Saturday, May 23, 2020 11:59 AM

# Covariance vs Correlation - Do you **really** know???

*Bite Size tip to understand ,so that next time u don't get confused !!*

*Covariance and Correlation are two terms which are exactly opposite to each other, both are used in statistics and regression analysis, covariance shows us how the two variables vary from each other whereas correlation shows us the relationship between the two variables and how are they related.*

*They are the two very important terms used in the field of statistics and probability describing the degree to which two random variables or sets of random variables tend to deviate from their expected values in the similar ways.*

*Both concepts describe the linear relationship between two numerical variables.*
*When conducting experiments and analyzing data, many people often confuse with the concepts of covariance and correlation .*

*In this article, you will learn the differences between the two and how to identify one over the other.*

*Before that lets understand the nitty-gritty.*

## Variance

In statistics, variance refers to the spread of a data set. It's a measurement used to identify how far each number in the data set is from the mean.

The larger the variance, the more spread in the data set.
A large variance means that the numbers in a set are far from the mean and each other. A small variance means that the numbers are closer together in value.
Variance is the average of the squared distances from each point to the mean.

**Formula to calculate variance**

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$s^2$ = sample variance
$x_i$ = value of $i^{th}$ element
$\bar{x}$ = sample mean
$n$ = sample size

## Standard Deviation

The standard deviation is a statistic that measures the dispersion of a dataset relative to its mean.
Lower standard deviation concludes that the values are very close to their average. Whereas higher values mean the values are far from the mean value.
Its value can never be negative.

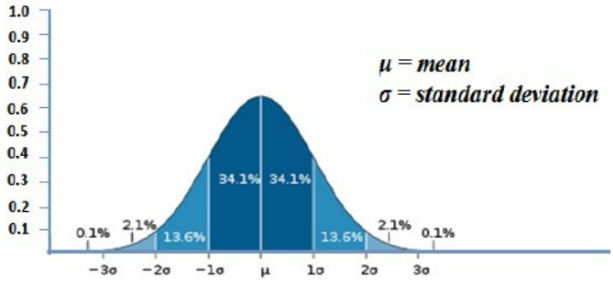
It is calculated as the square root of the variance by determining the variation between each data point relative to the mean.

**Formula to calculate Standard deviation**

$$s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{(n-1)}}$$

S = sample standard deviation
X = value of $i^{th}$ element
$\bar{X}$ = sample mean
n = sample size

- The below plot shows the normal distribution (or bell curve). Each colored band has a width of one standard deviation.
  For data that has a normal distribution, 68% of the data lies within one standard deviation of the mean



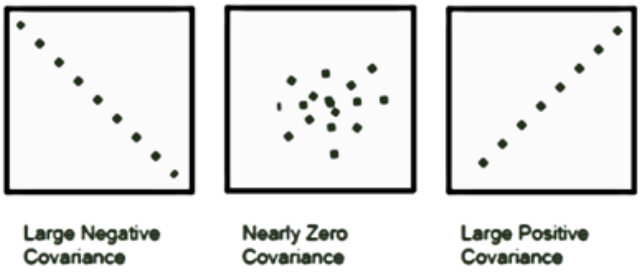

## Covariance

The covariance measures the strength of the linear relationship between two numerical variables(X and Y).
It can take any value between -infinity to +infinity, where the negative value represents the negative relationship whereas a positive value represents the positive relationship

Covariance provides insight into how two variables are related to one another. It is a measure of how much two random variables vary together. It's similar to variance, but where variance tells us how a single variable varies, Covariance tells us how two variables vary together.

A large covariance can mean a strong relationship between variables.

A positive covariance means that the two variables at hand are positively related, and they move in the same direction.
A negative covariance means that the variables are inversely related, or that they move in opposite directions.




A population includes all of the elements from a set of data. A subset of the population is called a sample.

## Population Covariance Formula

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{N}$$

## Sample covariance Formula

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-y)}{N-1}$$

**Notations in covariance formulas**

- $x_i$ = data value of x
- $y_i$ = data value of y
- $\bar{x}$ = mean of x
- $\bar{y}$ = mean of y
- N = number of data values.

# Correlation

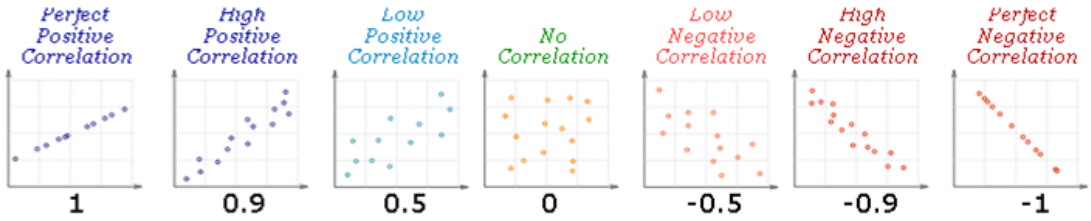
**Correlation** refers to a process for establishing whether or not relationships exist between two variables.
Whereas the strength of the linear association between two variables is quantified by the ***correlation coefficient***.

Correlation shows the relation between two variables and Correlation coefficient shows the measure of correlation. It is usually denoted by r and its value ranges from -1 for a negative correlation to +1 for positive correlation.

**The formula for computing the correlation coefficient-**

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2}\sqrt{(Y-\bar{Y})^2}}$$

Where, $\bar{X}$ - mean of X variable
$\bar{Y}$ - mean of Y variable




(Perfect in this case means that if the point were plotted in a scatter plot all the point could be connected in a straight line.)

- Correlation is **Negative** when one value **decreases** as the other increases
- Correlation is **Positive** when the values **increase** together, and
- 1 is a perfect positive correlation
- 0 is no correlation
- -1 is a perfect negative correlation

- 

## Comparison between Covariance and Correlation

| Covariance | Correlation |
|---|---|
| Covariance is used to measure the direction of relationship between two random variables. | Correlation is used to measure the strength of relationship between two random variables. |
| Covariance is affected by change in scales. | Correlation is not affected by change in scales or multiplication by a constant. |
| Covariance can take up any value from (-) infinity to (+) infinity | Correlation coeffident is a dimensionless metric and its value varies from (-1) to (+1) |
| Correlation can be deduced from covariance | Correlation provides a measure of covariance on a standard scale. It is deduced by dividing the calculated covariance with standard deviation. |
| Correlation is dimensionless, i.e. it is a unit-free measure of the relationship between variables. | In covariance the value is obtained by the product of units of the two variables. |

## Conclusion

Correlation and covariance are very closely related to each other and yet they differ a lot. Covariance defines the type of relation, but correlation defines not only the type but also the strength of this relationship.
Correlation is said to be as special case of covariance which can be obtained when the data is standardized. Now, when it comes to choose between covariance and correlation the later stands to be the first choice as it remains unaffected by the change in dimensions, location and scale, and can also be used to make a comparison between two pairs of variables.
since it is limited to a range of -1 to +1, it is useful to draw comparisons between variables across domains.
However, an important limitation is that both these concepts measure the only linear relationship.