BASICS THAT EVERYONE IN THE FIELD OF DATA SCIENCE SHOULD KNOW

Clearly explained: Pearson V/S Spearman Correlation Coefficient

Learn more about WHEN to use which coefficient in this post

Juhi Ramzai
Jun 25 · 5 min read ★

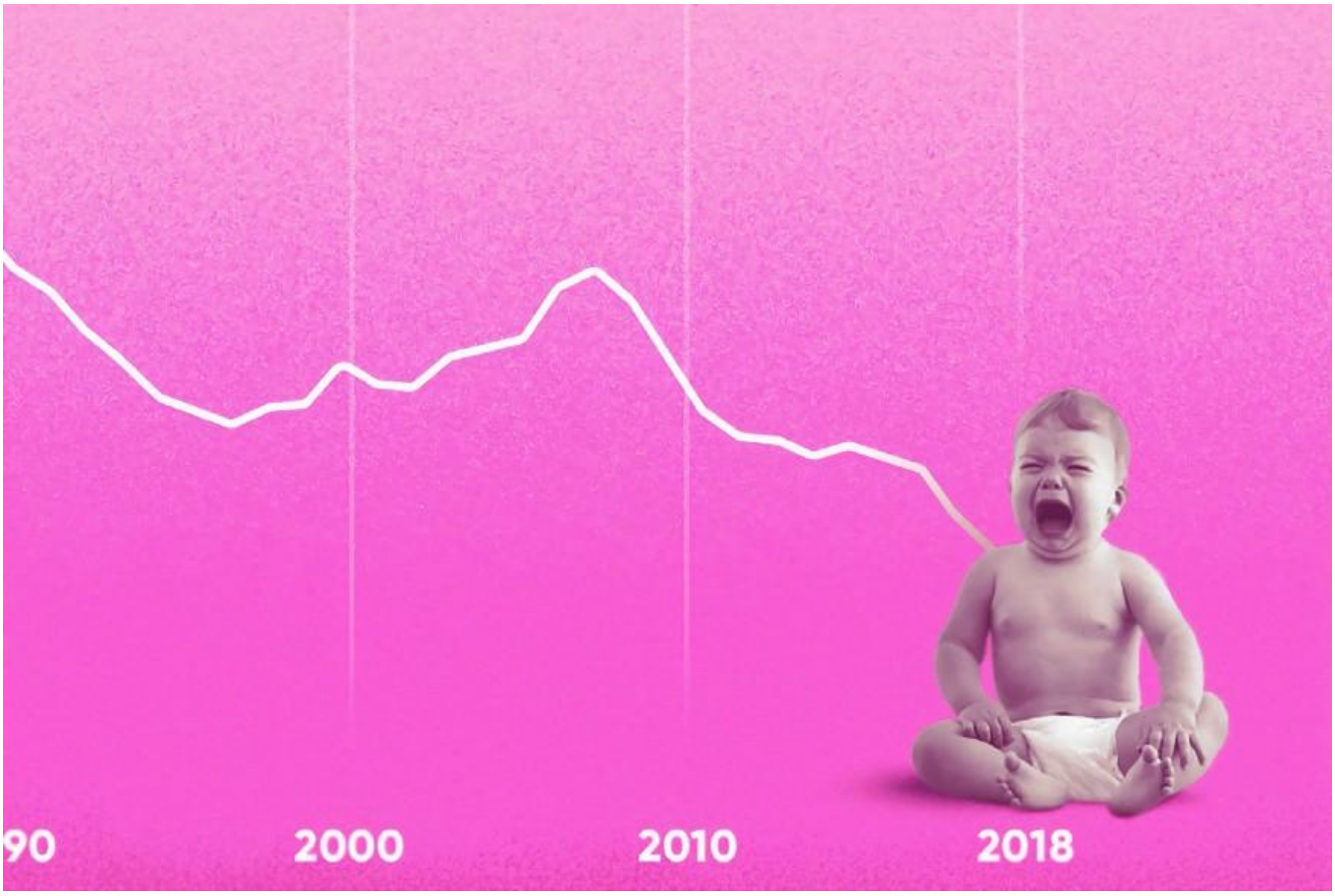


Photo by Morning Brew on Unsplash

I recently came across a scenario where I educated myself about the difference between the Pearson and Spearman correlation coefficient. I felt that is one piece of

information that a lot of people in the data science fraternity on the medium can make use of. I'll explain thoroughly the difference between the two and the exact scenarios where the use of each one is suitable. Read on!

Contents of this post:

1. **Definition of Correlation**
2. **Comparative analysis between Pearson and Spearman correlation coefficients**

. . .

Definition of Correlation

Correlation is the degree to which two variables are linearly related. This is an important step in bi-variate data analysis. In the broadest sense **correlation** is actually any statistical relationship, whether causal or not, between two random variables in bivariate data.

> *An important rule to remember is that Correlation doesn't imply causation*

Let's understand through two examples as to what it actually implies.

1. The consumption of ice-cream increases during the summer months. There is a strong correlation between the sales of ice-cream units. In this particular example, we see there is a causal relationship also as the extreme summers do push the sale of ice-creams up.
2. Ice-creams sales also have a strong correlation with shark attacks. Now as we can see very clearly here, the shark attacks are most definitely not caused due to ice-creams. So, there is no causation here.

Hence, we can understand that the correlation doesn't ALWAYS imply causation!

What is the Correlation Coefficient?

The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A correlation of -1.0 shows a perfect negative correlation, while a correlation of

1.0 shows a perfect positive correlation. A correlation of 0.0 shows no linear relationship between the movement of the two variables.

. . .

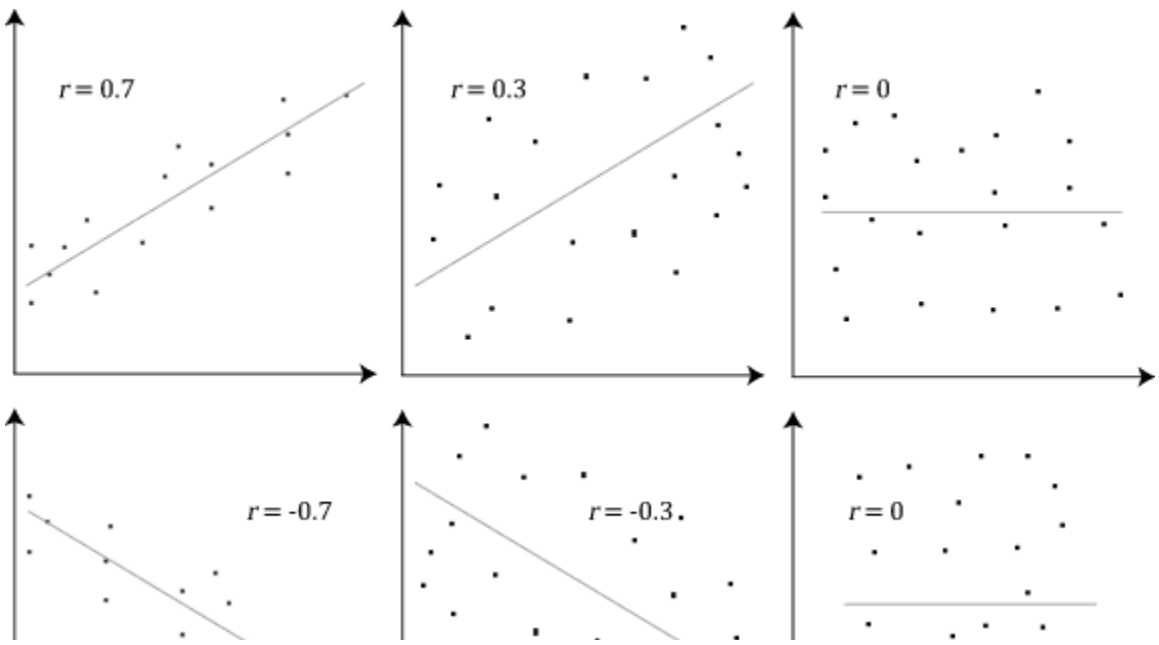
2 Important Correlation Coefficients — Pearson & Spearman

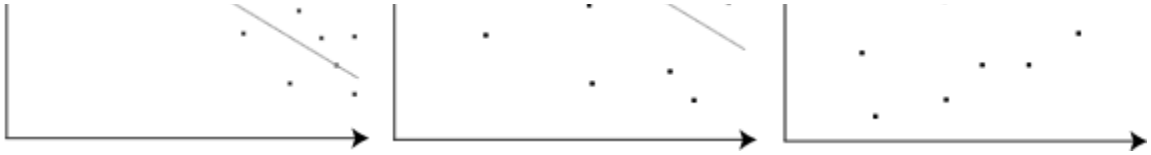
1. Pearson Correlation Coefficient

Wikipedia Definition: In statistics, the Pearson correlation coefficient also referred to as Pearson's *r* or the bivariate correlation is a statistic that measures the **linear correlation** between two variables *X* and *Y*. It has a value between +1 and −1. A value of +1 is a total positive linear correlation, 0 is no linear correlation, and −1 is a total negative linear correlation.

> ***Important Inference to keep in mind:*** *The Pearson correlation can evaluate ONLY a linear relationship between two continuous variables (A relationship is linear only when a change in one variable is associated with a proportional change in the other variable)*
>
> ***Example use case:*** *We can use the Pearson correlation to evaluate whether an increase in age leads to an increase in blood pressure.*

Below is an example of how the Pearson correlation coefficient (r) varies with the **strength and the direction of the relationship** between the two variables. Note that when no linear relationship could be established (refer to graphs in the third column), the Pearson coefficient yields a value of zero.





Source: Wikipedia

2. Spearman Correlation Coefficient

Wikipedia Definition: In statistics, Spearman's rank correlation coefficient or Spearman's ρ, named after Charles Spearman is a **nonparametric measure of rank correlation** (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a **monotonic function**.

> ***Important Inference to keep in mind:*** *The Spearman correlation can evaluate a monotonic relationship between two variables — Continous or Ordinal and it is based on the ranked values for each variable rather than the raw data.*

What is a monotonic relationship?

A monotonic relationship is a relationship that does one of the following:

(1) as the value of one variable increases, so does the value of the other variable, OR,

(2) as the value of one variable increases, the other variable value decreases.

BUT, not exactly at a constant rate whereas in a linear relationship the rate of increase/decrease is constant.

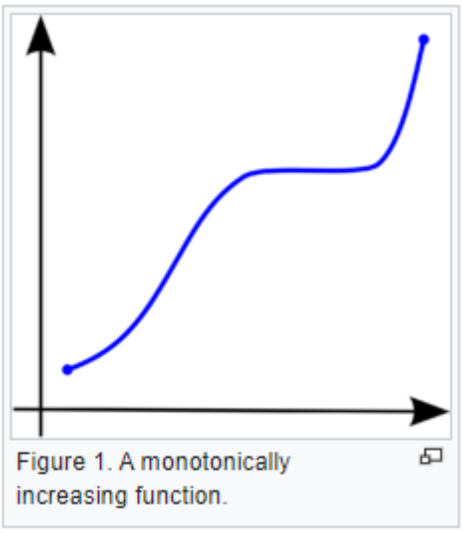


Figure 1. A monotonically increasing function.

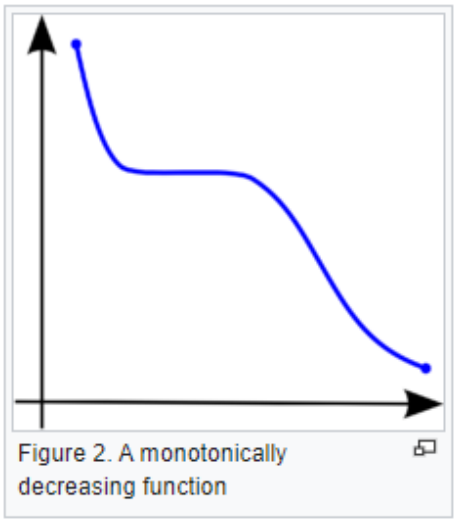


Figure 2. A monotonically decreasing function

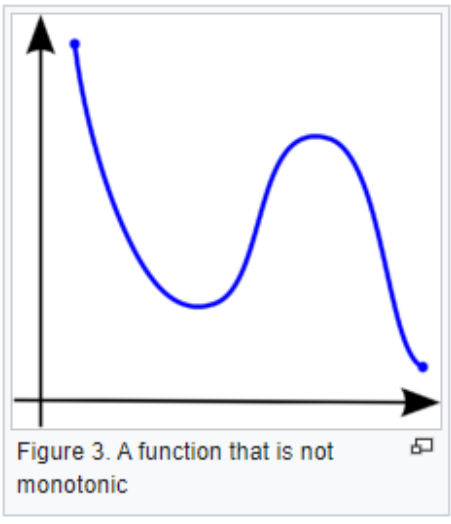


Figure 3. A function that is not monotonic

Source: Wikipedia

> ***Example use case:*** *Whether the order in which employees complete a test exercise is related to the number of months they have been employed or correlation between the IQ of a person with the number of hours spent in front of TV per week*

Comparison of Pearson and Spearman coefficients

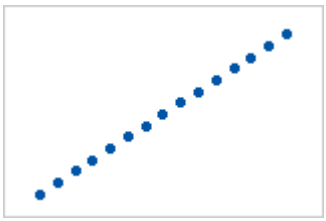
1. The fundamental difference between the two correlation coefficients is that the Pearson coefficient works with a linear relationship between the two variables whereas the Spearman Coefficient works with monotonic relationships as well.

2. One more difference is that Pearson works with raw data values of the variables whereas Spearman works with rank-ordered variables.

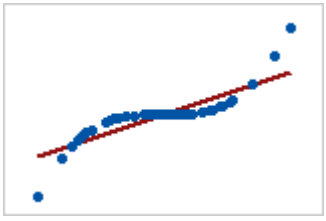
Now, if we feel that a scatterplot is visually indicating a "might be monotonic, might be linear" relationship, our best bet would be to apply Spearman and not Pearson. No harm would be done by switching to Spearman even if the data turned out to be perfectly linear. But, if it's not exactly linear and we use Pearson's coefficient then we'll miss out on the information that Spearman could capture.

Let's look at some examples which I found to be informative from this website:

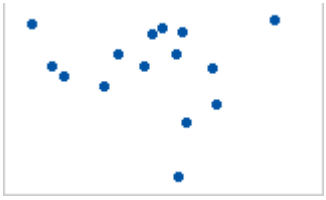
1. Pearson = +1, Spearman = +1



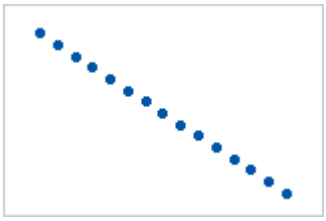
2. Pearson = +0.851, Spearman = +1 (This is a monotonically increasing relationship, thus Spearman is exactly 1)



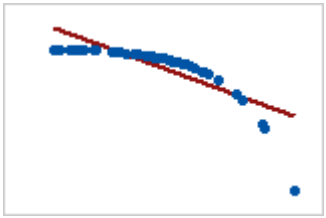
3. Pearson = −0.093, Spearman = −0.093



4. Pearson $= -1$, Spearman $= -1$



5. Pearson $= -0.799$, Spearman $= -1$ (This is a monotonically decreasing relationship, thus Spearman is exactly 1)



NOTE: Both of these coefficients cannot capture any other kind of non-linear relationships. Thus, if a scatterplot indicates a relationship that cannot be expressed by a linear or monotonic function, then both of these coefficients must not be used to determine the strength of the relationship between the variables.

. . .

Watch this space for more on Data Science, Machine Learning, and Statistics.

Happy Learning:)