**Correlation:**

Correlation encompasses a method employed to determine the connections between two variables. To gauge whether two variables possess a relationship, one common technique is to plot them on a graph known as a "scatter plot." While various measures of association exist for variables measured at the ordinal level or higher, correlation is the prevalent and widely adopted approach for such cases.

**Correlation in Statistics:**

This segment delves into the calculation and interpretation of correlation coefficients for both ordinal and interval-level scales. Correlation methods summarize the connection between two variables into a single figure termed the correlation coefficient. Represented by the symbol "r," the correlation coefficient's range spans from -1 to +1.

A correlation coefficient hovering around 0, while positive or negative, implies a minor or negligible relationship between the two variables. A correlation coefficient nearing +1 indicates a positive relationship, signifying that increases in one variable correspond with increases in the other.

Conversely, a correlation coefficient nearing -1 signifies a negative relationship, indicating that an increase in one variable corresponds with a decrease in the other. While correlation coefficients can be computed for ordinal, interval, or ratio-level variables, their significance diminishes for variables measured solely at a nominal scale.

For ordinal scales, Spearman's rank correlation coefficient (Spearman's rho) is used for calculation. On the other hand, for interval or ratio-level scales, Pearson's correlation coefficient (Pearson's r), commonly known as the correlation coefficient, is the most frequently employed measure.

**What Does Correlation Measure?**

In statistics, correlation investigates and gauges the direction and degree of a relationship between variables. It assesses the extent to which variables co-vary, indicating a change in one variable corresponds with changes in another. It's crucial to understand that correlation does not imply causation; it doesn't establish a cause-and-effect relationship.

For instance, if there's a correlation between variables X and Y, it signifies that when one variable's value changes in a certain direction (increase or decrease), the other variable's value also changes either in the same direction (positive change) or in the opposite direction (negative change). Moreover, if a correlation exists, it's linear in nature, meaning the relative movement of the two variables can be depicted using a straight line on a graph.

In summary, correlation provides insights into how two variables relate and vary together, but it does not indicate that changes in one variable directly cause changes in the other.

**Correlation Coefficient:**

The correlation coefficient, denoted as r, serves as a concise measure that characterizes the strength of the statistical connection between two variables at the interval or ratio level. The coefficient is normalized to a range between -1 and +1. When the value of r approaches 0, it suggests a limited relationship between the variables. Conversely, the farther away r is from 0 in either a positive or negative direction, the more pronounced the relationship between the two variables.

These two variables are typically represented by the symbols X and Y. To illustrate their relationship, the values of X and Y are portrayed through a scatter diagram, which plots various combinations of the two variables. The scatter diagram is presented initially, followed by the methodology for calculating Pearson’s r, a common type of correlation coefficient.

The examples provided often feature relatively small sample sizes. Subsequent instances showcase data derived from larger samples.

**Types of Correlation:**

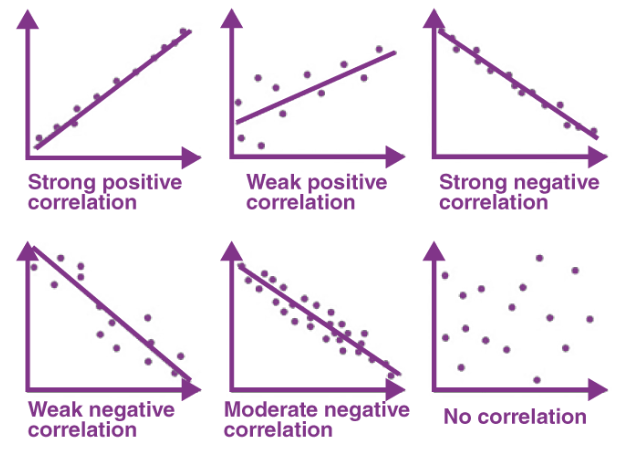
A scatter plot is a visual representation that illustrates the correlation between two attributes or variables. It conveys the degree of connection between the two variables. There are three main scenarios to consider when examining the relationship between two variables:

**1. Positive Correlation:** This occurs when the values of the two variables move in the same direction. An increase (or decrease) in the value of one variable is accompanied by a corresponding increase (or decrease) in the value of the other variable.

**2. Negative Correlation:** In this case, the values of the two variables move in opposite directions. An increase (or decrease) in the value of one variable is followed by a decrease (or increase) in the value of the other variable.

**3. No Correlation:** When there is no discernible linear relationship or connection between the two variables, it is termed as "no correlation." In this scenario, changes in one variable do not lead to any predictable changes in the other variable.

These three types of correlation provide insights into how the variables behave in relation to each other, helping us understand their mutual influence.



**Pearson Correlation Coefficient:**

The Pearson Correlation coefficient is the most widely used formula for assessing linear dependency between datasets.

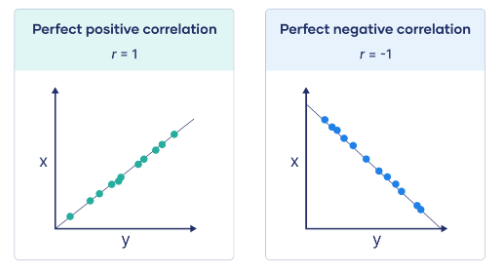
It is a way to measure linear correlation between the data points. The maximum value of pearson’s correlation is +1 and minimum value is -1. It can have a value 0 as well where there will be no correlation between the data points.

+ve correlation means when one variable changes other variable also changes in the same direction.

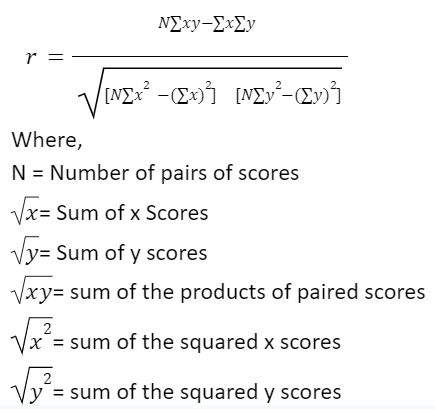
-ve correlation means when one variable changes other variable also changes in the opposite direction.

It is also known as a bivariate correlation. It summarizes the characteristics of the datasets. PCC is also known as an inferential stats which means that it can be used to test statistical hypothesis. We can think of PCC in this way also that how close the observations are to the best fit line. We can visualize the +e and –ve best fit as well by the slope of the line.

When the value of r is +1 the points will lie exactly on the best fit and same happens when r=-1.



The Pearson Correlation Coefficient formula calculates the strength and direction of the linear relationship between two datasets. It is given by:



The Pearson Correlation Coefficient formula allows us to determine the strength and direction of the linear relationship between two sets of data.

When to use Pearson’s Correlation Coefficient-

1. When both variables are quantitative.
2. The variables are normally distributed.
3. The data have no outliers.
4. The relationship between the variables are linear.

**Spearman’s rank correlation coefficient-**

Spearman’s rank correlation measures the strength and direction of association between two ranked variables. It basically gives the measure of monotonicity of the relation between two variables i.e. how well the relationship between two variables could be represented using a monotonic function.

**When to use Spearman’s rank correlation coefficient-**

We should use Spearman’s correlation coefficient when one or more of the following points are true-

1. The variables are ordinal.
2. The variables are not normally distributed.
3. The data has outliers.
4. The relationship between the variables is non- linear and monotonic.

The formula for Spearman’s rank coefficient is:

