**Statistics**

**Statistics:** Statistics is the science of collecting, organizing, summarizing, analyzing and interpreting data.

**Descriptive Statistics:** When performing *descriptive statistics* you collect, organize, summarize, and graphically present data; then you are able to make conclusions about said data.

**Inferential Statistics:** *Inferential statistics* are used when you want to make predictions and inferences about a larger group (a whole population) from data that was collected from a smaller group (a sample population)

**Why Statistics?** The practice of engineering is applying science to solve a problem. In engineering we’re used to solving a deterministic problem where our solution solves the problem all the time. E.g software written to dispense currency from an ATM machine. The solution is deterministic. We know each and every step how the machine will dispense currency. There are many problems where the solution is not deterministic. This is because either we don’t know enough about the problem or we don’t have enough computing power to model the problem. E.g how to classify whether a mail is spam or not. There is no single formula to determine a spam mail. It depends on the occurrence of certain words used together, length of email and other factors. Another example can be how to measure the happiness of humans. The solution to this problem will differ greatly from 1 person to another. For these problems, we need statistics.

**Data Mining:** It’s the process of automatically discovering useful information in large data repositories. The purpose of any **"data analysis"** is to derive meaningful information from it. One way to extract information from data is to study the variability in data points. The more is the variability, the more careful you have to study or explore the dataset, so that you can capture all of its meaning.

**Machine Learning:** Machine learning is a set of techniques, which help in dealing with vast data in the most intelligent fashion (by developing algorithms or set of logical rules) to derive actionable insights (delivering search for users in this case). **Machine Learning is the name given to generalizable algorithms that enable a computer to carry out a task by examining data rather than hard programming.** Its a subfield of computer science and artificial intelligence that focuses on developing systems that learn from data and help in making decisions and predictions based on that learning. ML enables computers to make data-driven decisions rather than being explicitly programmed to carry out a certain task. ML programs are also designed to learn and improve over time when exposed to new data. There are broadly to types of machine learning techniques:

* Supervised Learning: Regression, Classification
* Unsupervised Learning: Clustering

**Difference between Data Mining and Machine Learning:** Teaching someone how to dance is Machine Learning. And using someone to find best dance centers in the city is Data Mining.

**Reporting vs Analytics vs Advanced Analytics:**

**Reporting:** A report describes what events have happened in the business. It provides what is asked for and is typically standardized. A monthly sales summary report shows monthly sales by region.

**Analysis:** An analysis tries to answer why the events happened in the business have happened. E.g an analysis of sales summary report may show sales peaks on specific holidays or weekends. Basic Analytics involves slicing and dicing of data, monitoring large volumes of data in real time and anomaly detection

**Advanced Analytics:** Advanced analytics extends the insights provided by analytics by doing impact analysis on the business and prescribing the next steps which can be taken. It includes predictive modeling, text analytics and advanced data mining algorithms

**Forecasting vs Predictive Modeling:**

**Forecasting** is a process of estimating the future based on past events. It’s at a high level. E.g no of calls expected in a call center, no of passengers expected to travel from an airport next month etc

**Predictive modeling** is doing the prediction or estimation at a more granular level. E.g which customers are expected to buy the printer in next 30 days.

**Population:** Statisticians refer to an entire group that is being studied as a population. Each member of the population is called a unit, or subject. Different characterstics or attributes of a unit or subject are known as **variables**. Smaller, representative group from the population, is called a **sample.** An actual value of a population variable (for example, number of tortoises, average weight of all tortoises, etc.) is called a parameter. An estimate of a parameter derived from a sample is called a statistic.

**Types of Data:**

At the very basic level, data can be of 2 types: Quantitative or Qualitative

[**Quantitative variables**](http://statistics.berkeley.edu/~stark/SticiGui/Text/gloss.htm#quantitative) take numerical values whose "size" is meaningful. [Quantitative](http://statistics.berkeley.edu/~stark/SticiGui/Text/gloss.htm#quantitative) variables answer questions such as "how many?" or "how much?" For example, it makes sense to add, to subtract, and to compare two persons' weights, or two families' incomes. Quantitative variables typically have measurement units, such as pounds, dollars, years, volts, gallons, megabytes, inches, degrees, miles per hour, pounds per square inch, BTUs, and so on.

Some [variables](http://statistics.berkeley.edu/~stark/SticiGui/Text/gloss.htm#variable), such as social security numbers and zip codes, take numerical values, but are not quantitative: They are [**qualitative**](http://statistics.berkeley.edu/~stark/SticiGui/Text/gloss.htm#qualitative)**or**[**categorical**](http://statistics.berkeley.edu/~stark/SticiGui/Text/gloss.htm#categorical)**variables**. The sum of two zip codes or social security numbers is not meaningful. The average of a list of zip codes is not meaningful. [Qualitative](http://statistics.berkeley.edu/~stark/SticiGui/Text/gloss.htm#qualitative) and [categorical](http://statistics.berkeley.edu/~stark/SticiGui/Text/gloss.htm#categorical) variables typically do not have units. Qualitative or categorical variables—such as gender, hair color, or ethnicity—group individuals. Qualitative and categorical variables have neither a "size" nor, typically, a natural ordering to their values. They answer questions such as "which kind?" The values categorical and qualitative variables take are typically adjectives (for example, green, female, or tall). Arithmetic with [qualitative](http://statistics.berkeley.edu/~stark/SticiGui/Text/gloss.htm#qualitative) variables usually does not make sense, even if the variables take numerical values. [Categorical](http://statistics.berkeley.edu/~stark/SticiGui/Text/gloss.htm#categorical) variables divide individuals into categories, such as gender, ethnicity, age group, or whether or not the individual finished high school

**Levels of Measurement:** There are 4 levels of measurement: **Nominal, Ordinal, Interval and Ratio**.

* A **nominal** measurement is one in which the values of the variable are names. The names of the different species of Galapagos tortoises are an example of a nominal measurement. These variables are categorical. Nominal variables are organized into non-numeric categories that cannot be ranked or compared quantitatively.  So nominal level is used for qualitative variables. Appropriate mathematical operation:  counting the number of cases per category
* An **ordinal** measurement involves collecting information of which the order is somehow significant. E.g tracking student grades
* With **interval** measurement, the distance between any two values has a specific meaning. E.g difference in temperature. Addition and Subtraction can be done for interval variables but multiplication and division is not possible because zero is arbitrary in this level of measurement.
* A **ratio** measurement is the estimation of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind. The ratio between any two values has meaning, because the data includes an absolute zero value

**Measures of Location(Measures of Central Tendency):** Central tendency refers to the most typical value in a set of numbers

**Median** is the half-way point of data. The median is the number that divides the (ordered) data in half—the smallest number that is at least as big as half the data. At least half the data are equal to or smaller than the median, and at least half the data are equal to or greater than the median. If the distribution is skewed, median is typically used to describe the center. Median is appropriate for ordinal, interval and ratio variables but not for nominal variables.

**Mode:** The value that has highest frequency. Most frequently occurring value in the data set or the most popular value. It’s the only measure of central tendency that can be used with nominal variables. Though its also appropriate for ordinal, interval and ration variables.

**Mean:** The mean (more precisely, the arithmetic mean) is commonly called the average. It is the sum of the data, divided by the number of data. If there are outliers in data, mean can be strongly influenced. In such cases, median is more appropriate. The data point at Ratio level of measurement is a candidate for mean calculation

For [qualitative](http://statistics.berkeley.edu/~stark/SticiGui/Text/gloss.htm#qualitative) and [categorical](http://statistics.berkeley.edu/~stark/SticiGui/Text/gloss.htm#categorical) data, the mode makes sense, but the [mean](http://statistics.berkeley.edu/~stark/SticiGui/Text/gloss.htm#mean) and [median](http://statistics.berkeley.edu/~stark/SticiGui/Text/gloss.htm#median) do not

A perfectly normal distribution population has the same values for mean, median and mode.If you add a constant to every value, the mean, median and mode increase by the same constant. If you multiply every value by a constant. Then, the mean, median and mode will also be multiplied by that constant.

The pth percentile of a list of numbers is the smallest number that is atleast as large as p% of the list.

A bar chart gives 2 D data while Histogram is 1 D data

A pie chart is good for relative data and comparing.

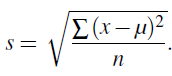
**Percentiles:** Assume that the [elements](http://stattrek.com/Help/Glossary.aspx?Target=Elements) in a data set are rank ordered from the smallest to the largest. The values that divide a rank-ordered set of elements into 100 equal parts are called **percentiles**.

**Quartiles:**

The median of a data set is located so that 50% of the data occurs to the left of the median (and 50% of the data occurs to the right of the median). There is no reason to restrict our attention to the 50% level. For example, we can find a point where 25% of the data occurs on its left (and 75% to its right). This point is known as the first "quartile"

**Measures of Dispersion(Spread):** These measure the extent of variability in data. Range, interquartile range and standard deviation are the three commonly used measures of dispersion.

* **Range:** Difference between the largest and smallest observation in the data.
* **Inter-quartile Range:** Difference between the 25th and 75th percentile. It describes the middle 50% of the observations.
* **Standard Deviation:** It is the measure of spread of data about the mean. It measures roughly how far off the entries are from their average. It tells us how the data is spread out. **The more the SD, the more spread out data is.** Since its simply a measure, it can’t be negative.



When you add a constant to a list of values, the average also adds up by constant but the SD doesn’t change. If you multiply by a constant, the new average and new SD also get multiplied by that constant.

* **Variance:** Mean Square of deviations from average. Or simply, it’s the square of Standard deviation.

Any point that is greater than the third quartile plus the IQR or less than the first quartile minus the IQR is considered an outlier.

**Significance of SD:** SD gives you an insight that how much your data is spread out. With the help of SD you can compare 2 datasets more effectively. If the average of 2 data sets is same, it does not means that the SD will be same. E.g 99,100,101 and 0 , 100 , 200 have same mean i.e 100 but they have different standard deviations. The SD of (99,100,101) is only 1 but the SD of (0,100,200) is 100 which is very large.

Lets say the average starting salary in a company is 80000$. Would you consider joining it? There may be few outliers which may have skewed the average. Additionally, if you know that SD is 2000$, you may consider joining it.

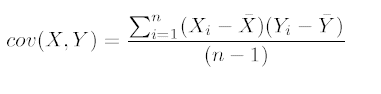
**Z Score:** A z-score is the measure of the number of standard deviations a particular data point is away from the mean i.e how many standard deviation away from mean is the observed value.  Its also called Z-value

Z = Deviation from mean/Standard Deviation

**Covariance:**

Variance and Standard Deviation only operate on 1 dimension so that you could only calculate the standard deviation for each dimension of the data set independently of the other dimensions. There should be a measure to find out how much the dimensions vary from the mean *with respect to each other.* Covariance is such a measure. Covariance is always measured between 2 dimensions.

If you calculate the covariance between one dimension and itself, you get the variance.

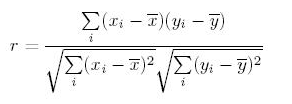


**Correlations** are mathematical relationships between variables. **Correlation Coefficient (r)** is a number between -1 and 1. It measures linear association i.e how tightly the points are clustered about a straight line. The correlation is said to be linear if the data points lye in an approximately straight line.

A correlation between two variables doesn’t necessarily mean that one caused the other or that they’re actually related in real life. A correlation between two variables means that there’s some sort of

mathematical relationship between the two. This means that when we plot the values on a chart, we can see a pattern and make predictions about what the missing values might be. What we dont know is whether there’s an actual relationship between the two variables, and we certainly don’t know whether one caused the other, or if there’s some other factor at work.

**Correlation = Covariance(X,Y) / SQRT( Var(X)\* Var(Y))**



Important things to remember about correlation:

* A non-dependency between two variable means a zero correlation. However the inverse is not true. A zero correlation can even have a perfect dependency.
* Correlation quantifies the linear dependence of two variables. It cannot capture non-linear relationship between two variables.
* Corelation is highly affected by outliers
* Corelation is not transitive. If X is correlated to Y and Y is correlated to Z then, its not necessary that X will be correlated to Z
* Corelation does not imply causation

**Populations and Samples:**

A**population**includes all of the[elements](http://stattrek.com/Help/Glossary.aspx?Target=element)from a set of data while A **sample** consists of one or more observations from the population. A measurable characteristic of a population, such as a [mean](http://stattrek.com/Help/Glossary.aspx?Target=Mean) or [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=Standard_deviation), is called a **parameter**; but a measurable characteristic of a sample is called a **statistic**. The mean of a population is defined with the symbol µ whereas the mean of a sample is defined x(bar)

The **error (or disturbance)** of an observed value is the deviation of the observed value from the true value of a quantity of interest (e.g population mean)

**Residual**of an observed value is the difference between the observed value and the estimated value of the quantity of interest (for example, a [sample mean](https://en.wikipedia.org/wiki/Sample_mean))

**Sampling:** The process of extracting samples from population is called sampling. 5 sampling methods are:

1. Random – every member of the population has an equally likely chance of being chosen
2. Systematic – there is a system to choosing the sample such as every 20th member
3. Convenience – Using a group that is convenient such as the first 25 people that walked into the room
4. Cluster – choosing random subgroups from the population and choosing everyone in those subgroups
5. Stratified – choosing random members out of every subgroup in the population

**Errors in Sampling:** Estimates derived from sampling are prone to inaccuracies. This cannot be avoided until and unless the entire population is used. The difference between the true parameter and the statistic obtained by sampling is called sampling error