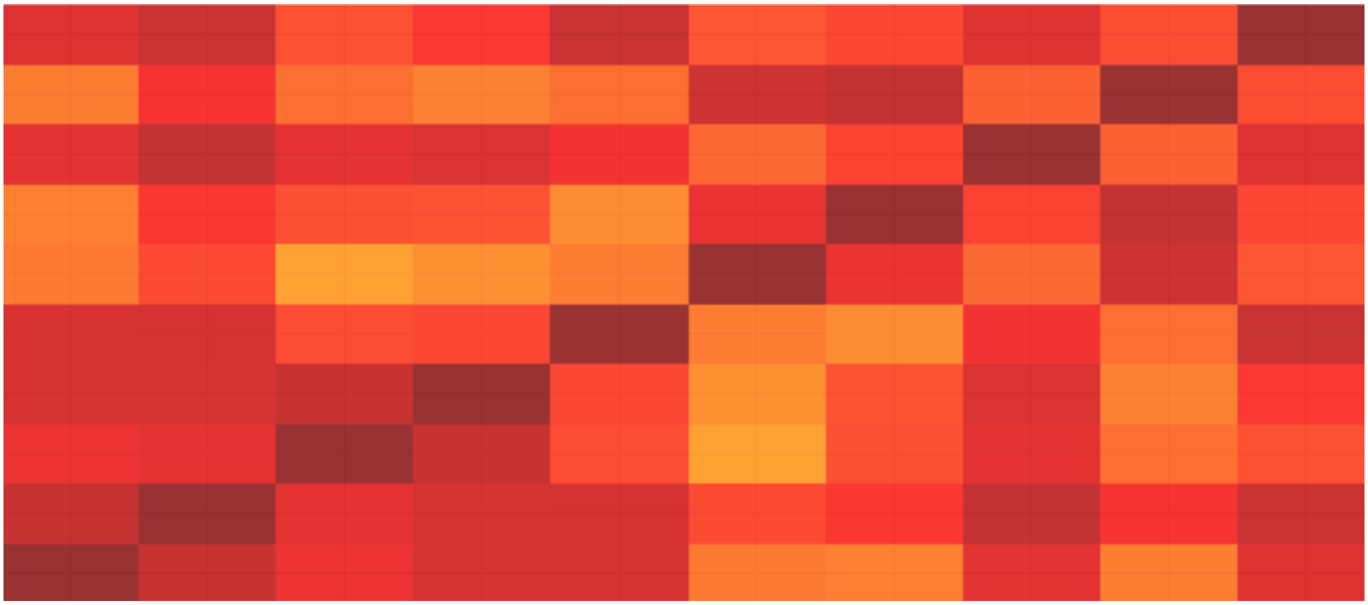
# Everything you need to know about interpreting correlations

Not all correlations are what they seem

Zakaria Jaadi Oct 15, 2019 · 8 min read



Correlation is the most widely used statistical measure to assess relationships among variables. However, correlation must be exercised cautiously; otherwise, it could lead to wrong interpretations and conclusions.

An example where correlation could be misleading, is when you are working with sample data. Because an apparent correlation in a sample is not necesseraly present in the population from which the sample came from and might be only due to chance

Also, while interpreting a relationship, one should be careful to not confound correlation and causality, because although a correlation demonstrates that a relationship exists between two variables, it does not automatically imply that one causes the other (cause-and-effect relationship).

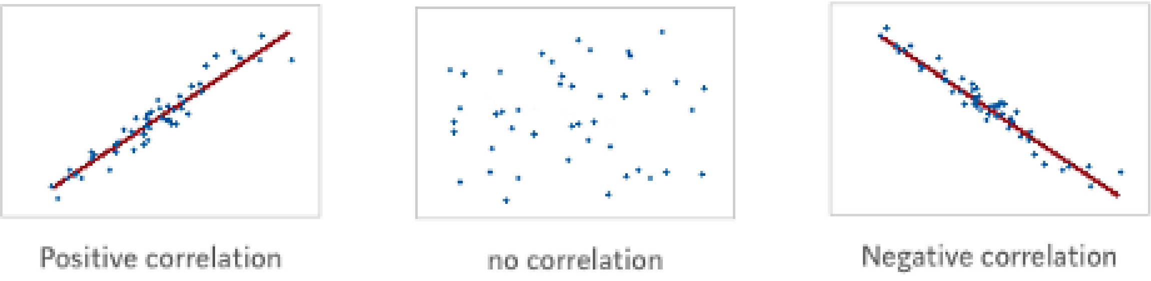
This post will define correlation, types of correlation, explain how to measure correlation using correlation coefficient, and especially how to assess the reliability of a linear correlation using a significance test. If you are familiar with correlation, you can skip the introduction.

## 1 — Introduction to correlation

Correlation is a statistical measure that describes how two variables are related and indicates that as one variable changes in value, the other variable tends to change in a specific direction. We can therefore pinpoint some real life correlations as income & expenditure, supply & demand, absence & grades decrease…etc.

Every correlation has a sign and a form, the sign could be positive, negative or neutral :

- **Positive correlation** : the two variables move in the same direction (i.e., one variable increases as the other increases. Or, one decreases as the other decreases).
- **Negative correlation** : the two variables move in opposite directions (i.e., one variable increases as the other decreases, and vice versa)
- **Neutral correlation** : the two variables show no relationship to one another.

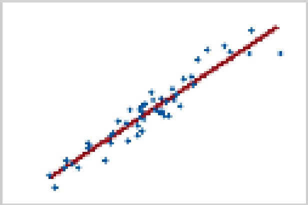


Concerning the form of a correlation , it could be linear, non-linear, or monotonic :

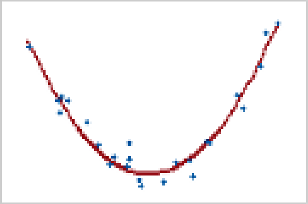
straight line).

- **Non-Linear correlation** : A correlation is non-linear when two variables don't change at a constant rate. In this case the relationship between the variables does not graph as a straight line, but as a curved pattern (parabola, hyperbola … etc).

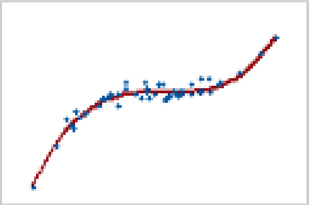
- **Monotonic correlation** : In a monotonic relationship, the variables tend to move in the same relative direction, but not necessarily at a constant rate. So all linear correlations are monotonic but the opposite is not always true, because we can have also monotonic non-linear relationships.



Linear correlation (Monotonic & Positive)



Non linear correlation



Monotonic non-linear correlation

## 2 — Correlation Coefficient

As we can see in the pictures above, drawing a scatter plot is very useful to eyeball the correlations that might exist between variables. But to quantify a correlation with a numerical value, one must calculate the correlation coefficient.

There are several types of correlation coefficients but the one that is most common is the Pearson correlation ***r***. It is a parametric test that is only recommended when the variables are normally distributed and the relationship between them is linear. Otherwise, non-parametric Kendall and Spearman correlation tests should be used.

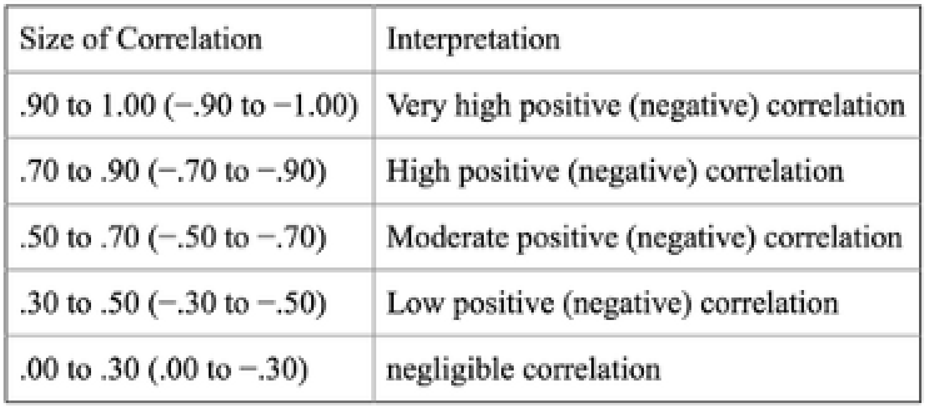
### Pearson's correlation coefficient

Pearson correlation (*r*) is used to measure strength and direction of a linear relationship between two variables. Mathematically this can be done by dividing the covariance of the two variables by the product of their standard deviations.

$$r = r_{xy} = \frac{cov(x, y)}{S_x * S_y}$$

The value of ***r*** ranges between -1 and 1. A correlation of -1 shows a perfect negative correlation, while a correlation of 1 shows a perfect positive correlation. A correlation of 0 shows no relationship between the movement of the two variables.

The table below demonstrates how to interpret the size (strength) of a correlation coefficient.



| Size of Correlation | Interpretation |
|---|---|
| .90 to 1.00 (−.90 to −1.00) | Very high positive (negative) correlation |
| .70 to .90 (−.70 to −.90) | High positive (negative) correlation |
| .50 to .70 (−.50 to −.70) | Moderate positive (negative) correlation |
| .30 to .50 (−.30 to −.50) | Low positive (negative) correlation |
| .00 to .30 (.00 to −.30) | negligible correlation |

credits : Parvez Ahammad

## 3 — Significance test

Quantifying a relationship between two variables using the correlation coefficient only tells half the story, because it measures the strength of a relationship in samples only. If we obtained a different sample, we would obtain different ***r*** values, and therefore potentially different conclusions.

So we want to draw conclusion about populations not just samples. To do so, we have to conduct a **statistical significance test**. The significance test tells us whether or not what we observe in the sample is expected to be true in the population, and can be conducted through a **hypothesis test**.

Hypothesis testing is a core part of what is known as statistical inference. Stastical inference is concerned with making inferences about a population based on a sample of the poplulation.

Before jumping into the hypothesis test, let’s sum up the above in the following formualtion.

- Say we have an n sized sample data with two variables x and y.
- The sample correlation coefficient (r) between x and y is **known** (can be computed using the formula above)
- The population correlation coefficient ρ (the greek letter “rho”) between x and y is **unknown** (because we only have sample data)
- **Goal**: We want to make an inference about the value of ρ based on r

### Performing the hypothesis test step by step

The hypothesis test will let us infer whether the value of the population correlation coefficient *ρ* is close to 0 or significantly different from 0. We decide this based on the sample correlation coefficient *r* and the sample size *n*.

- **ρ close to 0** : means there is not a significant linear correlation between x and y in the population.
- **ρ significantly different from 0 :** means there is a significant correlation between x and y in the population.

If the test shows that the population correlation coefficient *ρ is* close to zero, then we say there is insufficient statistical evidence that the correlation between the two variables is significant, i.e., the correlation occurred on account of chance coincidence in the sample and it’s not present in the entire population.

So without further ado, let’s see how we can run the test :

### Step 1: Hypotheses specification

We start by specifying the null and alternative hypotheses:

The alternative hypothesis is always what we are trying to prove, in our case, we try to prove that there is a significant correlation between x and y in the population (i.e. ρ ≠ *0)*.

linear correlation between x and y in the population (i.e. ρ = *0)*

- *Null hypothesis Ho:* ρ = *0*
- *Alternative hypothesis Ha*: ρ ≠ *0*

**Step 2: T-test**

T-test also called as **Student's T-test** is an inferential statistic that allows to test an assumption applicable to a population, or simply, it allows to use sample data to generalize an assumption to an entire population. In our case, it will help us find out if the sample correlation between x and y is repeatable for the entire population.

We calculate the value of the t-test using the following formula:

$$t = \frac{r * \sqrt{n-2}}{\sqrt{1-r^2}}$$

with :

- n is the sample size
- r is the sample correlation coefficient

The bigger the t-value, the more likely it is that the correlation is repeatable. but how big is "big enough" ? that's the job of the next step

**Step 3: P-value**

Every t-value has a p-value to go with it. A p-value is the probability that the null hypothesis is true. In our case, it represents the probability that the correlation between x and y in the sample data occurred by chance.

A p-value of 0.05 means that there is only 5% chance that results from your sample occurred due to chance. A p-value of 0.01 means that there is only 1% chance. So lower p-values are good, but how lower is "lower enough" ?.

to 0.05 (α =0.05) and find the P-value.

To find the p-value we need two things, the t-test value (from step2) and the number of degrees of freedom that can be computed as follows df=n-2 (with n is the size of the sample). Having these two values we can compute the p-value by:

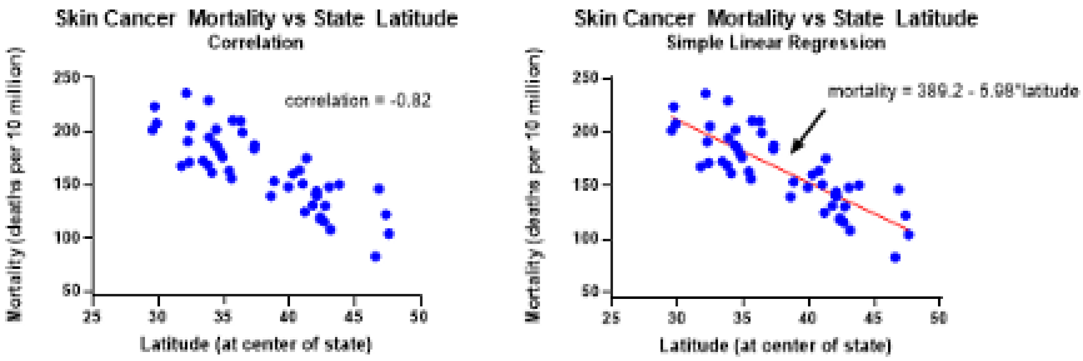
- Using a software
- Looking it up through the t-table

**Step 4: Decision**

Finally, we make a decision:

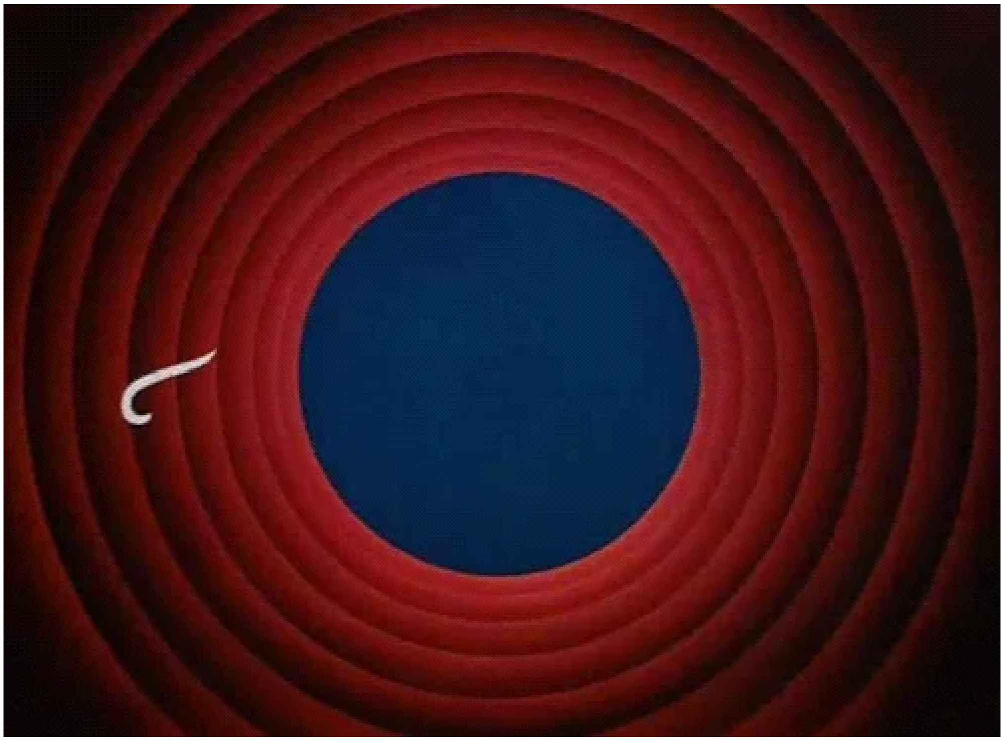
- If the *P*-value is smaller than the significance level (α =0.05), we REJECT the null hypothesis in favor of the alternative. We conclude that the correlation is **statically significant**. or in simple words “ we conclude that there is a linear relationship between x and y in the population at the α level ”
- If the *P*-value is bigger than the significance level (α =0.05), we fail to reject the null hypothesis. We conclude that the correlation is **not statically significant**. Or in other words “we conclude that there is not a significant linear correlation between x and y in the population”

## 3 — Correlation vs Regression



Credits: GraphPad

Correlation is a statistical measure that quantifies the direction and strength of the relationship between two numeric variables. On the other hand, Regression, is a statistical technique that predicts the value of the dependent variable Y based on the known value of the independent variable X through an equation of the form Y = a + bX.



• • •

## References :

- [**mintab.com**] : Linear, nonlinear, and monotonic relationships
- [**opentextbooks**]: Testing the Significance of the Correlation Coefficient
- [**janda.org**]: Significance of the Correlation Coefficient

Data Science Statistics Machine Learning AI Data