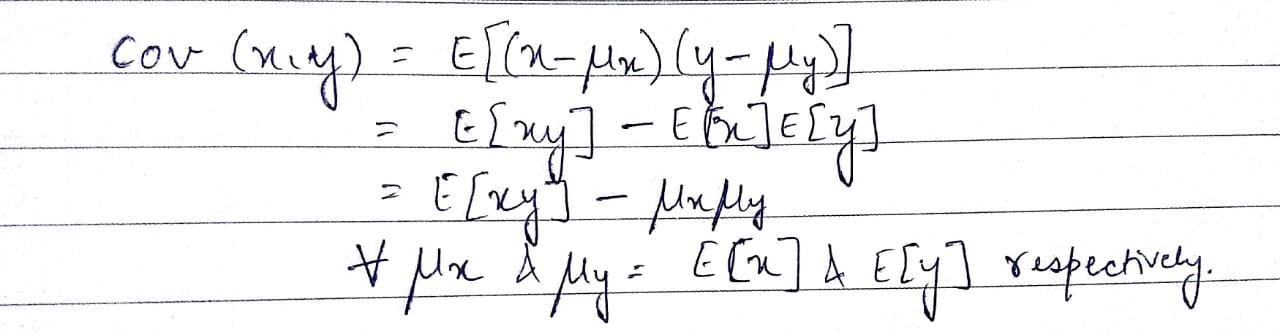
**Covariance and correlation:**

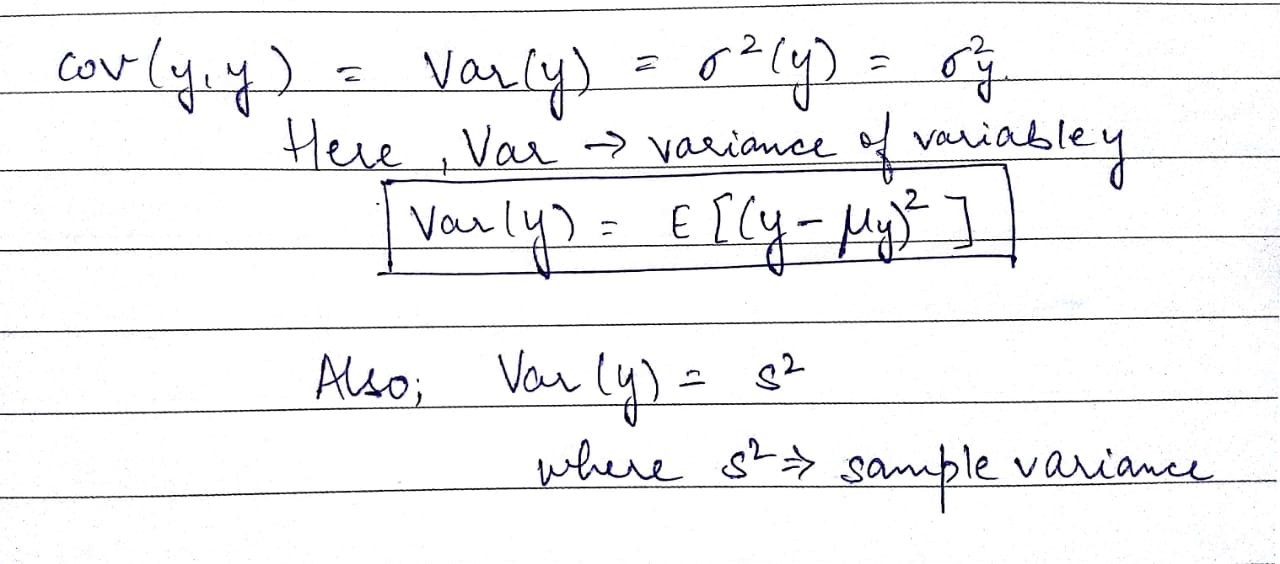
Covariance and correlation are two significantly used terms in the field of statistics and probability theory. Most articles and reading material on probability and statistics presume a basic understanding of terms like means, standard deviation, correlations, sample sizes and covariance. Let us demystify a couple of these terms today so that we can move ahead with the rest. The aim of the article is to define the terms: correlation and covariance matrices, differentiate between the two and understand the application of the two in the field of analytics and datasets.

**covariance**

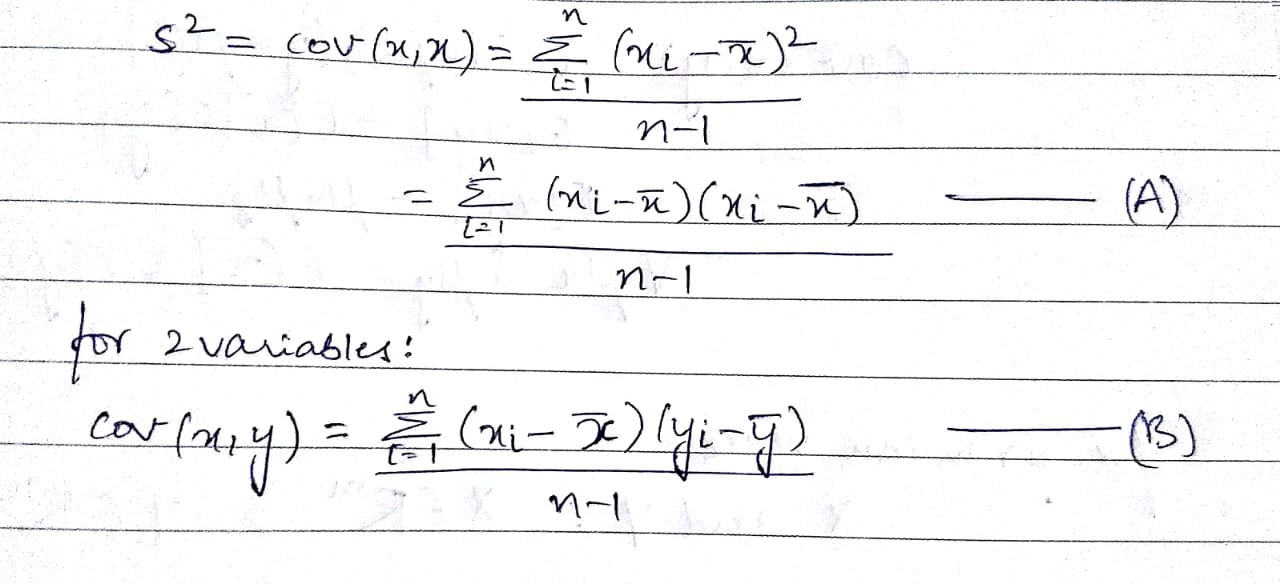
The covariance of two variables (x and y) can be represented as cov(x,y). If E[x] is the expected value or mean of a sample ‘x’, then cov(x,y) can be represented in the following way:



If we look at a single variable, say ‘y’, cov(y,y), the expression can be written in the following way:



Now as we see, in the image above, ‘s²’ or sampled variance is basically the covariance of a variable with itself. This term can also be defined in the following manner:

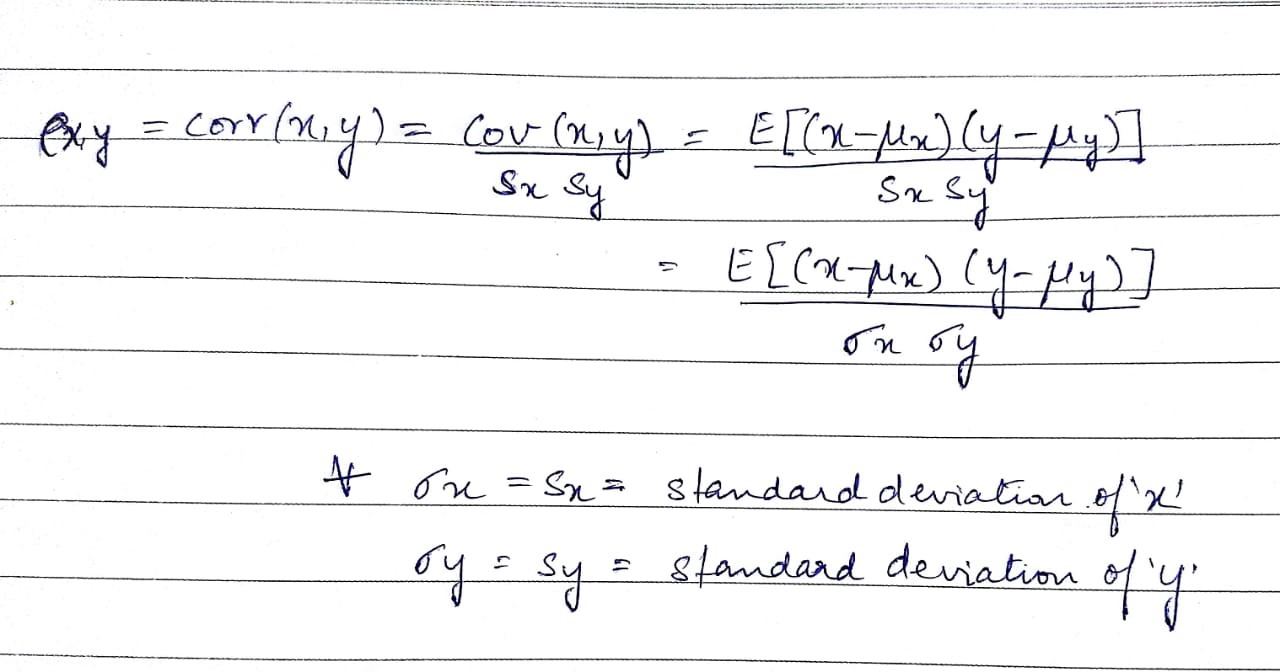


In the above formula, the numerator of the equation(A) is called the sum of squared deviations. In equation(B) with two variables x and y, it is called the sum of cross products. In the above formula, n is the number of samples in the data set. The value (n-1) indicates the degrees of freedom.

To explain what degrees of freedom are, let us just take an example. In a set of 3 numbers with the mean as 10 and two out of three variables as 5 and 15, there is only one possibility of the value that the third number can take up i.e. 10. With any set of 3 numbers with the same mean, for example: 12,8 and 10 or say 9,10 and 11, there is only one value for any 2 given values in the set. You can basically change the two values here and the third value fixes itself. The degree of freedom here is 2. Essentially, degrees of freedom is the number of independent data points that went into calculating the estimate. As we see in the example above, it is not necessarily equal to the number of items in the sample (n).

**Correlation**

The correlation coefficient is also known as the Pearson product-moment correlation coefficient, or Pearson’s correlation coefficient. As mentioned earlier, it is obtained by dividing the covariance of the two variables by the product of their standard deviations. The mathematical representation of the same can be shown in the following manner:



The values of the correlation coefficient can range from -1 to +1. The closer it is to +1 or -1, the more closely are the two variables are related. The positive sign signifies the direction of the correlation i.e. if one of the variables increases, the other variable is also supposed to increase.