**Coefficient of Variation**

The coefficient of variation (CV) is the ratio of the standard deviation to the mean. The lower the value of the coefficient of variation, the more precise the estimate.Below is the formula for how to calculate the coefficient of variation:

Coefficient of Variation = (Standard Deviation / Mean) \* 100

In finance, the coefficient of variation is used to measure the **risk per unit of return**. For example, investment X has the mean monthly return 0.5% with a standard deviation of 0.58%. Suppose we have another investment, say, Y with a 1.5% mean monthly return and standard deviation of 6%. Then,

1.CV=0.58/0.5=1.16

2.CV=6/1.5=4  
  
Interpretation: Therefore, investment Y is riskier than X.

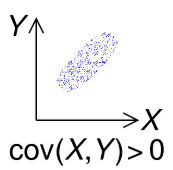
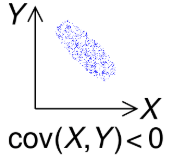
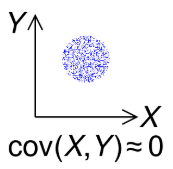
[**Covariance and correlation**](https://en.wikipedia.org/wiki/Covariance_and_correlation)

[Covariance and correlation](https://en.wikipedia.org/wiki/Covariance_and_correlation) are two mathematical concepts which are commonly used in statistics. When comparing data samples from different populations, covariance is used to determine how much two random variables vary together, whereas correlation is used to determine when a change in one variable can result in a change in another.

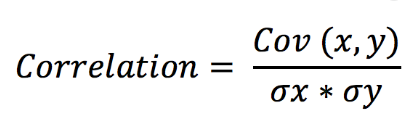
The value of covariance between 2 variables is achieved by taking the summation of the product of the differences from the means of the variables as follows:

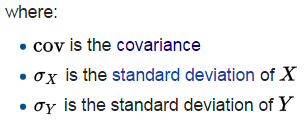
https://lh6.googleusercontent.com/RgC5ARNcZD9GkvbN9hAEMN6ECfya4exVW01rUy_Zi_k6pdGenwzREMjU-tqQGpJbSRSoJBMf7UGDYqlOj9gOsXz15s9HP3uomgiTFbsa0mXTkVEmT3i-OmW5f64wLCw704FITGoqpRp9q_u3GQ

Covariance is only useful to find the direction of the relationship between two variables and not the magnitude. Below are the plots which help us understand how the covariance between two variables would look in different directions.

Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables.





**Central limit theorem**

It states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement , then the distribution of the sample means will be approximately normally distributed.

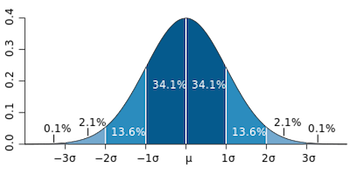
An essential component of the Central Limit Theorem is that the [average](https://www.statisticshowto.com/arithmetic-mean/)**of your sample means will be the population mean**. In other words, add up the means from all of your samples, find the average and that average will be your actual population mean. Similarly, if you find the average of all of the [standard deviations](https://www.statisticshowto.com/probability-and-statistics/standard-deviation/) in your [sample](https://www.statisticshowto.com/sample/), you’ll find the actual standard deviation for your population.

**Normal Distribution**

Data are said to be normally distributed if their frequency histogram is apporximated by a bell shaped curve. Properties of a normal distribution are given below

* [mean](https://www.mathsisfun.com/mean.html) = [median](https://www.mathsisfun.com/median.html) = [mode](https://www.mathsisfun.com/mode.html)
* symmetry about the center
* 50% of values less than the mean and 50% greater than the mean

But there are many cases where the data tends to be around a central value with no bias left or right, and it gets close to a "Normal Distribution" like this:



**Population vs Sample Data**

## When dealing with statistical data, it is important to distinguish between "population" data sets and "sample datasets.

## A population data set contains all members of a specified group (the entire list of possible data values).

Example: The population may be "ALL people living in the US."

**A sample data set contains a part, or a subset, of a population. The size of a sample is always less than the size of the population from which it is taken.**

Example: The sample may be "SOME people living in the US."

**Standard Error**

The standard error of the mean, also called the standard deviation of the mean, is a method used to estimate the standard deviation of a sampling distribution.

The standard error of the mean (SEM) can be expressed as:

## Standard Error of the Mean

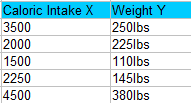
**σM= standard error of the mean**

**σ = the**[**standard deviation**](https://explorable.com/measurement-of-uncertainty-standard-deviation)**of the original distribution**

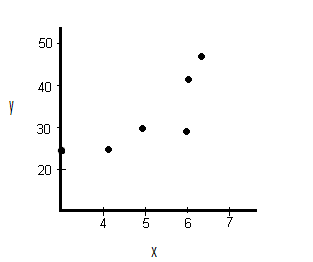
**N = the sample size**

**Bivariate analysis**

It means the analysis of bivariate data. It is one of the simplest forms of statistical analysis, used to find out if there is a relationship between two sets of values. It usually involves the [variables](https://www.statisticshowto.com/probability-and-statistics/types-of-variables/)X and Y.The results from bivariate analysis can be stored in a two-column data table. For example, you might want to find out the relationship between caloric intake and weight (of course, there is a pretty strong relationship between the two. You can read more [here](http://www.cdc.gov/healthyweight/calories/).). Caloric intake would be your [independent variable](https://www.statisticshowto.com/independent-variable-definition/), X and weight would be your [dependent variable](https://www.statisticshowto.com/dependent-variable-definition/), Y.



These give you a visual idea of the bivariate pattern that the variables follow.



**Z-score**

The **basic z score formula** for a [sample](https://www.statisticshowto.com/sample/)is:

**z = (x – μ) / σ**

For example, let’s say you have a test score of 190. The test has a mean (μ) of 150 and a [standard deviation](https://www.statisticshowto.com/probability-and-statistics/standard-deviation/) (σ) of 25. Assuming a [normal distribution](https://www.statisticshowto.com/probability-and-statistics/normal-distributions/), your z score would be:

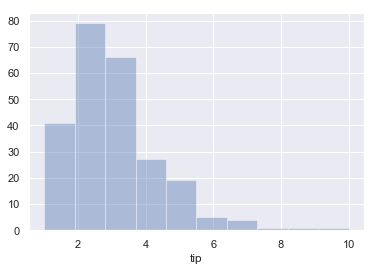
* z = (x – μ) / σ
* = (190 – 150) / 25 = 1.6.

**Data Visualization**

Data visualization is an important part of any data analysis. It helps us to recognize relations between variables and also to find which variables are significant or which variable can affect the predicted variable. [Seaborn](https://seaborn.pydata.org/tutorial.html) is a library built on [matplotlib](https://matplotlib.org/" \t "_blank). It’s easy to use and can work easily with Numpy and pandas data structures.

# Univariate plots

These plots are based on a single variable and show the frequency of uniques values of a given variable.

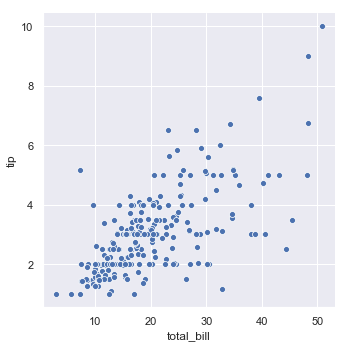


# Bivariate Plots

This type of plots is used when you need to find a relation between two variables and to find how the value of one variable changes the value of another variable. Different types of plots are used based on the data type of the variable.

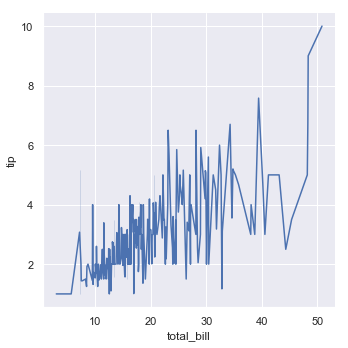
**Scatterplot**

It shows the relationship between two variables. So, if you need to find the correlation between two variables scatterplot can be used.



**Lineplot**

This plot is similar to the scatterplot but instead of dots, it displays the line joining all the dots by arranging the variable value represented on the x-axis.



**Box-and-whisker plot**

It shows the distribution of quantitative data in a way that facilitates comparisons between variables or across levels of a categorical variable. If the dataset is too large and the range of value is too big then it shows some values as outliers based on an inter-quartile function.

