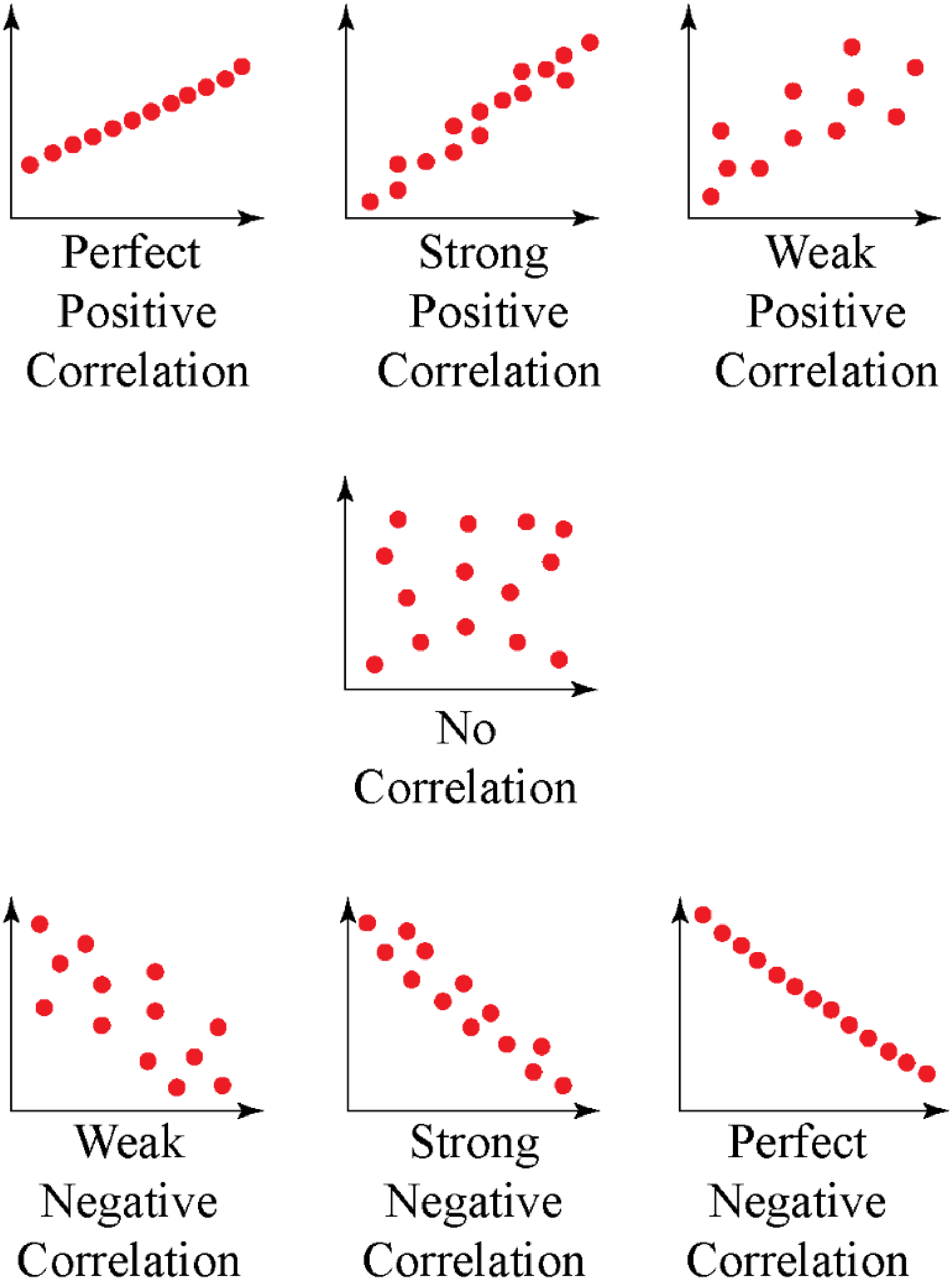
# Correlation coefficient

Correlation is a statistical measure that indicates how much two variables are related(if they change together at a costant rate).
Specifically, it indicates the strength of their relationship.



The stronger the relationship is, the closer to ±1 it gets.

## Pearson Correlation Coefficient

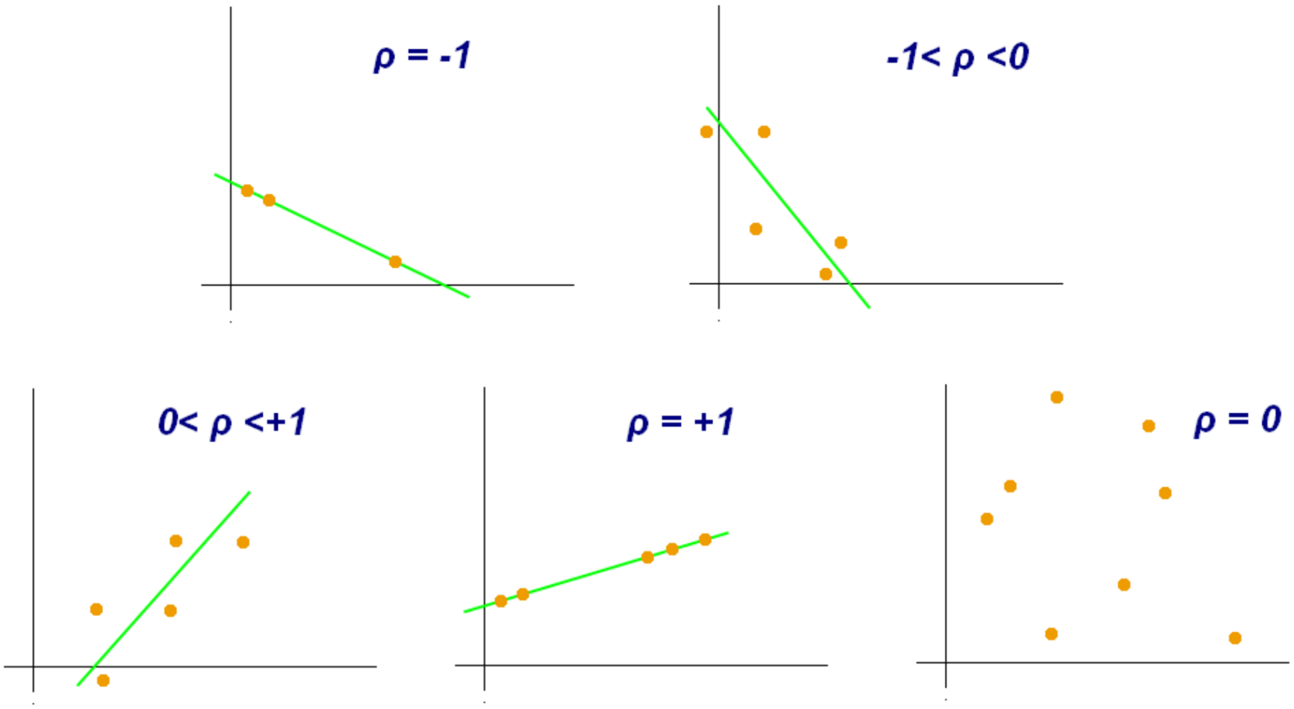
\Huge\rho_{X,Y} = \frac{ cov(X,Y) }{ \sigma_X \sigma_Y&nbsp;}

Where:

- **cov** – the covariance of the two series
- **σx** – the Standard Deviation of X
- **σy** – is the standard deviation of Y

Charateristics:

- The range of the result is [-1, +1].
- A value of exactly ±1 indicates that all data points lie on a line, a linear equation can describe the relationship between X and Y perfectly.
- The sign is determined by the slope: a value of +1 implies that all data points lie on a line for which Y increases as X increases, and vice versa for −1.
- 0 means that there is no linear dependency between variables



### Covariance vs. Correlation

Covariance and correlation both primarily assess the relationship between variables.

**Covariance** measures the total variation of two random variables from their expected values. Using covariance, we can only gauge the direction of the relationship (whether the variables tend to move in tandem or show an inverse relationship). However, it does not indicate the strength of the relationship, nor the dependency between the variables.

On the other hand, **correlation** measures the strength of the relationship between variables. Correlation is the scaled measure of covariance. It is dimensionless. In other words, the correlation coefficient is always a pure value and not measured in any units.

### Why are we dividing covariance by the product of the standard deviations?

BECAUSE, the covariance gives just the sum of the areas, which could be an arbitrary big number, there are no limits.
BUT, if we divide by the standard deviations, we are giving normalizing/standardizing this value.

Both the covariance and the standard deviation are divided by n, so in the equation they cancel it out.
Since the standard deviation, without n, basically takes the sum of all the all the differences between the point and the mean (squared), it will simply be a sum of all the variances of all points.

So, simply, the product of the standard deviations is the maximum value that the covariance can assume.

So if we multiply those two measures, we are basically forming an area, that is guaranteed to be bigger than the covariance, since the covariance can have negative areas and the standard deviations can't.

So if covariance is equal to the product of standard deviations, it means that all points lie on the same line, because all the areas of all the points are as big as they can get.