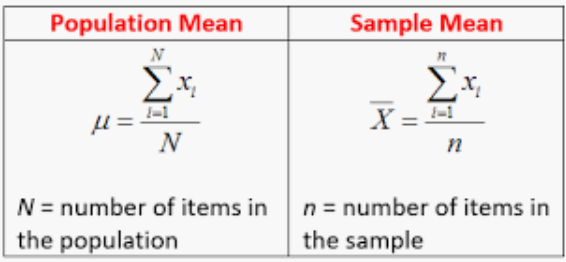
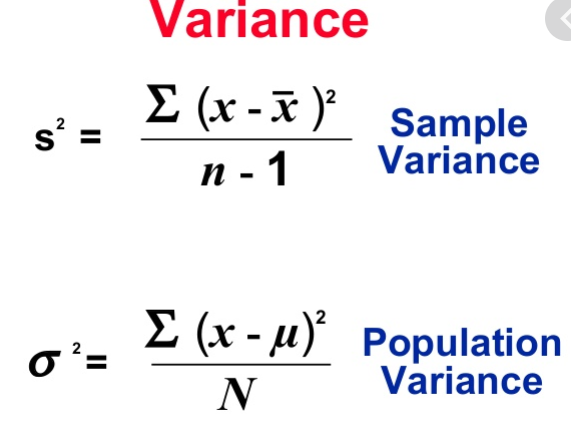
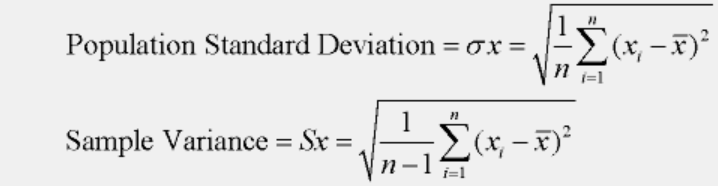
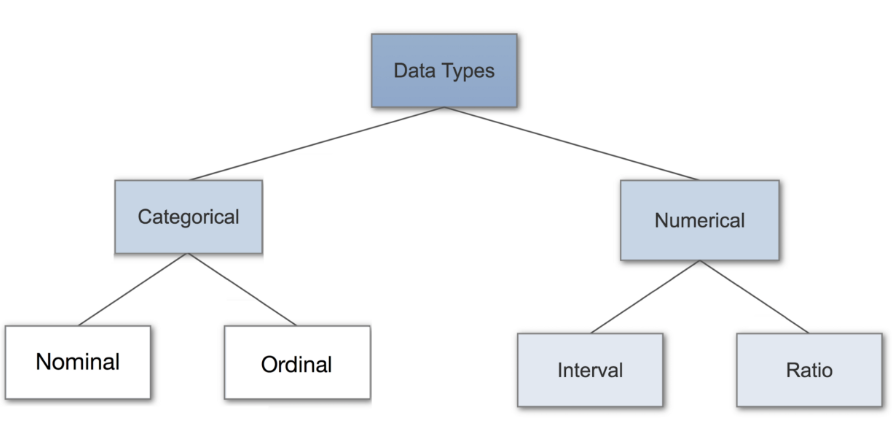
**Population (Parameters) and sample (Statistics) Symbols and formulas:**





**Different types of data:**

****

# 1. Categorical Data

Categorical data represents characteristics. Therefore, it can represent things like a person’s gender, language etc. Categorical data can also take on numerical values (Example: 1 for female and 0 for male). Note that those numbers don’t have mathematical meaning.

**1.1.** **Nominal Data:** Nominal values represent discrete units and are used to label variables, that have no quantitative value. Just think of them as „labels“. Note that nominal data that has no order. Therefore if you would change the order of its values, the meaning would not change. Such as sex, languages, etc.

**1.2 Ordinal Data:** Ordinal values represent discrete and ordered units. It is therefore nearly the same as nominal data, except that it’s ordering matters. Such as year, level of educations, etc. Note that the difference between Elementary and High School is different than the difference between High School and College. This is the main limitation of ordinal data, the differences between the values is not really known. Because of that, ordinal scales are usually used to measure non-numeric features like happiness, customer satisfaction and so on.

**2. Numerical data**

# 2.1. Discrete Data: We speak of discrete data if its values are distinct and separate. In other words: We speak of discrete data if the data can only take on certain values. This type of data ****can’t be measured but it can be counted****. It basically represents information that can be categorized into a classification. An example is the number of heads in 100 coin flips. You can check by asking the following two questions whether you are dealing with discrete data or not: Can you count it and can it be divided up into smaller and smaller parts?

# 2.2. Continuous Data: Continuous Data represents measurements and therefore their values ****can’t be counted but they can be measured****. An example would be the height of a person, which you can describe by using intervals on the real number line.

**2.3. Interval Data:** Interval values represent **ordered units that have the same difference**. Therefore we speak of interval data when we have a variable that contains numeric values that are ordered and where we know the exact differences between the values. The problem with interval values data is that they **don’t have a „true zero“**. That means in regards to our example, that there is no such thing as no temperature. With interval data, we can add and subtract, but we cannot multiply, divide or calculate ratios. Because there is no true zero, a lot of descriptive and inferential statistics can’t be applied.

**2.4. Ratio Data:** Ratio values are also ordered units that have the same difference. Ratio values are**the same as interval values, with the difference that they do have an absolute zero**. Good examples are height, weight, length etc.

# Statistical Methods

# 1. Nominal Data

When you are dealing with nominal data, you collect information through:

**1.1. Frequencies**: The Frequency is the rate at which something occurs over a period of time or within a dataset.

**1.2. Proportion**: You can easily calculate the proportion by dividing the frequency by the total number of events. (e.g how often something happened divided by how often it could happen)

**1.3. Percentage.**

**1.4. Visualisation Methods:** To visualise nominal data you can use a pie chart or a bar chart.

**Note:** In Data Science, you can use one hot encoding, to transform nominal data into a numeric feature.

# 2. Ordinal Data

When you are dealing with ordinal data, you can use the same methods like with nominal data, but you also have access to some additional tools. Therefore you can summarise your ordinal data with frequencies, proportions, percentages. And you can visualise it with pie and bar charts. Additionally, you can use percentiles, median, mode and the interquartile range to summarise your data.

**Note:** In Data Science, you can use one label encoding, to transform ordinal data into a numeric feature.

# 3. Continuous Data

When you are dealing with continuous data, you can use the most methods to describe your data. You can summaries your data using percentiles, median, interquartile range, mean, mode, standard deviation, and range.

**Visualization Methods:** To visualize continuous data, you can use a histogram or a box-plot. With a histogram (we use buckets or bins (range of data) to build histogram), you can check the central tendency, variability, modality, and kurtosis of a distribution. Note that a histogram can’t show you if you have any outliers. This is why we also use box-plots.

**Mean, Median and Mode**

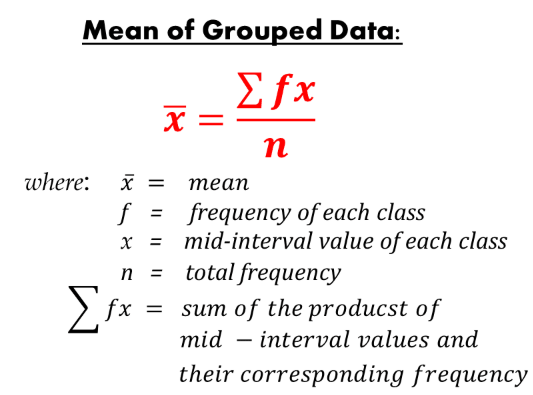
**1. Arithmetic Mean:** it is the average of all observation

**1.1. Trimmed Mean:** the mean can be influenced by extreme observations that are different from the rest of observation (*before jump to the trimmed mean first double check your data and make sure there is no data error or recording error*).

* For calculating the trimmed mean, 1) sort smallest to largest, 2) Remove the same number of observation or % from both ends 3) calculate the new mean.

**1.2. Mean of Group Frequencies:** sometimes you may not have all the original data. And only have a frequency count table or histogram for certain groups or bins of data. You can approximate the mean of your data by using midpoint method (*the higher the count in each bin and the narrower the bin range, the better the approximation*)

* **Midpoint method (upper+lower/2):** since we don’t know each individual data what we can do is the midpoint method. 1) multiply each midpoint by its frequency (weighting process), 2) sum all values from the first step, 3) divide the sum by total number of observations.



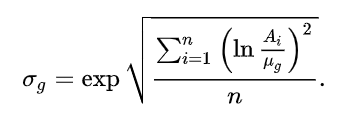
* **Variance and standard deviation of group frequency (*f*):**

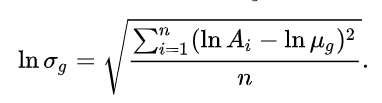
****

**1.3.** **Geometric Means:** as the name suggests, is used to calculate the **geometric mean** of a set of numbers. To recall, the **geometric mean** (or GM) is a type of **mean** that indicates the central tendency of a set of numbers by using the product of their values. It is defined as the nth root of the product of n numbers.  It should be noted that you cannot calculate the geometric mean from the arithmetic mean. In statistics, the geometric mean is well defined only for a positive set of real numbers. Example of using the formula for the geometric mean is to calculate the central frequency f0 of a bandwidth BW= f2–f1.



**Note1:** the geometrical mean mostly uses in Finance and business. But any major that use growth rate can use geometric mean.

**Note 2**: we can use the geometric mean for average growth rate over a period (e.g. compound annual growth rate in finance); using **Natural Logarithm Method**.



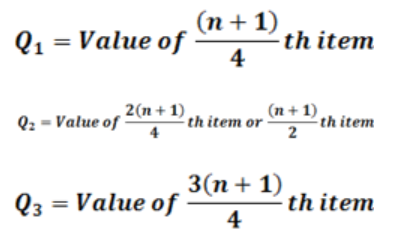
**Note 3:** the geometric mean is suitable for multiplicative process. All values must be positive. And any rate of change over sequential periods should not be used; meaning always keep periods standard (days, weeks, trials, etc.). In Excel you can use **GEOMEAN** formula.

**2. Median:** the middle number of ascending sorted data

**3. Mode:** the observation that occurs most often in the data. A dataset can have one mode, multiple modes, or none.

**Percentiles & Quartiles:**

## **1. Quartiles:** The values which divide an array (a set of data arranged in ascending or descending order) into four equal parts are called Quartiles. The first, second and third quartiles are denoted by Q1, Q2,Q3 respectively. The first and third quartiles are also called the lower and upper quartiles respectively. The second quartile represents the median, the middle value.



* **Grouped Quartiles:** Q2 is our Median.



*l* = lower class boundary of the class containing the [clip_image038](https://econtutorials.com/wp-content/uploads/2015/07/clip_image0382.png), i.e. the class corresponding to the cumulative frequency in which n/4 or 3n/4 lies

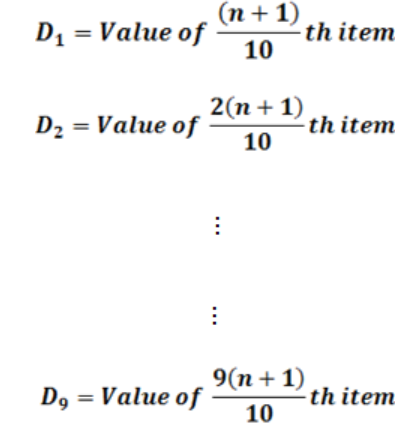
*h* = class interval size of the class containing[clip_image040](https://econtutorials.com/wp-content/uploads/2015/07/clip_image0402.png).

*f*= frequency of the class containing [clip_image038[1]](https://econtutorials.com/wp-content/uploads/2015/07/clip_image03811.png).

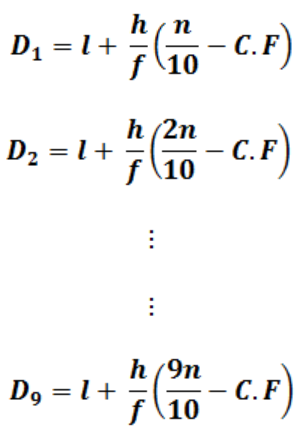
*n*= number of values, or the total frequency.

*C.F* = cumulative frequency of the class preceding the class containing [clip_image038[2]](https://econtutorials.com/wp-content/uploads/2015/07/clip_image03821.png).

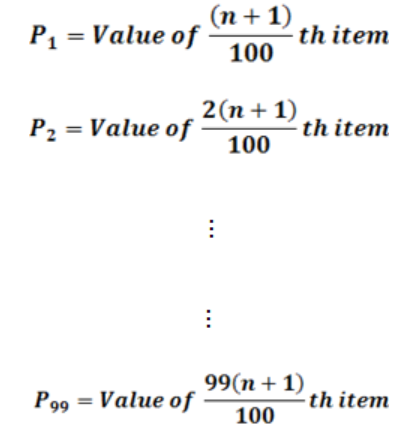
**2. Deciles:** The values which divide an array into ten equal parts are called deciles. The first, second,…… ninth deciles by [clip_image064](https://econtutorials.com/wp-content/uploads/2015/07/clip_image0641.png) respectively. The fifth decile ([clip_image066](https://econtutorials.com/wp-content/uploads/2015/07/clip_image0661.png) corresponds to median. The second, fourth, sixth and eighth deciles which collectively divide the data into five equal parts are called **quintiles**.

****

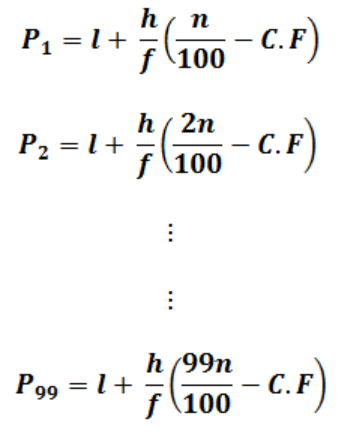
* **Grouped Deciles:**



**3. Percentiles:** The values which divide an array into one hundred equal parts are called percentiles. The first, second,……. Ninety-ninth percentile are denoted by [clip_image130](https://econtutorials.com/wp-content/uploads/2015/07/clip_image1301.png) The 50th percentile ([clip_image132](https://econtutorials.com/wp-content/uploads/2015/07/clip_image1321.png)) corresponds to the median. The 25th percentile [clip_image134](https://econtutorials.com/wp-content/uploads/2015/07/clip_image1342.png) corresponds to the first quartile and the 75th percentile [clip_image136](https://econtutorials.com/wp-content/uploads/2015/07/clip_image1361.png) corresponds to the third quartile.

****

* **Grouped Percentiles:**

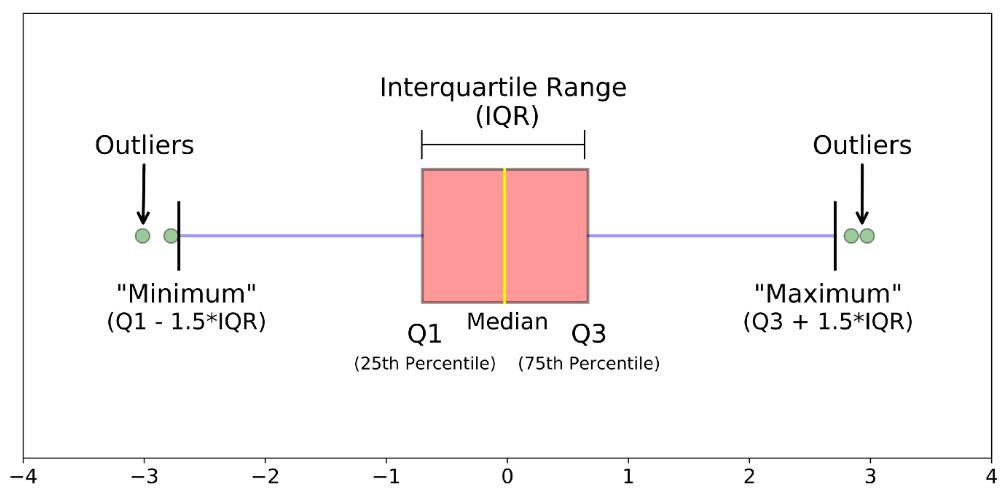
****

**V.imp Note 1:** for finding the location of data based on its percentile you can use the following formula. **LP = (p/100) \* (n+1).** Where **L** is a location in a sorted data. **P** is the percentile you are looking for, and **n** is the total number of observations. In Excel (=PERCENTILE.EXE())

**V.imp Note 2:** for finding the percentile of an observation we can use the following formula **P=(x+o.5y)/n**. where, **x** is the number of observations up to but not including the value of the nth value. **Y** is the count of observations equal to the value of desired observation. **N** is the total number of observations. Then round up to the next whole number.

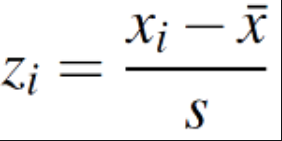
**IQR and Box Plots**

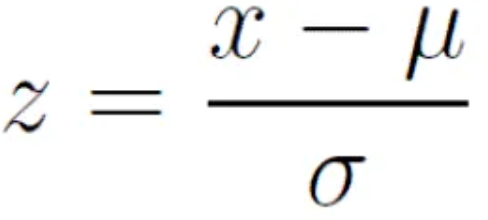
The "interquartile range", abbreviated "IQR", is just the width of the box in the box-and-whisker plot. That is, IQR = Q3 – Q1 . The IQR can be used as a measure of how spread-out the values are.Statistics assumes that your values are clustered around some central value. The IQR tells how spread out the "middle" values are; it can also be used to tell when some of the other values are "too far" from the central value. These "too far away" points are called "outliers", because they "lie outside" the range in which we expect them.The IQR is the length of the box in your box-and-whisker plot. An outlier is any value that lies more than one and a half times the length of the box from either end of the box.

****

**Z-Score**

**Why Z-Score?** Sometimes we may have two sets of data that have same mean but with different distribution, in this case the ”mean” alone can’t help us so we use z-score to understand the variation of our data.

Z-score is a measure of distance from the mean. We set the mean to “ZERO” like a starting point because it is ZERO distance from itself. We measure distance from the mean by the number of STD away.



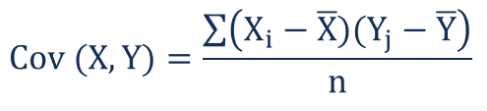
**Coefficient of Variant:** is a relative measure of variability (**Std/Mean\*100**). It measure the Std relative to mean “How large is Std relative to mean?”. Since it’s a percentage, it is unit independent and helps us to compare dataset that have different mean and Std.

**Covariance and Correlation**

**1. Covariance:** In mathematics and [statistics](https://corporatefinanceinstitute.com/resources/knowledge/basic-statistics-concepts/), covariance is a measure of the relationship between two random variables. It tells us how ‘two variables behave as PAIR?’ The metric evaluates how much – to what extent – the variables change together. In other words, it is essentially a measure of the variance between two variables. However, the metric does not assess the dependency between variables. Unlike the correlation coefficient, covariance is measured in units. The units are computed by multiplying the units of the two variables. The variance can take any positive or negative values. The values are interpreted as follows:

* **Positive covariance**: Indicates that two variables tend to move in the same direction.
* **Negative covariance**: Reveals that two variables tend to move in inverse directions.

**Note:** covariance only shows the direction of two variables not the strength of their relationship. For interpreting the relationship between two variables using covariance; only look at its ***SIGN*** (+ or -), ***IGNORE*** the ***SIZE*** of the number

****

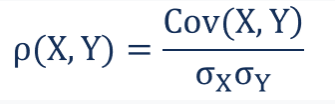
**1.1 Covariance Matrix:**



**Note 1 (V.V.IMP):** the first thing we should do to compare relationship of variables is to run **Matrix of Scatter Plots**. A figure that plots each variable against every other variable. We should search for clear linear relationship between two variables. This would help us to find out what we should expect from our analysis later on.

**Note 2:** Excel computes covariance using Population covariance formula which has a dominator of n instead of n-1 (sample). Hence, we should match the Excel result to proper statistical software by multiplying each cell by (n/n-1).

**2. Correlation:** Covariance provides only the direction of relationship between two variables while correlation provides directions and strength. Correlation always between -1 and +1 and its scale is *independent* of the scale of the variables.



**Note 1:** before going crazy computing correlations look at a scatterplot of your data. What pattern (if any) does it exhibits?

**Note 2:** Correlation is only applicable to LINEAR relationship.

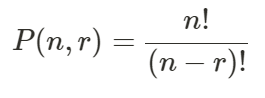
**Note 3:** Correlation is **Not** Causation.

**Note 4:** Correlation strength doesn’t necessarily mean the correlation is statistically significant; related to sample size.

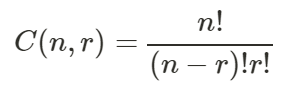
**2.1. Testing the significance of correlation (rule of thumb):** if |p(x,y)| >= 2/sqrt (n) then the relationship exist.

**Introduction to Probability**

**1. Premutation (Individual order):** the number of different ways that a certain number of objects can be *arranged in order* from a larger number of objects. If there are **n** objects, how many different ways can we make ordered list of size **r (Order Matters!)**



**2. Combinations (Unordered group or set):** the number of different ways that a certain number of objects **As a Group** can be selected for larger number of objects. Often said “**n** choose **r**”. in other words, if there are **n** objects, how many different ways can we select **group or sets** of size **r**.



**2.1. Combination-Nearly Normal:** as we increase our group size and then calculate the frequency of each possible combination where 0 =< r =< n our histogram begins to look like a normal (bell) curve. That is where we make jump from discrete random variables like C (n,r) where X=r to contains random variables; histogram to continuous function (bell curve).

* What we think of as the “normal curve” is actually a derivation or byproduct of counting and its associated probabilities
* Using the mean and Std of continuous random variable along with the fact that we know the area under the normal curve (=1), we can find out all kind of interesting stuff.

**3. Probability of Multiple Random Variables:** In majority of cases, we are likely to work with many random variables. For example, given a table of data, such as in excel, each row represents a separate observation or event, and each column represents a separate random variable. Variables may be either discrete, meaning that they take on a finite set of values, or continuous, meaning they take on a real or numerical value. As such, we are interested in the probability across two or more random variables. This is complicated as there are many ways that random variables can interact, which, in turn, impacts their probabilities. This can be simplified by reducing the discussion to just two random variables (X=A, Y=B), although the principles generalize to multiple variables.

* **Joint Probability**: Probability of events *A* and *B ->* P(A and B) = P(A|B) x P(B) = P(B|A) \* P(A)
* **Marginal Probability**: Probability of event X=*A* given variable *Y (*P(X=A) = sum P(X=A, Y=yi) for all y*)*.
* **Conditional Probability**: Probability of event *A* given event *B (P(A and B)/P(B))*.

**Discrete Probability Distribution**

**1. Random Variables:** is a variable takes on ***numerical value*** as a result of a random experiment or measurement; associate a numerical value with each possible outcome. Two things to keep separate in mind: 1) the random variable itself (random). 2) the possible outcomes or values the random variable can take (not random).

**1.1. Discrete Vs. Continuous Random Variable:** a very important distinction needs to be made between discrete and continuous random variables.

* A **Discrete random variable** has a finite number of values or an infinite sequence of values (0,1,2,…) **AND** the difference between the outcomes are meaningful (e.g. Dice). The probability distribution for a random variable X describe how probabilities are assigned to each outcome for the random variables. The discrete probability is always a number between 0 and 1. And the sum of all probabilities should be equal to 1.
* A **Continuous random variable** has a nearly infinite number of outcomes that ***can’t*** be easily counted **AND** the difference between the outcomes are not meaningful (e.g. with average income, the difference between 40,000 and 40,001 is not meaningful).

**1.2. Conditions for Discrete Probabilities:**

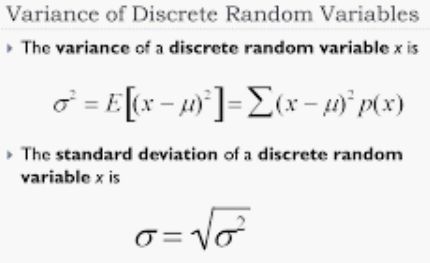
* The probability function P(x) denotes the associated probability for each outcome of the random variables
* 0 ≤ P(x) ≤ 1
* Σ P(x) = 1

**1.3. Probability Distribution:** describes how probabilities are assigned to each outcome for the random variable.

**1.2. Expected Value:** is simply the mean of random variables; the average expected outcome. It does not have to be a value the discrete variable can assume.

**E (X) = µ = ∑ X\*P(X) (a weighted average)**

**1.3. Variance of Discrete Random Variable:** though expected value tells us the mean of random variable; often times we need to know the variability, or how spread out the random variable is from the mean. We can use variance and Std to learn about how **dispersed** it is relative to its mean. This information can be used to calculate other statistics, compare datesets, and inform other conclusion about our data.



**2. The Binomial Distribution:** the **binomial distribution** with parameters ***n*** and ***p*** is the [discrete probability distribution](https://en.wikipedia.org/wiki/Discrete_probability_distribution) of the number of successes in a sequence of *n* [independent](https://en.wikipedia.org/wiki/Statistical_independence) [experiments](https://en.wikipedia.org/wiki/Experiment_(probability_theory)), each asking a [yes–no question](https://en.wikipedia.org/wiki/Yes%E2%80%93no_question)

**2.1. Binomial Experiment**: has below characteristics:

* The process consists of a sequence of ***n*** trials.
* Only two exclusive outcomes are possible in each trial. One outcome is called a “success ([yes](https://en.wikipedia.org/wiki/Yes_and_no)/[true](https://en.wikipedia.org/wiki/Truth_value)/[one](https://en.wikipedia.org/wiki/One))” and the other a “failure ([no](https://en.wikipedia.org/wiki/Yes_and_no)/[false](https://en.wikipedia.org/wiki/False_(logic))/[zero](https://en.wikipedia.org/wiki/Zero))”.
  + “success” and “failure” can be counterintuitive at times if taken literally.
* The probability of success (***p***) does not change from trial to trail. The probability of failure is ***q=1-p*** and is also fixed from trial to trial.
* The trials are independent; the outcomes of previous trials not influence future trials.
* The sum of all possible outcomes should be equal to 1.
* The probability of any given outcome is a combination of both the number of trials and the success rate **Bionmcdf: P(x) = C (n, x) Px (1-P)n-x**

**2.2 Binomial Mean (Expected Value) and Standard Deviation:**

* **µ = n.p**
* 

**2.3. Extra information about Binomial Distribution:** A single success/failure experiment is also called a [Bernoulli trial](https://en.wikipedia.org/wiki/Bernoulli_trial) or Bernoulli experiment and a sequence of outcomes is called a [Bernoulli process](https://en.wikipedia.org/wiki/Bernoulli_process); for a single trial, i.e., *n* = 1, the binomial distribution is a [Bernoulli distribution](https://en.wikipedia.org/wiki/Bernoulli_distribution). The binomial distribution is the basis for the popular [binomial test](https://en.wikipedia.org/wiki/Binomial_test) of [statistical significance](https://en.wikipedia.org/wiki/Statistical_significance). The binomial distribution is frequently used to model the number of successes in a sample of size *n* drawn [with replacement](https://en.wikipedia.org/wiki/With_replacement) from a population of size *N*. If the sampling is carried out without replacement, the draws are not independent and so the resulting distribution is a [hypergeometric distribution](https://en.wikipedia.org/wiki/Hypergeometric_distribution), not a binomial one. However, for *N* much larger than *n*, the binomial distribution remains a good approximation, and is widely used.

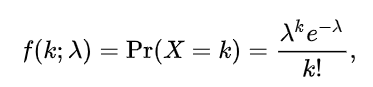
**Note:** In many cases we can use the reverse way which is easier to calculate (1 – the amount we don’t look for). In excel you can use =BINOMDIST (Number of suceesses; Total number of trials; probability of success; True/False). True will calculate Binomcdf and False will calculate Binompdf.

**How to interpret the probability of binomial distribution (v.v.imp):** the probability of a scenario give us the level of error in our hypothesis (assumption) based on random chance variation. If the probability cover more than 95% the result is not by accident and you fail to reject your hypothesis.

**3. Poisson Distribution:**  is a [**discrete probability distribution**](https://en.wikipedia.org/wiki/Discrete_probability_distribution) that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and [**independently**](https://en.wikipedia.org/wiki/Statistical_independence) of the time since the last event.[[1]](https://en.wikipedia.org/wiki/Poisson_distribution#cite_note-Haight1967-1) The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

For instance, an individual keeping track of the amount of mail they receive each day may notice that they receive an average number of 4 letters per day. If receiving any particular piece of mail does not affect the arrival times of future pieces of mail, i.e., if pieces of mail from a wide range of sources arrive independently of one another, then a reasonable assumption is that the number of pieces of mail received in a day obeys a Poisson distribution.[[2]](https://en.wikipedia.org/wiki/Poisson_distribution#cite_note-Brooks2007-2) Other examples that may follow a Poisson distribution include the number of phone calls received by a call center per hour.

**Landa = (number of occurrences)/interval** ; 



**3.1. Poisson characteristics:**

* Discrete outcomes (x=0,1,2,3, …)
* The number of occurrence in each interval can range from 0 to infinity.
* Describe the distribution of infrequent (rare) events
* Each event is independent of other events.
* Describes discrete events over an interval (time, distance, etc.)
* E(X) is assumed to be constant throughout the Experience

**Note:** In excel you can use =POISSON:DIST (X=K; Mean=Landa; True/False). True will calculate possioncdf and False will calculate possionpdf.

**Continuous Probability Distribution**

## **1. Uniform (rectangle) Distribution:** fixed number of outcomes and probability of each outcome is the same. **P(x) = 1 / n equal outcomes.**

## Hence if n moves toward infinity (in dataset such as temperature, distance, income, mass, etc.). The probability moves toward zero.

* So think of a bar chart where the Hight of an infinite number of bars is zero! We call this ***continuous data.*** The probability of any specific outcome is zero.
* Instead of the probability of any specific outcome (which is zero), we can only find the probability over ***Specific interval***.

E(X) = Mean = (a+b)/2; Variance = δ2 = (b-a)2 /12; Std = δ = (b-a)/√12;

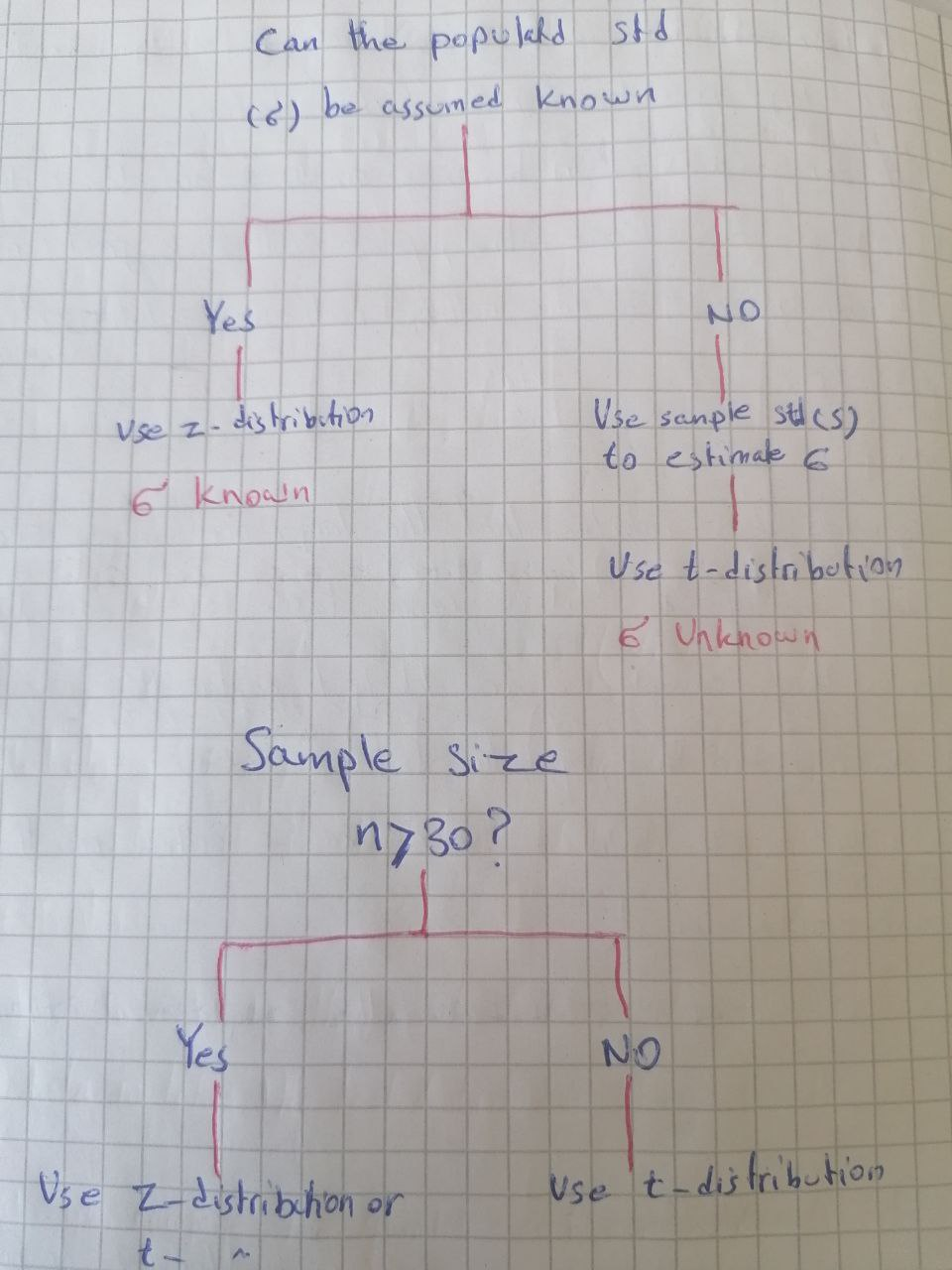
P(X) = (X2-X1)/(b-a)

**2. Normal (bell) Distribution:** the average (mean) tends to be very frequent while measures away from the mean are less frequent. µ (mean) can be any numerical value; slides distribution side-to-side. As much as δ is smaller we have narrower and taller tails and vice versa.

**2.1. The standard normal curve (Z-Distribution):** in this distribution µ = 0; δ= 1; and the area under the curve is 1. (when our sample size (n>30 or we know the δ of our population).

* In Excel **=NORM.DIST(x(z); µ(0); δ(1); TTUE/FALSE)**

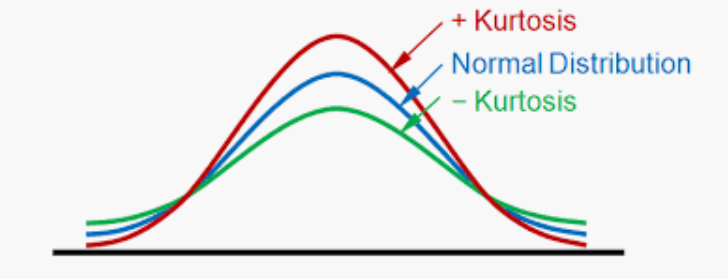
**2.2. The T-Distribution:** when our sample size is n=<30 and/or we don’t know our population’s variance/Std we use ***t-dist*** instead of ***z-dist.***

* The t-dist allows us to use small samples but in expenses of higher margin of error.
* It takes sample sizes into account using **df=n-1**; there is a different t-dist for any given **df**.
* The bell curve is squishier and fatter (tails) the smaller the n.
* However as n>30 and definitely n>100, the t-dist and z-dist become indistinguishable.
* 

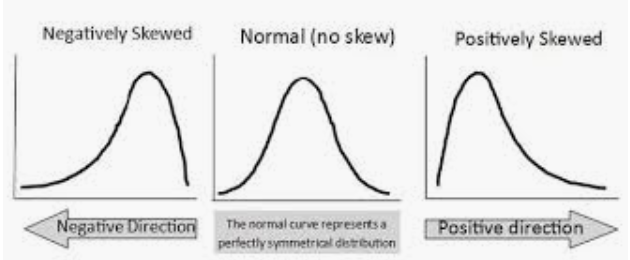
**V.V.Imp Note:** Before using normal distribution, first and foremost look at your data graphically first, before any analysis. Get to know the data. Look for patterns, potential problems, initial relationship, etc.

**2.3. Graphical Data Exploration:** by using a few simple visual tools, we can learn a tremendous amount of information about our data. Our data may have excess skew (lopsided), kurtosis (very fat tails), be bi-modal (two humps like camel), or follow a distribution other than normal distribution. Yet many statistical techniques **ASSUME** the data fits a normal distribution.

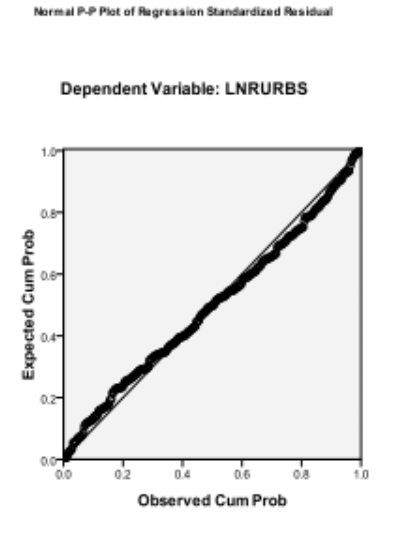
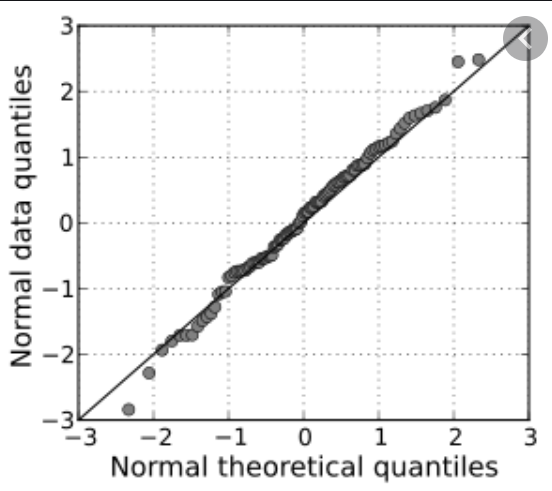
* **Graphical Tools for determining “normal” data:** 1) Histograms; 2) Steam and Leaf Plots; 3) Box Plots; 4) P-P Plots; 5) Q-Q Plots.
* **Excess Kurtosis:** More probability than expected in the tails of distribution due to the extreme values away from the mean. Probability (values) are pushed away from the mean and out towards the tails.



* **Excess Skewness:** More probability than expected is on one side of the distribution versus. Others; lopsided.



* **P-P Plots:** in a P-P plot, we compare the cumulative probability of our empirical data with an “**ideal test**” distribution (here: normal dist). Question to ask: **Do the points fall in straight line?** If our data matches the test distribution they should.
* **Q-Q Plots:** we compare the quantiles of our empirical data with the ideal**.** Question to ask: **Do the points fall in straight line?** If our data matches the test distribution they should.



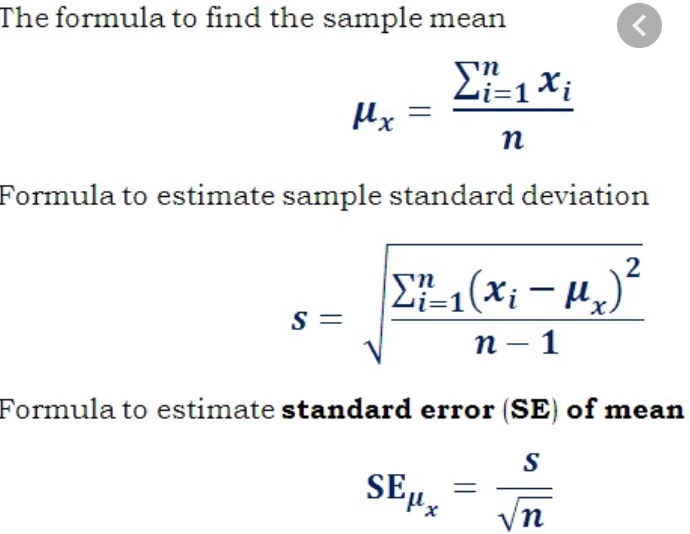
Q-Q Plot

**Sampling and Sampling Distribution:**

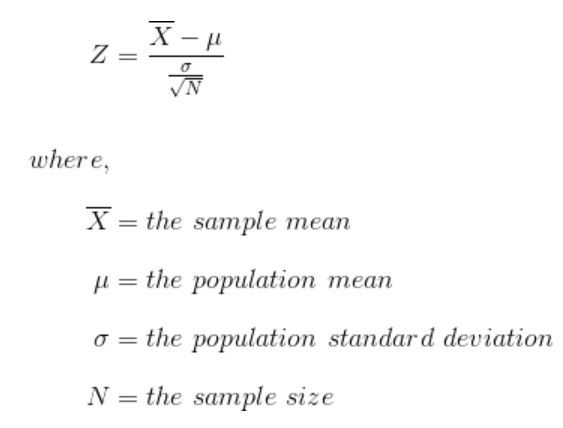
**1. Point Estimators (Inferential Statistics):** The largest part of statistics is about taking a sample from the population and using those as an estimation of the overall population. The idea of point estimation is to calculate a single value(statistic) by using sample data from the population. For example, it can be inference about the quality of some products based on sample data observations.

* When we don’t know parameters of the distribution, such as mean or standard deviation, we can make an estimation about these parameters by calculation **point estimators, such as the sample mean or sample standard deviation**. And we should remember that point estimates are never perfect since we are working only with a relatively small sample of the population. This is commonly referred to as the **margin of error**. And this error component is expressed as a **confidence interval (function of sample size and degree of “confidence”).**

**1.1 Standard (margin) Error of the Mean: is equal to Std of our sample.**

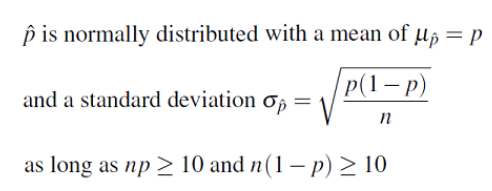
****

* As ***n*** increases the **SE** becomes smaller which means the standard error of the mean of the population moves toward zero.
* The **SE** is the ***same*** for all samples of same size.
* Std (s) allows us to calculate z-scores and hence the are (probability) under the curve for certain regions.
* Most often we don’t know the population Std therefore we have to estimate it and make necessary adjustment.
* Keep in mind that SE = Std of sampling distribution. Hence based on this fact we can draw a normal dist using z-score (if we know δ; µ; or n>30. Otherwise we use t-dist).
* For probability calculation; first calculate **z-value** using below formula, and then follow =NORM.DIST function in Excel.

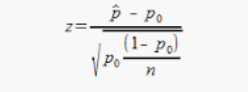


**2. Sample Proportions:** The ***sample proportion (p bar)*** is similar to ***sample mean (x bar)***. It is just one of the numerous samples that could be taken from population. If someone else went out and did the sampling again, it is likely there would be different result. However, sample proportions are different than sample mean ***(x bar)*** in important ways:

1. **P bar** represent what is essentially a nominal (usually binary) variables; Yes/No, … .
2. It is a count/frequency of a nominal variable of interest
3. In that sense it is related to *binomial distribution*.



* Expected value of the (p^) is equal to the mean of all potential values of (p^) is equal to the population proportion (p). **E(p^) = p^ = x/n**. where **x** is the number of observations that have what you are interested in. **n** is total sample observations.
* We refer to the Std of the sample proportion as the standard error of the sample proportion.
* However, the standard error of p^ is dependent on having finite or infinite population of interest. Yet in almost all cases, we are assuming the population is “infinite” or very large in relation to the sample size (np >= 10 and n(1-p) >= 10) or (n/N =< 0.05).
* For probability calculation; first calculate **z-value** using below formula, and then follow =NORM.DIST function in Excel. Where p^ is the probability we looking for, p0=x/n is the expected value or mean of the sample proportion.



* V.imp note: the number of sample doesn’t have anything to do with Std because **p** and **n** are the same for any number of samples. But increasing the number of samples reduces ***the standard error of the mean***, which means as much as we increase the number of sample, we move closer to population mean.

**Confidence Interval (C.I) Estimation**

**1. C.I estimation:** “confident” and “interval” always come in the form of plus/minus. Because we use a sample of population; there will be always error. We can express that error using an interval estimate.

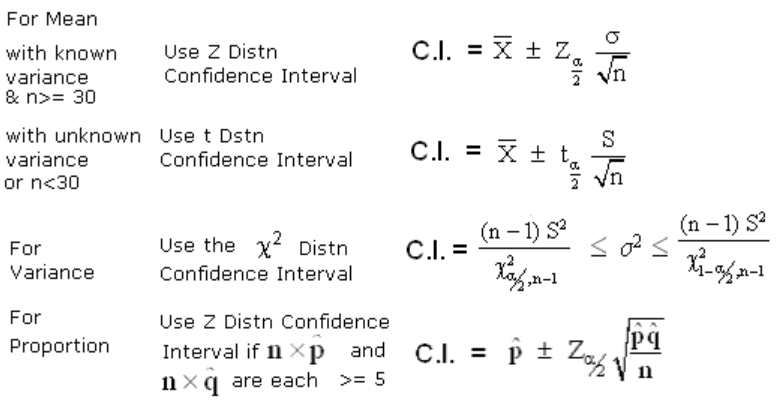
**1.1. C.I. and Standard of Error of Mean (SEM):** to find the SEM we need to know 1) population Std (δ) and 2) the sample size.

* Most often we don’t know the population Std (δ) and hence have to estimate it. For any population Std; increasing the sample size (**n**) would lead to the reduction of SEM.
* Hence C.I. is affected by **Std**, **n**, and **level of “confidence”** we are satisfied with.
* **95% Probability interval** (95% of all sample means (x bar) are in here). Therefore we have 5% left evenly divided between both tails called alpha (α=0.05 or 5%, 2.5% for each tail).
* If we know the population std (**δ**) we treat sampling distribution as we would a standard normal curve (**z-score**). By doing so we can assign z-score to the upper and to the lower boundary of the 95% interval. If **n<30** and/or **δ** we use **t-distribution**.
* Don’t forget that the Std of sampling distribution is equal to the standard error of the mean. δ x = δ/√n. if n is small theδ x is larger and distribution is wider.

**1.2. Interpretation of C.I:** the randomness lies in the elements chosen for the sample; not the population mean. It is ***merely*** the probability of obtaining a representative sample.

* The sample mean (x bar) is either plus/minus interval of the population mean (µ), or it is not.
* **Very Common Error**: The C.I is **NOT** the probability that the population mean lies within the interval.
* **C.I with δ unknown:** use t-table instead of z-table. Remember each sample size has its own t-table based on its degree of freedom (df=n-1). The t-dist has smaller mean and fatter tails (larger Std). in this case **Sx = S/√n**. Where **Sx**is standard error and **S** is the Std of sample.
* **Margin of Error (E):** Ponit estimate (X bar or P^) +- Margin of error ( look formulas below)

## **Decision Tree for selecting What Formula to use:**



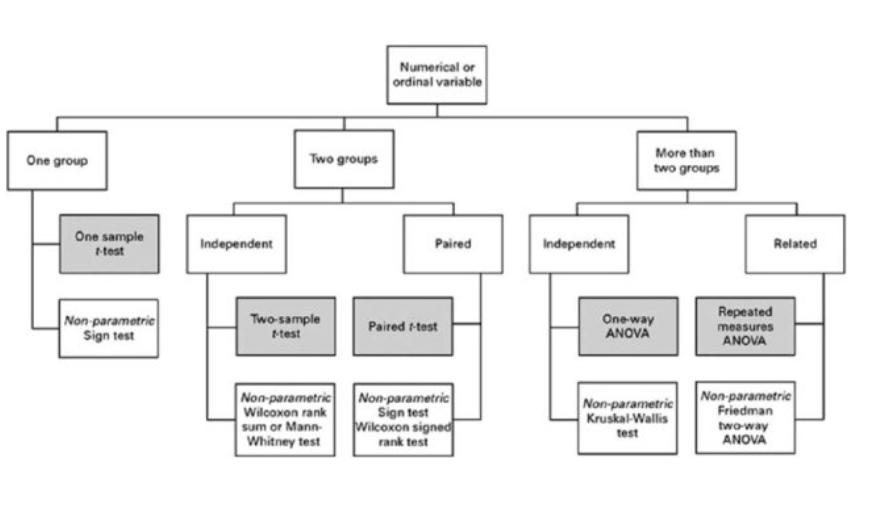
**1.3. How to calculate n by having E:** if we ignore (-) sign we can use E formula.

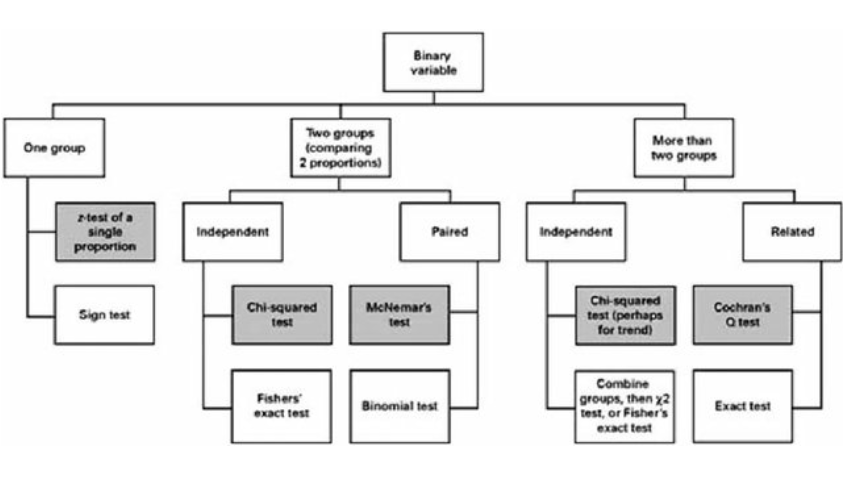
1. We choose **E**, our margin of error.
2. We choose our confidence probability boundary **(Z α/2)**
3. We are given or estimate population Std **(δ)**
4. Solve for n **n = ((δ. Z α/2)/E)2**

**A note about δ:**

1. Estimate δfrom previous studies using the same population of interest.
2. Conduct a pilot study to select a preliminary sample. Use the sample Std (s) from pilot study.
3. Use a judgment or “best guess” for δ. A common “guess” is the data range or (max-min/4)

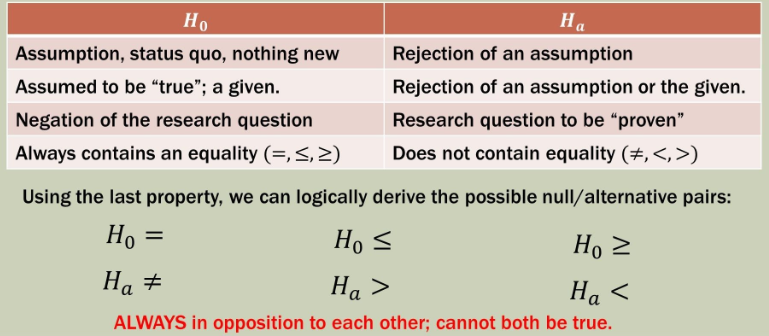
**Statistical analyses tests when the values are numerical (continuous) or ordinal or binary:**

****

****

* **Hypothesis Testing:**

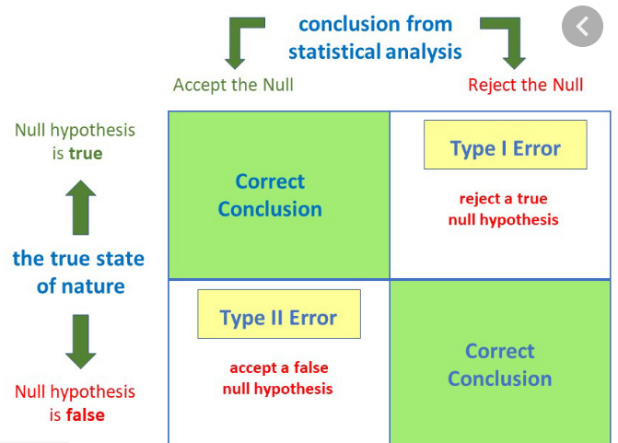
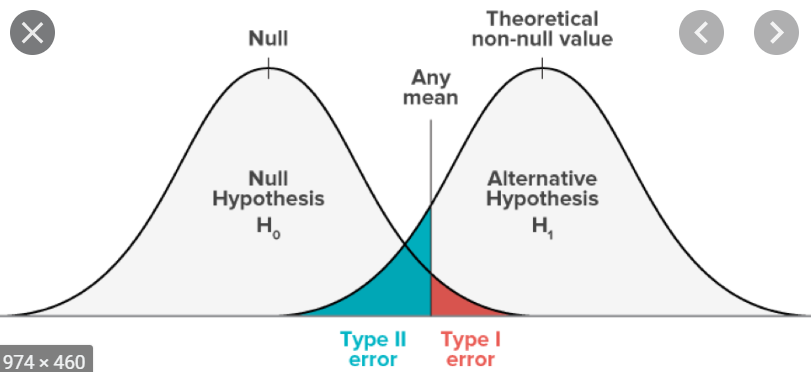
**1. Hypothesis Formulation:**

****

**Very important note 1:** if we*fail to reject the H0***,** it doesn’t mean we have proven the H0 is “*TRUE*”.

**Note 2:** statistics is never 100% certain; but it states its limitation explicitly (95% etc.)

**1.1 Type I and Type II Errors:**

****

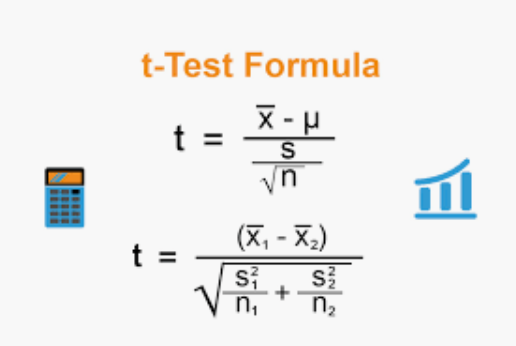
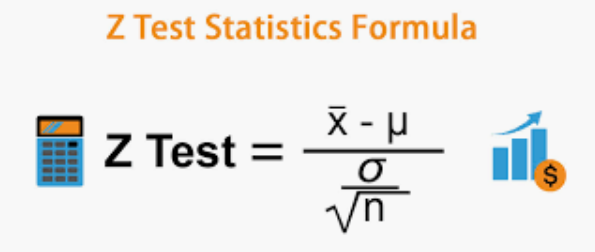
**Very. Imp Note:** in the majority of cases we would have type I and type II errors simply because of *the laws of chance*. But there may be some cases that we will end up with these errors because are sampling methods were flawed. And id our sampling is flawed so do our assumption (H0).

**1.2. Rejection regions**:

* Two-tailed rejection region: **H0 : µ0 = µ and Ha : µ0 not= µ**
* One-tailed (lower) rejection region: **H0 : µ0 >= µ and Ha : µ0 < µ**
* One-tailed (upper) rejection region: **H0 : µ0 =< µ and Ha : µ0 > µ**

**Note 1:** if **H0** is correct then *α percent* of the sample means should be in the nonrejection region (in case of one-tailed region *α/2 percent*).

**Note2:** the critical value is determined by α and if we are using the z-test or t-test. If our *z/t test statistics is smaller than our Z(T)α or Z(T)α/2*we are *failed to reject* **H0**



**Note 3:** As *α* decreases so does the chance of ***Type I Error***. The critical value to reject the **H0** moves outwards thus “capturing” more sample means.

**Note 4:** However, moving outward of the critical values may also “capture” a mean from a different population off the side. We would fail to reject **H0** when indeed we should. Thus, the chance of ***Type II Error*** increases when *α* decreases.

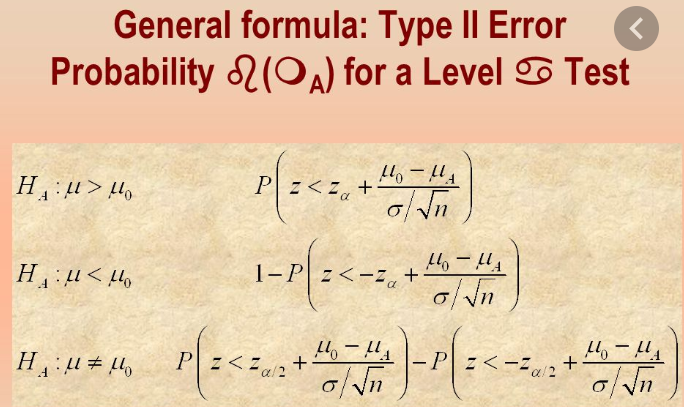
**2. Single Sample Hypothesis Test procedure:**

1. Start with a well-developed, clear research problem or question
2. Establish Hypotheses, both Ho and Ha
3. Determine appropriate statistical test and sampling distribution (z-dist or t-dist)
4. Choose Type I Error (α)
5. State the decision rule (z-statistics or t-statistics): if **δ** (std) of population is known and/or **δ** is unknown but **n>=30** we use z-statistics otherwise use t-statistics
6. Gather sample data (always check the sample data for normality)
7. Calculate test statistics
8. State statistical conclusion
9. Make decision or inference based on conclusion

**2.1. The P-value Method:** Based on our **α** we know that **α percent** of our area (probability) is in the upper/ lower (both) tail past our **z(t)-critic**. In the **P-Value Method**, we ask how much area (probability) is above our test statistic (using z(t) table). Hence if the probability of our test statistics **is greater than** our **α** then we can **reject** the **Ho**. This is often referred to as the ***observed significance level***.

* In excel we can use **=NORM.S.DIST (z- test statistics; TRUE)** to calculate the z-score probability of our test statistics and the comparing it with **α** to make a conclusion about **Ho**
* **=NORM.DIST(x, µ, δ, True)** for probability of single observation
* For calculating z-critic (α) we use **=NORM.S.INV (α)**
* In excel we can use **=T.Dist.RT(LT) (df; t-test statistics)** to calculate the t-score probability of our test statistics and the comparing it with **α** to make a conclusion about **Ho**
* For calculating **t-critic for Lower-Tail =T.INV (α; df = n-1); t-critic for Upper-Tail =T.INV (1-α; df = n-1); t-critic for Two-Tails =T.INV (α/2; df = n-1)**

**2.2 Calculating Type II error:**

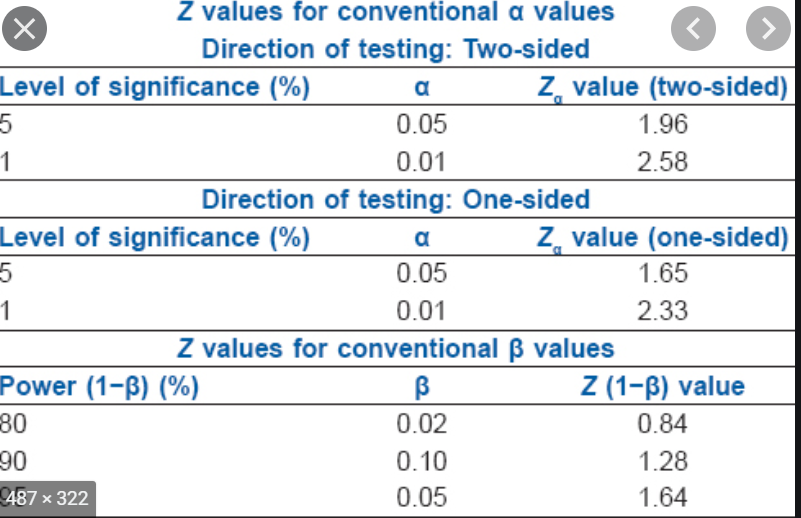


**2.3. Controlling Type II Error using sample size:**

* When we are looking to differentiate or compare two distributions, there are **two primary characteristics** we can analyze; **1)** **their means** (the distance between them) and **2)** **their variance** (means remain constant).
* It is possible to establish ahead of time the acceptable amount of type II error in the same fashion Type I error is established.
* To achieve this however we must locate two distribution in the proper location/alignment.
* Since the population means and variance are given in the problem, we will be forces to “manipulate” the sample size to achieve a certain standard error.
* We will find the appropriate sample size to shape the variance in a manner that brings the distribution into alignment at the proper Type I and Type II error locations.
* A larger sample size, given the same **δ,** will decrease the standard error making the interval between the marks smaller (**δ/√n decreases).**

**n = (δ2 (Zα +|Zβ|)2)/(µ0 - µa) 2**

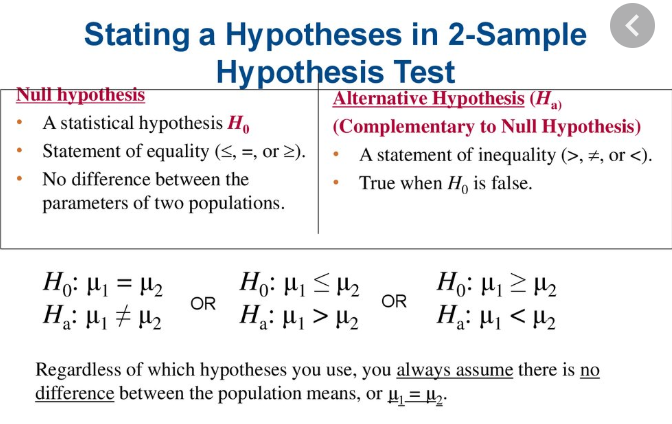
* **Zβ**: Value given area of **β** in the **upper tail** of the standard z-dist although in curve Zβ is in lower tail but for finding **n** we treat it as in upper tail (always positive value)

****

**Comparing Two Populations Means**

**1. Two populations, Z-Test with Hypothesis:**

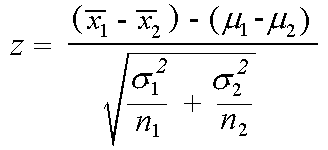
**Step 1:** hypothesis formats and regions

****

**Two-tailed or Upper-tailed or Lower-tailed**

**Step2: Statistic test and Decision rule:**

Given samples from two normal populations of size *n1* and *n2* with unknown means  and  and known standard deviations  and , the test statistic comparing the means is known as the ***two-sample z statistic***



which has the [standard normal distribution](http://www.stat.yale.edu/Courses/1997-98/101/normal.htm) (*N(0,1)*).

The null hypothesis always assumes that the means are equal, while the alternative hypothesis may be one-sided or two-sided.

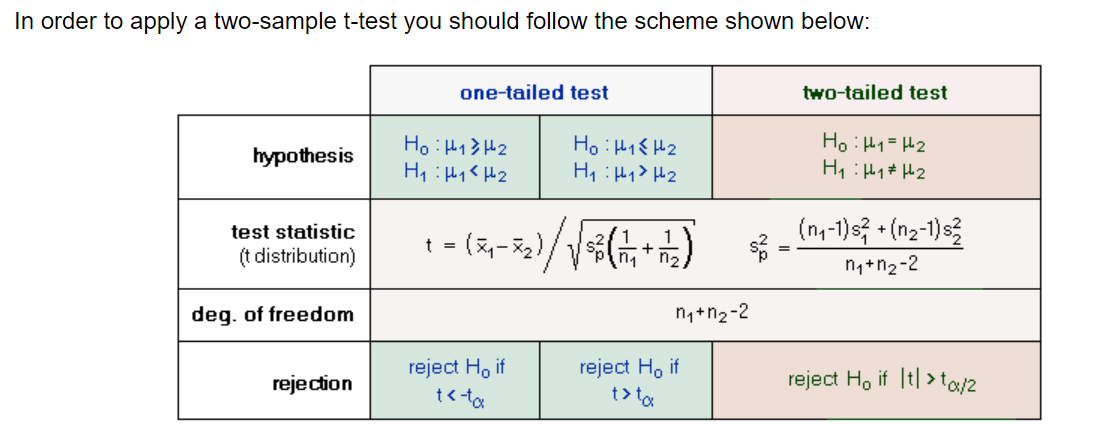
**2. Two populations, T-Test with Hypothesis:**

When the sample size is small, the assumption of the [central limit theorem](http://www.statistics4u.com/fundstat_eng/cc_central_limit.html) does not hold, since the estimates of  σ2 become unreliable. One therefore has to resort to the [t-distribution](http://www.statistics4u.com/fundstat_eng/cc_distri_student_t.html). The t-test requires some constraints to be fulfilled:

* the [variances](http://www.statistics4u.com/fundstat_eng/cc_variance.html) have to be equal
* the [samples](http://www.statistics4u.com/fundstat_eng/cc_sample.html) have to be independent of each other
* the samples have to follow a [normal distribution](http://www.statistics4u.com/fundstat_eng/cc_distri_normal.html)

Since we assume that σ12 and σ22 are equal, we can compute a pooled variance sp2. The rational for pooling the variances is to obtain a better estimate. The pooled variance is a weighted sum of variances. So when n1 equals n2, sp2 is just the average of the individual variances. The overall [degrees of freedom](http://www.statistics4u.com/fundstat_eng/cc_degree_of_freedom.html) is the sum of the individual degrees of freedom for the two samples:

df = df1 +df2 = (n1-1) + (n2-1) = n1+ n2 - 2.



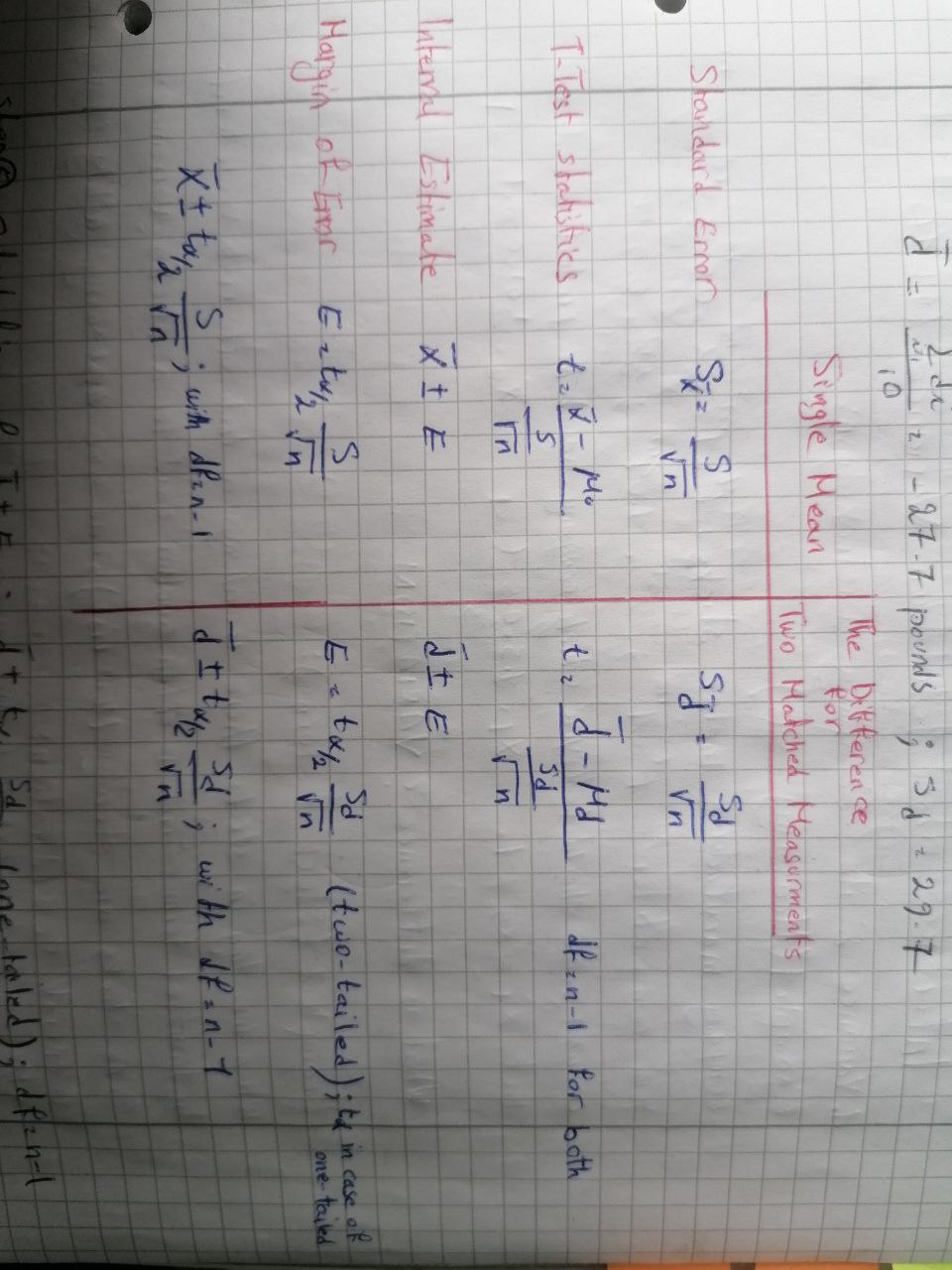
**3. Two populations, Matched Sample T-Test:**

**Step 1:** Find the difference for each subject; (di = xiafter - xibefore)

**Step 2:** Calculate d bar; (d bar = (∑di) / n)

**Step 3:** Research Hypothesis (Ho : d bar = 0 (two-tailed), Ho: d bar < 0 (Lower tail), Ho: d bar > 0 (upper tail))

**Step 4:** Test Statistic and decision rule



**Step 5:** Calculation of Interval Estimate (d bar +- E)

**Step 6:** Conclusion either reject Ho or fail to reject Ho (if d bar +-E does not include ***zero*** then reject Ho and vice versa)

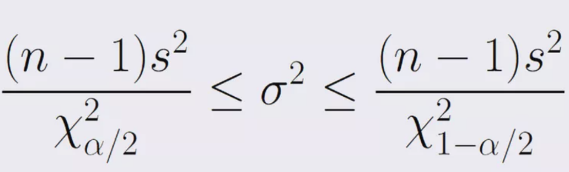
**Inferences about Population Variance**

**1. Variance and its Sampling Distribution**

* In many cases we may have two sample with the same means but difference variance. Hence it is very important to also analyze any differences in variance of our samples.
* In financial terms, “**risk**” is often another word for a **stock’s variance**. Some stocks are **steady** (low risk) and some swing **wildly** (high risk).
* Also analyzing samples’ variance is a cornerstone of **statistical quality control**. **Higher variance** could mean inconsistent production or **out of control processes**. Yet an **extremely low variance** could mean a **stuck sensor or human error in documentations**.
* **Sampling Distribution of x bar:** take many samples from population and then calculate the means (x bar) for all samples, those sample means will follow the **normal curve** when placed in their own distribution.

**1.1. Sampling Distribution of variance (S2):** however when we take many samples of same sample size from a normal population and the find those sample variance, then sample variance **DO NOT** follow the normal curve when placed in their own distribution.

* They follow **Chi-square (ꭙ2) distribution,** with **df = ⱴ = n-1.** There is unique **ꭙ2** for each **df**.
* The curve is asymptotic; it never touches the x-axis.
* The Chi-square “1” is at left side and “0” on the right-side. Therefore, the cumulative probability runs from right to left. However, the probability are found in Chi-square table in a same manner with normal table (t-table).
* At 95% C.I the critical value of **ꭙ2 (left-tail): 0.975 / ꭙ2 (right-tail): 0.025.** So, for finding our critical value we just need to use the above-mentioned probability to find the critical value for each df in Chi-square tables.

**1.2. Confidence Interval for Variance:** whenever a random sample of size n is selected from a normal population, the C.I for Variance of that sample is equal to:

**And ꭙ2 = (n-1)(S2) / δ2**

* Remember that **S** and **S2** of our sample should fit in C.I for **δ** and **δ2** that we find out. Meaning that we already have our S and C.I (from above formula) we just need to find out that whether S or S2 fit in that C.I or not. If not, there is something wrong and you need to check everything once again.

**Important Note 1:** Interval estimate for is **δ** very sensitive to the overall population being normally distributed. **So, ALWAYS FIRST check your data for normal-dist (if not the formula doesn’t work, and your analysis worth horseshitt!!!!).**

**Important Note 2:** with n<10 strange things happen, so always n>=10.

**Important Note 3:** The interval estimates for **S** follow chi-square dist. The interval estimates for **δ** the does not. The **δ** interval is an **“extra”** step.

**Important Note 4:** As n increases a) the interval estimates narrows (Ceteris Paribus) and b) distribution of the **δ2** becomes more and more like the normal-dist; with very large n.

**Important Note 5:** As compared to 95% C.I, a 99% C.I will widen the interval while 90% C.I narrow it. Think of the size of the **“NET”** needed to catch variances.

**Important Note 6:** We are estimating the **SPREAD** of the distribution as we are estimating the **“waistline”** of the population distribution.

**1.3. Hypothesis Test for Variance:**

**Left-Tail (Lower): Ho: δ2 >= δ02; Ha: δ2 < δ02. If ꭙ2 >= ꭙ21-α** fail to reject.

**Right-Tail (Upper): Ho: δ2 =< δ02; Ha: δ2 > δ02. If ꭙ2 =< ꭙ2α** fail to reject**.**

**Two-tailed: Ho: δ2 = δ02; Ha: δ2 != δ02. If ꭙ21-α /2=<ꭙ2 =< ꭙ2α /2** fail to reject.

* The most common cases are two-tailed and right-tail hypothesis.
* Keep in mind that we are saying nothing about the MEANS of these processes. It could be the same, or it could be way off the goal.

**Quality/performance Matrix:**

|  |  |  |
| --- | --- | --- |
|  | **Mean, x bar, on target** | |
| **Variance on target** | **Yes** | **No** |
| **Yes** | Good! Keep monitoring | Fix Process Mean |
| **No** | Fix Process Variation | BAD!! Crisis Meeting |

**2. F-Ratio Test for Two Equal Variance:** We use the F-distribution test when we want to find out whether the variance of two samples are equal.

* To conduct this analysis for quality of **S2**, we need new technique (F-ratio). Remember that we are not comparing the sample variance to a hypothesized variance; we are comparing **S1**and **S2**together.
* The easiest way to compare the relative size of two measurement is by using F-ratio
* **F = S2x/ S2y**; Where **Sx** is larger sample variance and **Sy** is smaller sample variance
* For calculation the **F-critic** either use online calculator or in excel **= F.INV.RT (α; df1 = n1 – 1; df2 = n2 -1)**

**2.1. Hypothesis Testing:**

**Ho: δ21 = δ22; Ha: δ21 != δ22**

**If F-ratio > F-critic, then reject Ho**

**Goodness of Fit and Independent Tests**

With the Goodness of Fit we want to make sure that the random variation is due to the chance alone. In other words we are interested to find out whether the variation beyond what we expect due to the chance alone or not. For doing so we can use **a) data visualization b) chi-square test**

**1. Data Visualization:**

* **Simple line graph:** shows the data trend and variation.
* **Stacked bar chart:** When to use stacked column charts. Stacked column charts work well when the focus of the chart is to compare the totals and one part of the totals. It's hard for readers to compare columns that don't start at the same baseline
* **Stacked bar% chart:** described by percentage of total 100%. Help us to find relative percent of each observation.
* **Stacked area chart:** A stacked area chart is the extension of a basic area chart to display the evolution of the value of several groups on the same graphic. The values of each group are displayed on top of each other. stacked area graph are appropriate to study the evolution of the whole and the relative proportions of each group. Indeed, the top of the areas allows to visualize how the whole behaves, like for a classic area chart
* **Stacked area% chart:** described by percentage of total 100%. Help us to find relative percent of each observation.
* **Spider (radar) diagram:** A radar chart is a [graphical method](https://en.wikipedia.org/wiki/List_of_graphical_methods) of displaying [multivariate](https://en.wikipedia.org/wiki/Multivariate_statistics) [data](https://en.wikipedia.org/wiki/Data) in the form of a two-dimensional [chart](https://en.wikipedia.org/wiki/Chart) of three or more quantitative variables represented on axes starting from the same point. One application of radar charts is the control of [quality improvement](https://en.wikipedia.org/wiki/Quality_management) to display the [performance metrics](https://en.wikipedia.org/wiki/Performance_metric) of any ongoing program. They are also used in sports to chart players' strengths and weaknesses, where they are usually called radar charts.

**2. Chi-square Test**

* It helps us understand the **relationship between two categorical variables** such as grade level, sex, age, year, etc.
* Chi-square involve **the frequency** of events; the counts
* It helps us to compare what we actually observed with what we expected oftentimes using population data or theoretical data **(observed Vs. expected)**.
* Chi-square assist us in determining the role of random chance variation between our categorical variables **(Ho: Is variance by chance alone or not?)**
* We use chi-square dist and critical value to accept or reject our **Ho**.
* If **χ2** > **χ-critic** then we reject the **Ho.**
* In excel we can use **χ-critic = CHIINV(P-Value (error); df =n-1) and χ2 = ∑ (O – E)2 / E**
* **Very Important Note:** in most cases we don’t have the expected value of all of our data. But if we have the column/row/grand total of our categorical values then we can calculate the expected value for each observation using **Expected count in each cell = (Row total x Column Total) / Grand total**
* **Analysis of Variance (ANOVA)**
* ANOVA can help us to compare multiple populations’ variance rather than just only two (F-ratio/matched population) and even solo groups of population.
* To this point we have been comparing two populations.
  + Independent samples t-test (random)
  + Matched-sample t-test (paired)
* With ANOVA we are able to compare mean/variance of more than two populations and/or subgroups of a population.
* In ANOVA, analyzing the means and find out whether all means came from same population is very important practice and is called **variability Among/Between the sample means**.
* If we scratch our sample variance, we would increase its SPREAD; in ANOVA we call it **variability Around/Within the distribution**.
* Hence ANOVA or (Analysis of Variance) is a **Variability Ratio**:

**F =** (**variability Among/Between the sample means**) / (**variability Around/Within the distribution)**

**F = (Distance from overall Mean) / each width/spread within internal spread)**

**Variance Between + Variance Within = Total Variance**

**NOTE:** this is called “partitioning” – Spreading total variance into its component parts.

**Interpretation of ANOVA:** if the variability **Between** the means (distance from overall mean) in the numerator is relatively large compare to the variance **Within** the samples (internal spread) in the denomination, the ratio will be much larger than **1**. The samples then most likely **DO NOT** come from a common population; **reject Ho: µ1 = µ2 = µ3.**

* **Large / Small = reject Ho:** At least one mean is outlier and each distribution is narrow; distinct from each other.
* **Similar / Similar = Fail to reject Ho:** Means are fairly close to overall mean and/or distribution overlap a bit (hard to distinguish).
* **Small / Large = Fail to reject Ho:** The means are very close to overall mean and/or distribution “melt” together.

**1. One-Way (Single-Factor) ANOVA:** In experimental contextis also known as “Completly Randomized Design”. **In excel: Data Analysis ANOVA: Single Factor**

**Step 1:** Calculate the column and overall mean

**Step 2:** Calculate Sum of Squares Total; **SST = ∑ (Xi – overall mean)2.**

**Step 3:** Calculate Sum of Squares Columns; **SSC = ∑ (Column mean – Overall mean)2.**

**Step 4:** Calculate Sum of Squares Error; SSE = **∑∑ (Xi – Columni mean)2**

**Step 5:** Partitioning SST**; SST = SSC + SSE.**

**Step 6:** Calculate F-ratio; **F-ratio = MSC / MSE**, where **MSC = SSC / (df column = C-1)** and **MSE = SSE / (df** **error = N – C).**

**Step 7:** if **F > F-critic** then reject Ho. For calculating F-critic **= F.INV.RT (α; dfC; dfE).** The **P-Value** give us the significance of our test.

**2. Two-Way ANOVA without Replication (Two-Factor Without Replication):** in experimental context it is also known as “Randomized Block Design”.

* In one-way ANOVA, we selected a random sample for each column. A two-way ANOVA allows us to “account for variation” at the Row level due to some other **factor or grouping.**
* This allows greater focus on **Column or Group differences** making it easier to **detect group differences.**
* We are attempting to **minimize the ERROR variance** by saying: “now, some of the Error variance is actually due to the **variance in ROWS**”.
* By adding blocks of factors to the **ROWS**, we can **“subtract out”** that **ROW variance** from the **overall ERROR** variance.
* So, we will have 4 types of **SS/Source of variance**. **SST = SSC + SSB (Blocks/Rows) + SSE.**

**Very Important Note:** Assigning or allocating a certain proportion of variation to specific factors or variable is at the heart of **simple and multiple linear Regression**. In fact, the explanatory power of linear regression is actually based on **ANOVA**. It’s all tied together.

**Step 1:** Replication the first 4 steps of One-Way ANOVA

**Step 2:** Calculation the variance of Rows: **SSB = ∑ (Row mean – Overall mean)2.**

**Step 3:** Partitioning SST; **SST = SSC + SSE + SSB or SSE = SST – (SSC + SSB)**

**Step 4:** Calculating Blocks F-Ratio; **F-ratioR = MSB / MSE**, where **MSB = SSB / (df Rows = B-1)** and **MSE = SSE / (df** **error = (B-1)(C-1))**

**Step 5:** Calculate F-ratio; **F-ratio = MSC / MSE**, where **MSC = SSC / (df column = C-1)** and **MSE = SSE / (df** **error = (B-1)(C-1)).**

**Step 6:** if **F > F-critic** then reject Ho. For calculating F-critic Rows **= F.INV.RT (α; dfB; dfE)** and For calculating F-critic Column **= F.INV.RT (α; dfC; dfE).** The **P-Value** give us the significance of our test.

**Note 1:** In excel; **Data Analysis ANOVA: Two Factor w/o Replication**

**Note 2:** In Two-way ANOVA; **dfT = n-1**, **dfC = C-1, dfE =B-1**

**3. Two-Way ANOVA with Replication (Two-Factor with Replication):** in experimental context it is also known as “Balanced Design”.

* In this type of ANOVA, we have 1) two factors and 2) multiple, equal numbers of measurements for each factor combination (hence “balanced”).
* Each cell has its own **µ (x bar), δ2 (S2),** since it contains multiple measurement. Both **xbar** and **S2**are important
* Each factor level will also have its own mean and variance.
* There will be an overall/grand mean.
* A **graph of marginal means** is very helpful for visualizing the data. On a marginal means graph, usually the factor of interest or the factor with the most level is on the x-axis and the dependent variable (what is being measured) is at y-axis.
  + The second factor makes up the series lines by graphing the dependent variable value for each level/category of first factor.
  + Look for crossing or non-parallel lines on the graph.

**Very Important Note:** Always look at significance of interaction, if it’s significant, then the main factor’s significance can’t be analyzed; the factors are too **intertwined (factors are interdependent).**

* Check the **p-value (sig.)** of our variable as well as their interaction. If **p-value < α** then our variable effect is statistically significan**t.**
* The first **p-value** we checkis **the sig. of interactions.** If it’s significant we **STOP** right there because it means our variables are **INTERDEPENDENT.**

**3.1. Interaction: Considering different factors**

* What we are really talking about then is factoring in, or taking into account, several piece of information.
* The outcome is **influenced by other factors**. Also those factor may **influence each other (very important).**
* If it appears that factors works for some samples but not for others we can say that factors have **interaction** with each other.
* An **interaction** occurs when the effect of one factor changes for different levels of the other factor.
* On a marginal means graph, as a general rule, we look to see if the lines **cross** or **“would cross”.**
* If interaction is sig. the individual factors can’t be analyzed.

**4. ANOVA Post Hoc Test (Fisher’s LSD):** it is also known as **after the fact / Multiple Comparison**. There are several Post Hoc test and Fisher’s LSD (Least Significant Differences) is one of them.

* The ANOVA procedure only tell us if all the population means are equal (are likely to come from the same population).
* If the F-test is significant, we don’t know **WHERE** the differences located.
* However, it’s necessary to compare each population pair:
  + For counting all possible pairs use **c (n,2)**
* There are several multiple comparison procedures such as: Tukey HSD, Scheffers’ Method, Bon ferroin, etc. and depending on the data, one method may be more appropriate (for more info: <https://www.statisticshowto.com/post-hoc/>)
* For formulas and procedures: <https://www.slideshare.net/789667/310-spring2012-chapter14-anova>

**Non-Parametric Models**

**1. Introduction:** most statistical methods we are familiar with make assumptions about probability distribution of the population under analysis as normal distribution. From this assumption we can develop sampling distributions which from it we can derive sample statistics. However, in non-parametric models we don’t need to have this assumption (the distribution does not need to be normal).

* While parametric models mostly require quantitative data, non-parametric methods allow us to work with **qualitative (nominal/ordinal) data**.
* Most of the time, even quantitative data is converted to **nominal/ordinal data** for use with non-parametric methods; the most common type being **ranked observation**.
* In non-parametric methods we are interested in **MEDIAN** compare to MEAN in parametric methods.

**2. Sign Test:**

* When we remove the numerical value and search for Median’s location, what we have is essentially a **binomial distribution problem**.
* A binomial has two (binary) possibilities; hence “bi-“nomial.
* We turned the quantitative measures (e.g. salary) into two discreate categories; **below hypothesized median (+)** and **above hypothesized median (-).**
* Hence, we can use binomial distribution to test our hypothesis about the median.

**Note:** Any observation that are the same as hypothesized median are **REMOVED** and the test proceeds as usual (one or more observations will deduct from the total number of observations).

**2.1. Sign Test procedure for Median:**

**Step 1:** Hypothesize about the median.

**Step 2:** Rank orders visually (optional but suggested).

**Step 3:** Remove any value equal to median.

**Step 4:** See which observation fall below/above the hypothesized median

**Step 5:** (Evantually) use the binomial distribution to test our hypothesis.

**Important Note:** The median is a measure of **LOCATON**

**2.2. Plus Sign Probability:** if our sample size is greater than 20, we can use normal approximation of the binomial distribution. Above around **n = 20**, the two distributions converge.

**H0: p = 0.5; Ha: p != 0.5; µ = 0.5n; std = δ = √o.25n**

**p-value = p (x =< c); where c is the number of observations below our Median’s value**

**P-value = BINOM.DIST (c; n; p; True)**

**If p-value < α then reject H0**

**3. Mann-Whitney-Wilcoxon Rank Sum Test**