Covariance

Getting the covariance of two series is a way of calculating the relationship between two variables and it will tell us how much two random variables vary together.

The formula is:

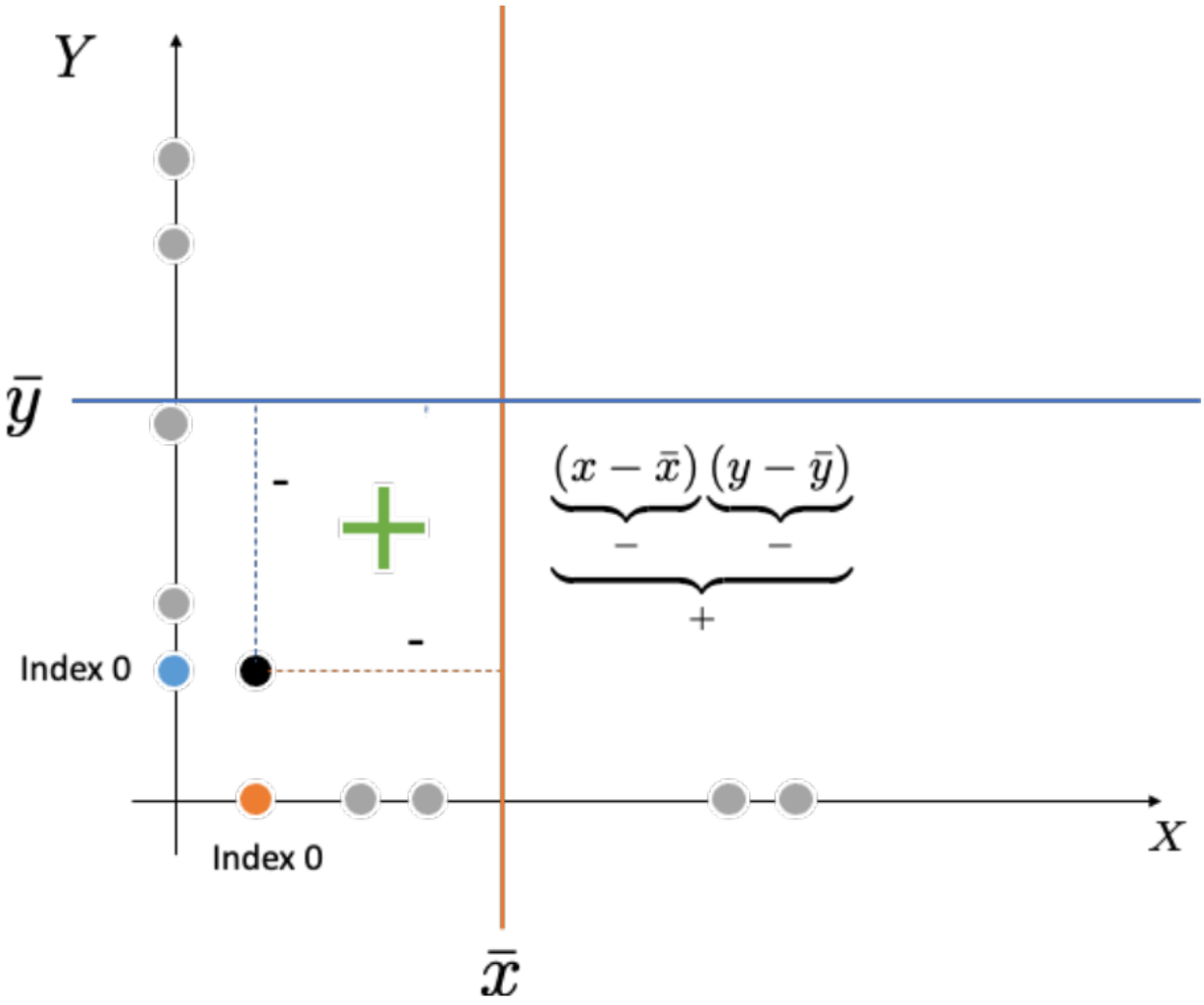
$$\operatorname{cov}(X, Y) = \sum (X - \mu_X)(Y - \mu_Y)$$

Where:

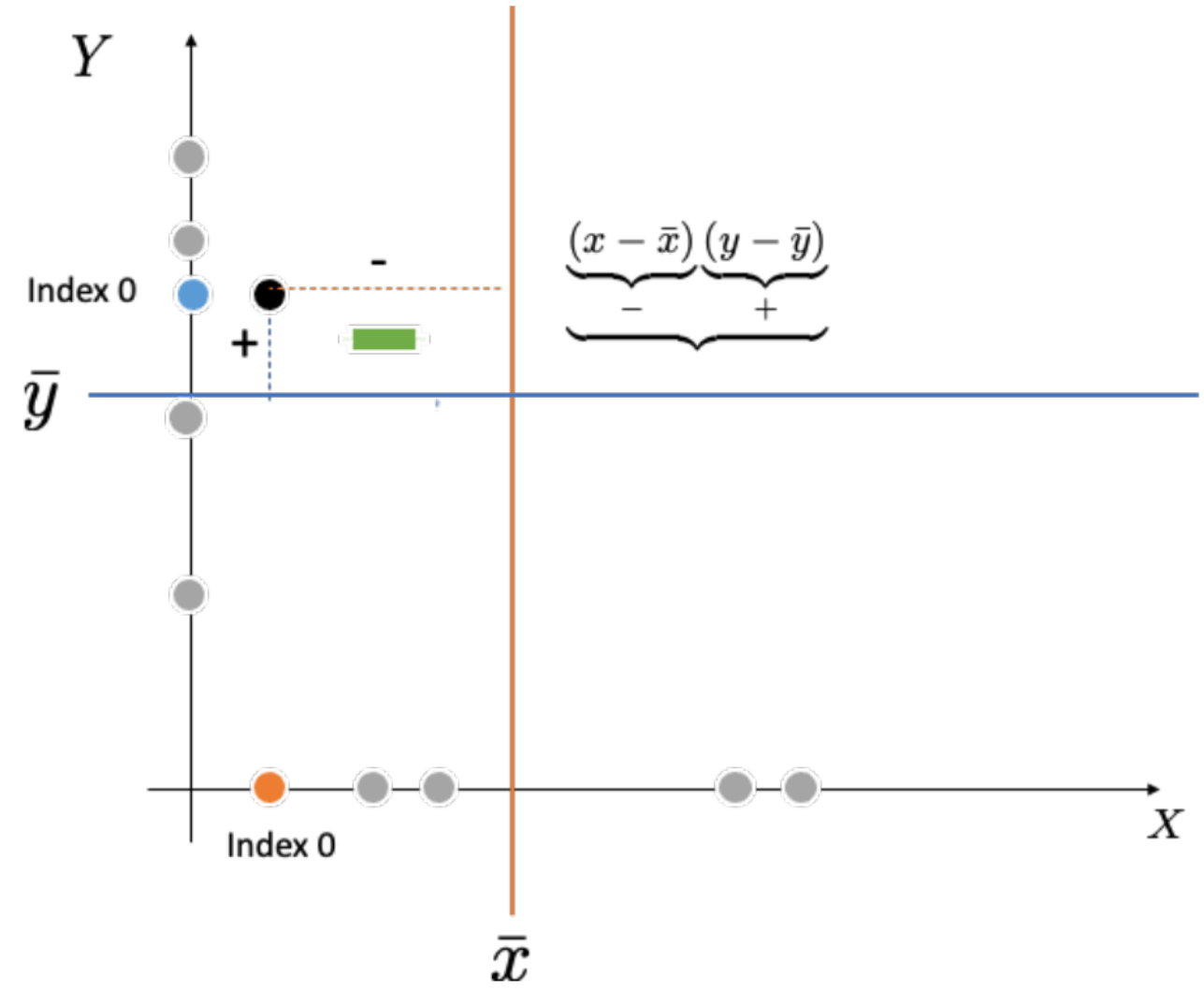
- **Xi** – the values of the X-variable
- **Yj** – the values of the Y-variable
- **μX** – the Mean (average) of the X-variable
- **μY** – the mean (average) of the Y-variable
- **n** – the number of data points

Why?
We are basically trying to get the direction of the values in respect to the mean of X and Y, that act as new axes.

Pearson Correlation Coefficient Geometry



Basically, we are summing all the areas of the squares that are formed with the new axes. The areas are the product of the differences of a point and the means.



One side of the square could be of negative value. For this reason we can have negative areas, in the upper-left and lower-right quadrant.
This is the reason why the covariance can be negative.

Obviously, if the points are all over and not going in a specific direction, the sum will be zero because the areas with different signs will eliminate each other.

On the contrary, if we have a very marked and specific direction in the data, the covariance will be 1. The more sparse the data is, the closer to 0 it gets because of all the different sign areas.

Final remark, we are not getting any information about the strength of the relation, only about its direction. We will see that this will be accomplished with the Pearson Correlation Coefficient, which will indicate direction and strength of the relation.