Instituto Tecnológico de Tijuana

Tecnológico Nacional de México

Subdirección Académica

Departamento de Sistemas Computacionales

Agosto – diciembre 2020

Datos Masivos

Unidad I

Correlación de Pearson

Manzano Guzmán Jesua 16212033

Romero Hernández José Christian

15 de octubre de 2020

Correlación de Pearson

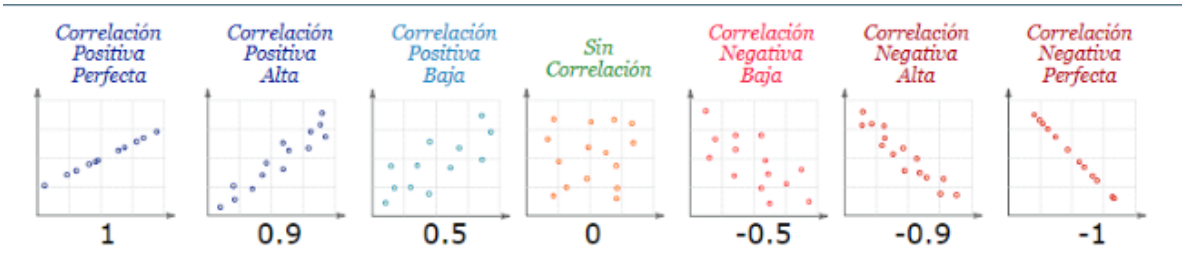
Pearson's correlation coefficient is a type of coefficient used in descriptive statistics. Specifically, it is used in descriptive statistics applied to the study of two variables.

Pearson's correlation coefficient is used to study the relationship (or correlation) between two quantitative random variables (minimum interval scale); for example, the relationship between weight and height.

It is a measure that gives us information about the intensity and direction of the relationship. In other words, it is an index that measures the degree of covariation between different linearly related variables.

We must be clear about the difference between relationship, correlation or covariation between two variables (joint variation) and causality (also called forecast, prediction or regression), since they are different concepts.

Pearson's correlation coefficient comprises values between -1 and +1. Thus, depending on its value, it will have one meaning or another.



If the Pearson correlation coefficient is equal to 1 or -1, we can consider that the correlation that exists between the variables studied is perfect.

If the coefficient is greater than 0, the correlation is positive (“More, more, and less less). On the other hand, if it is less than 0 (negative), the correlation is negative (“A more, less, and a less, more). Finally, if the coefficient is equal to 0, we can only say that there is no linear relationship between the variables, but there may be some other type of relationship.

How is it calculated?

Supposing that two random variables X and Y are being studied on a population; Pearson's correlation coefficient is symbolized by the letter $\rho_{x,y}$ and the expression that allows us to calculate it is:

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

Where:

• σ XY is the covariance of (X, Y)

• σ X is the standard deviation of the variable X

• σ Y is the standard deviation of the variable Y

In an analogous way we can calculate this coefficient on a sample statistic, denoted as rxy a:

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

Conclusion

Pearson's correlation allows us to identify the type of relationship that two variables have within a sample of data or population. If the correlation corresponds to 1 it means that there is a "perfect" relationship, which tells us that the larger the value of one of the variables, the greater the value of the other variable, if the correlation is 0 it means that there is no a relationship between the data of the two variables and if the relationship is -1 means that the greater the value of one of the variables, the lower the value of the other.

References

https://es.wikipedia.org/wiki/Coeficiente_de_correlaci%C3%B3n_de_Pearson

https://www.uv.es/webgid/Descriptiva/31_coeficiente_de_pearson.html

https://psicologiaymente.com/miscelanea/coeficiente-correlacion-pearson

https://www.questionpro.com/blog/es/coeficiente-de-correlacion-de-pearson/

https://personal.us.es/vararey/adatos2/correlacion.pdf