**Instituto Tecnológico de Tijuana**

**Ingeniería en Sistemas Computacionales**


**Investigación II:**

Correlación de Pearson

**Materia:** Datos Masivos

**Unidad:** Unidad I

**Facilitador:**

José Christian Romero Sánchez

**Alumno:** Hernández Negrete Juan Carlos

**Fecha:**

Tijuana Baja California a 16 de Octubre del 2020.

# PEARSON CORRELATION

Pearson's Correlation Coefficient is a measure of the correspondence or linear relationship between two random quantitative variables. In simpler words, it can be defined as an index used to measure the degree of relationship between two variables, both quantitative.

Having two variables, the correlation facilitates estimations of the value of one of them, with knowledge of the value of the other variable.

This coefficient is a measure that indicates the relative situation of the events with respect to the two variables, that is, it represents the numerical expression that indicates the degree of correspondence or relationship that exists between the 2 variables. These numbers vary between limits of +1 and -1

### How is it calculated?

To have a guide that allows:

- Establish the contiguous variation of the two variables
- Compare the different cases with each other

To do this, the Pearson correlation coefficient is used, defined as the covariance that occurs between two standardized variables and is calculated with the following expression:

$$r_{xy} = \frac{\sum Z_X Z_y}{N}$$

### How does that interpret Pearson's correlation coefficient?

Its dimension indicates the level of association between the variables.

- When it is less than zero (r <0) It is said that there is negative correlation: The variables are correlated in an inverse sense.

High values in one of the variables usually correspond to low values in the other variable and vice versa. The closer the value is to -1 said correlation coefficient, the more evident the extreme covariation will be.

If r = -1, we speak of a perfect negative correlation, which supposes an absolute determination between both variables, in a direct sense a perfect linear relationship with a negative slope coexists.

- When it is greater than zero (r> 0) It is said that there is a positive correlation: Both variables are correlated in a direct sense.

High values in one of the variables correspond to high values in the other variable, and in an inverse situation the same happens with low values. The closer to +1 the correlation coefficient is, the more evident the covariation will be.

• If r = 1 We speak of a perfect positive correlation, which implies an absolute determination between the variables, in a direct sense a perfect linear relationship with a positive slope coexists).

When it is equal to zero (r = 0) The variables are said to be incorrectly related, it is not possible to establish some sense of covariation.

There is no linear relationship, but this does not necessarily imply that the variables are independent, and non-linear relationships may exist between the variables.

When the two variables are independent they are said to be uncorrelated, although the reciprocity result is not necessarily true.

## Examine the linear relationship between variables (Pearson)

Pearson's correlation coefficient is used to examine the strength and direction of the linear relationship between two continuous variables.

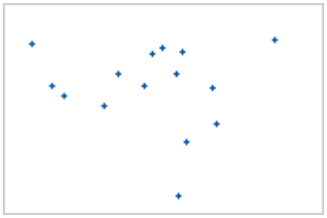
**Strength**

- The value of the correlation coefficient can vary from −1 to +1. The higher the absolute value of the coefficient, the stronger the relationship between the variables.
- For Pearson's correlation, an absolute value of 1 indicates a perfect linear relationship. A correlation close to 0 indicates that there is no linear relationship between the variables.

**Direction**

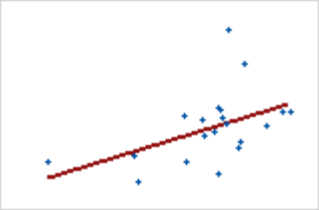
- The sign of the coefficient indicates the direction of the relationship. If both variables tend to increase or decrease at the same time, the coefficient is positive and the line representing the correlation slopes upward. If one variable tends to increase while the other decreases, the coefficient is negative and the line representing the correlation slopes downward.

The following graphs show data with specific values of the correlation coefficient to illustrate different patterns in the strength and direction of relationships between variables



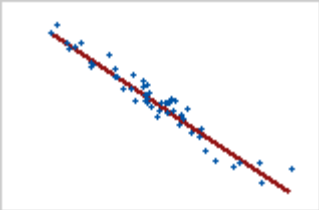
No relationship: Pearson r = 0

The points are located randomly on the graph, which means that there is no linear relationship between the variables.



**Moderate positive relationship: Pearson r = 0.476**

Some points are close to the line, but other points are far from it, indicating that there is only a moderate linear relationship between the variables.



**Large positive elation: Pearson r = 0.93**

The points are located near the line, indicating that there is a strong linear relationship between the variables. The relationship is positive because as one variable increases, the other variable also increases.

## Determine if the correlation coefficient is significant

To determine if the correlation between the variables is significant, compare the p-value with its level of significance. Generally, a significance level (denoted as α or alpha) of 0.05 works well. An α of 0.05 indicates that the risk of concluding that there is a correlation, when in fact it is not, is 5%. The p-value indicates whether the correlation coefficient is significantly different from 0. (A coefficient of 0 indicates that there is no linear relationship).

**P-value ≤ α: The correlation is statistically significant**

- If the p-value is less than or equal to the significance level, then you can conclude that the correlation is different from 0.

**P-value> α: The correlation is not statistically significant**

- If the p-value is greater than the significance level, then you cannot conclude that the correlation is different from 0.

## Advantages and disadvantages of the Pearson correlation coefficient

Among the main advantages of Pearson's correlation coefficient are:

- The value is independent of whatever unit is used to measure the variables.
- If the sample is large, the accuracy of the estimate is more likely.

Some of the disadvantages of the correlation coefficient are:

- It is necessary both variables be measured at a continuous quantitative level.
- The distribution of the variables must be similar to the normal curve.

# References


- Matias Riquelme. (11 de Mayo del 2019). ¿Qué es y cómo se interpreta el coeficiente de correlación de Pearson?. 15 de Octubre del 2020, de Web y Empresas Sitio web: https://www.webyempresas.com/coeficiente-de-correlacion-de-pearson/

- Minitab. (2019). Interpretar los resultados clave para Correlación. 2020, de Minitab Sitio web: https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/basic-statistics/how-to/correlation/interpret-the-results/key-results/

- QuestionPro. (2020). ¿Qué es el coeficiente de correlación de Pearson?. 2020, de QuestionPro Sitio web: https://www.questionpro.com/blog/es/coeficiente-de-correlacion-de-pearson/