**Statistics and its Types**

**Statistics** simply means numerical data, and is field of math that generally deals with collection of data, tabulation, and interpretation of numerical data.

It is actually a form of mathematical analysis that uses different quantitative models to produce a set of experimental data or studies of real life.

It is an area of applied mathematics concern with data collection analysis, interpretation, and presentation.

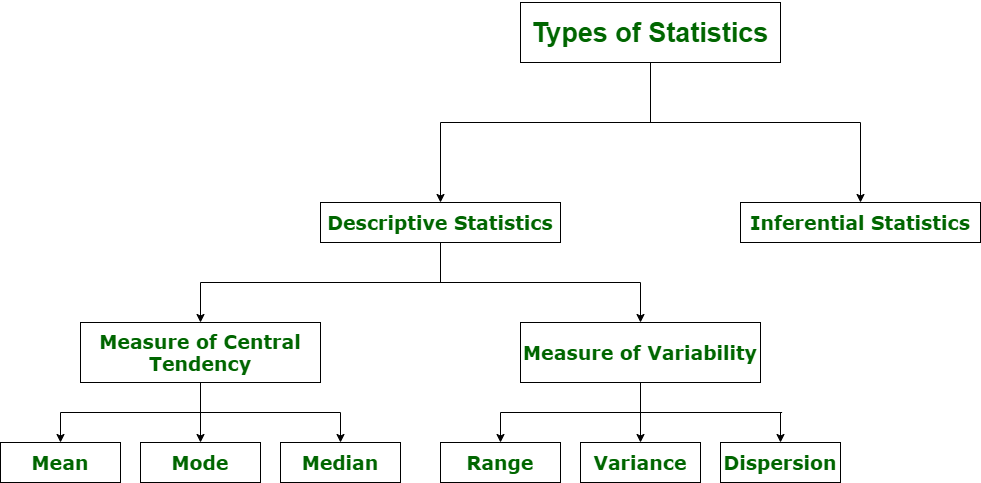
Statistics deals with how data can be used to solve complex problems. Some people consider statistics to be a distinct mathematical science rather than a branch of mathematics.

Statistics makes work easy and simple and provides a clear and clean picture of work you do on a regular basis.

**Basic terminology of Statistics :**

* **Population –**  
  It is actually a collection of set of individuals or objects or events whose properties are to be analyzed.
* **Sample –**  
  It is the subset of a population.

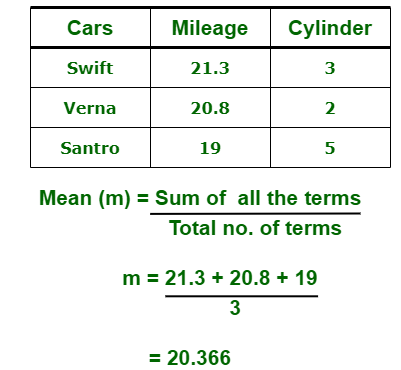
Types of Statistics :

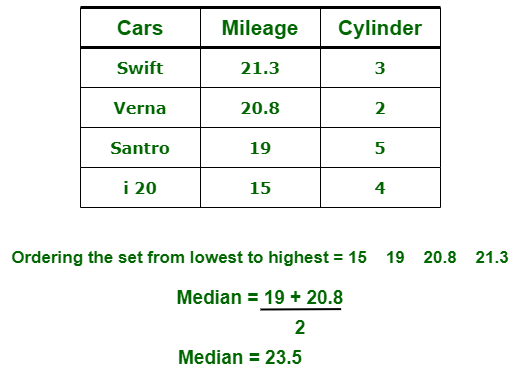


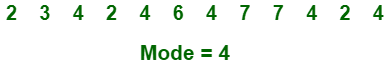
**1. Descriptive Statistics :**

1. Descriptive statistics uses data that provides a description of the population either through numerical calculation or graph or table.
2. It provides a graphical summary of data. It is simply used for summarizing objects, etc. There are two categories in this as following below.

* **(a). Measure of central tendency –**Measure of central tendency is also known as summary statistics that is used to represents the center point or a particular value of a data set or sample set.  
  In statistics, there are three common measures of central tendency as shown below:
  + **(i) Mean:**It is measure of average of all value in a sample set.  
    For example,



* + **(ii) Median :**It is measure of central value of a sample set. In these, data set is ordered from lowest to highest value and then finds exact middle.
  + For example,
  + **(iii) Mode :**It is value most frequently arrived in sample set. The value repeated most of time in central set is actually mode.  
    For example,



* **(b). Measure of Variability –**  
  Measure of Variability is also known as measure of dispersion and used to describe variability in a sample or population. In statistics, there are three common measures of variability as shown below:
  + **(i) Range :**It is given measure of how to spread apart values in sample set or data set.

Range = Maximum value - Minimum value

* + **(ii) Variance :**It simply describes how much a random variable defers from expected value and it is also computed as square of deviation.

**S2=** ∑ni=1 [(xi - ͞x)2 ÷ n]

In these formula, **n** represent total data points, **͞x** represent mean of data points and **xi** represent individual data points.

* + **(iii) Dispersion :**It is measure of dispersion of set of data from its mean.

**σ=** √ (1÷n) ∑ni=1 (xi - μ)2

**2. Inferential Statistics :**

Inferential Statistics makes inference and prediction about population based on a sample of data taken from population. It generalizes a large dataset and applies probabilities to draw a conclusion. It is simply used for explaining meaning of descriptive stats. It is simply used to analyze, interpret result, and draw conclusion. Inferential Statistics is mainly related to and associated with hypothesis testing whose main target is to reject null hypothesis.

Hypothesis testing is a type of inferential procedure that takes help of sample data to evaluate and assess credibility of a hypothesis about a population. Inferential statistics are generally used to determine how strong relationship is within sample. But it is very difficult to obtain a population list and draw a random sample.

Inferential statistics can be done with help of various steps as given below:

1. Obtain and start with a theory.
2. Generate a research hypothesis.
3. Operationalize or use variables
4. Identify or find out population to which we can apply study material.
5. Generate or form a null hypothesis for these population.
6. Collect and gather a sample of children from population and simply run study.
7. Then, perform all tests of statistical to clarify if obtained characteristics of sample are sufficiently different from what would be expected under null hypothesis so that we can be able to find and reject null hypothesis.

**Types of inferential statistics –**Various types of inferential statistics are used widely nowadays and are very easy to interpret. These are given below:

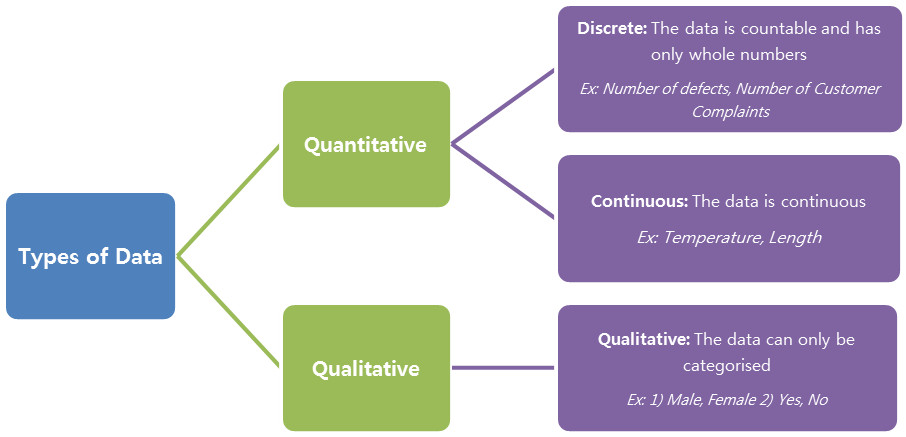
* One sample test of difference/One sample hypothesis test
* Confidence Interval
* Contingency Tables and Chi-Square Statistic
* T-test or Anova
* Pearson Correlation
* Bi-variate Regression
* Multi-variate Regression

Attention reader! Don’t stop learning now. Get hold of all the important DSA concepts with the [**DSA Self Paced Course**](https://practice.geeksforgeeks.org/courses/dsa-self-paced?utm_source=geeksforgeeks&utm_medium=article&utm_campaign=gfg_article_dsa_content_bottom) at a student-friendly price and become industry ready.

## Types of Data

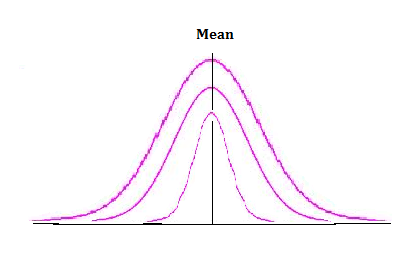
There are two broad categories of data: Qualitative and Quantitative data. The name itself indicates that the former cannot be quantified in numbers whereas the later can be.

The below tree will explain the types of data simply and clearly:

[](http://www.whatissixsigma.net/wp-content/uploads/2014/03/Basic-Statistics-Types-of-Data.png)

## Normal Distribution

Any population has numerous data points spread in different ranges. There are a lot of distributions that exist when they are plotted in a frequency plot. Normal distribution is one of them and the most frequently used one also. A Normal Distribution can be characterised by two values: Mean and Standard deviation. The Normal distribution looks like a bell shaped curve. The peak central value represents the mean and the tapered ends represent the maximum and minimum values. The size of standard deviation determines the width of the curve. Higher is the SD, the curve looks broad. Lower the SD, the narrow is the curve, for the same Mean.

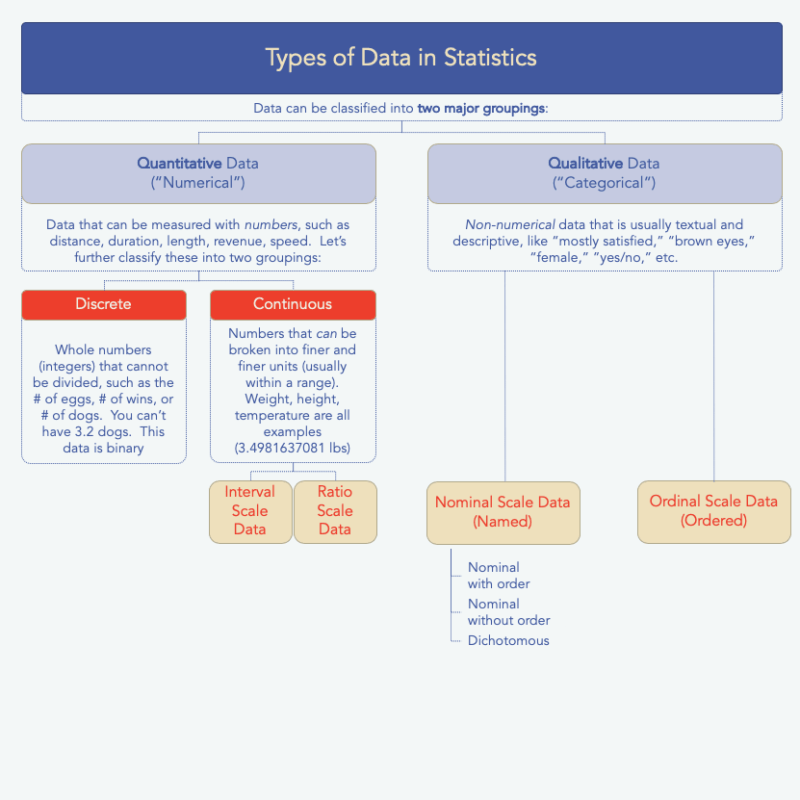
[](http://www.whatissixsigma.net/wp-content/uploads/2014/03/Basic-Statistics-Mean.png)The above diagram explains how the data points are spread differently for the same mean and different standard deviations.

The entire concept of Six Sigma (99.9997% of values falling within six standard deviations), Control charts, Process Variation reduction, and many others revolve around this Normal curve. Thus it becomes very much necessary to understand the normal distribution.

Knowing the types of data, how to represent the data in descriptive statistics and Normal distribution lay the foundation for Statistical data analysis.

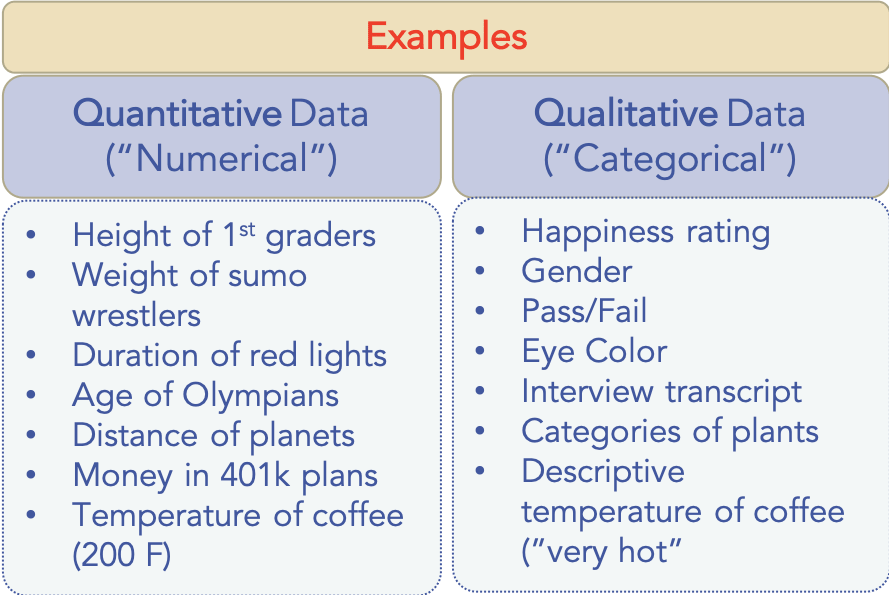
**Data Types in Statistics**

Statistical data can be classified in several ways. Here we provide an overview of the major **types of data in statistics**. This article is intended to provide a 30,000 foot overview, while a deeper dive on some of these categories can be found on other pages.

An Overview of the Categories of Data

**Quantitative vs. Qualitative Data**

Our first characterization is whether the data is numerical or not. This one is pretty simple to understand. If you’ve got numerical data, you’ve got **quantitative** data. If you don’t have numbers, you’ve got **qualitative** data (non-numerical, or also called categorical data). See the sidebar for some examples.



Since “quantitative” and “qualitative” are almost identical in spelling, it’s easy to mix them up. The way I remember is that qualitative has an “l” it in and I always think of “l for letters.” Hope that helps you too.

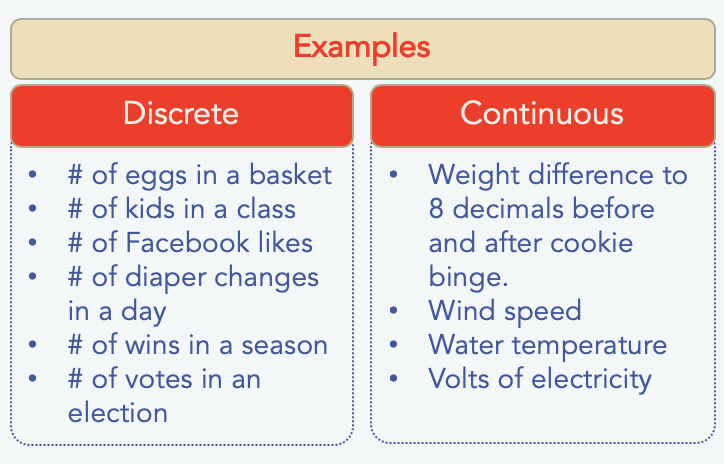
By the way, I’ve written an entire post on [quantitative vs. qualitative research](https://www.mymarketresearchmethods.com/quantitative-vs-qualitative-research-whats-the-difference/), which of course covers the ways to obtain such data. Take a look there for more info.

**Discrete vs Continuous Data**

Within quantitative (or numerical) data, there exists a sub-level of data types: **discrete** data or **continuous** data. I think you’ll find this characterization very simple to understand as well.

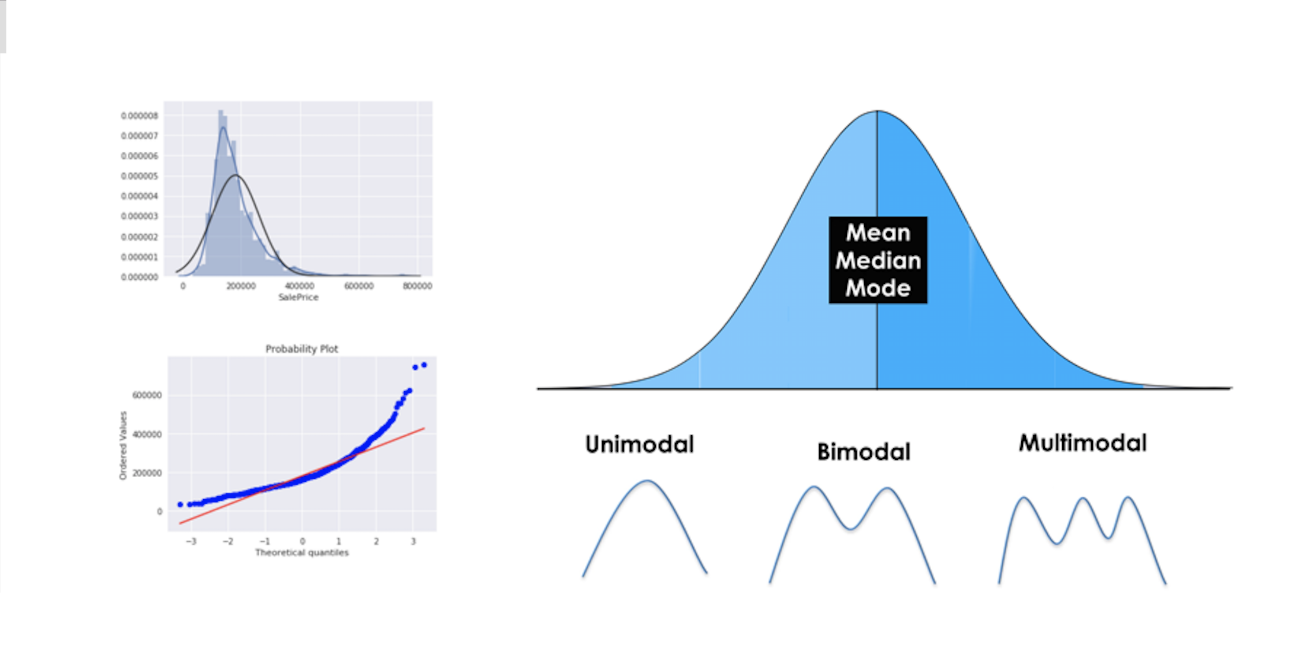
**Discrete** data is a whole number (integer) and it can’t be subdivided into smaller and smaller parts. Classic examples are the # of people in a classroom, # of brothers in a family, etc. You can’t have 30.5 people in the class and you can’t have 1.5 brothers. There are plenty of other examples as well, like the # of wins a rugby team gets in a season or the number of fish inside a fish tank.

**Continuous** data *continues* on and on and on, at least that’s how I think about it. These are data that can be broken into smaller and smaller units. My weight can be infinitely measured using precision equipment and doesn’t have to stop at 3 meters. Usain Bolt’s 100m time is another example of continuous data, because it could be measured to the nanosecond and beyond.



Continuous data can be further categorized into a couple of types: interval and ratio. These are classifications of data scales and you can read all about them in their own post about the [types of measurement scales in statistics](https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/). In fact, you’ll find some data scales there that apply to qualitative data as well, which takes us back to the top-level category.

# Intro to Descriptive Statistics



**Descriptive Statistical Analysis helps you to understand your data and is a very important part of Machine Learning. This is due to Machine Learning being all about making predictions. On the other hand, statistics is all about drawing conclusions from data, which is a necessary initial step. In this post you will learn about the most important descriptive statistical concepts. They will help you understand better what your data is trying to tell you, which will result in an overall better machine learning model and understanding.**

**Table of Contents:**

* Introduction
* Normal Distribution
* Central Tendency (mean, mode, median)
* Measures of Variability (range, interquartile range)
* Variance and Standard Deviation
* Modality
* Skewness
* Kurtosis
* Summary

# Introduction

Doing a descriptive statistical analysis of your dataset is absolutely crucial. A lot of people skip this part and therefore lose a lot of valuable insights about their data, which often leads to wrong conclusions. Take your time and carefully run descriptive statistics and make sure that the data meets the requirements to do further analysis.

But first of all, we should go over what statistics really is:

**Statistics is a branch of mathematics that deals with collecting, interpreting, organization and interpretation of data.**

Within statistics, there are two main categories:

**1. Descriptive Statistics:**In Descriptive Statistics your are describing, presenting, summarizing and organizing your data (population), either through numerical calculations or graphs or tables.

**2. Inferential statistics:**Inferential Statistics are produced by more complex mathematical calculations, and allow us to infer trends and make assumptions and predictions about a population based on a study of a sample taken from it.

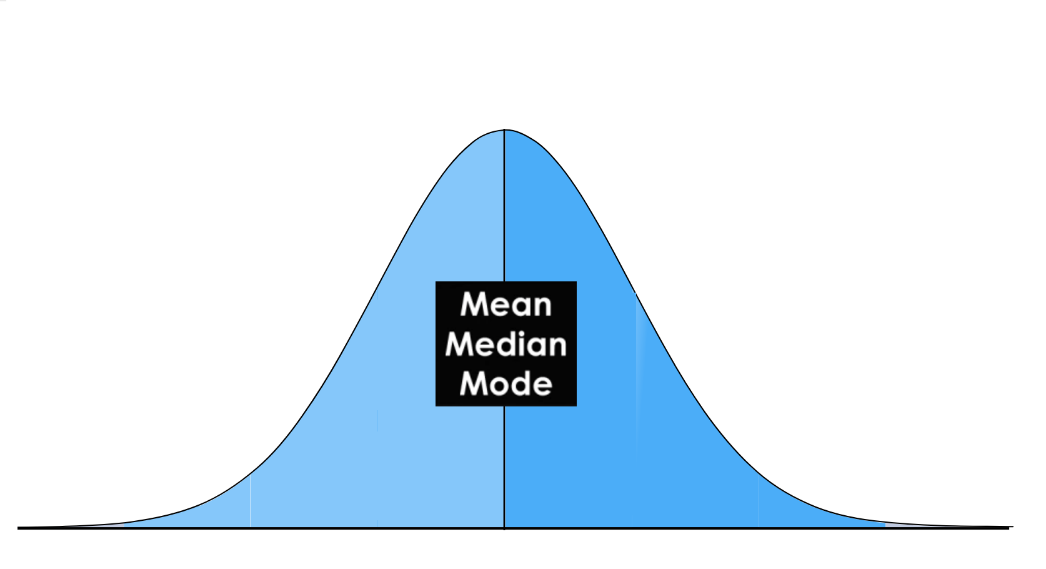
# Normal Distribution

The normal distribution is one of the most important concepts in statistics since nearly all statistical tests require normally distributed data. It basically describes how large samples of data look like when they are plotted. It is sometimes called the “bell curve“ or the “Gaussian curve“.

Inferential statistics and the calculation of probabilities require that a normal distribution is given. This basically means, that if your data is not normally distributed, you need to be very careful what statistical tests you apply to it since they could lead to wrong conclusions.

**A normal Distribution is given if your data is symmetrical, bell-shaped, centered and unimodal.**

In a perfect normal distribution, each side is an exact mirror of the other. It should look like the distribution on the picture below:



You can see on the picture that the distribution is bell-shaped, which simply means that it is not heavily peaked. Unimodal means that there is only one peak.

# Central Tendency

In statistics we have to deal with the mean, mode and the median. These are also called the „Central Tendency“. These are just three different kinds of „averages” and certainly the most popular ones.

**The mean is simply the average** and considered the most reliable measure of central tendency for making assumptions about a population from a single sample. Central tendency determines the tendency for the values of your data to cluster around its mean, mode, or median. The mean is computed by the sum of all values, divided by the number of values.

**The mode is the value or category that occurs most often within the data.** Therefore a dataset has no mode, if no number is repeated or if no category is the same. It is possible that a dataset has more than one mode, but I will cover this in the „Modality“ section below.  
The mode is also the only measure of central tendency that can be used for categorical variables since you can’t compute for example the average for the variable „gender“. You simply report categorical variables as numbers and percentages.

**The median is the “middle” value or midpoint in your data** and is also called the „50th percentile“. Note that the median is much less affected by outliers and skewed data than the mean. I will explain this with an example: Imagine you have a dataset of housing prizes that range mostly from $100,000 to $300,000 but contains a few houses that are worth more than 3 million Dollars. These expensive houses will heavily effect then mean since it is the sum of all values, divided by the number of values. The median will not be heavily affected by these outliers since it is only the “middle” value of all data points. Therefore the median is a much more suited statistic, to report about your data.

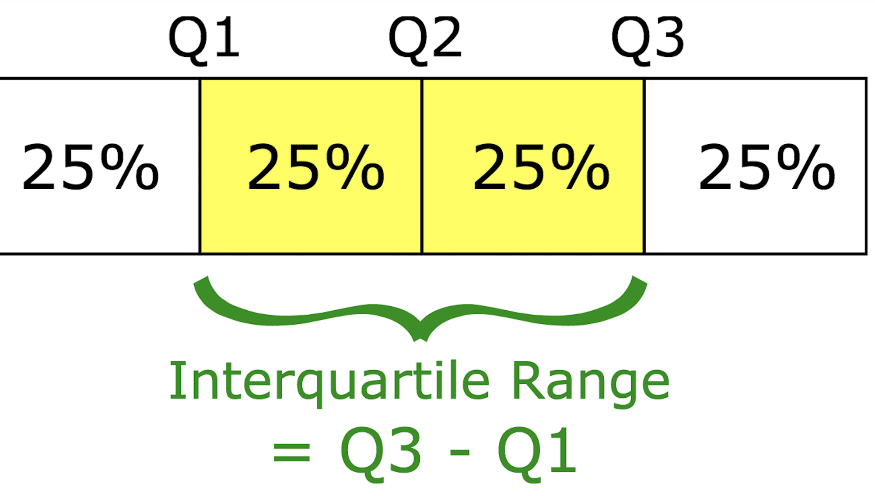
In a normal distribution, these measures all fall at the same midline point. This means that the mean, mode and median are all equal.

# Measures of Variability

The most popular variability measures are the range, interquartile range (IQR), variance, and standard deviation. These are used to measure the amount of spread or variability within your data.

**The** **range describes the difference between the largest and the smallest points in your data.**

The interquartile range (IQR) is a measure of statistical dispersion between upper (75th) and lower (25th) quartiles.



**While the range measures where the beginning and end of your datapoint are, the interquartile range is a measure of where the majority of the values lie.**

The difference between the standard deviation and the variance is often a little bit hard to grasp for beginners, but I will explain it thoroughly below.

# Variance and Standard Deviation

The Standard Deviation and the Variance also measure, like the Range and IQR, how spread apart our data is (e.g the dispersion). Therefore they are both derived from the mean.

The variance is computed by finding the difference between every data point and the mean, squaring them, summing them up and then taking the average of those numbers.

The squares are used during the calculation because they weight outliers more heavily than points that are near to the mean. This prevents that differences above the mean neutralize those below the mean.

The problem with Variance is that because of the squaring, it is not in the same unit of measurement as the original data.

Let’s say you are dealing with a dataset that contains centimeter values. Your variance would be in squared centimeters and therefore not the best measurement.  
  
This is why the Standard Deviation is used more often because it is in the original unit. It is simply the square root of the variance and because of that, it is returned to the original unit of measurement.

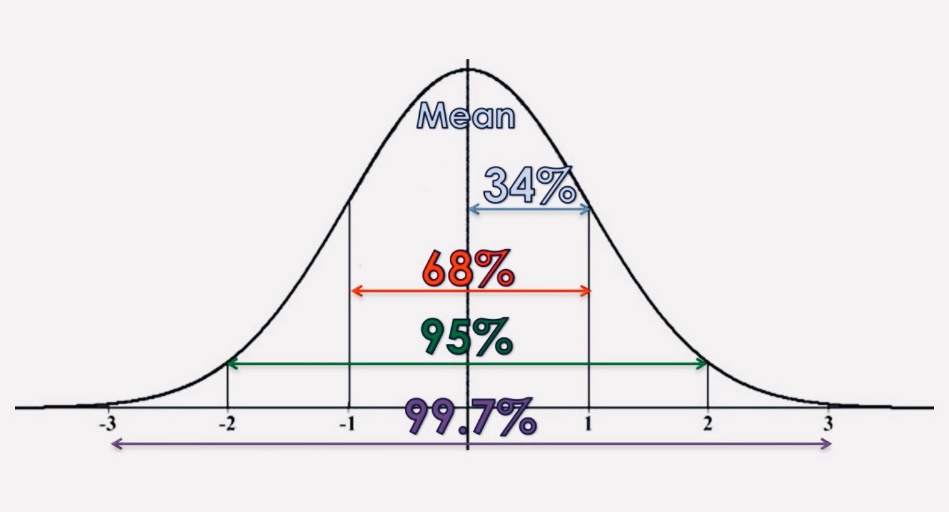
Let’s look at an example that illustrates the difference between variance and standard deviation:

Imagine a data set that contains centimeter values between 1 and 15, which results in a mean of 8. Squaring the difference between each data point and the mean and averaging the squares renders a variance of 18.67 (squared centimeters), while the standard deviation is 4.3 centimeters.

When you have a low standard deviation, your data points tend to be close to the mean. A high standard deviation means that your data points are spread out over a wide range.

Standard deviation is best used when data is unimodal. In a normal distribution, approximately 34% of the data points are lying between the mean and one standard deviation above or below the mean. Since a normal distribution is symmetrical, 68% of the data points fall between one standard deviation above and one standard deviation below the mean. Approximately 95% fall between two standard deviations below the mean and two standard deviations above the mean. And approximately 99.7% fall between three standard deviations above and three standard deviations below the mean.

The picture below illustrates that perfectly.

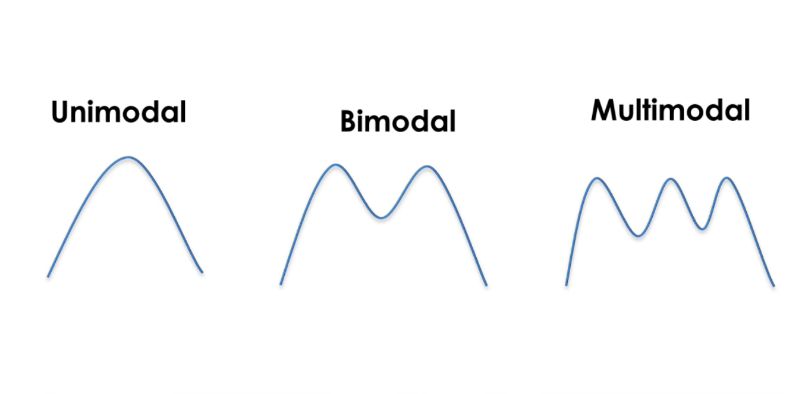


With the so-called „Z-Score“, you can check how many standard deviations below (or above) the mean, a specific data point lies. With pandas you can just use the [„std()“ function](https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.std.html). To better understand the concept of a normal distribution, we will now discuss the concepts of modality, symmetry and peakedness.

# Modality

**The modality of a distribution is determined by the number of peaks it contains.** Most distributions have only one peak but it is possible that you encounter distributions with two or more peaks.

The picture below shows visual examples of the three types of modality:



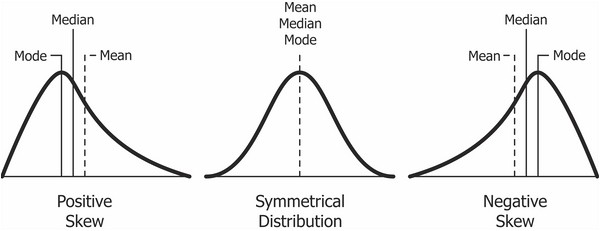
Unimodal means that the distribution has only one peak, which means it has only one frequently occurring score, clustered at the top. A bimodal distribution has two values that occur frequently (two peaks) and a multimodal has two or several frequently occurring values.

# Skewness

**Skewness is a measurement of the symmetry of a distribution.**

Therefore it describes how much a distribution differs from a normal distribution, either to the left or to the right. The skewness value can be either positive, negative or zero. Note that a perfect normal distribution would have a skewness of zero because the mean equals the median.

Below you can see an illustration of the different types of skewness:



**We speak of a positives skew if the data is piled up to the left**, which leaves the tail pointing to the right.

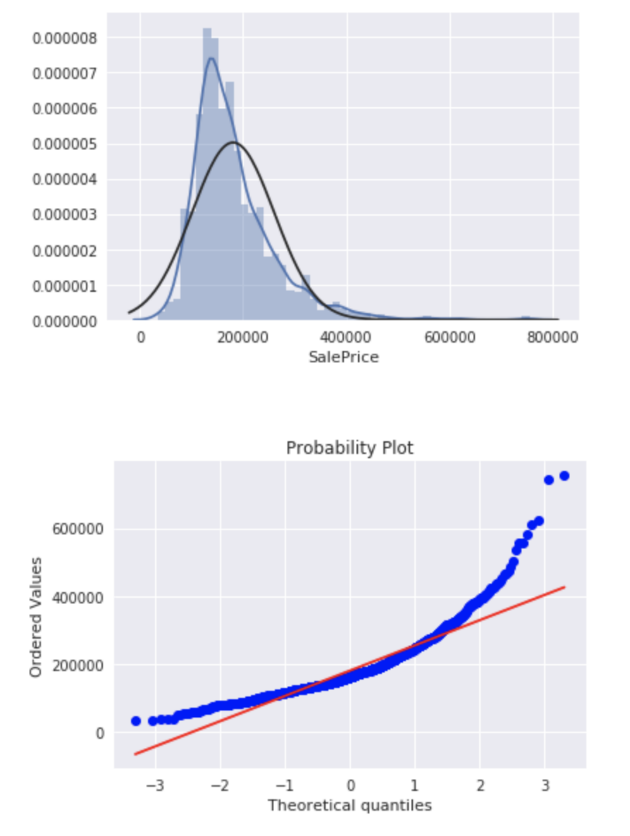
**A negative skew occurs if the data is piled up to the right**, which leaves the tail pointing to the left. Note that positive skews are more frequent than negative ones.

A good measurement for the skewness of a distribution is Pearson’s skewness coefficient that provides a quick estimation of a distributions symmetry. To compute the skewness in pandas you can just use the [„skew()“ function](https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.skew.html).

# Kurtosis

Kurtosis measures whether your dataset is heavy-tailed or light-tailed compared to a normal distribution. Data sets with high kurtosis have heavy tails and more outliers and data sets with low kurtosis tend to have light tails and fewer outliers. Note that a histogram is an effective way to show both the skewness and kurtosis of a data set because you can easily spot if something is wrong with your data. A probability plot is also a great tool because a normal distribution would just follow the straight line.

You can see both for a positively skewed dataset in the image below:



A good way to mathematically measure the kurtosis of a distribution is fishers measurement of kurtosis.

Now we will discuss the three most common types of kurtosis.

A normal distribution is called **mesokurtic** and has kurtosis of or around zero. A **platykurtic** distribution has negative kurtosis and tails are very thin compared to the normal distribution. **Leptokurtic** distributions have kurtosis greater than 3 and the fat tails mean that the distribution produces more extreme values and that it has a relatively small standard deviation.

If you already recognized that a distribution is skewed, you don’t need to calculate it’s kurtosis, since the distribution is already not normal. In pandas you can view the kurtosis simply by calling the [„kurtosis()“ function](https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.kurtosis.html).