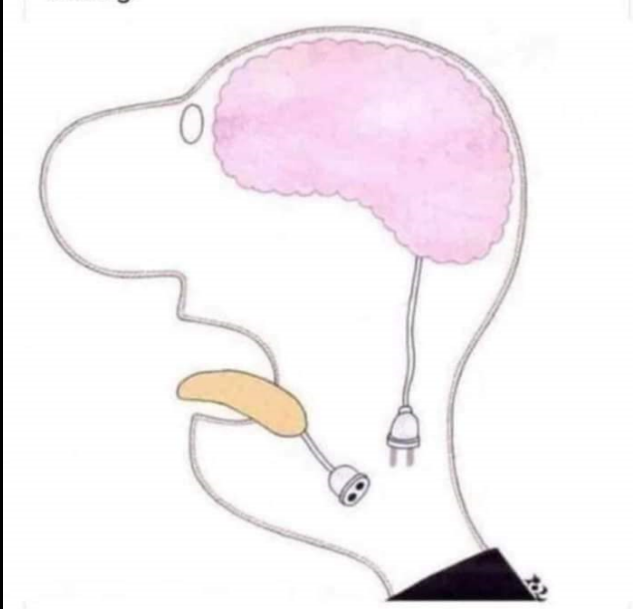
**Correlation & Covariance**

Mind: Even though I had good thoughts but you speak bad words no one values.

Tongue: Even though I speak good words but if you think it in wrong sense no one values.



Mind: So the only solution for this is we should make sure we are “Correlated” before you start talking to avoid all miscommunication and misunderstandings.

Tongue: What do you mean by **Correlation**???

Mind: a mutual relationship or connection between two or more things.

For more info**:**

This blog explains about most important and commonly used statistical topic correlation. We are going to learn about what is meant by covariance and difference between correlation and covariance. Below given index are the learning objectives of this blog.

**Learning Objectives from this blog:**

* What is Correlation?
* What is Covariance?
* What is the difference between Correlation and Covariance?
* What is the relationship between the Correlation and Covariance?

**Correlation:**

Correlation can be defined as a statistical tool that defines the relationship between two variables. For example, correlation may be used to define the relationship between the price of a product and its quantity demanded. It describes the relation between two variables but do not describe the cause-effect association. It only gives an understanding as to the direction and intensity of relation between two variables. Correlation can be of two types:

#### 1) Positive Correlation

Two variables are positively correlated when they move together in the same direction. For example quantity supplied increases as the price increases. This is because sellers find it profitable to sell when the prices are high, so they will sell more. Thus, we can call price and quantity supplied to be positively correlated.

#### 2) Negative Correlation

Two variables are negatively correlated if they move in opposite directions. For instance, as the price of increases, the quantity demanded declines as the product becomes more expensive relative to when the price had not increased. Thus, we can say that price and quantity demanded are negatively correlated.

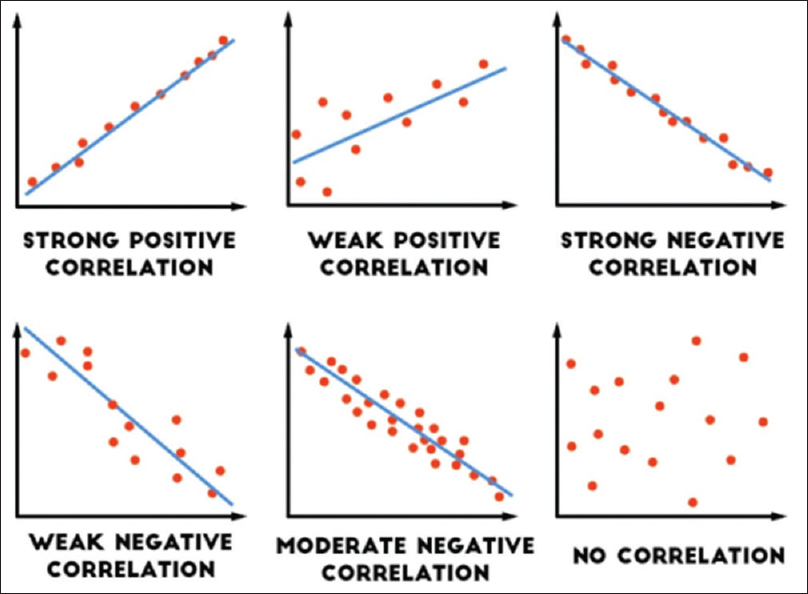
## ****Correlation Estimation****

There are different ways of calculating or estimating the correlation between variables.

### 1] Scatter Diagram

This is a diagrammatic method of correlation estimation. Here, the x-values are depicted on the horizontal axis and y-values are on the vertical axis. The degree of closeness between the coordinates would indicate the correlation. If the trend is downward sloping from right to left, it means negative correlation.

If it is upward sloping from left to right, it is upward sloping. And if the points are scattered all around with no trend, then the variables are said to be uncorrelated*.*



### 2] Karl Pearson’s Correlation Coefficient

This is a quantitative method of correlation estimation. The formula is:

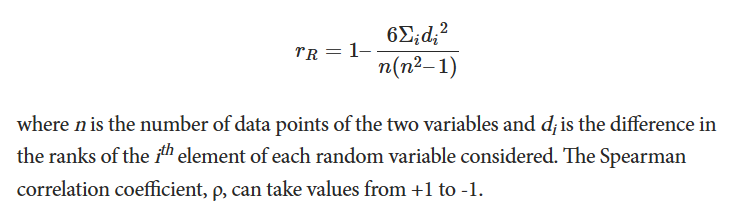
**r = cov(x,y) / σx . σy**

where, cov(x,y) is the covariance among the two variables. And, σx and σy are the standard deviations of x and y variables, respectively. Properties of ‘r’:

* It has no unit.
* It lies between -1 and +1, which indicates perfect negative or perfect positive correlation, respectively.
* A value of zero means no correlation.
* A high value of ‘r’ indicates strong linear relationship, and vice versa.
* A positive value indicates positive correlation.
* The value of ‘r’ is unaffected by a change of origin or change of scale.

### 3] Spearman’s Rank Correlation

This method is a non-parametric method of correlation estimation. It is used to determine the correlation between the ranks of different variables. These variables are usually qualitative in nature, such as beauty, honesty, intelligence, wisdom, etc. Ranks such as first, second, third, etc. are assigned and then the correlation is calculated using a formula.



Where n is the number of data points of the two variables and di is the difference in the ranks of the ith element of each random variable considered. The Spearman correlation coefficient, ῤ, values range from +1 to -1.

thht

### Degree of Correlation

|  |  |
| --- | --- |
| +1 | Perfect Positive |
| -1 | Perfect Negative |
| 0 | Uncorrelated |

In other cases (positive or negative), if the value of ‘r’ is 0.50, it is called moderate correlation. When it lies between 0.50- 0.75, the degree of correlation is high and when it lies between 0.25 -0.50, the degree of correlation is low.

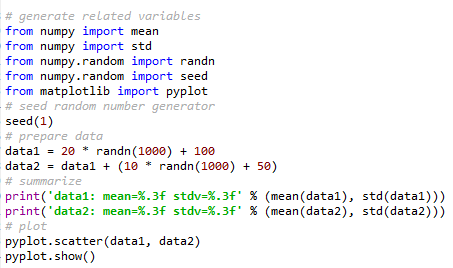
## Test Dataset

Before we look at correlation methods, let’s define a dataset we can use to test the methods.

We will generate 1,000 samples of two variables with a strong positive correlation. The first variable will be random numbers drawn from a Gaussian distribution with a mean of 100 and a standard deviation of 20. The second variable will be values from the first variable with Gaussian noise added with a mean of a 50 and a standard deviation of 10.

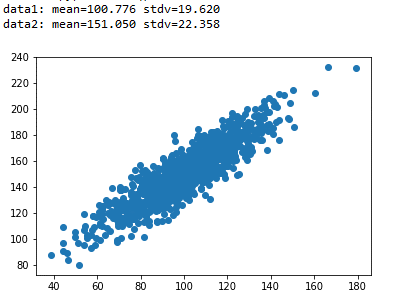
We will use the randn() function to generate random Gaussian values with a mean of 0 and a standard deviation of 1, then multiply the results by our own standard deviation and add the mean to alter the values into the preferred dimension

The pseudorandom number generator is seeded to ensure that we get the same sample of numbers each time the code is run.



Output:

A scatter plot of the two variables is created. Because we can observe a relationship among the two variables. This is clear when we review the generated scatter plot where we can see an increasing trend.



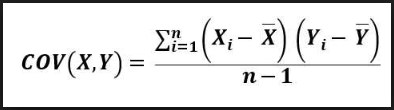
Before we look at calculating some correlation scores, we must first look at an important statistical building block, called covariance.

## Covariance

Variables can be related by a linear relationship. This is a relationship that is consistently additive across the two data samples.

This relationship can be summarized between two variables, called the covariance. It is calculated as the average of the product between the values from each sample, where the values haven been centered (had their mean subtracted).

Simple Covariance formula:



Xᵢ= Observation point of variable X

x̅= Mean of all observations(X)

Yᵢ= Observation point of variable Y

ȳ = Mean of all observations(Y)

n= Number of observations

Below we have sample dataset of ice cream sales according to the temperature

|  |  |
| --- | --- |
| **Temperature** (in centigrade) (**X)** | **Ice**-**Cream sales (Y)** |
| 28 | 2500 |
| 26 | 2200 |
| 32 | 2900 |
| 38 | 4000 |
| 31 | 2400 |

Step 1: Find the Mean of each column X and Y separately

Mean of X, x̅ = =31

Mean of Y, Ȳ = =2800

Step 2: Subtract current value with the mean value of the specific column

Covariance (X, Y) = SXY = = 2931.25

The calculation of the sample covariance is as follows:

  
The use of the mean in the calculation suggests the need for each data sample to have a Gaussian or Gaussian-like distribution.

The sign of the covariance can be interpreted as whether the two variables change in the same direction (positive) or change in different directions (negative). A covariance value of zero illustrates that the two variables are perfectly independent.

The cov() NumPy function can be used to calculate a covariance matrix between two or more variables.

  
  
The diagonal of the matrix contains the covariance between each variable and itself. The other values in the matrix represent the covariance between the two variables; in this case, the remaining two values are the same given that we are calculating the covariance for only two variables.

Correlation formula:

You can calculate the **sample correlation** **coefficient** between *X* and *Y* directly from the sample covariance with the following formula for the same dataset:

image3.png

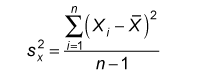
* rXY = sample correlation between X and Y
* SXY = sample covariance between X and Y = 2931.25 (from covariance calculation)
* SX = sample standard deviation of X
* SY = sample standard deviation of Y

Mean of X, x̅ = =31

Mean of Y, Ȳ = =2800

Computing sample covariance for X

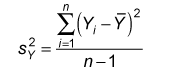
|  |  |  |
| --- | --- | --- |
| **Temperature** (in centigrade) (**X)** | **Xi - x̅** | **(Xi - x̅)** 2 |
| 28 | 28-31= -3 | (-3) 2 =9 |
| 26 | 26-31= -5 | (-5) 2 =25 |
| 32 | 32-31= 1 | 1. 2 =1 |
| 38 | 38-31= 7 | (7) 2 =49 |
| 31 | 31-31= 0 | 1. 2 =0 |
|  | **SUM** | **84** |



= = 21

Sample standard deviation of X = = 4.5

|  |  |  |
| --- | --- | --- |
| **Ice**-**Cream sales (Y)** | **Yi -** Ȳ | **(Yi -** Ȳ**)** 2 |
| 2500 | 2500-2800=300 | (300) 2 =90000 |
| 2200 | 2200-2800=600 | (600) 2 =360000 |
| 2900 | 2900-2800=100 | (100) 2 =10000 |
| 4000 | 4000-2800=1200 | (1200) 2 =1440000 |
| 2400 | 2400-2800=400 | (400) 2 =160000 |
|  | **SUM** | **2060000** |



= = 515000

Sample standard deviation of Y = = 717.63

image3.png

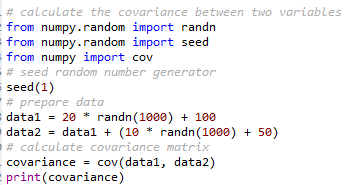
= = = 0.9

The positive result show that Temperature and icecream sales follow strong positive correlation. Which indicates that two variables show tendency to move in same direction.

The formula used to compute the sample correlation coefficient ensures that its value ranges between –1 and 1.

We can calculate the covariance matrix for the two variables in our test problem.

Example:



Output:



The covariance and covariance matrix are used widely within statistics and multivariate analysis to characterize the relationships between two or more variables.

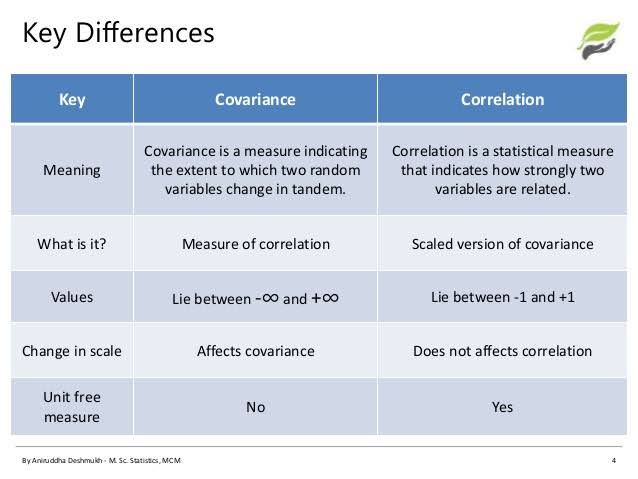
Running the example calculates and prints the covariance matrix.

Because the dataset was affected with each variable drawn from a Gaussian distribution and the variables linearly correlated, covariance is a reasonable method for describing the relationship.

The covariance between the two variables is 389.75. We can see that it is positive, suggesting the variables change in the same direction as we expect.

**Difference between covariance and correlation:**

Both conditions measure the relationship and the dependency among two variables. **Covariance** demonstrates the direction of the linear relation among variables. whereas **Correlation** demonstrates both the strength and direction of the linear relationship among two variables.



**Conclusion:**

The correlation coefficient shows how strong the linear relationship between two variables are. If the correlation is positive, that means both the variables are moving in same direction. Negative correlation indicates, when one of the variable increases the other decreases. If correlation is +/- 0.8 and above, high degree of correlation or the association between the dependent variables are strong. correlation between +/- 0.5 to+/\_0.8, sufficient degree of correlation and less than +/-0.5, weak correlation.