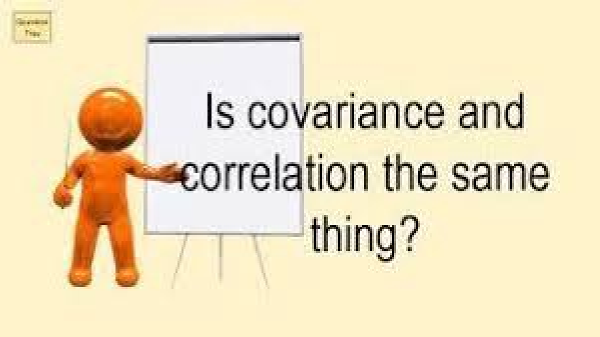
Getting the Basics of Correlation & Covariance

Seema Singh Follow
Feb 24, 2019 · 6 min read ★

Correlation is one of the widely used statistical concepts. This blog post tries to answer what correlation is, why it is so helpful, what relationship correlation and covariance share and some of ways to calculate correlation.



What is Correlation?

Correlation, statistical technique which determines how one variables moves/changes in relation with the other variable. It gives us the idea about the degree of the relationship of the two variables. It's a bi-variate analysis measure which describes the association between different variables. In most of the business it's useful to express one subject in terms of its relationship with others.

For example: Sales might increase if lot of money is spent on product marketing.

Why it is useful?

1. If two variables are closely correlated, then we can predict one variable from the other.

2. Correlation plays a vital role in locating the important variables on which other variables depend.

3. It’s used as the foundation for various modeling techniques.

4. Proper correlation analysis leads to better understanding of data.

5. Correlation contribute towards the understanding of causal relationship(if any).

Relationship of Correlation and Covariance

Before diving more into correlation, let’s get the understanding of covariance.

Covariance: The prefix ‘Co’ defines some kind of joint action and variance refers to the change or variation. So it says, two variables are related based on how these variables change in relation with each other.

But wait, is covariance same as correlation?

As covariance says something on same lines as correlation, correlation takes a step further than covariance and also tells us about the strength of the relationship.

Both can be positive or negative. Covariance is positive if one increases other also increases and negative if one increases other decreases.

Covariance is calculated as

Covariance formula

X_i= Observation point of variable X

$\bar{x}$= Mean of all observations(X)

Y_i= Observation point of variable Y

$\bar{y}$ = Mean of all observations(Y)

n= Number of observations

Decoding the covariance formula: Covariance between two variables x and y is the sum of the products of the differences of each item and their respective means divided by the number of items in the dataset minus one..

Getting better understanding with a simple example of sample data:

Following data shows the number of customers with their corresponding temperature.

Example to understand correlation and covariance

First find means of both the variables, subtract each of the item with its respective mean and multiply it together as follows

Mean of X, $\bar{x}$ = (97+86+89+84+94+74)/6 = 524/6= 87.333

Mean of Y, $\bar{Y}$ = (14+11+9+9+15+7)/6 = 65/6= 10.833

$COV(x, y) = 112.33/(6–1) = 112.33/5 = 22.46$

The covariance between the temperature and customers is 22.46. Since the covariance is positive, temperature and number of customers have a positive relationship. As temperature rises, so does the number of customers.

But here there is no information about how strong the relationship is, and that's where correlation comes into the picture.

Correlation coefficient is the term used to refer the result of any correlation measurement methods.

So here, the sample Correlation coefficient is calculated as

Correlation formula

$COV(x, y)$ = covariance of the variables x and y
σx = sample standard deviation of variable x
σy = sample standard deviation of variable y

COV(x, y) = 22.46

σx = 331.28/5=66.25= 8.13

σy = 48.78/5=9.75=3.1

correlation = 22.46/(8.13x 3.1)= 22.46/25.20 =0.8

0.8 shows that strength of the correlation between temperature and number of customers is very strong.

Sample correlation coefficient can be used to estimate the population correlation coefficient.

Different methods exist to calculate correlation coefficient between two subjects. Some of the methods are:

1. Pearson Correlation Coefficient

It captures the strength and direction of the linear association between two continuous variables. It tries to draw the line of best fit through the data points of two variables. Pearson correlation coefficient indicates how far these data points are away from the line of best fit. The relationship is linear only when the change in one variable is proportional to the change in another variable.

Pearson Correlation Coefficient calculated as

r = Pearson Correlation Coefficient

n = number of observations

∑xy = sum of the products of x and y values

∑x = sum of x values

∑y = sum of y values

∑x2= sum of squared x values

∑y2= sum of squared y values

Open in Colab (https://colab.research.google.com/gist/seema200/d351baa7893a21dda4e8ccc3e712671d/untitled3.ipynb)

In [0]:

```
import pandas as pd
import numpy as np

np.random.seed(50)

a = pd.DataFrame({'A':np.random.randn(10),
                  'B':np.random.randn(10),
                  'C':np.random.randn(10),
                  'D':np.random.randn(10),
                  'E':np.random.randn(10)})

b = pd.DataFrame({'A':np.random.randn(10)})
```

In [3]:

```
a.head()
```

Out[3]:

	A	B	C	D	E
0	-1.560352	0.126338	-0.393956	1.095210	-0.200416
1	-0.030978	0.862194	0.148116	1.338409	0.743038
2	-0.620928	0.696737	-0.412234	-1.368982	0.857362

correlation.ipynb hosted with ❤ by **GitHub** **view raw**

Spearman's Correlation Coefficient

It tries to determine the strength and the direction of the monotonic relationship which exists between two ordinal or continuous variables. In a monotonic relationship two variables tend to change together but not with the constant rate. It's calculated on the ranked values of the variables rather than on the raw data.

Monotonic and non- monotonic relationships are shown below:

Spearman rank correlation coefficient

ρ= Spearman rank correlation coefficient
di= the difference between the ranks of corresponding variables
n= number of observations

Comparison: Pearson and Spearman correlation coefficient

Pearson and Spearman correlation coefficient can take values from -1 to 1.

(i) If one variable increases with the other variable at the consistent rate then Pearson coefficient would be 1, which results in a perfect line. In this case Spearman coefficient would also be 1.

(ii) If one variable increases with the other variable but not with the consistent rate then Pearson coefficient would be positive but less than 1. In this case Spearman coefficient would be still 1.

(iii) If the relationship is random then both the coefficients would be near 0.

(iv) If the relationship between the variables is a perfect line but with a decreasing relationship then both the coefficients would be -1.

(v) If the relationship between two variables is such that one variable decreases when the other increases but not with the consistent rate, then Pearson coefficient would be negative but greater than -1. Spearman coefficient would be -1 in this case.

When to use what?

Pearson correlation describes linear relationships and spearman correlation describes monotonic relationships. A scatter plot would be helpful to visualize the data and understand which correlation coefficient should be used. Other way of doing is to apply both the methods and check which is performing well. For instance if results show spearman correlation coefficient is greater than Pearson coefficient, it means our data has monotonic relationships and not linear.

Also, correlation does not imply causation. Read here why.

More Reads:

1. https://365datascience.com/covariance-linear-correlation-coefficient/#close
2. https://www.wallstreetmojo.com/correlation-vs-covariance/

. . .

Thanks for reading!

Data Science Correlation Causation Statistics Science