Covariance and correlation and variance

Covariance and correlation are essential tools in various fields, such as statistics, data science, machine learning, and data analysis. They serve as useful measures for determining the relationship between two variables.
These concepts are particularly significant in artificial intelligence and machine learning, as they are frequently employed in linear regression and neural networks to model and predict the relationship between variables.

However, they have different properties and may be used in different contexts depending on the research question and the data being analyzed.

Variance- Variance is a measure of the variability of the dataset. Variability is the spread from the average of the dataset.

$$\text{Population Variance: } \mathrm{Var}(x) = \frac{\sum_{i=1}^{N}\left(X_i-\mu\right)^2}{N}$$

$$\text{Sample Variance: } \mathrm{Var}(x) = \frac{\sum_{i=1}^{n}\left(X_i-\bar{x}\right)^2}{n-1}$$

Variance can have only positive values. The higher the value of variance is the higher the variability will be of the data.

Covariance- Covariance measures how the two variables are varying together and also the degree to which the deviation of one variable X from its mean is related to the deviation of another variable Y from its mean.

$$\text{Population Covariance: } \mathrm{Cov}(x,\ y) = \frac{\sum_{i=1}^{N}\left(X_i-\mu_x\right)\left(Y_i-\mu_Y\right)}{N}$$

$$\text{Sample Covariance: } \mathrm{Cov}(x,\ y) = \frac{\sum_{i=1}^{N}\left(X_i-\bar{x}\right)\left(Y_i-\bar{y}\right)}{n-1}$$

Covariance can have positive and negative both values. The values lie between –infinite to +infinite.
Positive covariance indicates that these two variables move in the same direction and the negative covariance means that the two variables move in the opposite directions. Zero covariance means there is no relation between the variables.

Correlation- While covariance measures how the variables are varying together but correlation measures how strongly the variables are related to each other and measures both the direction and strength of the relation between the variables.

$$\text{Population Correlation: } \rho_{xy} = \frac{\mathrm{cov}(X,\ Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}} = \frac{\sum_{i=1}^{N}\left(X_i-\mu_x\right)\left(X_i-\mu_x\right)}{\sqrt{\sum_{i=1}^{N}\left(X_i-\mu_x\right)^2}\sqrt{\sum_{i=1}^{N}\left(Y_i-\mu_y\right)^2}}$$

$$\text{Sample Correlation: } r_{xy} = \frac{\mathrm{cov}\left(x,\ y\right)}{\sqrt{\mathrm{var}(x)}\sqrt{\mathrm{var}(y)}} = \frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(X_i-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}}$$

The values of correlation vary from -1 to +1.
When the value of correlation is close to 1 it means that the variables are moving by almost same exact percentage and in the same direction. When the value of correlation is close to -1 it means that the variables are moving by almost same exact percentage but in the opposite direction.

Covariance vs correlation

Understanding the relationship between two variables is a critical aspect of data analysis. Two commonly used measures of relationships are correlation and covariance.

While both provide useful insights into the relationship between two variables, they have distinct properties and may be used in different contexts.

Correlation and covariance are sensitive to outliers, so checking for outliers is important before calculating these measures.

While correlation measures the linear relationship between two variables, it may not capture the full extent of the relationship if it is not linear. In cases where the relationship is not linear, other measures, such as nonparametric correlation coefficients or nonlinear regression, may be more appropriate.

Sometimes, a high correlation coefficient may not necessarily imply causality between the two variables. The correlation only measures the association between two variables, and other factors may affect the relationship between the two variables.

Covariance and correlation can be calculated using different methods, such as raw data, deviations from the mean, or data ranks. The choice of method can affect the resulting correlation or covariance coefficient.

Example

Suppose we have two variables, X and Y, and we want to measure their relationship. We calculate the covariance and correlation coefficients and obtain the following results:

Covariance: 500
Correlation: 0.8

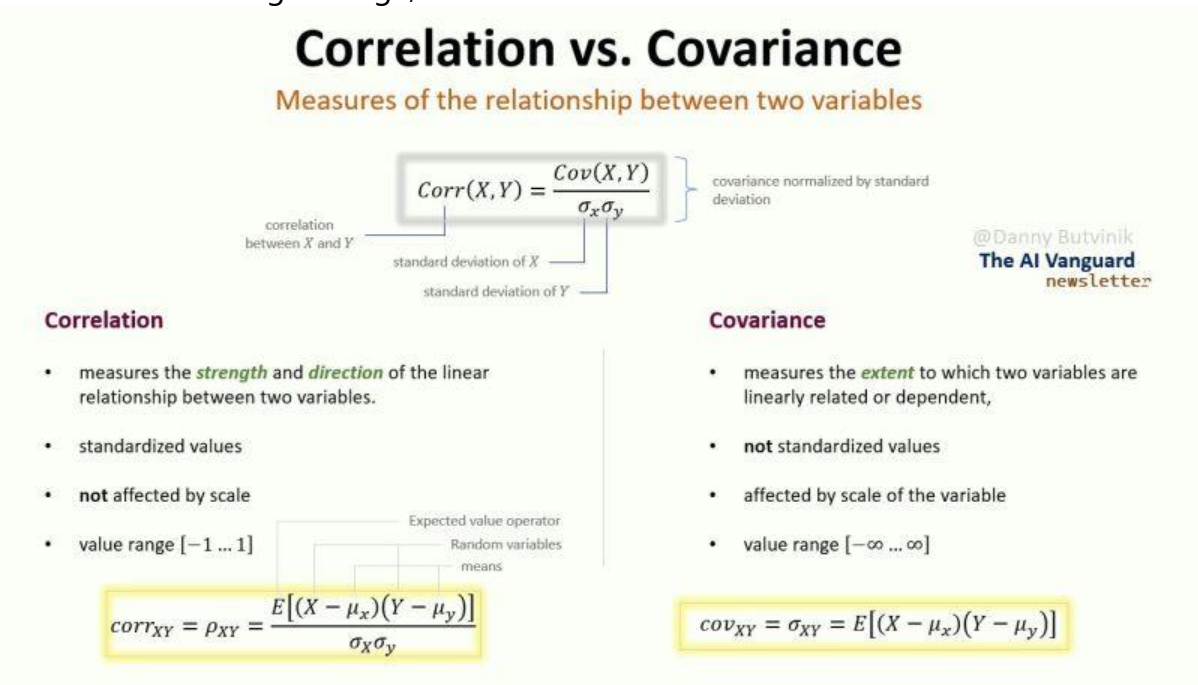
At first glance, X and Y appear to have a strong, positive relationship. However, upon further inspection, we find one outlier in the data driving the results. After removing the outlier, we recalculate the covariance and correlation coefficients and obtain the following results:

Covariance: 200
Correlation: 0.6

We can see that the correlation coefficient decreased, indicating a weaker relationship between X and Y, and the covariance decreased even more. This illustrates the importance of checking for outliers and the potential for spurious correlations when working with real-world data.

It's crucial to carefully examine the data and ensure that no outliers or other elements that might skew the results are responsible for the findings.

Activate to view larger image,



Correlation is preferred over covariance because of the following reasons:

Measurement units- Correlation is a unit free measure which takes the value from -1 to +1. It makes it easier to interpret than covariance.

Change in scale- Covariance will be affected by scaling the variables. For example, if we multiply one variable by a constant value and multiply another variable by a different constant value, then the covariance will change. However, correlation will not change in this case**.**