CORRELATION AND PEARSON'S COEFFICIENT

(Alba González)

CORRELATION

WHAT IS CORRELATION?

Correlation is a statistical measure that quantifies the degree of linear association between two variables, indicating how they change together at a constant rate. This tool is frequently used to describe simple relationships without making a statement about cause and effect. Correlations describe data moving together; thus, they are useful for describing simple relationships among data.

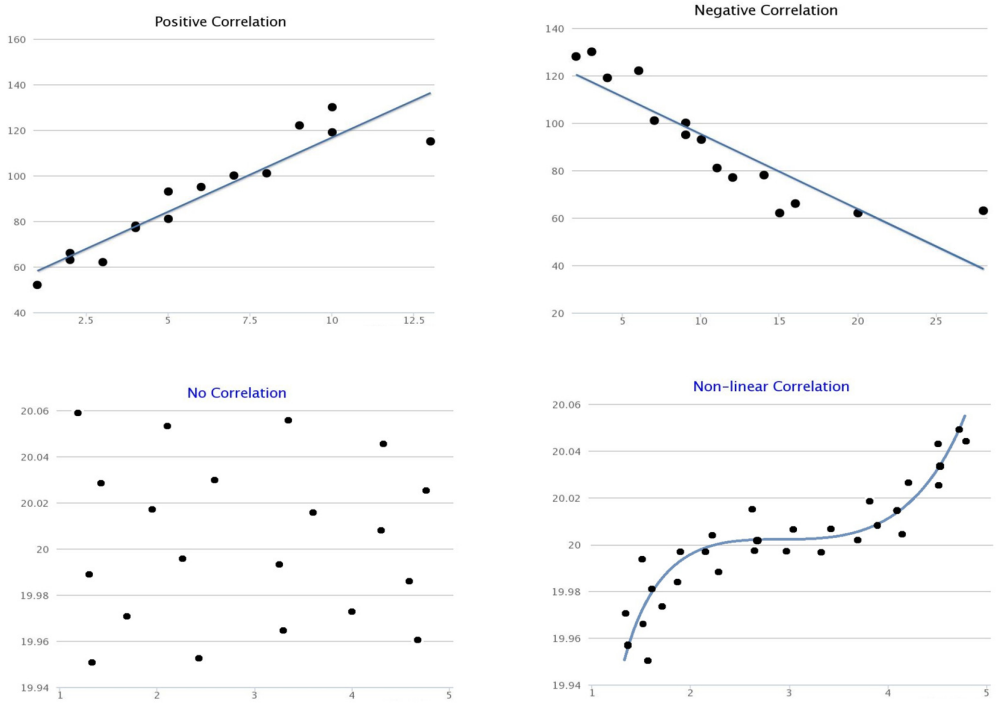
TYPES OF CORRELATION

There is a **positive linear correlation** when the variable on the x-axis increases as the variable on the y-axis increases. For example, most of the time there is an increase in between a person's education level and their family income.

A **negative linear correlation** is found when one variable increases as the other variable decreases. This is shown by a downwards-sloping straight regression line. For example, the longer time it takes a worker to reach their workplace, the lower the job satisfaction is.

No correlation implies that there is no pattern that can be detected between the variables. For example, the amount of ice cream sold at the number of shark attacks.

There is a **non-linear correlation** when there is a relationship between variables, but the relationship is not linear (straight), although this is not the focus of this project. For example, the implantation of



new technology might follow an S-shaped growth pattern – slow adaptation onset, followed by a rapid increase, and a plateau state when it is fully established.

CORRELATION COEFFICIENT

HOW IS CORRELATION MEASURED?

It is measured by the correlation coefficient, r, which is a parameter that quantifies the strength and direction of a linear relationship between two variables. It ranges from -1 to +1, the closer to these values, the stronger the positive or negative relationship. If the correlation coefficient is 0, it means that there is no relationship.

Size of r	Interpretation
0.9 to 1.00	Very high correlation
0.70 to 1.89	High correlation
0.50 to 0.69	Moderate correlation
0.30 to 0.49	Low correlation
0.00 to 0.29	Little if any correlation

SCATTERPLOTS

We can use scatterplots to visualize correlations. The correlation coefficient, r, shows how close the point in the scatterplot comes to a linear relationship; the stronger the relationship or bigger the r values, the closer to the line in which we want to fit the data.

PEARSON'S COEFFICIENT

Pearson's r measures the strength and direction (decreasing or increasing, depending on the sign) of a linear relationship between two variables X and Y can be defined as:

$$r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2(y_i-\bar{y})^2}}$$

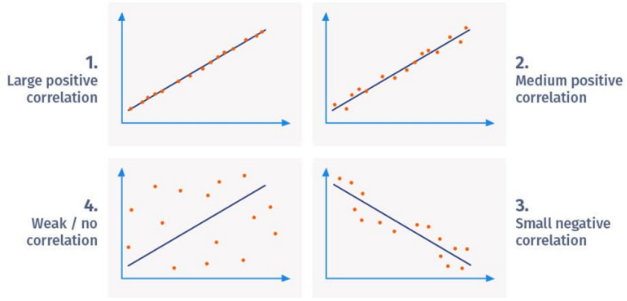
Where:

- X_i and Y_i are individual data points.
- $\bar{X}$ and $\bar{Y}$ are the means of two variables

Strength refers to how one variable will change due to the change in the other. The closer to +1 and -1, the stronger the relationship. In a scatterplot, the values will lie closer to the line.

Direction indicates a positive linear or negative linear relationship between variables. In the scatterplot, if the slope goes up is positive, and if it goes down then it is negative.

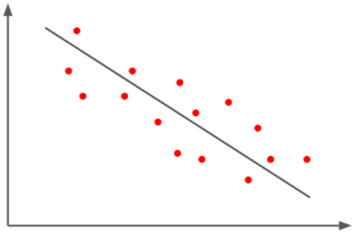
INTERPRETATION



Strength of Association	Coefficient, r: Positive	Coefficient, r: Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to 1.0

The closer the scatterplots lie next to the line, the stronger the relationship between variables. The further they move from the line, the weaker the relationship.

For example:



This scatterplot corresponds to a small negative correlation, as the values do not lie close to the straight line. The change in one variable is inversely proportional to the change in the other variable, as the slope is negative.

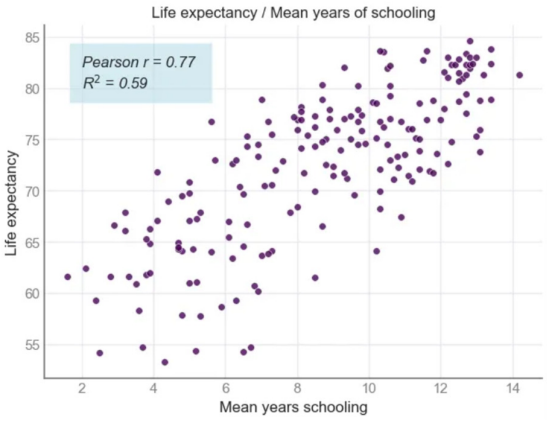
It is important to understand that the negative correlation should not be mistaken with no correlation, for instance, if the Pearson coefficient is -0.9 it indicates a higher correlation than +0.7, and a correlation of +0.8 is not better than -0.8.

PROPERTIES

LINK WITH THE COEFFICIENT OF DETERMINATION

The square of the Pearsons' correlation coefficient, R^2, also known as the coefficient of determination, explains the percentage of variation of one variable that is explained by the variations of the other variable.

Looking at the scatterplot:



The Pearson correlation coefficient (r) is 0.767, this is a close value to +1, so we can conclude that there is a strong correlation between life expectancy and years of schooling. The R^2 is 0.59, which is the result of 0.77^2. For a statistical point of view, a linear regression model can predict 59% of the variations of life expectancy based on the schooling durations.

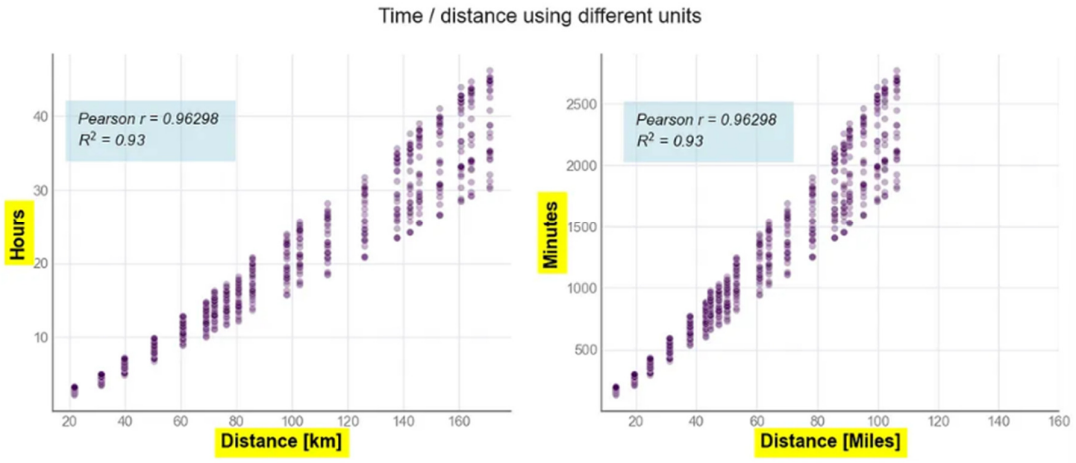
SYMMETRY

This coefficient is symmetric, meaning that either x or y can be expressed as a function of the other; thus, flipping the axis will not affect the Pearson r.

Using the previous example, the mean years of schooling can be plotted against the life expectancy and the value of r and R^2 remains unchanged.

INSENSITIVITY TO SCALE AND LOCATION

The Pearson's coefficient does not have any specific unit, meaning it lacks dimension. Therefore, multiplying x or y by a negative/positive number, or adding/subtracting, it does not have any impact on the outcome; there will be different location of the values in the scatterplot, but the correlation remains constant.



In these scatterplots, the x-axis values are different. Miles were obtained by dividing by 1.61, and minutes by multiplying by 60. The overall result shows that the slope and the values are different, but the correlation remains constant.

CONDITIONS FOR EXISTENCE

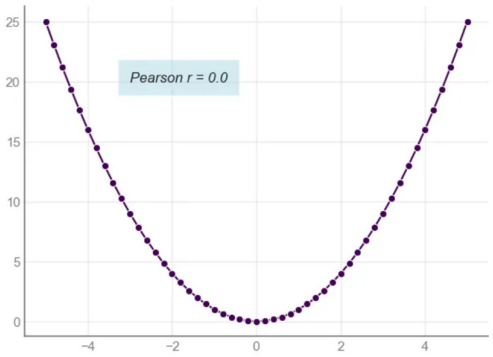
To apply the Pearson correlation coefficient, the following condition must be satisfied:

- The variables must be measured at the interval or ratio level; thus, x and y variables are quantitative and are expressed as real numbers.
- The data should be organized in paired observations and shown in a 2-column value: x value with its corresponding y value.
- The variance and covariance of x and y must be defined, and the variances must be non-null.

PROBLEMS WITH PEARSON'S CORRELATION COEFFICIENT

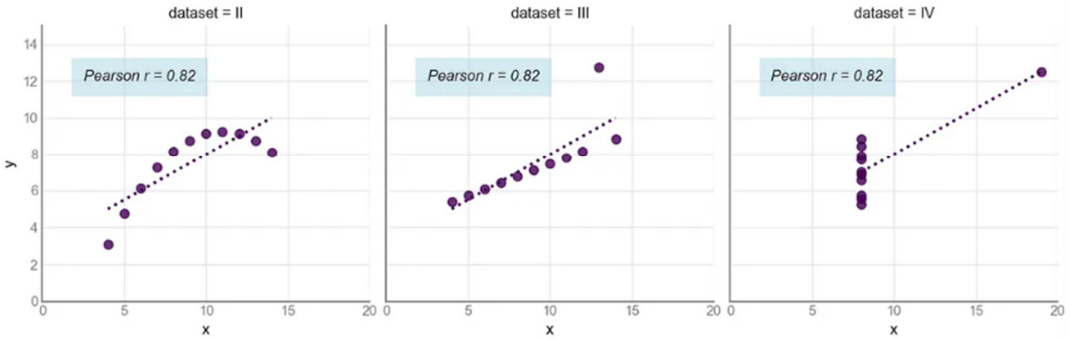
LINEARITY ASSUMPTION

Pearson correlation assumes a linear correlation between variables, thus, if the relationship is non-linear, the coefficient will not accurately represent the association. This is seen when the correlation coefficient is weak, we may conclude that there is no relationship between the two variables, however, the issue may be that the relationship is non-linear.



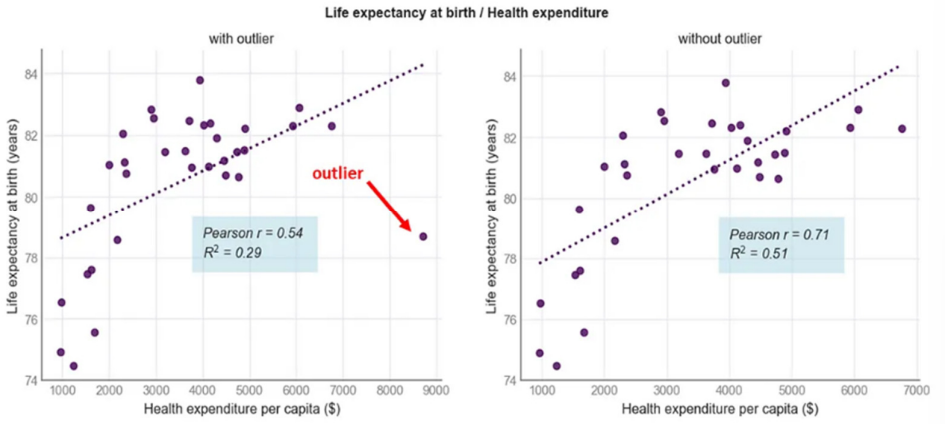
HIGH CORRELATION DOES NOT IMPLY LINEAR RELATIONSHIP

If we obtain an r=0.82, most of the people will plot a straight regression line. But it might represent something like this:



SENSITIVE TO OUTLIERS

To observe this, there is an example measuring the relationship between life expectancy and health expenditure.



Pearson's coefficient is 0.54 with the outlier, while the coefficient without outlier is 0.71. Therefore, visually, the scatterplot values fit better to the line. Meaning that one single outlier changes noticeably the outcome of the analysis.

This leads to the following question, should outliers be excluded from the analysis?

It will depend on the context and what the outlier represents, or the sample size – if it is large enough, outliers are expected to be seen in the analysis. Therefore, keep the outliers if they represent data of the population studied, and remove them if they appear due to experimental or measurement errors, or if there is any significant reason why they should be excluded.

In the context of the example, is there a good reason to remove the outlier? Not really, there is no sign that indicates the removal of the outlier; in fact, it seems more relevant to consider.

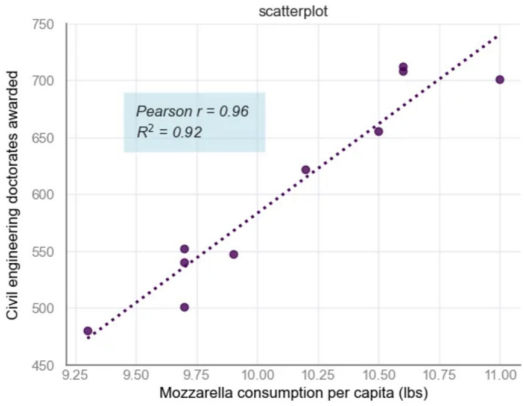
ESTIMATOR BIAS

The Pearson's correlation coefficient can slightly underestimate the absolute value for a population, especially on small sample sizes. This produces a bias, noticeable around the absolute range value (0.5-0.7).

CORRELATION IS NOT CAUSATION

Observing a high correlation between two variables does not imply a causal relationship where the value of one variable directly influences the other. In some situations, even when this Pearson r is high, the correlation may be coincidental, this is known as spurious correlations.

In this example, the consumption of mozzarella cheese per capita and the number of Civil engineering doctorates awarded in the US. In these scatterplots the correlation is remarkably high, but we can safely assume that there is no casualty at play.



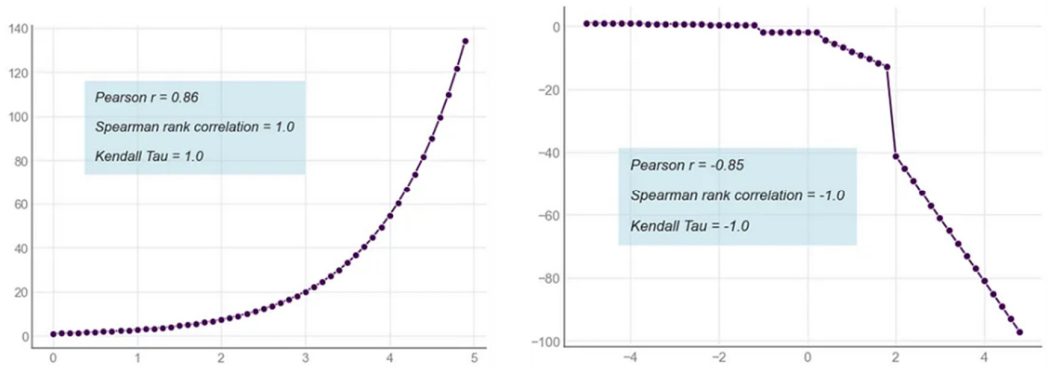
ALTERNATIVES TO THE PEARSON CORRELATION COEFFICIENT

The **Spearman rank correlation** coefficient serves as an alternative to the previous coefficient mentioned, the Pearson correlation coefficient. The spearman rank is the Pearson correlation coefficient calculated between the ranks of the x and t values. For instance, we set the smallest value of the x axis as 1, the next value 2, etc., and in the y axis we do the same. This means that the Spearman rank correlation is just the Pearson correlation between the two new list of values.

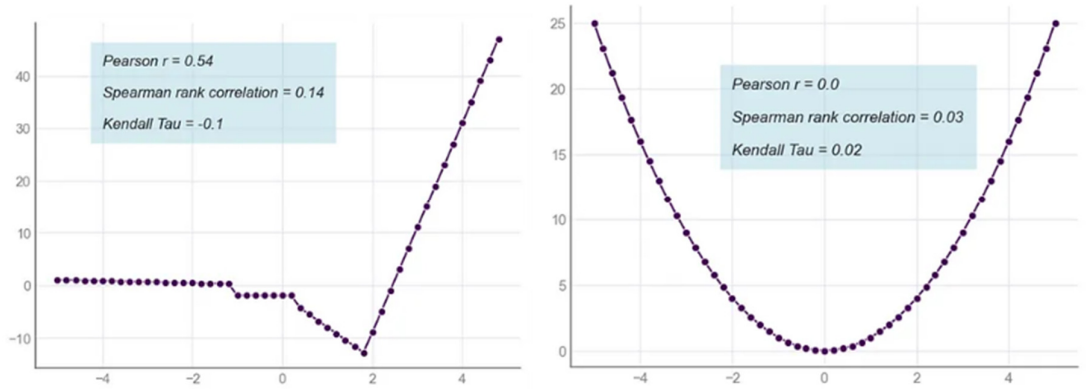
We also find the **Kendall Tau** which is another rank-based correlation. In this case, it is a non-parametric procedure, thus, the data obtained does not have to be normally distributed, compared to Pearson's correlation which is parametric. This alternative is preferred over Spearman's when there is truly little data.

Both alternatives are useful for assessing increases in a y variable according to changes in a x variable, being the main advantage compared to Pearson's, is that we can see the correlation even when the relationship is not linear. When the values of one variable consistently move in the same direction as the other, this phenomenon is known as monotonic relationships.

In the following examples we can compare the alternatives and the Pearson's correlation:



The Spearman rank correlation and Kendall Tau values are 1, meaning a perfectly monotonic relationship when positive, and the same would happen for a value of -1. In contrast, the Pearson correlation coefficient is 0.86 in the first plot and 0.85 in the second sample as it is affected by the non-linear character of the relationship between x and y. But what would occur when the relationship is not monotonic?



As predicted, the values of Spearman rank and Kendall Tau are influenced by the non-monotonic relationship between x and y. Furthermore, similar to Pearson's coefficient, they cannot detect the association between the two variables in the second plot.

REFERENCES

- https://www.jmp.com/en_au/statistics-knowledge-portal/what-is-correlation.html#:~:text=Correlation%20is%20a%20statistical%20measure,statement%20about%20cause%20and%20effect.
- https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/types-of-correlation.html
- https://www.questionpro.com/blog/pearson-correlation-coefficient/#what_is_the_pearson_correlation_coefficient?
- https://medium.com/@anthony.demeusy/pearson-correlation-methodology-limitations-alternatives-part-3-alternatives-cc2a56f7ad1f