**Correlation**

The correlation is one of the most common and most useful statistics. A correlation is a single number that describes the degree of relationship between two variables.

It is a statistical method which enables the researcher to find whether two variables are related and to what extent they are related. Correlation is considered as the sympathetic movement of two or more variables. We can observe this when a change in one particular variable is accompanied by changes in other variables as well, and this happens either in the same or opposite direction, then the resultant variables are said to be correlated.

The word Correlation is made of **Co-** (meaning "together"), and **Relation**

 Correlation is **Positive** when the values **increase** together, and

 Correlation is **Negative** when one value **decreases** as the other increases

**Correlation are of three types:**

* Positive Correlation
* Negative Correlation
* No correlation

In correlation, when values of one variable increase with the increase in another variable ,it is supposed to be a **positive correlation**. On the other hand, if the values of one variable decrease with the decrease in another variable, then it would be a **negative correlation**. There might be the case when there is no change in a variable with any change in another variable. In this case, it is defined as **no correlation** between the two.

Correlation can have a value:

* 1 is a perfect positive correlation
* 0 is no correlation (the values don't seem linked at all)
* -1 is a perfect negative correlation

The value shows **how good the correlation is** (not how steep the line is), and if it is positive or negative.

*Positive Correlation*

#### Positive correlation occurs when an increase in one variable increases the value in another.

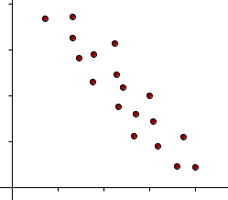
#### The line corresponding to the scatter plot is an increasing line.

#### Positive Correlation

#### Negative Correlation

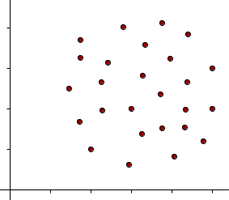
Negative correlation occurs when an increase in one variable decreases the value of another.

The line corresponding to the scatter plot is a decreasing line.



#### No Correlation

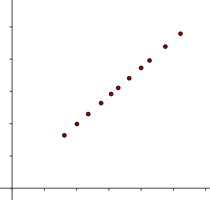
No correlation occurs when there is no linear dependency between the variables.



#### Perfect Correlation

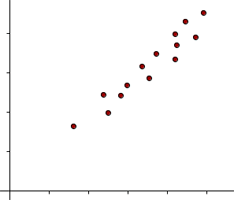
Perfect correlation occurs when there is a funcional dependency between the variables.

In this case all the points are in a straight line.



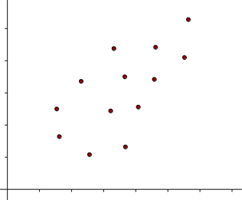
#### Strong Correlation

A correlation is stronger the closer the points are located to one another on the line.



#### Weak Correlation

A correlation is weaker the farther apart the points are located to one another on the line.



**Correlation Symbol**

Symbol of correlation = r

## Correlation Formula

The formula for correlation is as follows,  
  
**Correlation ( r )=**

Where,  
x and y are the variables.

b = the slope of the regression line is also called as the regression coefficient

a = intercept point of the regression line which is in the y-axis.

N= Number of values or elements

X = First Score

Y = Second Score

∑X∑Y = Sum of the product of the first and Second Scores  
  
∑X = Sum of First Scores  
  
∑Y = Sum of Second Scores  
  
∑= Sum of square first scores.  
  
∑ = Sum of square second scores.

**Example:**

Let's assume that we want to look at the relationship between two variables, height (in inches) and self esteem.

|  |  |  |
| --- | --- | --- |
| **Person** | **Height** | **Self Esteem** |
| 1 | 68 | 4.1 |
| 2 | 71 | 4.6 |
| 3 | 62 | 3.8 |
| 4 | 75 | 4.4 |
| 5 | 58 | 3.2 |
| 6 | 60 | 3.1 |
| 7 | 67 | 3.8 |
| 8 | 68 | 4.1 |
| 9 | 71 | 4.3 |
| 10 | 69 | 3.7 |
| 11 | 68 | 3.5 |
| 12 | 67 | 3.2 |
| 13 | 63 | 3.7 |
| 14 | 62 | 3.3 |
| 15 | 60 | 3.4 |
| 16 | 63 | 4.0 |
| 17 | 65 | 4.1 |
| 18 | 67 | 3.8 |
| 19 | 63 | 3.4 |
| 20 | 61 | 3.6 |

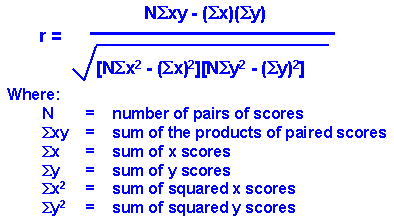
Here, we have to find the relation between the height and selfesteem.

By observing the values, we can clearly say that the relation between height and self esteem is positive correlation because with the increase of height value self esteem value increases.

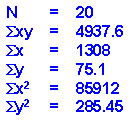
**Hypothesis statement:** correlation coefficient ( r ) is equal to zero.

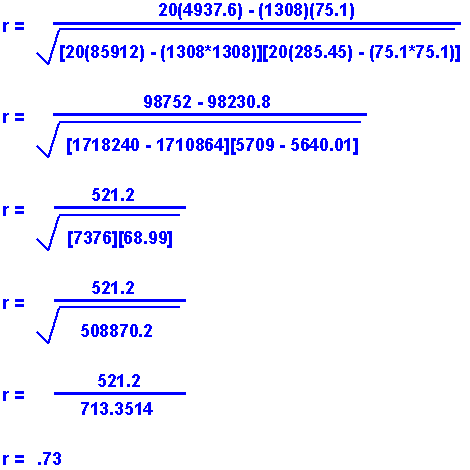
**Alternate Hypothesis**: Correlation coefficient ( r ) is not equal to zero.

**Formula:**



From the given data,





From the result, r is not equal to zero.

So, Reject the Null hypothesis.

**Coefficient of correlation**:

Coefficient of correlation, r, called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables. It also called as Pearson product moment correlation coefficient. The algebraic method of measuring the correlation is called the coefficient of correlation. 

**There are mainly three coefficients of correlation**

1. Karl Pearson’s Coefficient of correlation
2. Pearson’s rank correlation coefficient
3. Concurrent correlation

**Karl Pearson’s Coefficient of correlation**  
  
The most important algebraic method of measuring correlation is Karl Pearson’s Coefficient of correlation or Pearson’s coefficient of Correlation. It has widely used application in Statistics. It is denoted by r.   
  
**The formula is given by**

**r=**

**Interpretation of Karl Pearson’s Coefficient of correlation**  
  
Karl Pearson’s Coefficient of correlation denoted by r is the degree of correlation between two variables. r takes values between –1 and 1  
  
When r is –1, we say there is perfect negative correlation.   
When r is a value between –1 and 0, we say that there is a negative correlation  
When r is 0, we say there is no correlation  
When r is a value between 0 and 1, we say there is a positive correlation  
When r is 1, we say there is a perfect positive correlation.

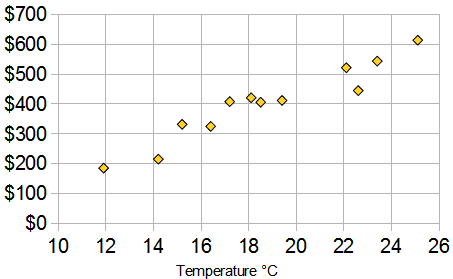
**Example: Ice Cream Sales**

The local ice cream shop keeps track of how much ice cream they sell versus the temperature on that day, here are their figures for the last 12 days:

**Ice Cream Sales vs Temperature**

|  |  |
| --- | --- |
| **Temperature °C** | **Ice Cream Sales** |
| 14.2° | $215 |
| 16.4° | $325 |
| 11.9° | $185 |
| 15.2° | $332 |
| 18.5° | $406 |
| 22.1° | $522 |
| 19.4° | $412 |
| 25.1° | $614 |
| 23.4° | $544 |
| 18.1° | $421 |
| 22.6° | $445 |
| 17.2° | $408 |

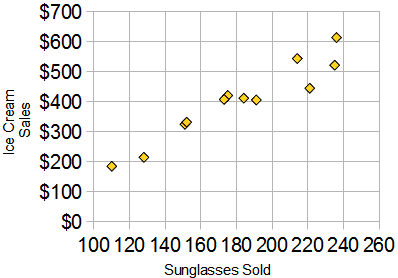
And here is the same data as a [Scatter Plot](https://www.mathsisfun.com/data/scatter-xy-plots.html):



"Correlation Is Not Causation" ... which says that a correlation does **not** mean that one thing causes the other (there could be other reasons the data has a good correlation).

### Example: Sunglasses vs Ice Cream

Our Ice Cream shop finds how many sunglasses were sold by a big store for each day and compares them to their ice cream sales:



The correlation between Sunglasses and Ice Cream sales is high

Using R

Let the two variables be x and y.

The correlation between those two variables is :

Cor(x,y)

This command gives the coefficient of correlation as output. By analysing the coefficient, we can check whether the relation is positive, negative or no correlation.

To find any relation between two variables, first we have to check whether those two variables significant or not. If and only they are significant, then there exist some relation between those two variables.

How to check whether two variables significant or not?

In R, we have a command to check the significance of variables.

Cor.test(x,y)

Finding the correlation coefficient can be done in three methods:

1. Pearson’s method
2. Spearman’s method
3. Kendall method

The default method that is applied is Pearson method.

If we wanted to find in other method, then the command is:

Cor(x,y,method=”method\_name”)

We can also test it in different methods.

if you use a dataset instead of separate variables you will return a matrix of all the pairwise correlation coefficients, then

cor(dataset, method = "method\_name")

Here, method\_name describes the name of the method.

If the correlation coefficient is close to 1, it would indicate that the variables are positively linearly related and the scatter plot falls almost along a straight line with positive slope. For -1, it indicates that the variables are negatively linearly related and the scatter plot almost falls along a straight line with negative slope. And for zero, it would indicate a weak linear relationship between the variables.

Example:

person<-1:20

height<-c(68,71,62,75,58,60,67,68,71,69,68,67,63,62,60,63,65,67,63,61)

selfesteem<-c(4.1,4.6,3.8,4.4,3.2,3.1,3.8,4.1,4.3,3.7,3.5,3.2,3.7,3.3,3.4,4.0,4.1,3.8,3.4,3.6)

analysis<-data.frame(person,height,selfesteem)

Here we have to calculate the relationship between height and selfesteem.

**Null hypothesis:** correlation coefficient ( r ) is equal to zero.

**Alternate Hypothesis:** correlation coefficient is not equal to zero.

cor(height,selfesteem)

output:

[1] 0.7306357

cor.test(height,selfesteem)

Pearson's product-moment correlation

data: height and selfesteem

t = 4.5401, df = 18, p-value = 0.0002536

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.4257781 0.8865242

sample estimates:

cor

0.7306357

From the result, reject the null hypothesis.

Here, the relation between the height and selfesteem exists only when they are significant.

If p value less than significant level value, then those variables are said to be significant.

The t value is calculated using the degree of freedom and coefficient of correlation.

And the p value is calculated using the t value and degree of freedom value.

( or )

We can also check by plotting the scatter plots.

plot(x,y)🡪plots the points on the graph

abline(lm(y ~ x)🡪draws a line joining the points that are plotted above.

If all the points fall on the straight line then the correlation is perfect correlation.

(or)

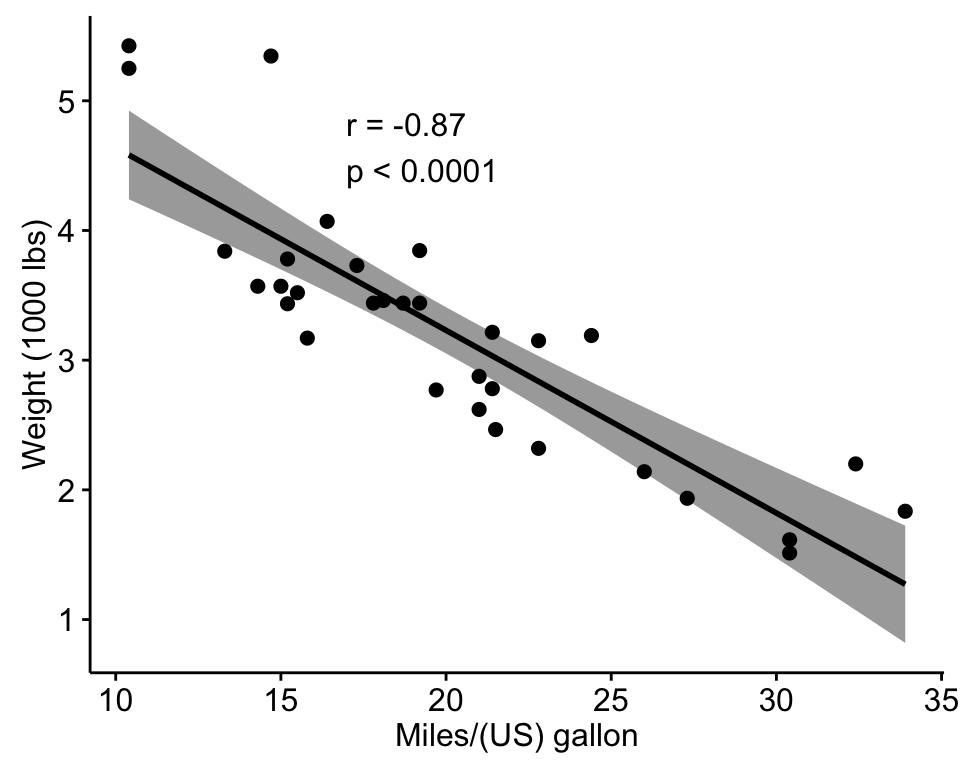
We can also do this by installing a package ggpubr.

Example:

Install.package(ggpubr)

Library(ggpubr)

ggscatter(my\_data, x = "mpg", y = "wt", add = "reg.line", conf.int = TRUE, cor.coef = TRUE, cor.method = "pearson", xlab = "Miles/(US) gallon", ylab = "Weight (1000 lbs)")



**R functions**

* **cor()** computes the **correlation coefficient**
* **cor.test()** test for association/correlation between paired samples. It returns both the **correlation coefficient** and the **significance level**(or p-value) of the correlation .