# INTRODUCTION TO STATISTICS

Statistics is branch of mathematical dealing with collecting, organizing and analyzing data in such a way that meaningful conclusions can be drawn from them. In general, its investigations and analysis fall into two broad categories called descriptive and inferential statistics.

* **Descriptive statistics** deals with the processing of data without attempting to draw any inferences from it. The data are presented in the form of tables and graphs. The characteristics of the data are described in simple terms. Events that are dealt with include everyday happenings such as accidents, prices of goods, business, incomes, epidemics, sports data, population data.
* **Inferential statistics** is a scientific discipline that uses mathematical tools to make forecasts and projections by analyzing the given data. This is of use to people employed in such fields as engineering, economics, biology, the social sciences, business, agriculture and communications.



# DATA

Data is information. One-way data is data in which we record information regarding one or more individuals. If we can fetch the required data by answering a single question then it’s an example of one-way data. The data is recorded in the form of a data table.



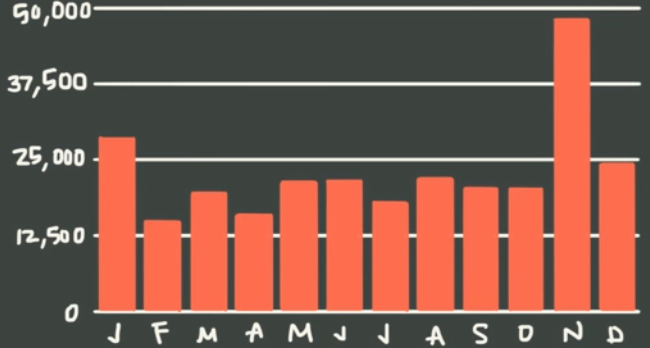
One-way data can be visualized using bar graph, line graph, ogives (accumulated values) and pie chart.

When we have more variables (dependent variable) than individuals (independent variable) then it’s always better to flip / transpose the table.



Comparison bar graph for one-way data would look something like below

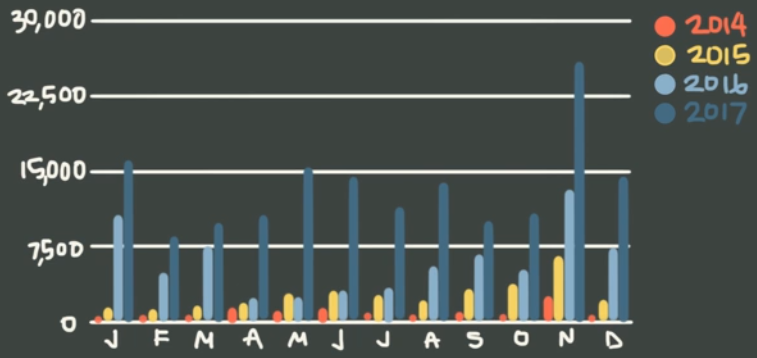




A two-way data, we have two independent variables and one or more dependent variables.



Comparison bar graph for a two-way data table would look something like this



Venn diagrams can be used to visualize two-way table



# VISUALIZING DATA

One Categorical variable are visualized using:

* Frequency distribution table (count for each categorical type)
* pie chart
* bar chart / column chart (x axis is categorical and y axis is numerical)
* pareto chart (80-20 rule)

Two Variables are visualized using:

* Side by side bar chart
* Cross table

One Variable numerical variables are visualized using:

* Frequency distribution table
* Histogram (both x and y axis are numerical)

Two Variables are visualized using:

* Scatter Plot

Divide the data into interval, interval is calculated using formula [highest - lowest]/ desired intervals. For example, we have data from 1 - 100 and we want to have 5 intervals so we should have (100-1)/5 ~ 20 elements in each interval.

# VARIABLE TYPES IN STATS

Variables are of two types:

1. Categorical
2. Numerical

Numerical data can be divided into:

* Continuous
* Discrete

MEASUREMENT LEVELS / GROUPS

1. Qualitative (categorical)
2. Quantitative (numeric)

Qualitative data / variables can be measured

* Nominal (no order) e.g. car types, gender etc.
* Ordinal (ordered) customer rating, rank etc.

Quantitative data / variables can be measured

* Ratio (have true zero for example temp in kelvin, age).
* Interval (they don’t have true zero e.g. temperature in c, time).

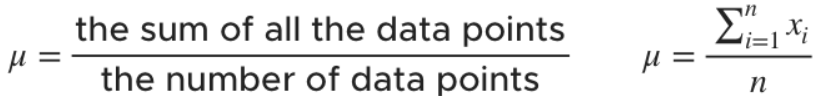
# MEASURE OF CENTRAL TENDENCY

To calculate centre of the data in sample/ population we have mainly three measures:

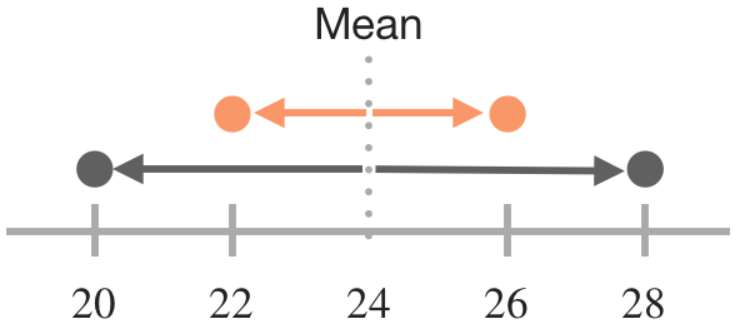
1. Mean
2. Median
3. Mode

## MEAN

Also known as average or arithmetic mean. The formula to calculate the mean is as below:



If we try to calculate the distance of points on the left and right of the mean with the mean value and then sum them up, the total value on the right will be equal to the total value on the left.

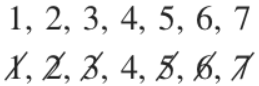


Mean is affected by outliers hence we need other measures of central tendency.

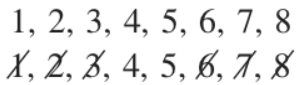
## MEDIAN

middle of the ordered data set).

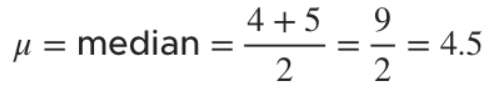
To calculate median, we arrange the data set in sorted order from min to max and then take the number or pair of numbers in the middle. For example, in the below set, Median is 4



But if we have even number of items in the set then



Then to find the median of the dataset, we find the mean of the two data points in the middle



## MODE

The Mode of the dataset is the value that occurs most often. Consider below example, in the given dataset Mode is 4



Consider another set as below:



In the above dataset both 4 and 6 occurs twice and hence we don’t have a clear winner. In such situation either we can say that the set do not have a mode or the set has two modes and hence it’s called **bi-modal**. If a data set has more than two modes then we say that dataset is **multi-modal**.

# MEASURE OF SPREAD

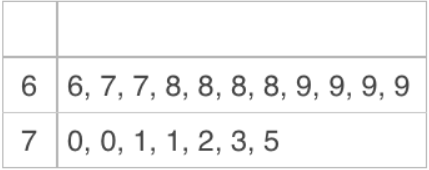
We looked at measure of central tendency which represents the middle of a dataset. But central tendency isn’t the only thing we’re interested in when it comes to data.

We also want to know about spread which is how and by how much our data set is spread out around its centre we also call measure of spread measure of dispersion / variability or scatter.

## RANGE

In statistics, the range of a dataset is the difference between the largest and smallest values. It is measured in the same units as the data. Since it only depends on two of the observations, it is most useful in representing the dispersion of small data sets.

For example, in the below steam and leaf plot



The lowest score is 66 and the highest score is 75. Therefore, the range of the dataset is



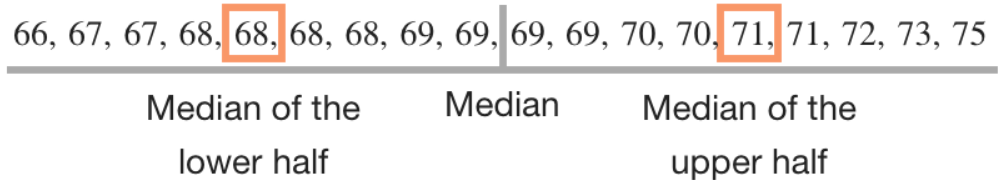
## INTER QUARTILE RANGE (IQR)

In descriptive statistics, the interquartile range (IQR) is also called the mid spread or middle 50% or technically H-spread.

We divide the dataset into quarters using the medians in the data. We cut the data into half at the median and then find the median of each of the halves and split halves again at those points. Each quarter of the data created is called a quartile.

The interquartile range is the difference between the lowest number in the upper quartile and the highest number in the lower quartile i.e. it is calculated as the difference between 75th and 25th percentiles or between upper and lower quartiles i.e. IQR = Q3 - Q1.

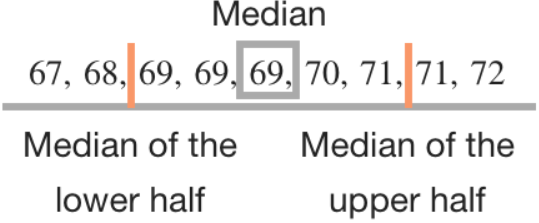
Let’s consider below example where our dataset has even number of observations:



Now that we have the medians of the lower half and upper half, we can find the IQR as below



Suppose our dataset has odd number of observation then we can calculate IQR as below



The IQR of the above dataset is



## VARIANCE

In probability theory and statistics, variance is the expectation of the squared deviation of a random variable from its mean.

Informally, it measures how far a set of (random) numbers are spread out from their average value.

Variance has a central role in statistics, where some ideas that use it include descriptive statistics, statistical inference, hypothesis testing, goodness of fit, and Monte Carlo sampling. Variance is an important tool in the sciences, where statistical analysis of data is common.

The variance is the square of the standard deviation, the second central moment of a distribution.

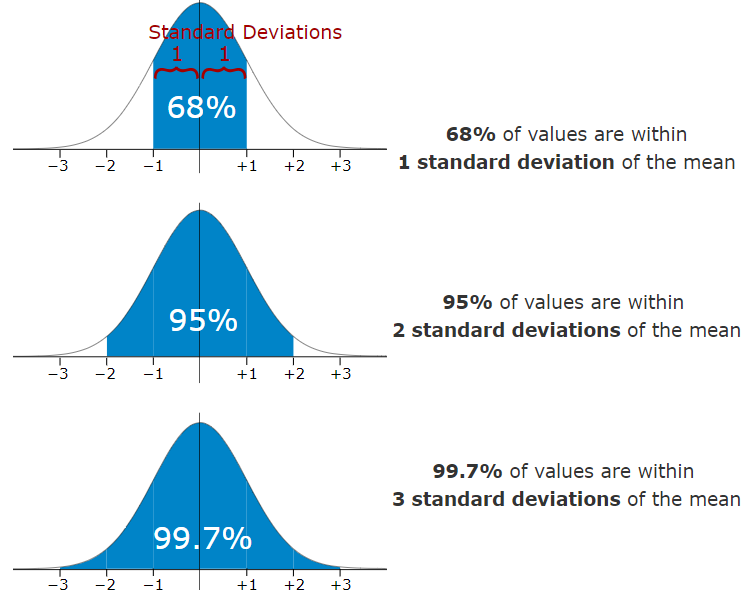
Sum of squared of deviation from mean / no of observations (n - 1 for sample). It’s not preferred as a measurement of variability coz usually it’s pretty large and hard to compare as the unit of measurement is squared.

## STANDARD DEVIATION

The Standard Deviation is a measure of how spread out numbers are. It’s the square root of variance.

It’s the most common measurement of variation for single dataset. Standard deviation is a measure that is used to quantify the amount of variation or dispersion of a set of data values.

* + A low standard deviation indicates that the data points tend to be close to the mean (also called the expected value) of the set.
  + While a high standard deviation indicates that the data points are spread out over a wider range of values.



It is good to know the standard deviation, because we can say that any value is:

* likely to be within 1 standard deviation (68 out of 100 should be)
* very likely to be within 2 standard deviations (95 out of 100 should be)
* almost certainly within 3 standard deviations (997 out of 1000 should be)

## COEFFICIENT OF VARIATION

In probability theory and statistics, the coefficient of variation (CV), also known as relative standard deviation (RSD), is a standardized measure of dispersion of a probability distribution or frequency distribution. It is often expressed as a percentage, and is defined as the ratio of the standard deviation to the mean.

This measure of dispersion is dimensionless. In other words, they have no units even if the variable itself has units.

* Distributions with CV < 1 are considered low-variance, while those with CV > 1 are considered high-variance.
* CV can also be used to compare two distributions as it’s a unit free measurement.

## MEAN ABSOLUTE DEVIATION

The mean absolute deviation of a dataset is the average distance between each data point and the mean. It gives us an idea about the variability in a dataset.

## MEDIAN ABSOLUTE DEVIATION (MAD)

The median absolute deviation of a dataset is the average distance between each data point and the median. We can use this measure of locate outliers in the dataset as median is not affected by outliers.

IMPORTANT: **Spread** for a distribution is till what point we have significant probabilities.

# CHANGING DATASET

What happens to central tendency and spread when we make changes to the dataset?

## SHIFTING DATA

Adding or subtracting

The Mean, Median and mode also gets changed as we shift the data. They get changed by the same amount.

Where as Range and IQR does not get affect and remains unchanged.

## SCALING DATA

# MEASURE OF ASYMMETRY (SKEWNESS)

To identify outliers:

1. **Right skewed** (outliers are located to the right) mean > median
2. **Left skewed** (outliers are located to the left) mean < median
3. **Zero skewness** (mean = median = mode)

# MEASURE OF RELATIONSHIP

Variables within a dataset can be related for lots of reasons. For example:

* One variable could cause or depend on the values of another variable.
* One variable could be lightly associated with another variable.
* Two variables could depend on a third unknown variable.

It can be useful in data analysis and modeling to better understand the relationships between variables. The statistical relationship between two variables is referred to as their correlation.

A correlation could be positive, meaning both variables move in the same direction, or negative, meaning that when one variable’s value increases, the other variables’ values decrease. Correlation can also be neural or zero, meaning that the variables are unrelated.

* Positive Correlation: both variables change in the same direction.
* Neutral Correlation: No relationship in the change of the variables.
* Negative Correlation: variables change in opposite directions.

The performance of some algorithms can deteriorate if two or more variables are tightly related, called multicollinearity. An example is linear regression, where one of the offending correlated variables should be removed in order to improve the skill of the model.

We may also be interested in the correlation between input variables with the output variable in order provide insight into which variables may or may not be relevant as input for developing a model.

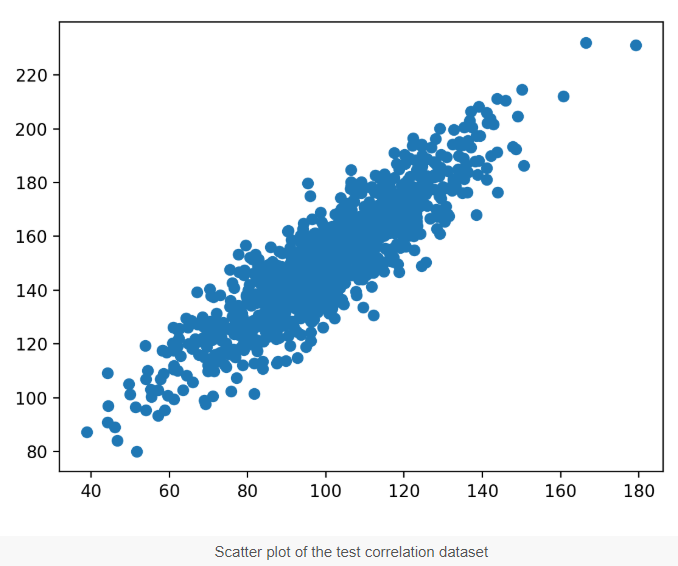
The structure of the relationship may be known, e.g. it may be linear, or we may have no idea whether a relationship exists between two variables or what structure it may take. Depending what is known about the relationship and the distribution of the variables, different correlation scores can be calculated.

## COVARIANCE

**Covariance** indicates the direction of the linear relationship between variables. Covariance tells us how much two variables disperse/ vary from the mean together.



It is calculated as the average of the product between the values from each sample, where the values haven been centered (had their mean subtracted).



The sign of the covariance can be interpreted as whether the two variables change in the same direction (positive) or change in different directions (negative). The magnitude of the covariance is not easily interpreted. A covariance value of zero indicates that both variables are completely independent. We can say two variables are correlated based on covariance value. It can have value ranging from -∞ and +∞.

## R (PEARSON’S CORRELATION)

Correlation measures both the strength and direction of the linear relationship between two variables. Correlation is a function of the covariance. Correlation values are standardized whereas, covariance values are not.



The Pearson’s correlation coefficient is calculated as the covariance of the two variables divided by the product of the standard deviation of each data sample.

The coefficient returns a value between -1 and 1 that represents the limits of correlation from a full negative correlation to a full positive correlation. A value of 0 means no correlation. The value must be interpreted, where often a value below -0.5 or above 0.5 indicates a notable correlation, and values below those values suggests a less notable correlation.

The Pearson’s correlation coefficient can be used to evaluate the relationship between more than two variables.

This can be done by calculating a matrix of the relationships between each pair of variables in the dataset. The result is a symmetric matrix called a correlation matrix with a value of 1.0 along the main diagonal as each column always perfectly correlates with itself.

## SPEARMAN’S CORRELATION

Two variables may be related by a nonlinear relationship, such that the relationship is stronger or weaker across the distribution of the variables. Furthermore, the two variables being considered may have a non-Gaussian distribution.

In this case, the Spearman’s correlation coefficient (named for Charles Spearman) can be used to summarize the strength between the two data samples.

This test of relationship can also be used if there is a linear relationship between the variables, but will have slightly less power (e.g. may result in lower coefficient scores).



A linear relationship between the variables is not assumed, although a monotonic relationship is assumed. This is a mathematical name for an increasing or decreasing relationship between the two variables.

If you are unsure of the distribution and possible relationships between two variables, Spearman correlation coefficient is a good tool to use.

# SAMPLE AND POPULATION

Below the definitions

## POPULATION

The term “population” is used in statistics to represent all possible measurements or outcomes that are of interest to us in a particular study.

## CENSUS

Census attempt to gather information from each and every unit of the population of interest.

## SAMPLE

The term “sample” refers to a portion of the population that is representative of the population from which it was selected.

Depending on the sampling method, a sample can have fewer observations than the population, the same number of observations, or more observations. More than one sample can be derived from the same population.

Now the question is why we use sample in statistics why don’t we go for census?

Why using a sample? Why not census?

* Less time consuming than a census.
* Less costly to administer than a census.
* Measuring the variable of interest may involve the destruction of the population unit.
* A population may be infinite.

## PARAMETERS AND STATISTICS

One goal of statistical inference is to estimate a population parameter from a sample statistic.

Parameters are

* Numerical characteristic of a population
* Constant (fixed) at any one moment
* Usually unknown

Statistics are

* Numerical summary of a sample
* Calculated from sample data (not constant)
* Used to estimate a parameter

## SAMPLING

A sampling method is a procedure for selecting sample elements from a population. Sampling is necessary to make inferences about a population. If sample is not representative it is biased — you cannot generalize to the population from your statistical data.

## POPULATION DISTRIBUTION

The population is the whole set of values, or individuals, you are interested in. For example, if you want to know the average height of the residents of India, that is your population, i.e. the population of India.

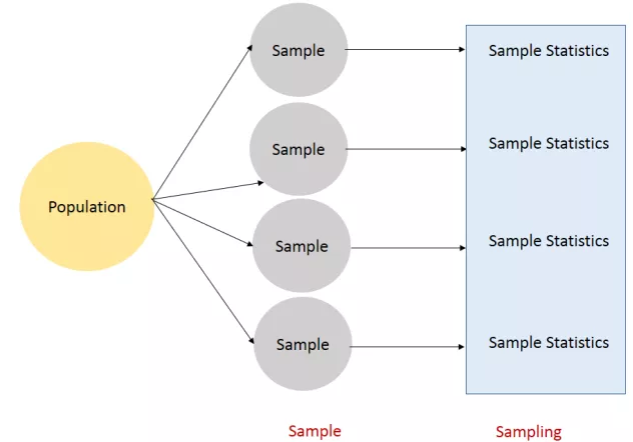
Population characteristic are mean (μ), Standard deviation (σ), proportion (P), median, percentiles etc. The value of a population characteristic is fixed. These characteristics are called population distribution. They are symbolized by Greek characters as they are population parameters.

## SAMPLE DISTRIBUTION

The sample is a subset of the population, and is the set of values you actually use in your estimation. Let’s think 1000 individual you have selected for your study to know about average height of the residents of India. This sample has some quantity computed from values e.g. mean (x), Standard deviation (s), sample proportion etc. This is called sample distribution. The mean and standard deviation are symbolized by Roman characters as they are sample statistics.

## SAMPLING DISTRIBUTION

Researchers often use a sample to draw inferences about the population that sample is from. To do that, they make use of a probability distribution that is very important in the world of statistics: the sampling distribution. It is theoretical distribution. The distribution of sample statistics is called sampling distribution.



## OUTLIER

An outlier is an observation that is numerically distant from the rest of the data.

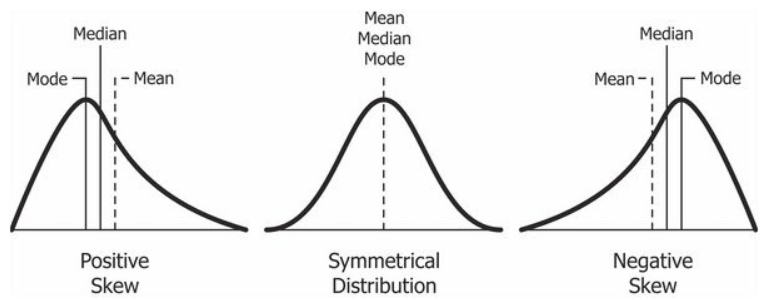
An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. Outliers can occur by chance in any distribution, but they are often indicative either of measurement error or that the population has a heavy-tailed distribution.

## SKEWNESS

It is the degree of distortion from the symmetrical bell curve or the normal distribution. It measures the lack of symmetry in data distribution.

It differentiates extreme values in one versus the other tail. A symmetrical distribution will have a skewness of 0.

There are two types of Skewness: Positive and Negative



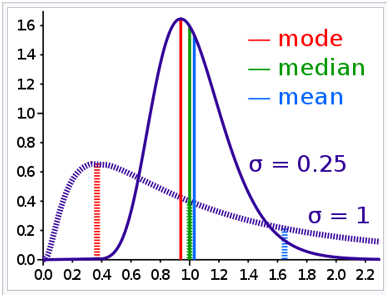
Positive Skewness means when the tail on the right side of the distribution is longer or fatter. The mean and median will be greater than the mode.

Negative Skewness is when the tail of the left side of the distribution is longer or fatter than the tail on the right side. The mean and median will be less than the mode.

The rule of thumb seems to be:

* If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.
* If the skewness is between -1 and -0.5(negatively skewed) or between 0.5 and 1(positively skewed), the data are moderately skewed.
* If the skewness is less than -1(negatively skewed) or greater than 1(positively skewed), the data are highly skewed.

Skewness is a descriptive statistic that can be used on conjunction with the histogram and the normal quantile plot to characterize the data or distribution.



## KURTOSIS

Kurtosis is all about the tails of the distribution — not the peakedness or flatness. It is used to describe the extreme values in one versus the other tail. It measures the tail-heaviness of the distribution. It is actually the measure of outliers present in the distribution.

High kurtosis in a data set is an indicator that data has heavy tails or outliers. If there is a high kurtosis, then, we need to investigate why do we have so many outliers. It indicates a lot of things, maybe wrong data entry or other things. Investigate!

Low kurtosis in a data set is an indicator that data has light tails or lack of outliers. If we get low kurtosis (too good to be true), then also we need to investigate and trim the dataset of unwanted results.

Most often, kurtosis is measured against the normal distribution.

* If the kurtosis is close to 0, then a normal distribution is often assumed. These are called mesokurtic distributions.
* If the kurtosis is less than zero, then the distribution is light tails and is called a platykurtic distribution.
* If the kurtosis is greater than zero, then the distribution has heavier tails and is called a leptokurtic distribution.

# UNIVARIATE ANALYSIS

Univariate analysis is perhaps the simplest form of statistical analysis. Like other forms of statistics, it can be inferential or descriptive. The key fact is that only one variable is involved.

Univariate analysis can yield misleading results in cases in which multivariate analysis is more appropriate.

## DESCRIPTIVE METHODS

Descriptive statistics describe a sample or population. They can be part of exploratory data analysis.

The appropriate statistic depends on the level of measurement.

* For nominal variables (categorical variable with no ordering), a frequency table and a listing of the mode(s) is sufficient.
* For ordinal (categorical variable with ordering) variables the median can be calculated as a measure of central tendency and the range (and variations of it) as a measure of dispersion.
* For interval level variables (with no true zero), the arithmetic mean (average) and standard deviation are added to the toolbox.
* For ratio level variables (with true zero), we add the geometric mean and harmonic mean as measures of central tendency and the coefficient of variation as a measure of dispersion.

## INFERENTIAL METHODS

Inferential methods allow us to infer from a sample to a population.

* For a nominal variable a one-way chi-square (goodness of fit) test can help determine if our sample matches that of some population.
* For interval and ratio level data, a one-sample t-test can let us infer whether the mean in our sample matches some proposed number (typically 0).

# STATISTICAL INFERENCE METHODS

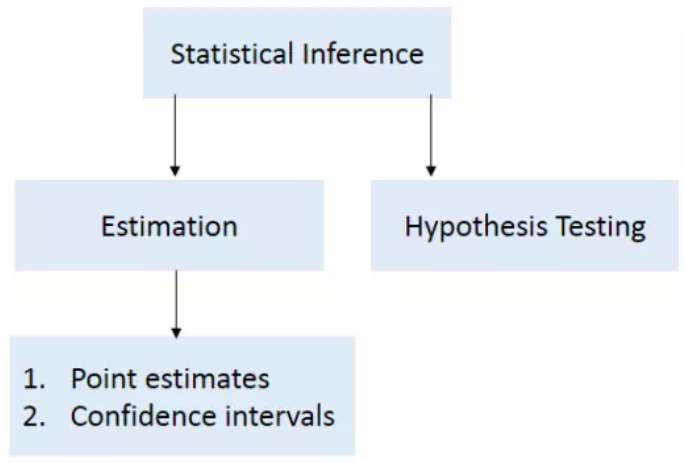
The key thing in statistical inference is, based on sample information draw conclusion about the population from where the sample was drawn.

There are two types of statistical inference methods:

1. We can estimate population parameters.
2. Test hypothesis about these parameters.

There are two ways to estimate the value of a population parameter:

* The first one is so called **point estimate**. It is a single number that is the best guess for the population parameters.
* And the second one is the **interval estimate**. It is a range of values within which we expect the parameters to fall around.



## POINT ESTIMATE

The statistic calculated from the sample is a point estimate of the corresponding population parameter. Point estimate serve as a "best guess" or "best estimate" of an unknown population parameter. For example:

* The sample average is a point estimate of the true population mean.
* The sample proportion is a point estimate of the population proportion.

The SE of the statistic provides a measure of the precision of the estimate

* A larger SE indicates a less precise point estimate
* A smaller SE indicates a more precise point estimate

**For example**: What is the average height of south Indian man?

We’re going to consider the south India as population and collected a simple random sample of 20,000 people from this population.

We want to estimate the population mean based on the sample. The most intuitive way to do this is to simply take the sample mean. That is, to estimate the average height of all south Indian people, take the average height for the sample. Let’s think all 20,000 samples that we collected the sample mean ¯x = 172.72 cm. Then the height 172.72 cm is called a point estimate of the population mean. If we can only choose one value to estimate the population mean, this is our best guess.

Suppose we take a new sample of another 30,000 people and recompute the mean; we will probably not get the exact same answer that we got first time. Point estimates generally vary from one sample to another and this sampling variation suggests our estimate may be close, but it may not be exactly equal to the parameter. So, the moral of the story is point estimates are not exact and we should not expect our estimate to be very good.

## CONFIDENCE INTERVALS

A Confidence Interval is an interval of numbers containing the most believable values for our Population Parameter. A point estimate gives a single value for a parameter. However, a point estimate is not perfect and usually there is some error in the estimate. Instead of giving just a point estimate of a parameter, it would be better to provide a range of values for the parameter. A plausible range of values for the population parameter is called a confidence interval.

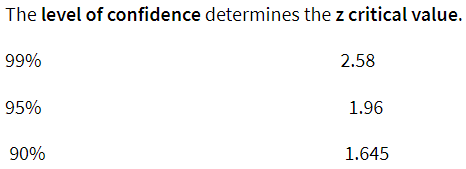
So basically, we try to build the confidence interval around the point estimate. The probability that this procedure produces an interval that contains the actual true parameter value is known as the Confidence Level and is generally chosen to be 0.9, 0.95 or 0.99.

Confidence Intervals (for a population mean) take the form:

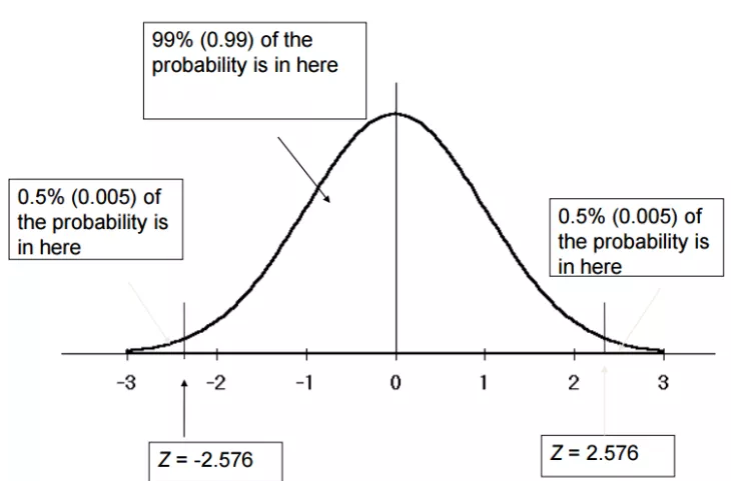
Point Estimate +/- Critical Value x Standard Error



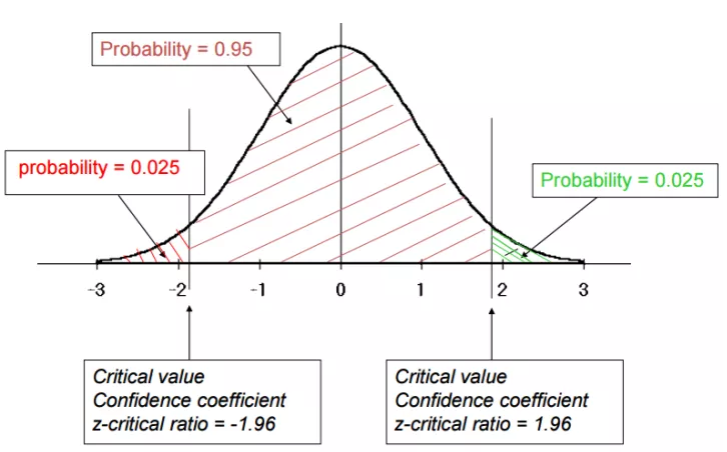
Z-scores are appropriate confidence coefficients for a confidence interval of the mean when the population standard deviation (σ) is known.



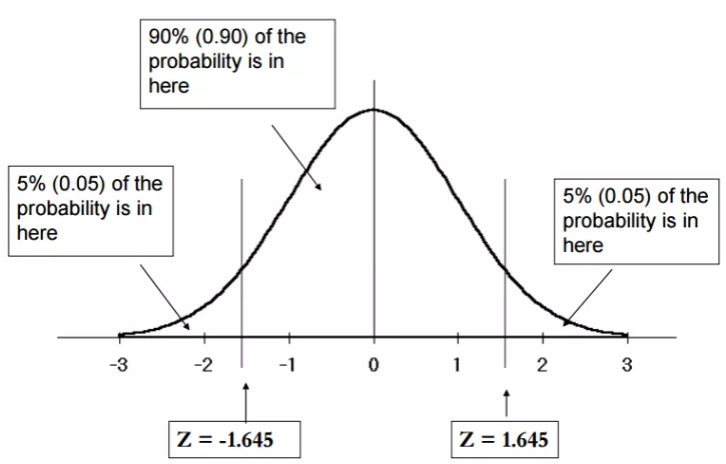
Confidence Coefficients for 99% Confidence Interval from standard normal distribution:



Confidence Coefficients for 95% Confidence Interval from standard normal distribution:

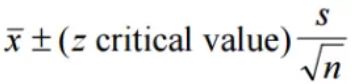


Confidence Coefficients for 90% Confidence Interval from standard normal distribution:



However, most of the time when the population mean is being estimated from sample data the population variance is unknown and must also be estimated from sample data. The sample standard deviation (s) provides an estimate of the population standard deviation (σ).

Since n is large the unknown σ can be replaced by the sample value s.



The standard error represents the standard deviation associated with the estimate, and roughly 95% of the time the estimate will be within 2 standard errors of the parameter.

If the interval spreads out 2 standard errors from the point estimate, we can be roughly 95% confident that we have captured the true parameter: point estimate ± 2 × SE. Similarly, we can construct 90% and 99% confidence interval using above z critical value.

## MARGIN OF ERROR

In a confidence interval, Z × SE is called the margin of error.

## CONDITIONS FOR CONFIDENCE INTERVAL FOR POPULATION MEAN

Some conditions need to be satisfied to use the above formula and to build the confidence interval. In fact, since this method is based on CLT it follows the same conditions for CLT.

* Independence: Sampled observations must be independent.
* Random sample/random assignment
* If sampling without replacement, then needs to be n < 10% of population.
* Sample size/ skew: Either the population distribution is normal, or if the population distribution is skewed, the sample size is large (rule of thumb: n > 30)

If sample size is less than 50 and population Standard deviation in not known then we use t-distribution instead of Z distribution.

# TYPE OF FUNCTIONS

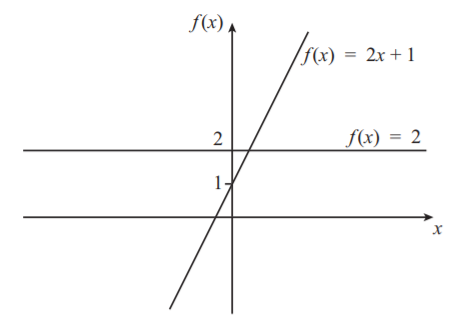
Below are few important functions that one should know:

## CONSTANT FUNCTION



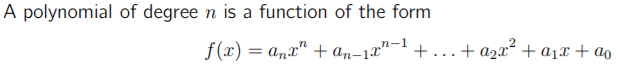
## LINEAR FUNCTION





It is important to notice that the graphs of constant functions and linear functions are always straight lines.

## POLYNOMIAL FUNCTION



Functions containing other operations such as square root are not polynomial functions:



## EXPONENTIAL FUNCTIONS

A function in which the variable appears as an exponent (power) is called an exponential function



## LOGARITHMIC FUNCTION

A function in which the variable appears as an argument of a logarithm is called a logarithmic function.

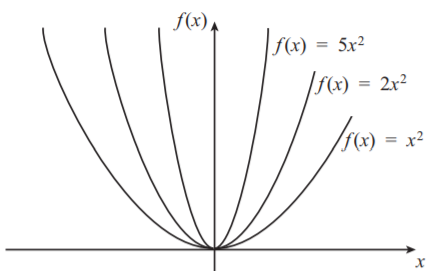


## CHANGING COEFFICIENTS OF POLYNOMIAL FUNCTIONS

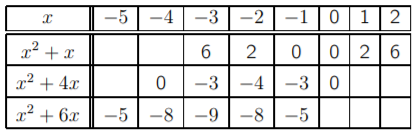
Impact of changing the coefficients of polynomial function can be studied by plotting the graphs of the equation.



You can see from the graph that, as the coefficient of x2 is increased, the graph is stretched vertically (that is, in the y direction).

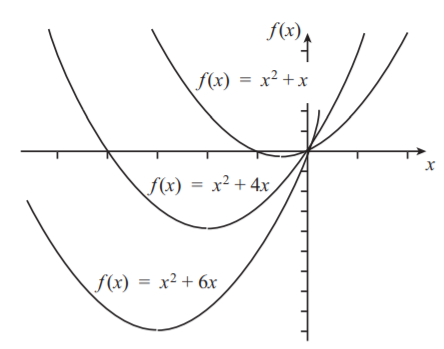


Now let us look at some other quadratic functions to see what happens when we vary the coefficient of x, rather than the coefficient of x2. We shall use a table of values in order to plot the graphs, but we shall fill in only those values near the turning points of the functions.

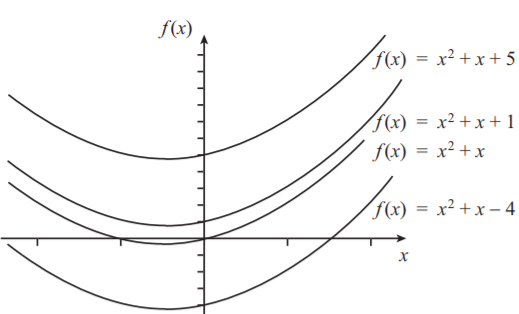


You can see the symmetry in each row of the table, demonstrating that we have concentrated on the region around the turning point of each function. We can now use these values to plot the graphs.

As you can see, increasing the positive coefficient of x in this polynomial moves the graph down and to the left.

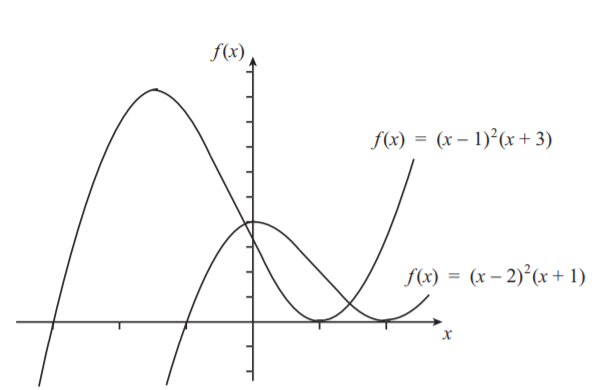


So now we know what happens when we vary the x2 coefficient, and what happens when we vary the x coefficient. But what happens when we vary the constant term at the end of our polynomial.



As we can see straight away, varying the constant term translates the x2 + x curve vertically. Furthermore, the value of the constant is the point at which the graph crosses the f(x) axis.

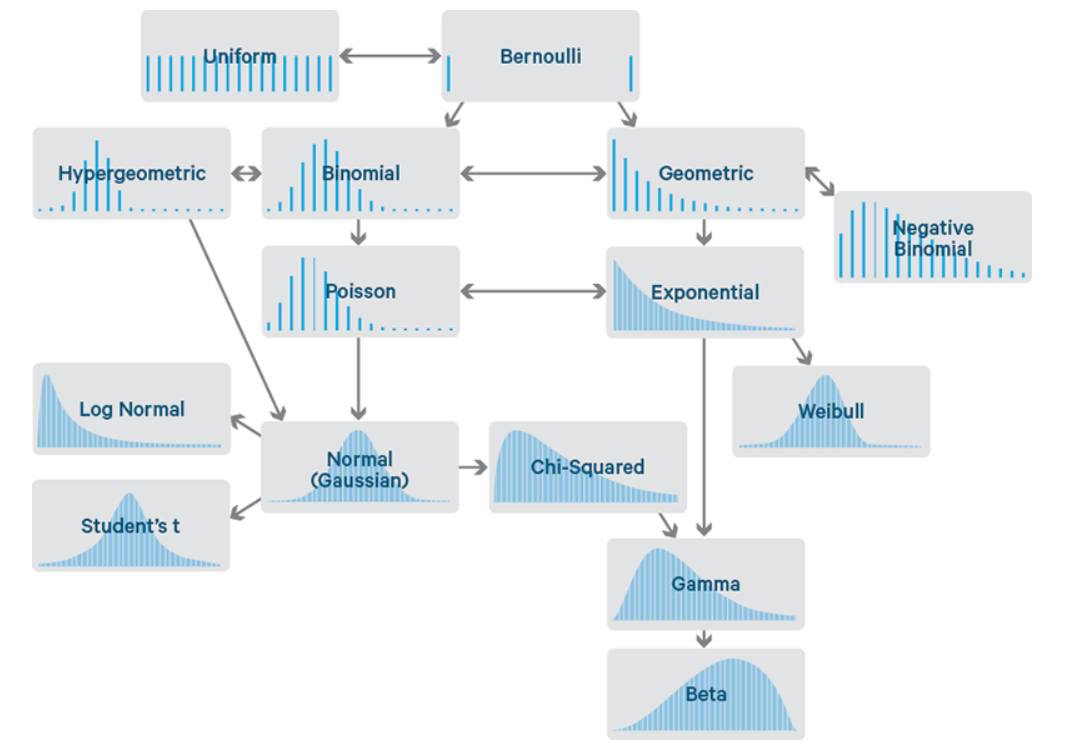
Also, roots of polynomial function have below graph:



# TYPE OF DISTRIBUTIONS

Below are some of the important distributions in statistics:

* The horizontal axis in each box is the set of possible numeric outcomes. The vertical axis describes the probability of outcomes.
* Some distributions are discrete, over outcomes that must be integers like 0 or 5. These appear as sparse lines, one for each outcome, where line height is the probability of that outcome.
* Some are continuous, for outcomes that can take on any real numeric value like -1.32 or 0.005. These appear as dense curves, where it’s areas under sections of the curve that give probabilities.
* The sums of the heights of lines, and areas under the curves, are always 1.

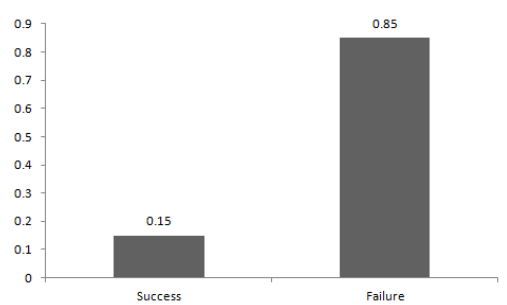


## BERNOULLI

A Bernoulli distribution has only two possible outcomes, namely 1 (success) and 0 (failure), and a single trial. So, the random variable X which has a Bernoulli distribution can take value 1 with the probability of success, say p, and the value 0 with the probability of failure, say q or 1-p.

For example, the occurrence of a head denotes success, and the occurrence of a tail denotes failure. Probability of getting a head = 0.5 = Probability of getting a tail since there are only two possible outcomes.

The probabilities of success and failure need not be equally likely, like the result of a fight between me and Undertaker. He is pretty much certain to win. So, in this case probability of my success is 0.15 while my failure is 0.85



Basically, expected value of any distribution is the mean of the distribution. The expected value of a random variable X from a Bernoulli distribution is found as follows:

E(X) = 1\*p + 0\*(1-p) = p

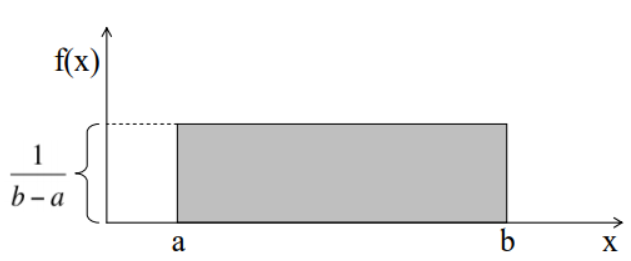
The variance of a random variable from a Bernoulli distribution is: p(1-p)

There are many examples of Bernoulli distribution such as whether it’s going to rain tomorrow or not where rain denotes success and no rain denotes failure and Winning (success) or losing (failure) the game.

## UNIFORM

When you roll a fair die, the outcomes are 1 to 6. The probabilities of getting these outcomes are equally likely and that is the basis of a uniform distribution. Unlike Bernoulli Distribution, all the n number of possible outcomes of a uniform distribution are equally likely.

You can see that the shape of the Uniform distribution curve is rectangular, the reason why Uniform distribution is called rectangular distribution.



For a Uniform Distribution, a and b are the parameters. F(x) is known as **Probability Density Function (PDF)**.

**Example**: The number of bouquets sold daily at a flower shop is uniformly distributed with a maximum of 40 and a minimum of 10. Try calculating the probability that the daily sales will fall between 15 and 30.

P(sales) = (30-15) \*(1/ (40-10)) = 15/30 = 0.5

Similarly, the probability the sales are greater than 20 is:

P(sales) = (40-20) \* (1/ (40-10)) = 20/30 = 0.67

The mean and variance of X following a uniform distribution is:

Mean(X) = (a + b)/2

Variance(X) = (b-a) ²/12

## BINOMIAL DISTRIBUTION

An experiment with only two possible outcomes repeated n number of times is called binomial.

A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose and where the probability of success and failure is same for all the trials is called a Binomial Distribution.

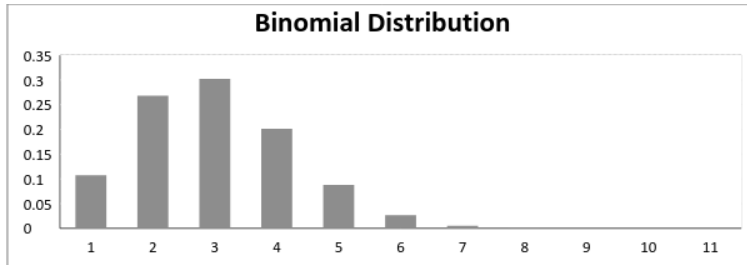
Each trial is independent since the outcome of the previous toss doesn’t determine or affect the outcome of the current toss.

The parameters of a binomial distribution are n and p where n is the total number of trials and p is the probability of success in each trial (probability of failure q = 1 - p).

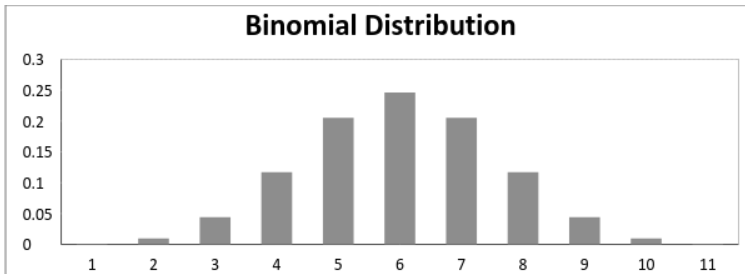
On the basis of the above explanation, the properties of a Binomial Distribution are

1. Each trial is independent.
2. There are only two possible outcomes in a trial- either a success or a failure.
3. A total number of n identical trials are conducted.
4. The probability of success and failure is same for all trials. (Trials are identical.)

A binomial distribution graph where the probability of success does not equal the probability of failure looks like



Now, when probability of success = probability of failure, in such a situation the graph of binomial distribution looks like



The mean and variance of a binomial distribution are given by:

Mean(µ) = n\*p

Variance = n\*p\*q

## NORMAL DISTRIBUTION

Normal distribution represents the behaviour of most of the situations in the universe. Any distribution is known as Normal distribution if it has the following characteristics:

* The mean, median and mode of the distribution coincide.
* The curve of the distribution is bell-shaped, unimodal and symmetrical about the line x=μ i.e. exactly half of the values are to the left of the centre and the other half to the right.
* The curve is concentrated in the centre and decreases on either side i.e. most observations are close to the mean.
* It can be determined entirely by the values of mean and std dev.
* The total area under the curve is 1.

A normal distribution is highly different from Binomial Distribution. However, if the number of trials approaches infinity then the shapes will be quite similar.

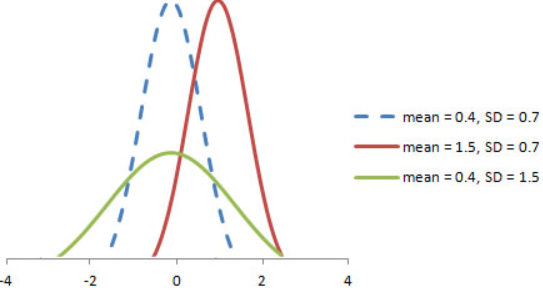
The mean and variance of a random variable X which is said to be normally distributed is given by:

Mean = µ

Variance = σ^2

Here, µ (mean) and σ (standard deviation) are the parameters. The graph of a random variable X ~ N (µ, σ) is shown below.

Note: Higher the std deviation more spread out the curve is.



A normal distribution is characterized by:

* Mean (middle value)
* Std deviation (Dispersion / spread)

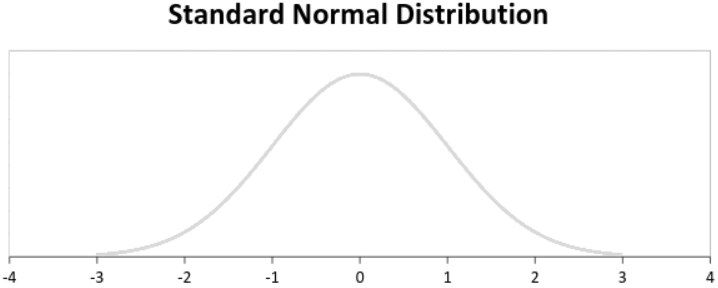
Any distribution with more than 30 samples will tend to converge to normal distribution.

The empirical 68-95-99.7 rule states that for a normal distribution:

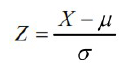
* 68.3% of the data will fall within 1 SD of mean
* 95.4% of the data will fall within 2 SD’s of the mean
* Almost all (99.7%) of the data will fall within 3 SD’s of the mean

## STANDARD NORMAL DISTRIBUTION

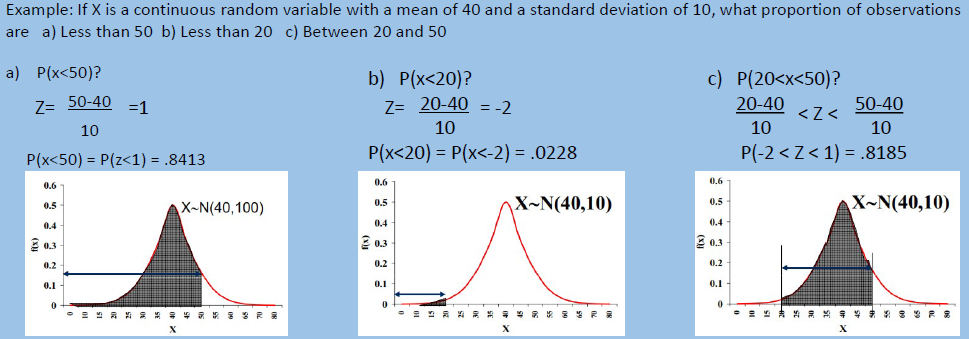
Standard Normal distribution is a special case of the Normal distribution which has a mean of 0 and a standard deviation of 1. It is also known as Z Distribution. Z distribution is used for population.

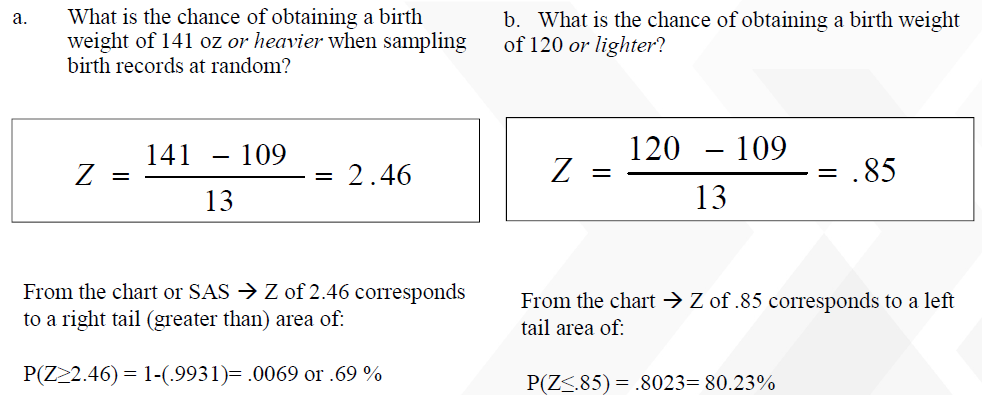


Any normal distribution can be converted to a Standard normal distribution through:



Z table always give left side probability.





# CENTRAL LIMIT THEOREM

It is always not possible to get the true information about the population. In this case we have to live with samples. For e.g. we don’t know the actual average income for India, but can estimate it based on a random sample picked from the Indian population.



If we take a similar second sample, it is extremely unlikely that the average calculated for the second sample will be the same as the average calculated for the first sample. In fact, statisticians know that repeated samples from the same population give different sample means.

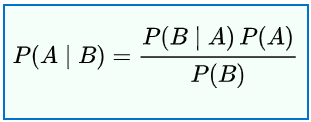
They have also proven that the distribution of these sample means will always be normally distributed, regardless of the shape of the parent population. This is known as the Central Limit Theorem.

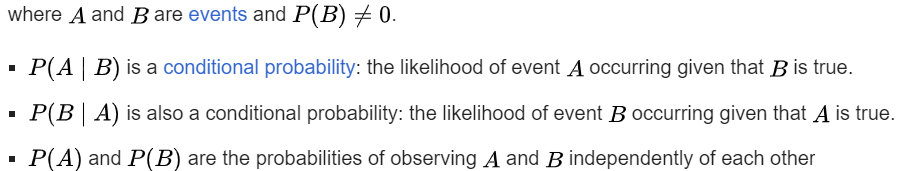
As the sample size n increases, the variance of the sampling distribution decreases. This is logical, because the larger the sample size, the closer we are to measuring the true population sample size, the closer we are to measuring the true population parameters.

# NAÏVE BAYES THEOREM

Bayes' theorem (alternatively Bayes' law or Bayes' rule) describes the probability of an event, based on prior knowledge of conditions that might be related to the event. For example, if cancer is related to age, then, using Bayes' theorem, a person's age can be used to more accurately assess the probability that they have cancer, compared to the assessment of the probability of cancer made without knowledge of the person's age.

Bayes' theorem is stated mathematically as the following equation

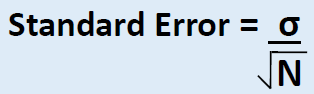




# STANDARD ERROR

Since all samples drawn from a population are similar BUT NOT the same as population, we calculate a Standard Error.

**Standard Error is the standard deviation of the sample means from the population mean**. Also, Standard Error ultimately converges to the Standard Deviation of the population.

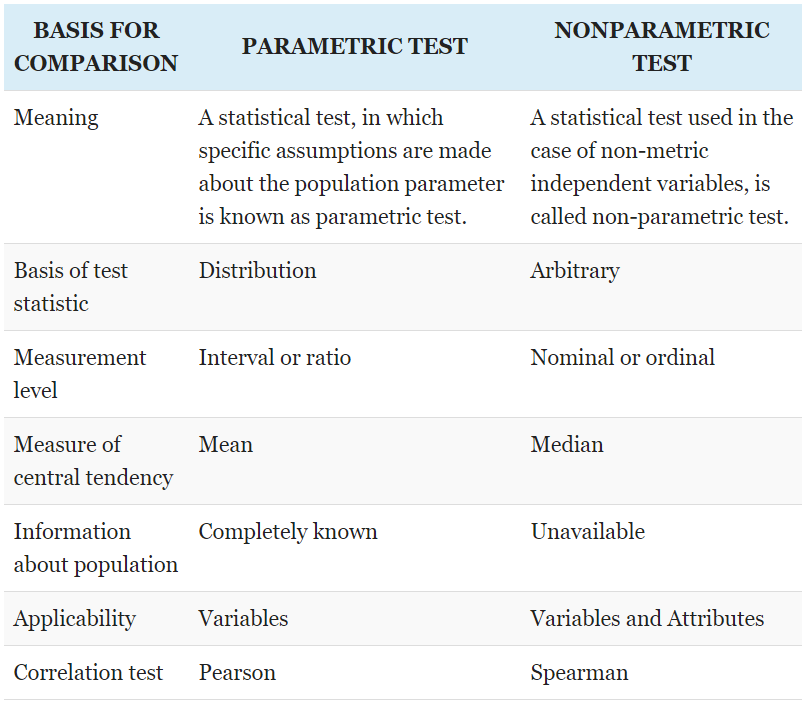


# PARAMETRIC AND NONPARAMETRIC TESTS

Parametric tests make certain assumptions about a data set namely that the data is drawn from a population with a specific (normal) distribution. Non-parametric tests make fewer assumptions about the data set.

Nonparametric tests are also called distribution-free tests because they don't assume that your data follow a specific distribution. You may have heard that you should use nonparametric tests when your data don't meet the assumptions of the parametric test, especially the assumption about normally distributed data.

The majority of elementary statistical methods are parametric, and parametric tests generally have higher statistical power.

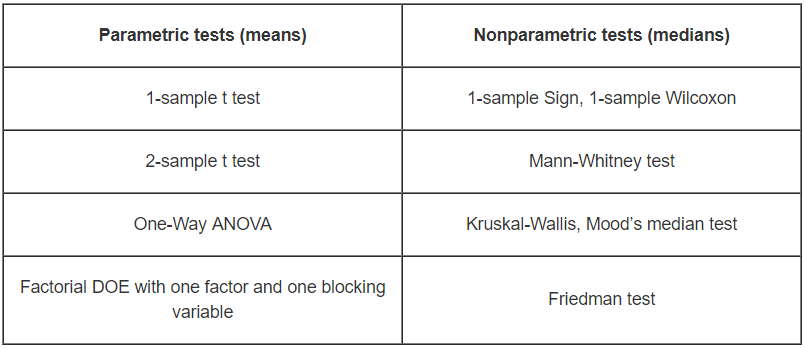


The decision often depends on whether the mean or median more accurately represents the center of your data’s distribution.

* If the mean accurately represents the center of your distribution and your sample size is large enough, consider a parametric test because they are more powerful.
* If the median better represents the center of your distribution, consider the nonparametric test even when you have a large sample.
* Finally, if you have a very small sample size, you might be stuck using a nonparametric test.

## HYPOTHESIS TESTS OF THE MEAN AND MEDIAN

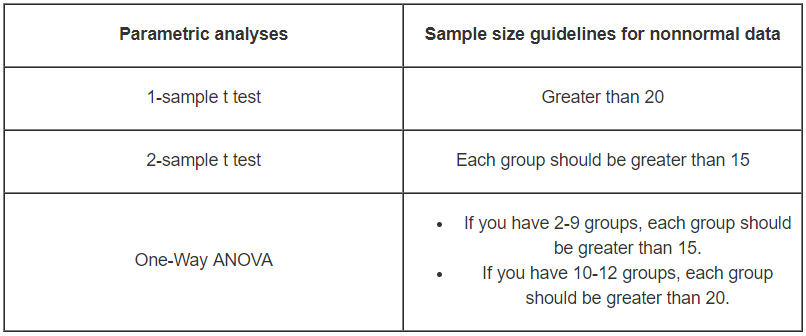
Nonparametric tests are like a parallel universe to parametric tests. The table shows related pairs of hypothesis tests



## REASONS TO USE PARAMETRIC TESTS

**Reason 1: Parametric tests can perform well with skewed and non-normal distributions.**

This may be a surprise but parametric tests can perform well with continuous data that are non- normal if you satisfy the sample size guidelines in the table below.



**Reason 2: Parametric tests can perform well when the spread of each group is different.**

While nonparametric tests don’t assume that your data follow a normal distribution, they do have other assumptions that can be hard to meet. For nonparametric tests that compare groups, a common assumption is that the data for all groups must have the same spread (dispersion). If your groups have a different spread, the nonparametric tests might not provide valid results.

On the other hand, if you use the 2-sample t test or One-Way ANOVA, you can simply mention equal\_var= False and you’re good to go even when the groups have different spreads.

**Reason 3: Statistical power.**

Parametric tests usually have more statistical power than nonparametric tests. Thus, you are more likely to detect a significant effect when one truly exists.

## REASONS TO USE NONPARAMETRIC TESTS

**Reason 1: Your area of study is better represented by the median.**

The fact that you can perform a parametric test with non-normal data doesn’t imply that the mean is the best measure of the central tendency for your data.

For example, the centre of a skewed distribution, like income, can be better measured by the median where 50% are above the median and 50% are below. If you add a few billionaires to a sample, the mathematical mean increases greatly even though the income for the typical person doesn’t change.

When your distribution is skewed enough, the mean is strongly affected by changes far out in the distribution’s tail whereas the median continues to more closely reflect the centre of the distribution.

**Reason 2: You have a very small sample size.**

If you don’t meet the sample size guidelines for the parametric tests and you are not confident that you have normally distributed data, you should use a nonparametric test. When you have a really small sample, you might not even be able to ascertain the distribution of your data because the distribution tests will lack sufficient power to provide meaningful results.

**Reason 3: You have ordinal data, ranked data, or outliers that you can’t remove.**

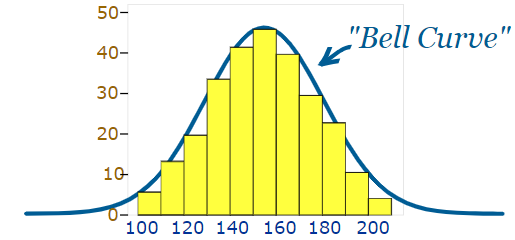
Typical parametric tests can only assess continuous data and the results can be significantly affected by outliers. Conversely, some nonparametric tests can handle ordinal data, ranked data, and not be seriously affected by outliers.

# TYPE OF DISTRIBUTIONS FOR HYPOTHESIS TESTING

Below are important distributions that are used to in hypothesis testing:

## NORMAL DISTRIBUTION

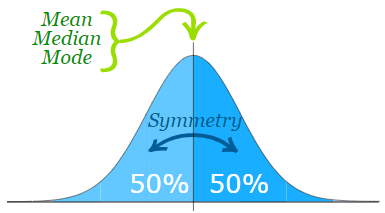
It is used specially in testing hypotheses about means or proportions of samples drawn from populations whose population standard deviations are known.



The "Bell Curve" is a Normal Distribution and the yellow histogram shows some data that follows it closely, but not perfectly (which is usual).

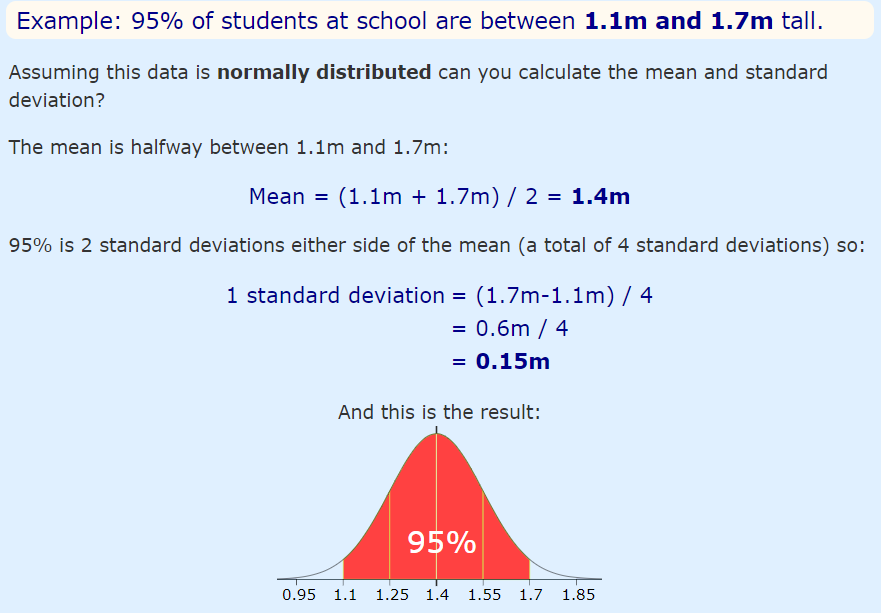
The Normal Distribution has:

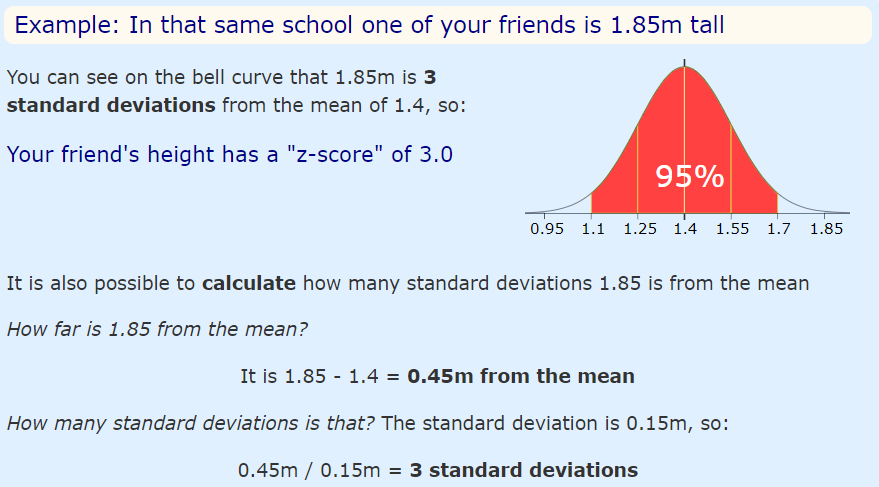
* mean = median = mode
* symmetry about the center
* 50% of values less than the mean
* and 50% greater than the mean



## STANDARD SCORE

The number of standard deviations from the mean is also called the "Standard Score", "sigma" or "z-score". Get used to those words!



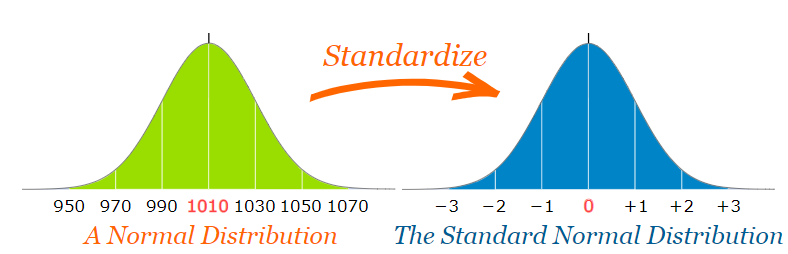


So, to convert a value to a Standard Score ("z-score"):

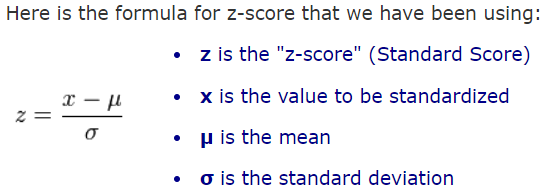
* First subtract the mean,
* Then divide by the standard deviation

## Z DITRIBUTION

A probability density function which has a normal distribution with a mean equal to zero and a standard deviation equal to one. We can take any Normal Distribution and convert it to The Standard Normal Distribution.

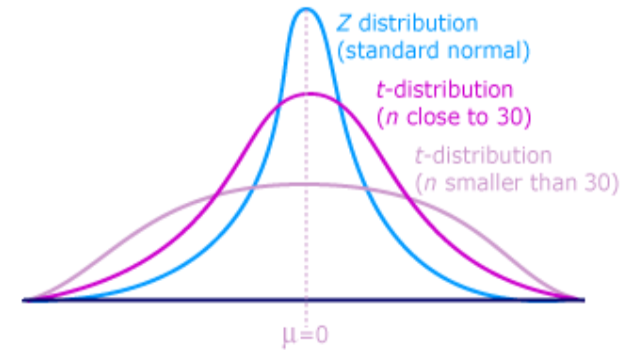


Z statistics can be calculated using below formula:



## T DISTRIBUTION

The t distribution (aka, Student’s t-distribution) is a probability distribution that is used to estimate population parameters when the sample size is small and/or when the population variance is unknown.



T – Statistics can be calculated using below formula:



where xbar is the sample mean, μ is the population mean, s is the standard deviation of the sample, and n is the sample size. The distribution of the t statistic is called the t distribution or the Student t distribution. DF for t test is n-1 where n is no of observations.

The t distribution allows us to conduct statistical analyses on certain data sets that are not appropriate for analysis, using the normal distribution. Example if the standard deviation of population is not known or when sample size is less (usually less than 30).

GENRAL CORRECT RULE: If σ is not known, then using t-distribution is correct. If σ is known, then using the normal distribution is correct.

It resembles the normal distribution and as the sample size increases the t-distribution looks more normally distributed with the values of means and standard deviation of 0 and 1 respectively.

PROPERTIES OF T-DISTRIBUTION

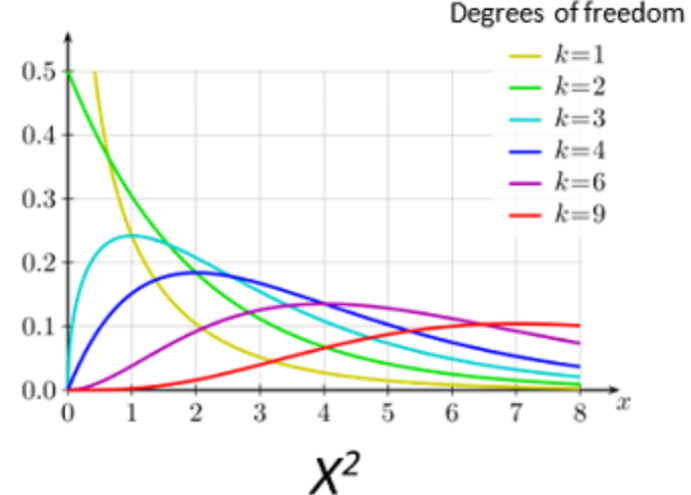
* Like, standard normal distribution the shape of the student distribution is also bell-shaped and symmetrical with mean zero.
* The student distribution ranges from –∞ to ∞ (infinity).
* The shape of the t-distribution changes with the change in the degrees of freedom.
* The variance is always greater than one and can be defined only when the degrees of freedom ν ≥ 3 and is given as: Var (t) = [ν/ν -2]
* It is less peaked at the centre and higher in tails, thus it assumes platykurtic shape.
* The t-distribution has a greater dispersion than the standard normal distribution. And as the sample size ‘n’ increases, it assumes the normal distribution. Here the sample size is said to be large when n ≥ 30.

APPLICATIONS OF THE T-DISTRIBUTION

* Test of the Hypothesis of the population mean.
* Test of Hypothesis of the difference between the two means (Two sample T Test).
* Test of Hypothesis of the difference between two means with dependent samples (Two Sample Paired T Test).
* Test of Hypothesis about the coefficient of correlation.

## CHI SQUARE DISTRIBUTION

The distribution of the chi-square statistic is called the chi-square distribution.



Suppose we conduct the following statistical experiment. We select a random sample of size n from a normal population, having a standard deviation equal to σ. We find that the standard deviation in our sample is equal to s. Given these data, we can define a statistic, called chi-square, using the following equation:



The distribution of the chi-square statistic is called the chi-square distribution.



Where:

Y0 is a constant that depends on the number of degrees of freedom

Χ2 is the chi-square statistic

v = n - 1 is the number of degrees of freedom

e is a constant equal to the base of the natural logarithm system (approximately 2.71828).

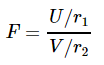
Y0 is defined, so that the area under the chi-square curve is equal to one.

The chi-square distribution has the following properties:

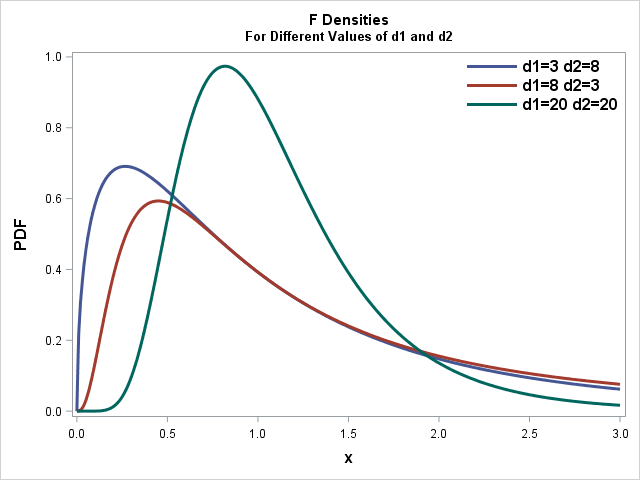
* The mean of the distribution is equal to the number of degrees of freedom: μ = v.
* The variance is equal to two times the number of degrees of freedom: σ2 = 2 \* v
* When the degrees of freedom are greater than or equal to 2, the maximum value for Y occurs when Χ2 = v - 2.
* As the degrees of freedom increase, the chi-square curve approaches a normal distribution.

## F DISTRIBUTION

If U and V are independent chi-square random variables with r1 and r2 degrees of freedom, respectively, then:



follows an F-distribution with r1 numerator degrees of freedom and r2 denominator degrees of freedom. We write F ~ F (r1, r2).



F-distributions are generally skewed. The shape of an F-distribution depends on the values of r1 and r2, the numerator and denominator degrees of freedom, respectively, as this picture pirated from your textbook illustrates:

# HYPOTHESIS TESTING

A hypothesis is an assumption about a population parameter. It is a statement about the population that may or may not be true. Hypothesis testing aims to make a statistical conclusion about accepting or not accepting the hypothesis.

So, a statistical hypothesis is a declaration or estimation concerning one or more populations. To prove that a hypothesis is true, or false, with absolute certainty, we would need absolute knowledge. That is, we would have to examine the entire population. Instead, hypothesis testing concerns on how to use a random sample to judge if it is evidence that supports or not the hypothesis.

Hypothesis testing is formulated in terms of two hypotheses:

* **H0**: the null hypothesis;
* **HA**: the alternate hypothesis

## NULL HYPOTHESIS (H0)

* Represents the status quo.
* The hypothesis that states there is no statistical significance between two variables in the hypothesis.
* Believed to be true unless there is overwhelming evidence to the contrary.
* It is the hypothesis the researcher is trying to disprove.

Example: It is hypothesised that flowers watered with lemonade will grow faster than flowers watered with plain water.

Null hypothesis: There is no statistically significant relationship between the type of water used and the growth of the flowers.

## ALTERNATIVE HYPOTHESIS (HA):

* Inverse of the null hypothesis.
* States that there is a statistical significance between two variables.
* Holds true if the null hypothesis is rejected.
* Usually what the researcher thinks is true and is testing

Null hypothesis: If one plant is fed lemonade for one month and another is fed plain water, there will be no difference in growth between the two plants.

Alternative Hypothesis: If one plant is fed lemonade for one month and another is fed plain water, the plant that is fed lemonade will grow more than the plant that is fed plain water.

**IMPORTANT:** Failure to reject H0 does not mean the null hypothesis is true. There is no formal outcome that says “accept H0.” It only means that we do not have sufficient evidence to support HA.

## P-VALUE

The p-value is a way of quantifying the strength of the evidence against the null hypothesis and in favor of the alternative. Formally the p-value is a conditional probability.

The p-value is the probability of observing data at least as favourable to the alternative hypothesis as our current data set, if the null hypothesis is true. We typically use a summary statistic of the data, here the sample mean, to help compute the p-value and evaluate the hypotheses.

The P-value is the actual area under the standard normal distribution curve (or other curve, depending on what statistical test is being used) representing the probability of a particular sample mean occurring if the null hypothesis is true.

* If P-value ≤ α, reject the null hypothesis.
* If P-value > α, do not reject the null hypothesis.

## ONE-TAILED TEST

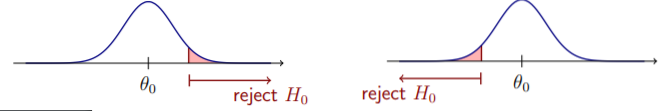
A one-tailed test (right or left) indicates that the null hypothesis should be rejected when the test value is in the critical region on one side of the mean.

## TWO-TAILED TEST

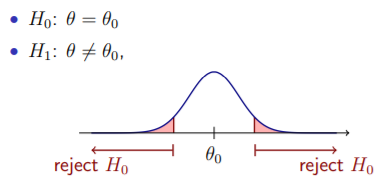
In a two-tailed test, the null hypothesis should be rejected when the test value is in either of the two critical regions.

Examples of one tailed test:





Example of two tailed test:



**Example1:** A medical researcher is interested in finding out whether a new medication will have any undesirable side effects. The researcher is particularly concerned with the pulse rate of the patients who take the medication.

* The hypotheses to test are whether the pulse rate will be different from the mean pulse rate of 82 beats per minute?
* H0: µ = 82 H1: µ ≠ 82
* **This is a two-tailed test.**

**Example2:** A chemist invents an additive to increase the life of an automobile battery. If the mean lifetime of the battery is 36 months, then his hypotheses are

* H0: µ ≤ 36 H1: µ > 36
* **This is a right-tailed test.**

**Example3:** A contractor wishes to lower heating bills by using a special type of insulation in houses. If the average of the monthly heating bills is $78, her hypotheses about heating costs will be

* H0: µ ≥ $78 H0: µ < $78
* This is a left-tailed test

## STATISTICAL TEST

A **statistical test** uses the data obtained from a sample to make a decision about whether or not the null hypothesis should be rejected.

## TEST VALUE

The numerical value obtained from a statistical test is called the **test value**.

## THE DECISION PROBLEM

How do we choose between H0 and HA? The standard procedure is to assume H0 is true - just as we presume innocent until proven guilty. Using probability theory, we try to determine whether there is sufficient evidence to declare H0 false.

We reject H0 only when the chance is small that H0 is true. Since our decisions are based on probability rather than certainty, we can make errors.

### TYPE I ERROR

We reject the null hypothesis when the null is true. The probability of Type I error = α. Put another way



Typical values chosen for α are .05 or .01. So, for example, if α = .05, there is a 5% chance that, when the null hypothesis is true, we will erroneously reject it.

### TYPE II ERROR

we accept the null hypothesis when it is not true. Probability of Type II error = ß. Put another way



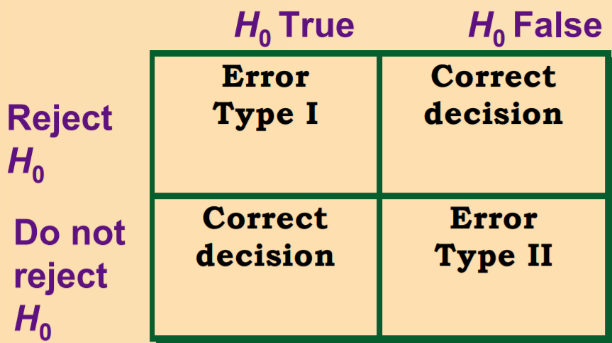
**EXAMPLE:** Type I and type II error:



Suppose µ really does equal 100. But, suppose the researcher accepts HA instead. A type I error has occurred. Or, suppose µ = 105 - but the researcher accepts H0. A type II error has occurred.

* α and ß are not independent of each other - as one increases, the other decreases. However, increases in N cause both to decrease, since sampling error is reduced.
* α and β should be as small as possible because they are probabilities of errors. They are rarely zero.

In the hypothesis-testing situation, there are four possible outcomes as shown below:



## LEVEL OF SIGNIFICANCE

The level of significance is the maximum probability of committing a type I error. This probability is symbolized by α (Greek letter alpha).

* P (type I error) = α.
* P (type II error) = β (Greek letter beta).
* Typical significance levels are: 0.10, 0.05, and 0.01.

**For example:** When α = 0.10, there is a 10% chance of rejecting a true null hypothesis.

## CRITICAL VALUE

The critical value(s) separates the critical region from the noncritical region. The symbol for critical value is C.V.

## CRITICAL REGION

The **critical or rejection region** is the range of values of the test value that indicates that there is a significant difference and that the null hypothesis should be rejected.

The **noncritical or nonrejection** region is the range of values of the test value that indicates that the difference was probably due to chance and that the null hypothesis should not be rejected.

## CONFIDENCE INTERVALS AND HYPOTHESIS TESTING

When the null hypothesis is rejected in a hypothesis-testing, the confidence interval for the mean using the same level of significance will not contain the hypothesized mean.

Likewise, when the null hypothesis is not rejected, the confidence interval computed using the same level of significance will contain the hypothesized mean.

## TEST OF HYPOTHESIS ABOUT COEFFICIENT OF CORRELATION

There are three cases of testing the hypothesis about the coefficient of correlation. These are:

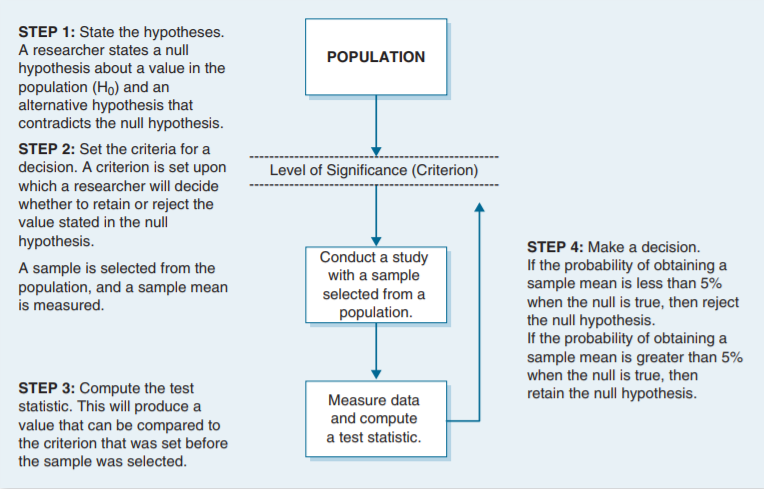
* When the population **coefficient of correlation is zero i.e. ρ = 0**. The coefficient of correlation measures the degree of relationship between the variables, and when ρ = 0, then there is no statistical relationship between the variables. To test the null hypothesis which assumes that there is no correlation between the population, it is necessary that the sample coefficient of correlation **‘r’** is known.
* When the Population **Coefficient of Correlation is equal to some other value**, other than zero, i.e. **ρ≠0**. In this case, the test based on t-distribution will not be correct and hence the hypothesis is tested using the **Fisher’s z- transformation**.
* When the hypothesis is tested for the difference between two Independent Correlation Coefficients: To test the hypothesis of two correlations derived from the two separate samples, then the difference of the two corresponding values of z is to be compared with the standard error of the difference.

The symbol for Pearson's correlation is **‘ρ’** when it is measured in the population and **‘r’** when it is measured in a sample.

## STEPS IN HYPOTHESIS TESTING

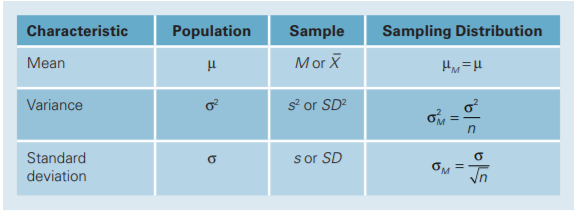
Five steps for hypothesis-testing:

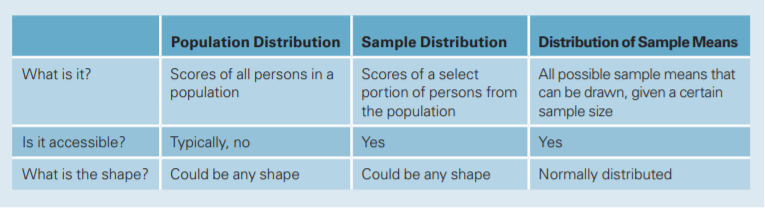
* State the hypotheses and identify the claim.
* Find the critical value(s).
* Compute the test value.
* Make the decision to reject or not reject the null hypothesis.
* Summarize the results.



## COMMON TERMS

Below are some common terms used in hypothesis testing





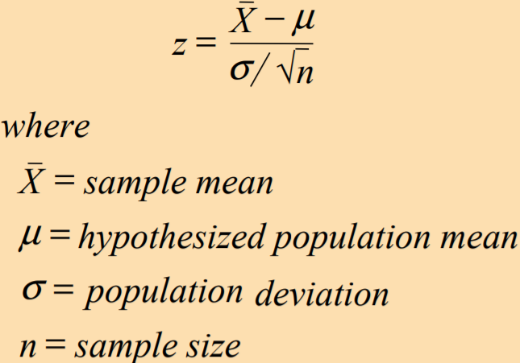
## HYPOTHESIS TESTING IN PYTHON

Below are some important points we should remember when we do hypothesis testing in Python:

* When we call scipy.stats.ttest\_ind(x, y), this makes a Hypothesis test on the value of x.mean()-y.mean(). Hence a negative value of t stats implies that x.mean > y.mean vice versa.
* By default a 2 tailed test is performed in Python.
* When null hypothesis is Ho: P1>=P2 and alternative hypothesis is Ha: P1<P2. In order to test it in Python, you write ttest\_ind(P2,P1)

# Z TEST

The z test is a statistical test for the mean of a population. It can be used when n ≥ 30, or when the population is normally distributed and std deviation (σ) is known.



Z-TEST can be performed in python using scipy.f\_oneway.

**Example1:** A researcher reports that the average salary of assistant professors is more than $42,000. A sample of 30 assistant professors has a mean salary of $43,260. At α = 0.05, test the claim that assistant professors earn more than $42,000 a year. The standard deviation of the population is $5230

**Step 1:** State the hypotheses and identify the claim.

H0: µ ≤ $42,000

H1: µ > $42,000 (claim)

**Step 2:** Find the critical value.

Since α = 0.05 and the test is a right-tailed test, the critical value is z = +1.65.

**Step 3:** Compute the test value.

z = [43,260 – 42,000]/ [5230/√30]

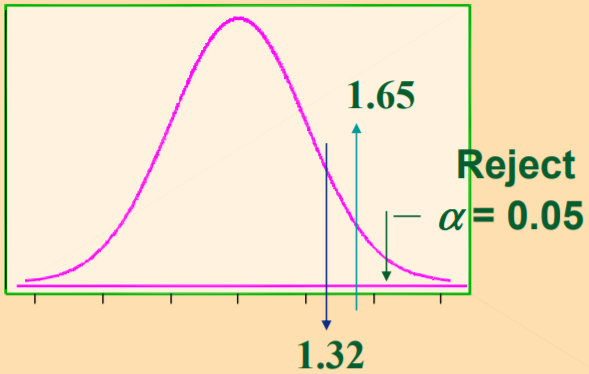
z = 1.32.

**Step 4:** Make the decision.

Since the test value, +1.32, is less than the critical value, +1.65, and not in the critical region, the decision is “Do not reject the null hypothesis.”

**Step 5:** Summarize the results.

There is not enough evidence to support the claim that assistant professors earn more on average than $42,000 a year.



**Example2:** A national magazine claims that the average college student watches less television than the general public. The national average is 29.4 hours per week, with a standard deviation of 2 hours. A sample of 30 college students has a mean of 27 hours. Is there enough evidence to support the claim at α = 0.01?

Step1: State the hypotheses and identify the claim.

H0: µ ≥ 29.4 hours

H1: µ < 29.4 (claim)

Step2: Find the critical value.

Since α = 0.01 and the test is a left-tailed test, the critical value is z = –2.33.

Step3: Compute the test value.

Z = [27 – 29.4]/ [2/√30]

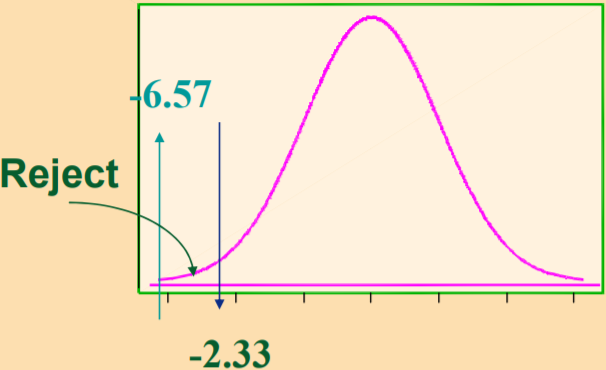
Z = -6.57

Step4: Make the decision.

Since the test value, – 6.57, falls in the critical region, the decision is to reject the null hypothesis.

Step5: Summarize the results.

There is enough evidence to support the claim that college students watch less television than the general public.



**Example3:** The Medical Rehabilitation Education Foundation reports that the average cost of rehabilitation for stroke victims is $24,672. To see if the average cost of rehabilitation is different at a large hospital, a researcher selected a random sample of 35 stroke victims and found that the average cost of their rehabilitation is $25,226. The standard deviation of the population is $3,251. At α = 0.01, can it be concluded that the average cost at a large hospital is different from $24,672?

Step 1: State the hypotheses and identify the claim.

H0: µ = $24,672

H1: µ ≠ $24,672 (claim)

Step 2: Find the critical values.

Since α = 0.01 and the test is a two-tailed test, the critical values are z = –2.58 and +2.58.

Step 3: Compute the test value.

z = [25,226 – 24,672] / [3,251/√35]

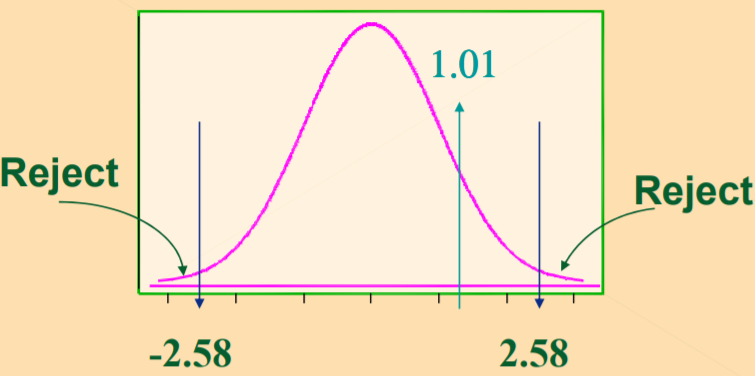
z = 1.01.

Step 4: Make the decision.

Do not reject the null hypothesis, since the test value falls in the noncritical region.

Step 5: Summarize the results.

There is not enough evidence to support the claim that the average cost of rehabilitation at the large hospital is different from $24,672.



## TWO SAMPLE Z-TEST

To test the difference between the means of two samples, the variance s² of the population is presumed to be known.

## TWO SAMPLE PAIRED Z-TEST

For carrying out two samples paired test:

* The two samples have to be of the same size.
* For the z-test, the variance s² of the population is presumed to be known.
* Where values are missing from certain observations, either the observation is removed from both samples or the missing values are estimated.

## Z-TEST IN PYTHON

By default, two tailed Z-Test is performed in Python. We can perform z test in python using scipy and statsmodels packages.

Z Test

* **F\_test** in scipy
* statsmodels.stats.weightstats.ztest

Two sample Independent Z-Test

* statsmodels.stats.weightstats.ztest

# T-TEST

When the population standard deviation is unknown and n < 30, the z test is inappropriate for testing hypotheses involving means. The t test is used in this case. The t-Distribution, also known as Student’s t-Distribution.

* Variances are equal: When the population variances, though unknown are taken as equal.
* Variances are Unequal.

## TWO SAMPLE (INDEPENDENT) T-TEST

In Testing hypothesis about the difference between two means drawn from the two systematic population whose variance is unknown, then t-test can be calculated in two ways:

By default, we assume variances to be unequal. The F test can be used to establish whether the variances are equal or not.

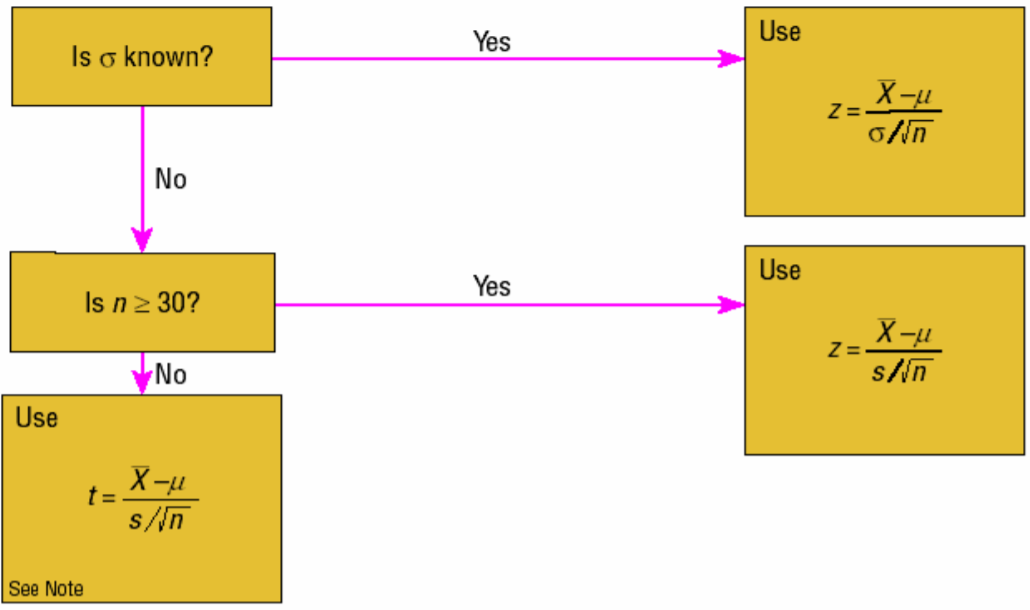
## TWO SAMPLE PAIRED T-TEST

In several situations, it is possible that the samples are drawn from the two populations that are dependent on each other.

Thus, the samples are said to be dependent, as each observation included in sample one is associated with the particular observation in the second sample. Hence, due to this property the t-test that will be used here is called the paired t-test.

The samples size of the two samples should be same.

This test is applied in the situations when before and after experiments are to be compared. Usually, two methods are adopted that are related to each other.



For Two sample paired T-Test:

* When the values are dependent, deploy a t test on the differences.
* Denote the differences with the symbol D, the mean of the population of differences, and the sample standard deviation of the differences.

IMPORTANT: This test is similar to a one sample t test, except it is done on the differences when the samples are dependent.

## T-TEST IN PYTHON

By default, two tailed Z-Test is performed in Python. We can perform T-TEST in python using

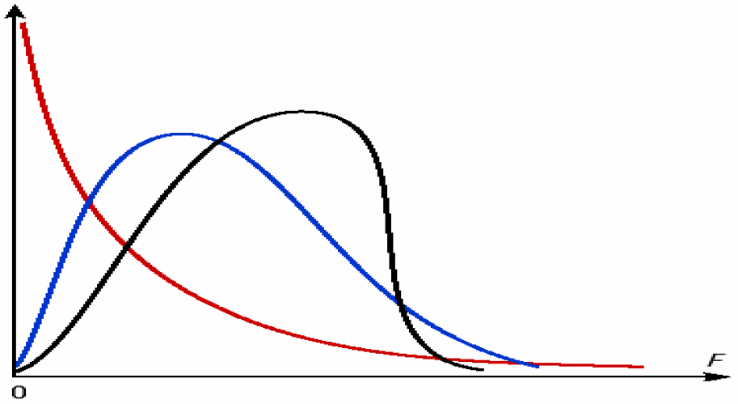
* **ttest\_1samp** using scipy

# F-TEST (ANOVA)

F – Test id used to test the difference between two Variances. For the comparison of two variances or standard deviations, an F test is used. The sampling distribution of the variances is called the F distribution.

## CHARACTERISTICS OF THE F DISTRIBUTION

* The values of F cannot be negative.
* The distribution is positively skewed.
* The mean value of F is approximately equal to 1.
* The F distribution is a family of curves based on the degrees of freedom of the variance of the numerator and denominator.



Assumptions for testing the difference between two variances:

* The populations from which the samples were obtained must be normally distributed.
* The samples must be independent of each other.

Note that the F-test is extremely sensitive to non-normality of X and Y, so you're probably better off doing a more robust test such as Levene's test or Bartlett's test unless you're reasonably sure that X and Y are distributed normally. These tests can be found in the scipy api.

# CHI SQUARE TEST

The chi-square independence test is a procedure for testing if two categorical variables are related in some population. Example: a scientist wants to know if education level and marital status are related for all people in some country.

The null hypothesis for a chi-square independence test is: **Two categorical variables are independent in some population.**

## CHI SQUARE TEST FOR INDEPENDENCE

The test is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables.

For example, in an election survey, voters might be classified by gender (male or female) and voting preference (Democrat, Republican, or Independent). We could use a chi-square test for independence to determine whether gender is related to voting preference.

When to Use Chi-Square Test for Independence

* The sampling method is simple random sampling.
* The variables under study are each categorical.
* If sample data are displayed in a contingency table, the expected frequency count for each cell of the table is at least 5.

Steps: This approach consists of four steps:

1. State the hypotheses.
2. Formulate an analysis plan
   1. Choose the Significance level.
   2. Choose the test method
3. Analyze sample data
   1. Degree of freedom DF = (r - 1) \* (c - 1)

where r is the number of levels for one categorical variable, and c is the number of levels for the other categorical variable.

* 1. Expected frequencies.
  2. Test statistic
  3. P-value

1. Interpret results.

## CHI SQUARE TEST FOR TESTING GOODNESS OF FIT

Chi square test for testing goodness of fit is used to decide whether there is any difference between the observed (experimental) value and the expected (theoretical) value.

The test is applied when you have one categorical variable from a single population. It is used to determine whether sample data are consistent with a hypothesized distribution.

## CHI SQUARE TEST FOR SINGLE VARIANCE

Chi square test for single variance is used to test a hypothesis on a specific value of the population variance. Statistically speaking, we test the null hypothesis H0: σ = σ0 against the research hypothesis H1: σ # σ0 where σ is the population mean and σ0 is a specific value of the population variance that we would like to test for acceptance.

In other words, this test enables us to test if the given sample has been drawn from a population with specific variance σ0. This is a small sample test to be used only if sample size is less than 30 in general.

## ASSUMPTIONS

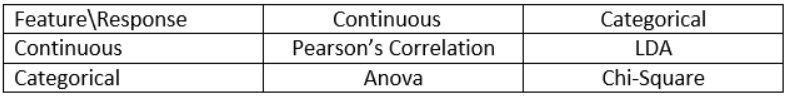
The Chi square test for single variance has an assumption that the population from which the sample has been is normal. This normality assumption need not hold for chi square goodness of fit test and test for independence of attributes.

However, while implementing these two tests, one has to ensure that expected frequency in any cell is not less than 5. If it is so, then it has to be pooled with the preceding or succeeding cell so that expected frequency of the pooled cell is at least 5.

<https://stattrek.com/regression/slope-test.aspx?Tutorial=AP>

# CORRELATION HYPOTHESIS TESTING

We can use below tests to find the correlation between variables based on their type.



# POWER OF A TEST

Power is the measure of a test's ability to accurately detect that the null hypothesis is false. Specifically, power is the probability that a test with the specified assumptions correctly rejects the null hypothesis when the alternate hypothesis is true.

# BARTLETT’S TEST

Bartlett’s test tests the null hypothesis that all input samples are from populations with equal variances. For samples from significantly non-normal populations, Levene’s test `levene` is more robust.



# LEVENE TEST

The Levene test tests the null hypothesis that all input samples are from populations with equal variances. Levene’s test is an alternative to Bartlett’s test bartlett in the case where there are significant deviations from normality.



# REFERENCES

Below are the references:

<https://www.mathsisfun.com/data/index.html>

<http://www2.univet.hu/users/jfodor/biomath/>

Important Hypothesis Test in Python

https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/

T test analysis: is it always correct to compare means?

<http://www.sthda.com/english/wiki/print.php?id=94>

Normality Test in Python

<https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/>

<https://seeing-theory.brown.edu/>