l.033 – 2.160 x 0.18055 to l.033 + 2.160 x 0.18055 = 0.643 to 1.422.

Regression lines give us useful information about the data they are collected from. They show how one variable changes on average with another, and they can be used to find out what one variable is likely to be when we know the other – provided that we ask this question within the limits of the scatter diagram. To project the line at either end – to extrapolate – is always risky because the relationship between x and y may change or some kind of cut off point may exist. For instance, a regression line might be drawn relating the chronological age of some children to their bone age, and it might be a straight line between, say, the ages of 5 and 10 years, but to project it up to the age of 30 would clearly lead to error. Computer packages will often produce the intercept from a regression equation, with no warning that it may be totally meaningless. Consider a regression of blood pressure against age in middle aged men. The regression coefficient is often positive, indicating that blood pressure increases with age. The intercept is often close to zero, but it would be wrong to conclude that this is a reliable estimate of the blood pressure in newly born male infants!

