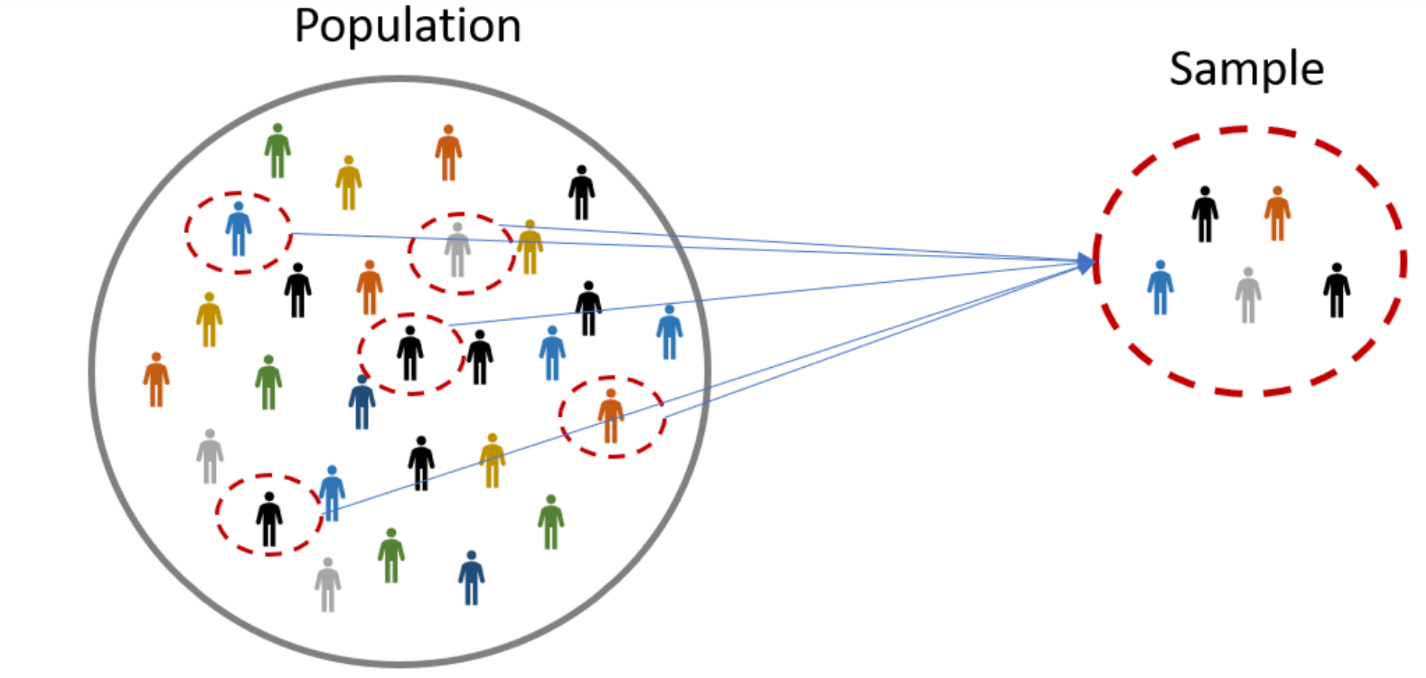
**Population(N)**

* Total number of observations in a dataset is called the population.
* It has parameters like mean, median, mode, standard deviation etc.
* The numbers we obtain when using a population are called parameters.

**Sample(n)**

* A random subset of a population is a sample.
* Samples are used when a dataset is large enough which makes it difficult to analyse it.
* The numbers we obtain when working with a sample are called statistics.
* A random sample is collected when each member of the sample is chosen from the population strictly by chance (Each member is equally likely to be chosen).
* **A representative sample is a subset of the population that accurately reflects the members of the entire population.**



**Measures of central tendency**

Mean, median and mode are measures that gives us some idea about the of centrality of data.

**Mean (or Average)**

* Denoted as **µ** for a population and‾**X** {\displaystyle {\bar {x}}} for a sample.
* Mean or average is the value obtained by dividing the sum of all the data by the total number of data points.
* It is not a robust statistics as it is affected by extreme values.
* Very large or very low value can imply that outliers exist in the data.

**Median**

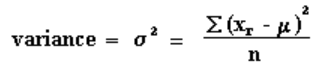
* It refers to the data situated at the middle position of the set.
* It splits the data in half and is also called 50th percentile.
* It is not affected by the outliers much.
* In a set with odd number of data points the median is the middlemost value while if the number of data points is even then it is the average of the two middle items.

**Mode**

* It is the value that occurs more frequently in a dataset.
* It is the only measure of central tendency that can be used for categorical variables.
* Sometimes two or more values can be the mode of a data. In such case, the data set is multimodal.
* Multimodal distribution is a statistical distribution of values with multiple peaks.

**Variance**

* It is the average squared deviation from mean.
* Denoted by **σ2**

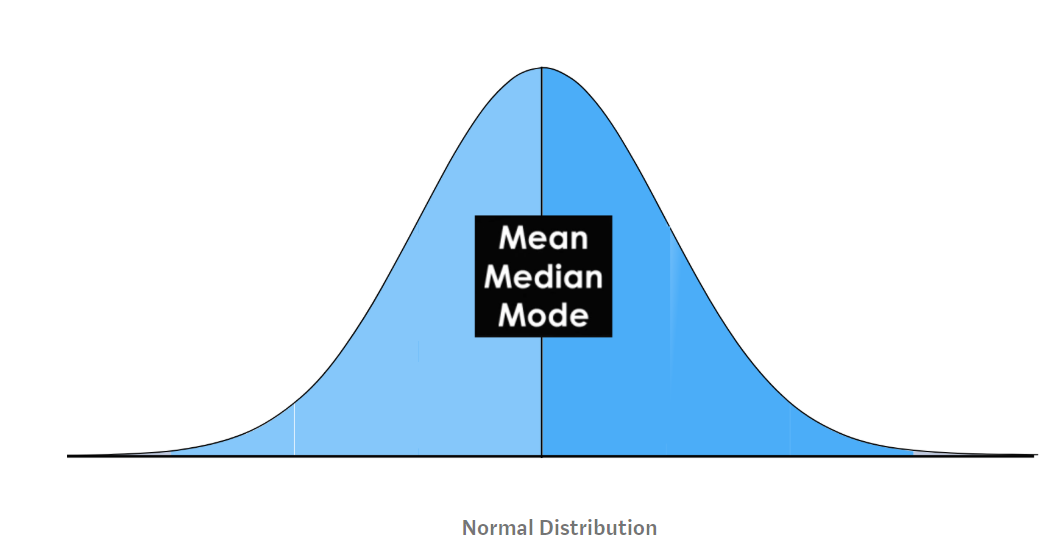


**Standard Deviation**

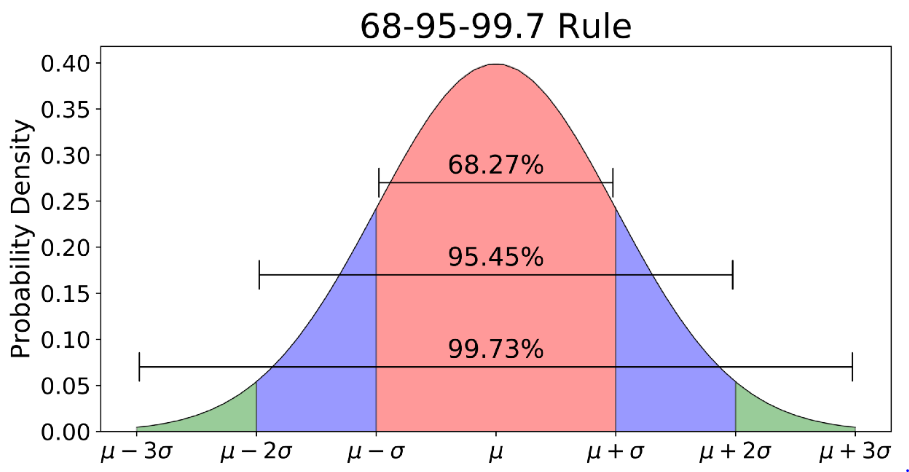
* It is the square root of variance.
* Denoted by **σ**
* Standard Deviation is used more often because it is in the original unit.
* When you have a low standard deviation, your data points tend to be close to the mean.
* A high standard deviation means that your data points are spread out over a wide range.

**Normal Distribution**

* It basically describes how large samples of data look like when they are plotted.
* It is also called the “bell curve” or the “Gaussian curve”.
* The mean, mode and median are all equal and fall at the same midline point.
* A normal distribution with a mean of 0 and a standard deviation of 1 is called a **standard normal distribution**. Area under the standard normal distribution curve would be 1.



* The normal distribution is commonly associated with the 68-95-99.7 rule.
* 68% of the data is within 1 standard deviation (σ) of the mean (μ).
* 95% of the data is within 2 standard deviations (σ) of the mean (μ).
* 99.7% of the data is within 3 standard deviations (σ) of the mean (μ).



**Percentile**

* Percentiles divide ordered data into **hundredths**.
* In a sorted dataset, a given percentile is the point at which that percent of the data is **less than** the point we are at.
* The 50th percentile is pretty much the median.
* Example, a student taking a difficult exam might earn a score of 75 percent. This means that he correctly answered every three out of four questions. A student who scores in the 75th percentile, however, has obtained a different result. This percentile means that the student earned a higher score than 75 percent of the other students who took the exam. In other words, the percentage score reflects how well the student did on the exam itself; the percentile score reflects how well he did in comparison to other students.