Mithilesh singh

HYPOTHESIS TESTING

Likelihood vs. Probability: What's the Difference?

likelihood and **probability**.

In contexts where numbers are not important, we can also talk about the likelihood of something happening.

Likelihood is used to talk in very general terms about whether there is a chance that something will happen or not.

Example:-

The treaty increases the likelihood that the cease-fire will hold.

There's every likelihood that he'll graduate university exam on time.

There's very little likelihood that the publisher will publish your book.

On the other hand,

Probability is , like chance, used to describe the level of how likely it is that something will happen.

Examples: -

What is the probability of winning a game?

What is the probability that result is distinction?

Here’s the difference in a nutshell:

- **Probability** refers to the chance that a particular outcome occurs based on the values of parameters in a model.

- **Likelihood** refers to how well a sample provides support for particular values of a parameter in a model.

When calculating the probability of some outcome, we assume the parameters in a model are trustworthy.

However, when we calculate likelihood we’re trying to determine if we can trust the parameters in a model based on the sample data that we’ve observed.

Example 1. Likelihood vs. Probability in Gambling

Suppose a casino claims that the probability of winning money on a certain slot machine is 40% for each turn.

If we take one turn , the **probability** that we will win money is 0.40.

Now suppose we take 100 turns and we win 42 times. We would conclude that the **likelihood** that the probability of winning in 40% of turns seems to be fair.

When calculating the probability of winning on a given turn, we simply assume that P(winning) =0.40 on a given turn.

However, when calculating the likelihood, we’re trying to determine if the model parameter P(winning) = 0.40 is actually correctly specified or Not.

In the example above, winning 42 times out of 100 makes us believe that a probability of winning 40% of the time seems reasonable.

Example 2: Likelihood vs. Probability in Coin Tosses

Suppose we have a coin that is assumed to be fair. If we flip the coin one time, the **probability** that it will land on heads is 0.5.

Now suppose we flip the coin 100 times and it only lands on heads 17 times. We would say that the **likelihood** that the coin is fair is quite low. If the coin was actually fair, we would expect it to land on heads much more often.

When calculating the probability of a coin landing on heads, we simply assume that P(heads) = 0.5 on a given toss.

However, when calculating the likelihood, we’re trying to determine if the model parameter (p = 0.5) is actually correctly specified or Not.

Hypothesis: -

Hypothesis Tests -→statistical analysis.

Hypothesis Testing is a **form of inferential statistics that allows us to draw conclusions about an entire population based on a representative sample**

In most cases, it is simply impossible to observe the entire population to understand its properties. The only alternative is to collect a random sample and then use statistics to analyze it

Suppose we toss a coin 10 times and we get 8 tails. Now we can start wondering whether the coin is fair.

So the question becomes, is getting 8 tails sufficient evidence to conclude that the coin is biased?

This is a question that's being addressed by what's called hypothesis tests.

some terminology: -

The null hypothesis which is sometimes written as

H_0 null hypothesis says that nothing extraordinary is going on.

→Null hypothesis is the current situation whatever it is,

→Null hypothesis is not interested in change.

So that's a very generic description.

In the case of coin tossing, nothing extraordinary simply means that the coin is fair.

So in other words,

null hypothesis H_0 → probability of getting tails is a 0.5.

alternative hypothesis H_1 → probability of getting tails is not equal to 0.5

The purpose of Statistical hypothesis tests is to determine whether the null hypothesis is likely to be true given sample data.

If there is evidence against the null hypothesis given in the data, we might reject the null hypothesis in favor of the alternative hypothesis: that means something interesting is going on, some changes will be there.

Significance level: -

Once we have the null and alternative hypothesis in hand, we should have a significance level (often denoted by the Greek letter α.).

The significance level is a probability threshold that determines when we can reject the null hypothesis.

After carrying out a test, if the probability of getting a result as extreme as the one you observe due to chance is lower than the significance level, we reject the null hypothesis in favor of the alternative.

This probability is known as the p-value.

One medical example.

Suppose a company develops a new drug to lower blood pressure, then it tests the drug with an experiment that involves 1,000 patients.

Remember, the null hypothesis always means nothing extraordinary is going on.

In this case, nothing extraordinary means that the drug has no special effect.

So, our null hypothesis, → no change in the blood pressure of the patients,

whereas the alternative hypothesis → the blood pressure drop.

So, if you are a scientist in this company,

your goal actually is to reject the null Hypothesis.

PROBABILITY VALUE (p-Value) approach: -

p-values – given the null hypothesis is compared with alpha value (alpha value or significant value of 0.05 or 0.01 are used, corresponding to a 5% chance or 1%)

The result of a test of significance is either “statistically significant” or “not statistically significant”; there are no shades of grey.

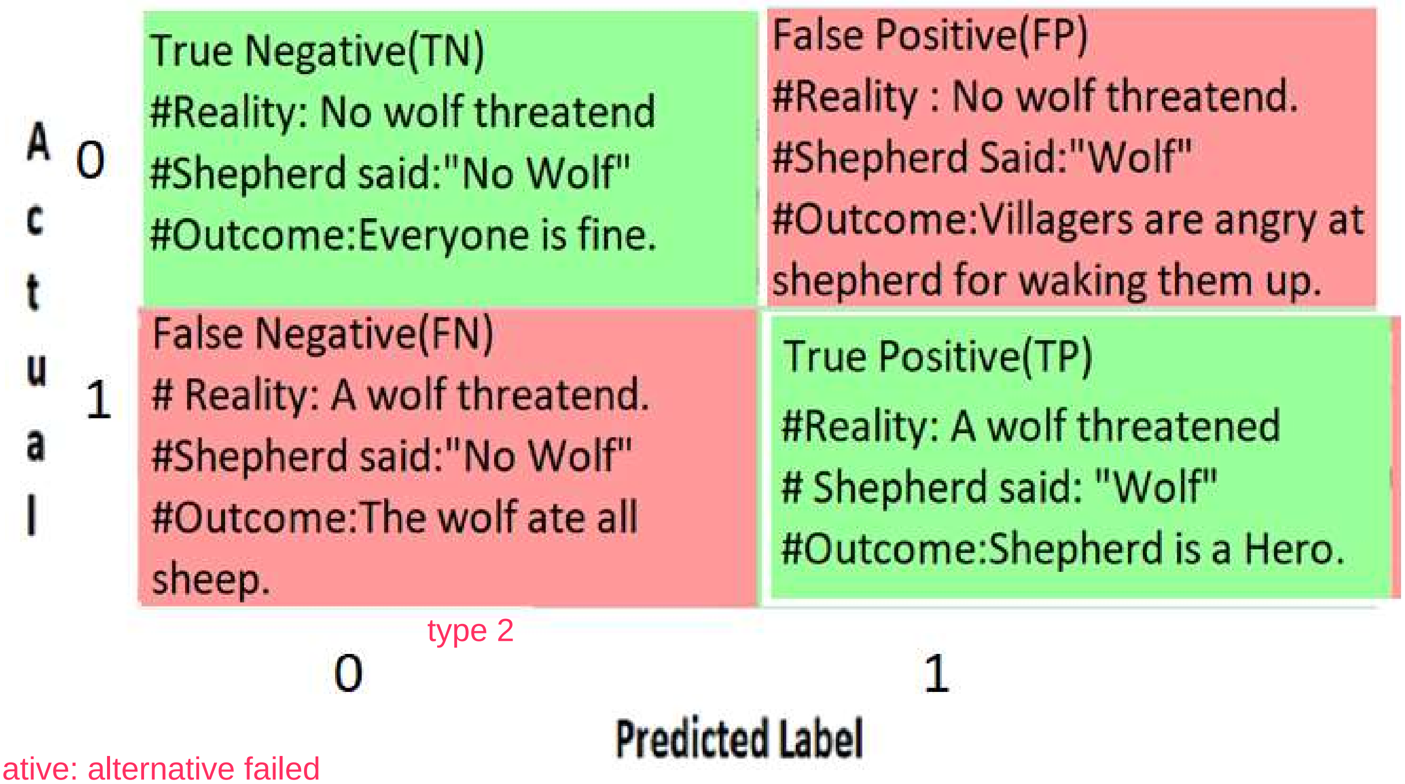
Understanding Type, I and Type II Error: -

Type I error describes a situation where you reject the null hypothesis when it is actually true. This type of error is also known as a "false positive". The type 1 error rate is equal to the significance level α, so setting a higher confidence level (and therefore lower alpha) reduces the chances of getting a false positive.

Type II error describes a situation where you fail to reject the null hypothesis when it is actually false. Type II error is also known as a "false negative". The higher your confidence level, the more likely you are to make a type II error.

CONFUSION MATRIX

type 1



Actual		Predicted Label 0	Predicted Label 1
	0	True Negative(TN) #Reality: No wolf threatend #Shepherd said:"No Wolf" #Outcome:Everyone is fine.	False Positive(FP) #Reality : No wolf threatend. #Shepherd Said:"Wolf" #Outcome:Villagers are angry at shepherd for waking them up.
	1	False Negative(FN) # Reality: A wolf threatend. #Shepherd said:"No Wolf" #Outcome:The wolf ate all sheep.	True Positive(TP) #Reality: A wolf threatened # Shepherd said: "Wolf" #Outcome:Shepherd is a Hero.

type 2

Negative: alternative failed
Positive : alternative sucess

Another example of MEDICAL TEST: -

- **True Negatives (TN): We predicted no, and they don't have the disease.**
- **True Positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease.**
- **False Negatives (FN): We predicted no, but they actually do have the disease. (Also known as a "Type II error.")**
- **False Positives (FP): We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")**

When performing a Hypothesis Test, two types of errors could occur:

this is major issue in approval of new product

- Type-I Error: reject the Null Hypothesis when it is actually true.

this is major issue when saying a covid patient is not a patient

- Type-II Error: accept the Null Hypothesis when it is actually false.

Types of Hypothesis Tests

Hypothesis Tests can be classified into two big families :

- **Parametric Tests**, if samples follow a normal distribution. In general, samples follow a normal distribution if their mean is 0 and variance is 1.

on scaling part standard
scalar(z score)

- **Non-Parametric Tests**, if samples do not follow a normal distribution.

on scaling min max scalar,robust scalar

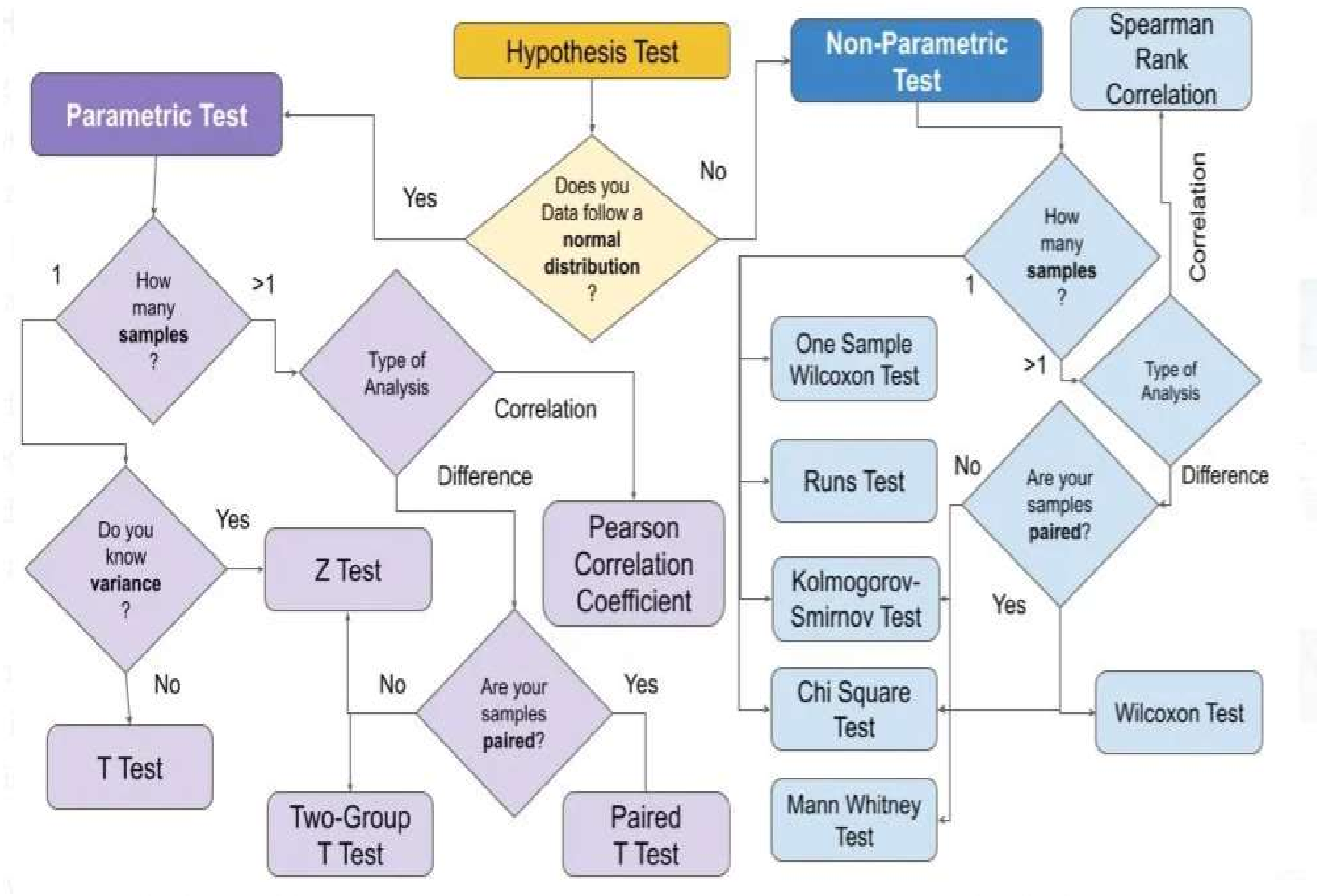
Depending on the number of samples to be compared, two families of Hypothesis Tests can be formulated:

- **One Sample**, if there is just one sample, which must be compared with a given value
- **Two Samples**, if there are two or more samples to be compared. In this case, possible tests include **correlation** and **difference** between samples. In both cases, samples can be paired or not. **Paired samples** are also called dependent samples, while not paired samples are also called

independent samples. In paired samples, natural or matched couplings occur.

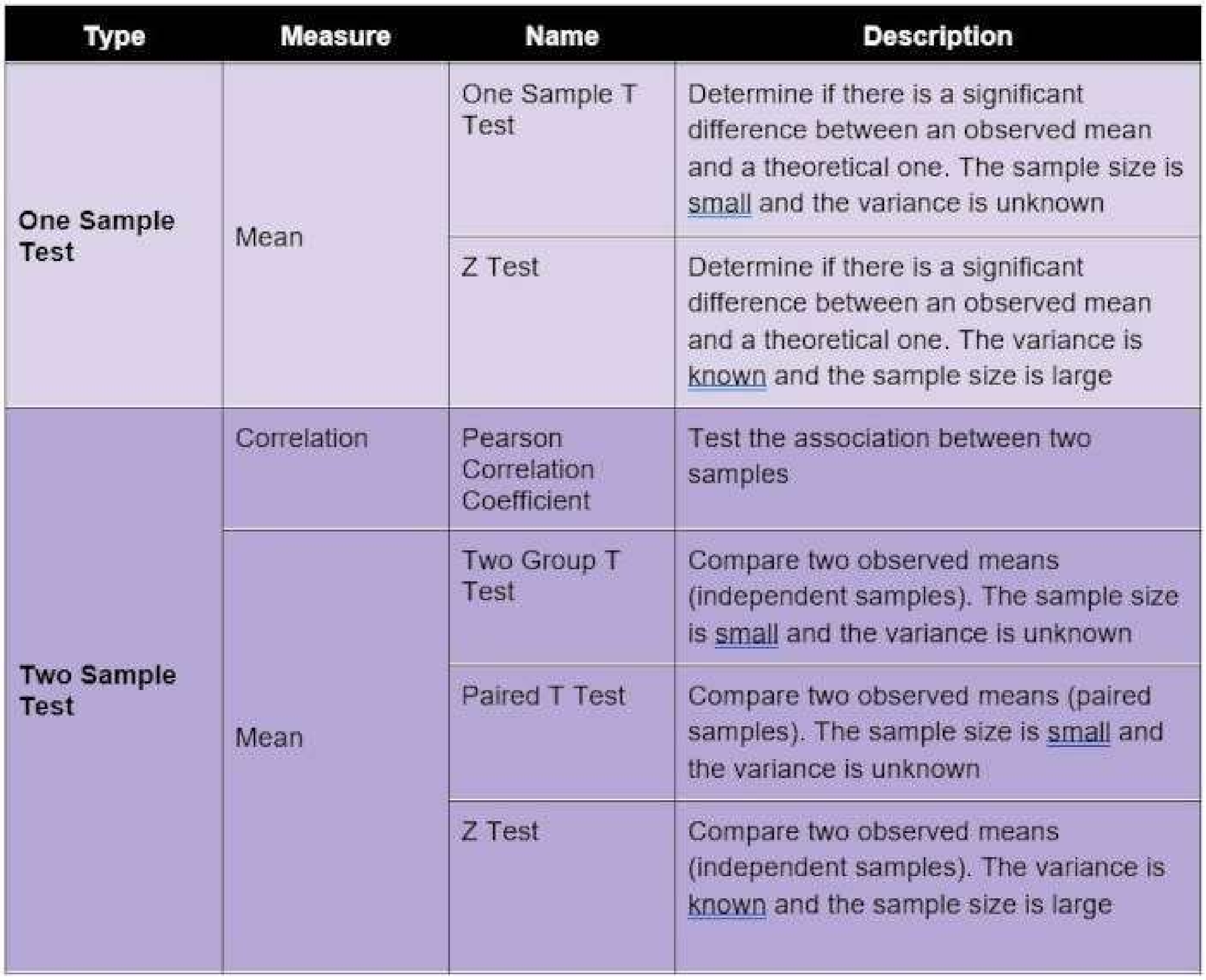
Usually, parametric tests have the corresponding non-parametric test, as well described in

The diagram featured at the top of this article reviews how to choose the right Hypothesis Test according to the sample.



Parametric Tests

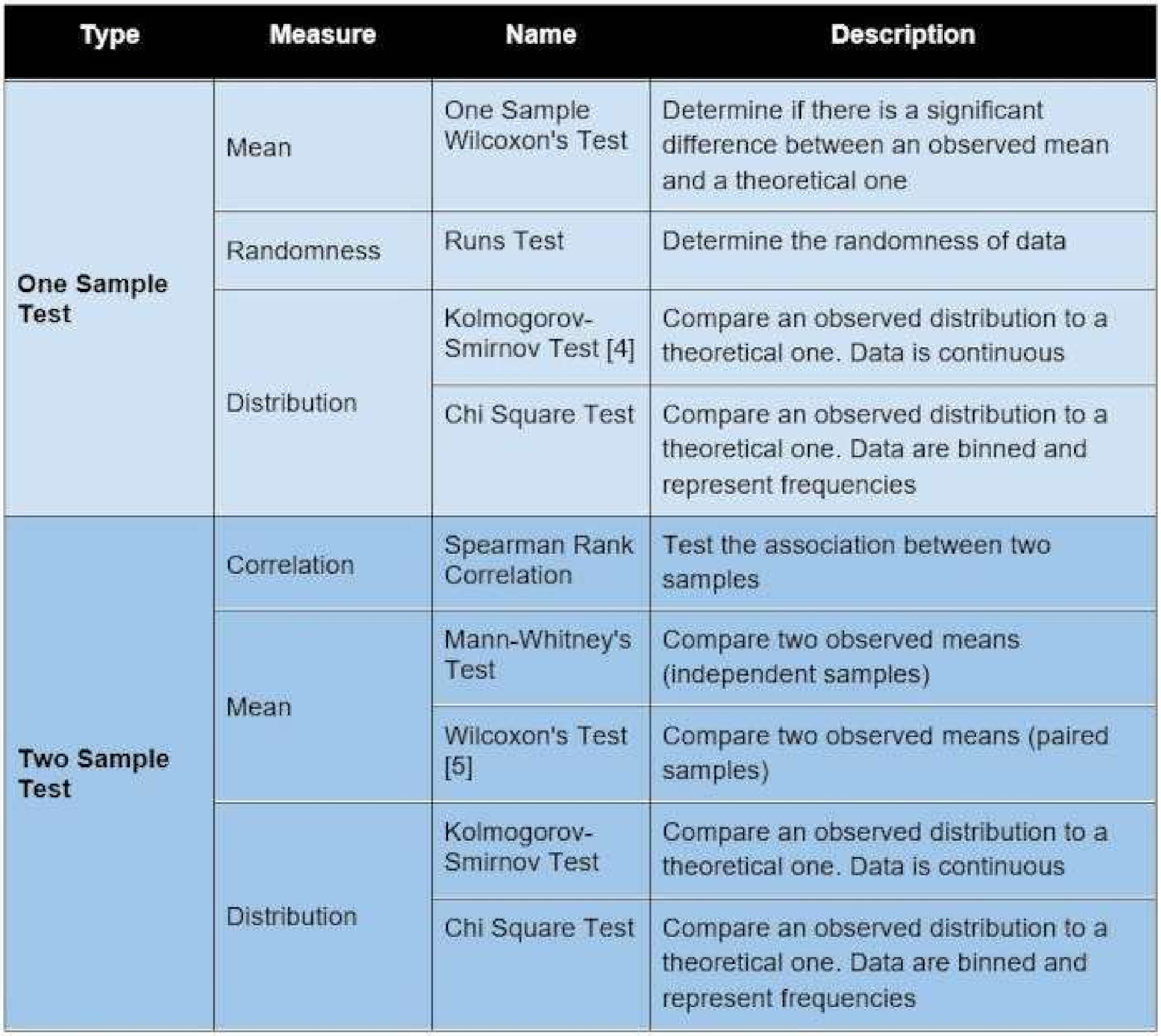
As already said, Parametric Tests assume a normal distribution in the data. The following table describes some of the most popular parametric tests and what they measure.



Type	Measure	Name	Description
One Sample Test	Mean	One Sample T Test	Determine if there is a significant difference between an observed mean and a theoretical one. The sample size is small and the variance is unknown
		Z Test	Determine if there is a significant difference between an observed mean and a theoretical one. The variance is known and the sample size is large
Two Sample Test	Correlation	Pearson Correlation Coefficient	Test the association between two samples
	Mean	Two Group T Test	Compare two observed means (independent samples). The sample size is small and the variance is unknown
		Paired T Test	Compare two observed means (paired samples). The sample size is small and the variance is unknown
		Z Test	Compare two observed means (independent samples). The variance is known and the sample size is large

Non-Parametric Tests

Non-Parametric Tests do not make any assumptions on distribution in the data. The following table describes some of the most popular non-parametric tests and what they measure.



Type	Measure	Name	Description
One Sample Test	Mean	One Sample Wilcoxon's Test	Determine if there is a significant difference between an observed mean and a theoretical one
	Randomness	Runs Test	Determine the randomness of data
	Distribution	Kolmogorov-Smirnov Test [4]	Compare an observed distribution to a theoretical one. Data is continuous
		Chi Square Test	Compare an observed distribution to a theoretical one. Data are binned and represent frequencies
Two Sample Test	Correlation	Spearman Rank Correlation	Test the association between two samples
	Mean	Mann-Whitney's Test	Compare two observed means (independent samples)
		Wilcoxon's Test [5]	Compare two observed means (paired samples)
	Distribution	Kolmogorov-Smirnov Test	Compare an observed distribution to a theoretical one. Data is continuous
		Chi Square Test	Compare an observed distribution to a theoretical one. Data are binned and represent frequencies

Difference between Parametric and Non Parametric statistics

- The parametric tests are based on assumptions using the data connected to the normal distribution used in the analysis. Knowledge on the parameters is very essential. Whereas on the other hand non-parametric test does not depend on any parameters.
- It does not work on assumptions.
- Unlike parametric test there is no requirement of information on population in non-parametric.
- Parameter test is applicable to only for variables whereas non-parametric test is applicable both for variables and attributes.
- Person's coefficient or co-relation is used to measure degree of association between two different variables in case of parametric tests whereas spearman's rank correlation is used in non-parametric tests.

As per science and statistics: -

- **Hypothesis in Science**: Provisional explanation that fits the evidence and can be confirmed or disproved.

- **Hypothesis in Statistics**: Probabilistic explanation about the presence of a relationship between observations.

Null Hypothesis (H0): Suggests no effect.

Alternate Hypothesis (H1): Suggests some effect.

Hypothesis Definition as per Machine Learning: -

Machine learning, specifically supervised learning, :- use available data to learn a function that best maps inputs to outputs.

Technically, this is a called function approximation, where we are approximating an unknown target function (that we assume exists) that can best map inputs to outputs on all possible observations from the problem domain.

An example of a model that approximates the target function and performs mappings of inputs to outputs is called a hypothesis in machine learning.

The choice of algorithm (e.g. neural network) and the configuration of the algorithm (e.g. network topology and hyper-parameters) define the space of possible hypothesis that the model may represent.

- Candidate model that approximates a target function for mapping examples of inputs to outputs.
- Considering different Independent variables .

  - **Null Hypothesis (H0)**: Suggests no significance or no effect no improvement.
  - **Alternate Hypothesis (H1)**: Suggests some effect or some significant

Some examples that we will use Hypothesis Test in machine learning are: -

- A test that assumes that data has a normal distribution.
- The Independent variables are significant.
- A test that assumes that two samples were drawn from the same underlying population distribution.
- An example of a model that approximates the target function and performs mappings of inputs to outputs is called a hypothesis in machine learning.

- The choice of algorithm (e.g. Linear Regression, KNN, neural network) and the configuration of the algorithm (e.g. network topology and hyper parameters) define the space of possible hypothesis that the model may represent.

Example: - Hypothesis Testing Population sample or two dataset

Hypothesis testing: - **testing of significance regarding a population parameter on the basis of sample**.

The sample is drawn from a population, its statistics are found and on the basis of such statistics it is seen whether the sample so drawn has come from the parent population with certain specified characteristics or not.

➔ The computed sample statistics may be differing from the hypothetical value of the population parameter. If the difference is small, it is considered that the small difference is arises due to sampling fluctuations and the Null hypothesis is accepted.

➔ If the difference is large it is it is considered that the large difference is arises not due to sampling fluctuations and the Null hypothesis is rejected.

➔ A hypothesis is a quantitative statement about the population. It may or may not be true. By testing the hypothesis, we can find out whether is deserves acceptance or Rejection.

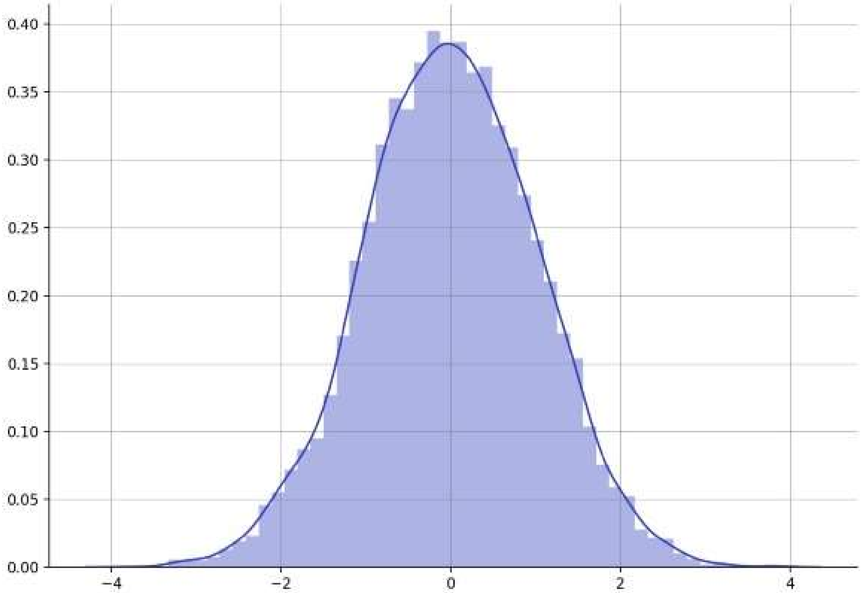
Hypothesis Testing to compare two datasets

Hypothesis Tests to compare two datasets, or a sample from a dataset. It is a **statistical inference method** we'll **draw a conclusion** —— about the characteristics of what we're comparing.

For example: -

we need to know *what distribution it follows*. Because, the different tests assume that data follows a specific distribution.

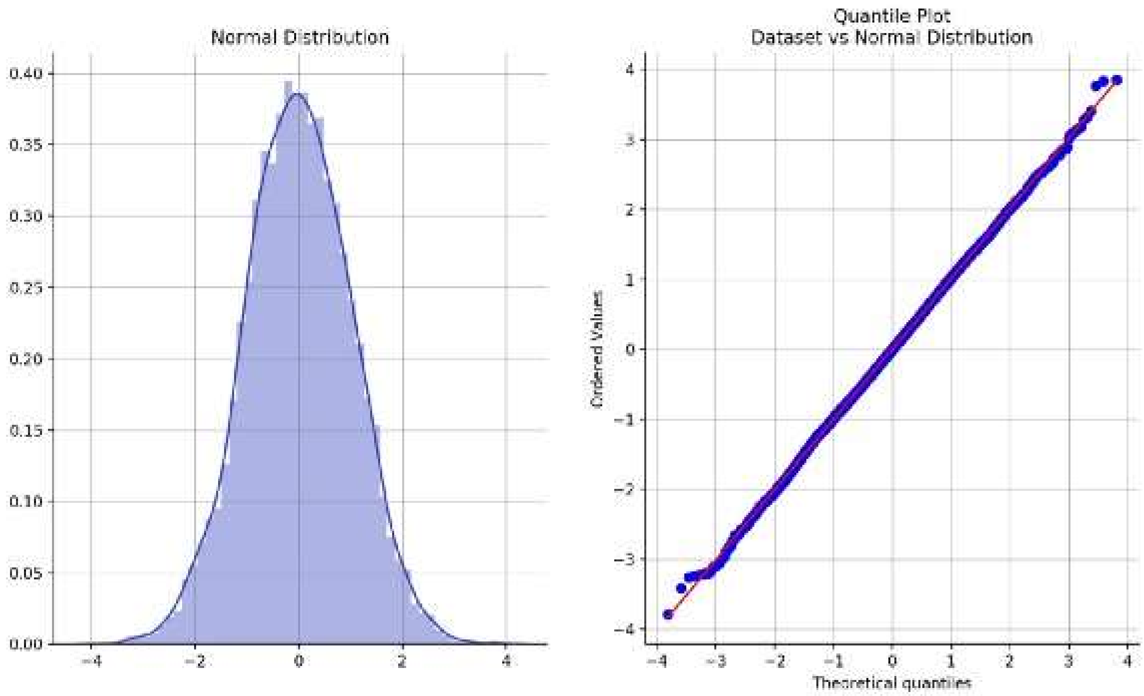
One of the most famous distributions is the so called *Bell Curve*, the Normal Distribution.



Example of a dataset that follows a Normal Distribution with mean 0 and standard deviation of 1

Do the Data follow a Normal Distribution?

Another method is : the Quantile-Quantile Plot, a.k.a., Q-Q Plot.



A dataset that follows a Normal Distribution and the Q-Q plot that compares it with the Normal Distribution

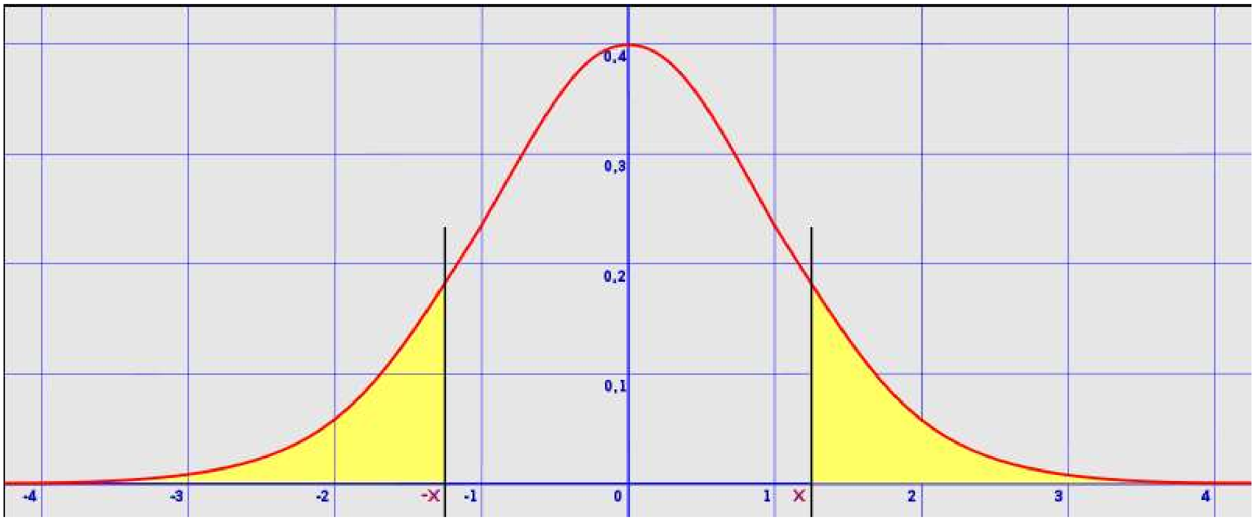
Q-Q plots helps visualize the quantiles of two probability distributions against one another.

The Q-Q plot intends to visually represent is that, if both datasets follow the same distribution, they'll roughly be aligned along the diagonal red line. The more the blue dots, deviate from the red diagonal line, , the bigger the difference between the two distributions.

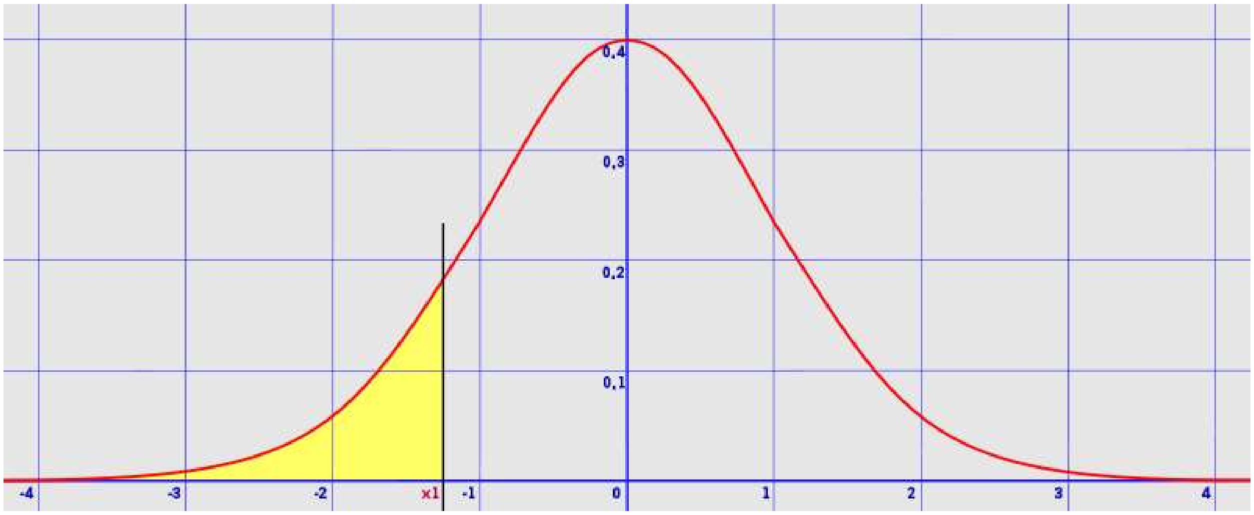
One tail test and Two tail test

If the test is a 2-tailed test, we divide the alpha by 2 to equally distribute the significance level on the lower and upper cut-off. In the case of a 1-tailed test, we keep the alpha as it is.

By looking at the figure



Two tailed test



One tailed test

If given a picture, we'll be able to tell if our test is one-tailed or two-tailed by comparing it to the image above.

However, most of the time you're given questions, not pictures.
So it's a matter of understanding the problem and picking out the important piece of information.

We're basically looking for keywords like equals, more than, or less than.

Example question #1: A government official claims that the dropout rate for local schools is 25%. Last year, 190 out of 603 students dropped out. Is there enough evidence to reject the government official's claim?

Example question #2: A government official claims that the dropout rate for local schools is less than 25%. Last year, 190 out of 603 students dropped out. Is there enough evidence to reject the government official's claim?

Example question #3: A government official claims that the dropout rate for local schools is greater than 25%. Last year, 190 out of 603 students dropped out. Is there enough evidence to reject the government official’s claim?

Step 1: Read the question.

Step 2: Rephrase the claim in the question with an equation.

In example question #1, Dropout rate = 25%
In example question #2, Dropout rate < 25%
In example question #3, Dropout rate > 25%.

Step 3: If the claim or statement has an equals sign in it, this is a two-tailed test.

If it has > or < it is a one-tailed test.
In the above examples, given specific wording like “greater than” or “less than.”

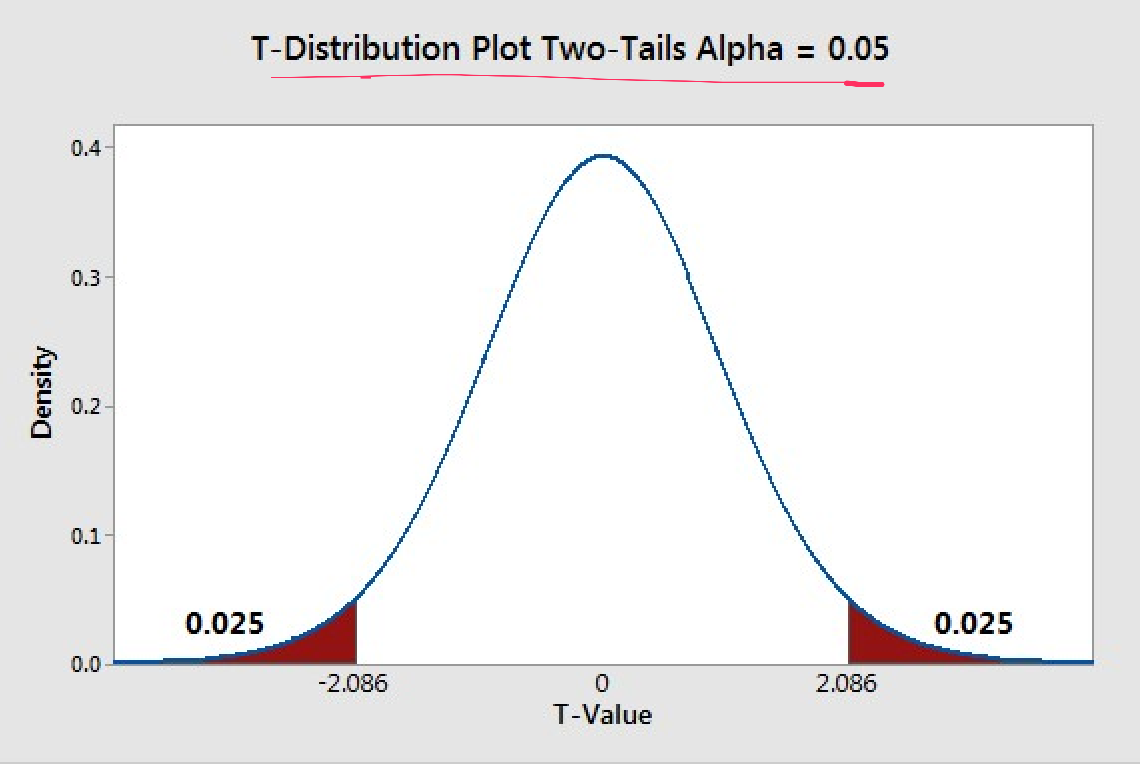
Pl note:-
The null hypothesis must always include an equals sign,

whether it be ≥, ≤, or just= ≥, ≤, or just =.
Usually, however, it's just = .
The alternative hypothesis is what we wish to show.

Important:-

- Alpha levels (also called "significance levels") are used in hypothesis tests; it is the probability of making the wrong decision when the null hypothesis is true.
- A one-tailed test has the entire 5% of the alpha level in one tail (in either the left, or the right tail).
- A two-tailed test splits your alpha level in half (as in the image to the left). if standard alpha level of 0.5 (5%). A two tailed test will have half of this (2.5%) in each tail.
- If this test statistic falls in the top 2.5% or bottom 2.5% of its probability distribution (in this case, the t-distribution), you would reject the null hypothesis.
- The terms "one tailed" and "two tailed" can more precisely be defined as referring to where your rejection regions are located.

Two tailed hypothesis test – non directional test because it has effect in both directions.



When a test statistics falls in any critical region , we can reject the null hypothesis.

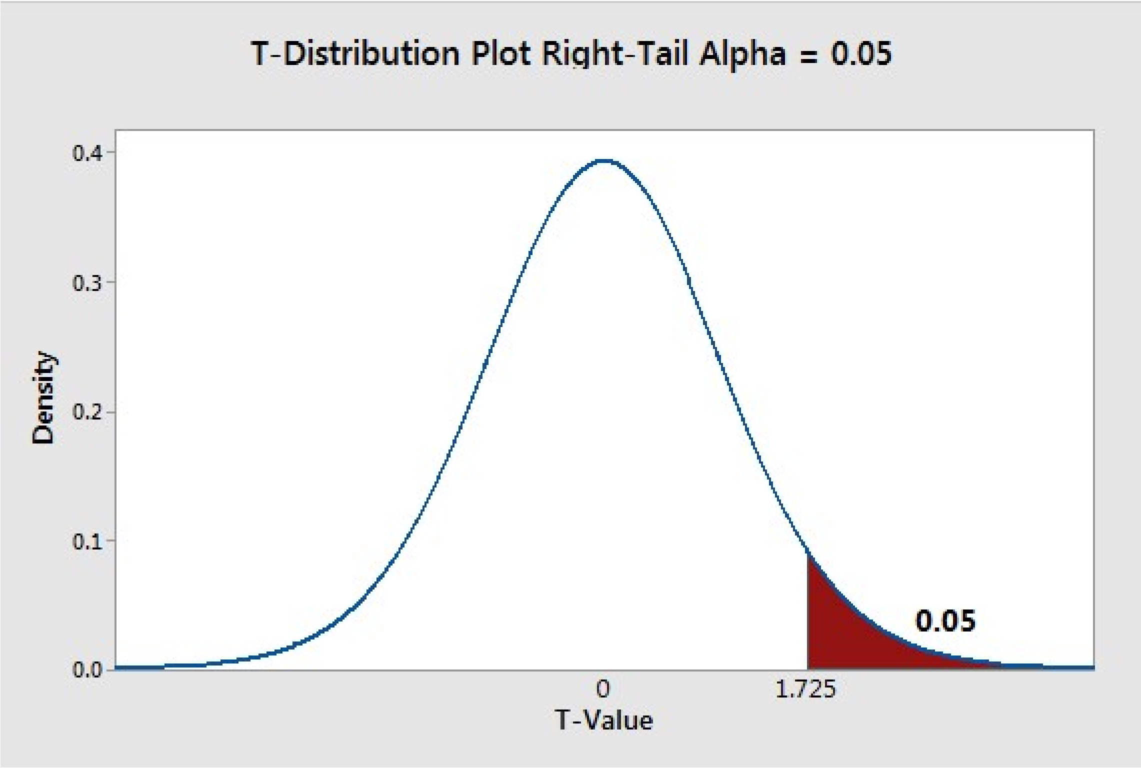
In a two-tailed test, the generic null and alternative hypotheses are the following:

- **Null**: The effect equals zero.
- **Alternative**: The effect does not equal zero.

Example of 2 tailed test: - Suppose we compare the mean strength of parts from a supplier to a target value (strength=100). We use a two-tailed test because we care whether the mean strength of the new part is greater than or less than the target value.

Two-tailed tests are standard in scientific research ,manufacturing industries, where discovering any type of effect.

One-Tailed Hypothesis Tests—directional test, because we test the effect in one direction. It can be either left side of right side.

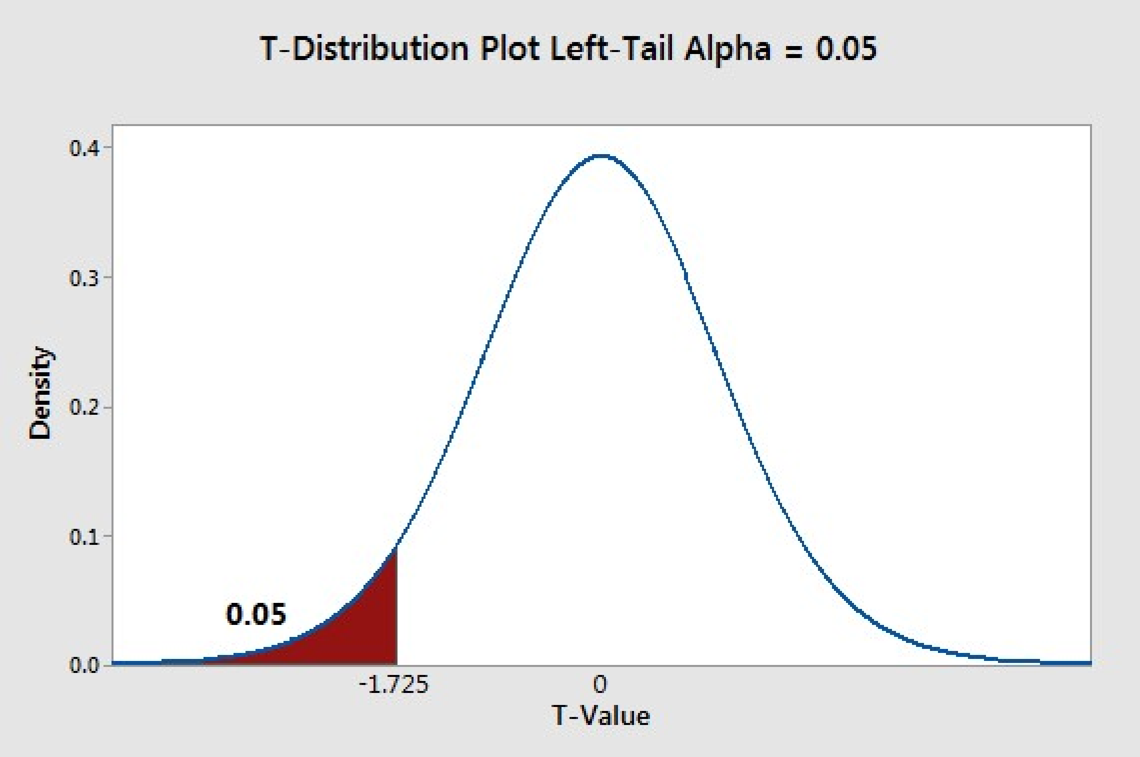


In a two-tailed test, the generic null and alternative hypotheses are the following:

- **Null**: The effect is less than or equals to zero.
- **Alternative**: The effect greater than zero.

Or:

- **Null**: The effect is greater than or equal to zero.
- **Alternative**: The effect is less than zero.



Example of a one-tailed

Suppose a new supplier has approached—we have already a old supplier

In last case we compared the mean strength of parts from the old supplier () to a target value (strength=100).

Imagine that we are considering a new parts supplier.

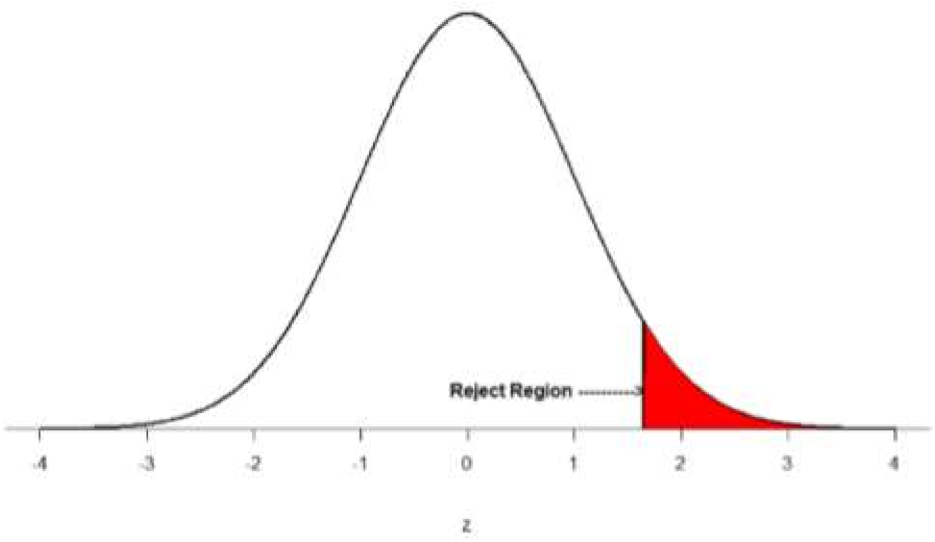
We will use them only if the mean strength of new and old supplier parts is greater than our target value.

We are not interested to test whether their parts are equally strong. Because if the new supplier part is equally strong or less strong than the target value—than we'd just stick with our current supplier.

Z score , Critical values, p-values, and significance level

Values of z which fall in the tails of the standard normal distribution represent unlikely values. That is, the proportion of the area under the curve as or more extreme than z is very small as we get into the tails of the distribution.

Our significance level corresponds to the area under the tail that is exactly equal to alpha(α): if we use our normal criterion of α= .05, then 5% of the area under the curve becomes what we call the rejection region (also called the critical region) of the distribution.



One Tail rejection region

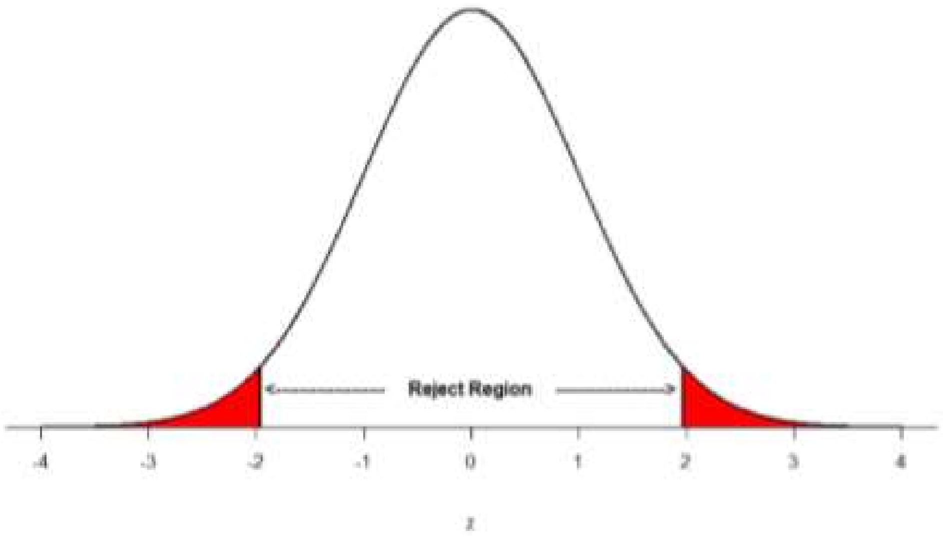
The shaded rejection region takes us 5% of the area under the curve.

The rejection region is bounded by a specific z-value, as is any area under the curve. In hypothesis testing, the value corresponding to a specific rejection region is called the critical value $zcrit$ ("z-critical") or $z*$ (hence the other name "critical region").

Refer the Z normal table, we will find that the z-score corresponding to 5% of the area under the curve is Approximately equal to 1.645 if we go to the right and -1.645 if we go to the left.

Suppose, however, that we want to do a non-directional test means TWO TAIL TEST . We need to put the critical region in both tails, but we don't want to increase the overall size of the rejection region .To do this, we simply split it in half so that an equal proportion of the area under the curve falls in each tail's rejection region.

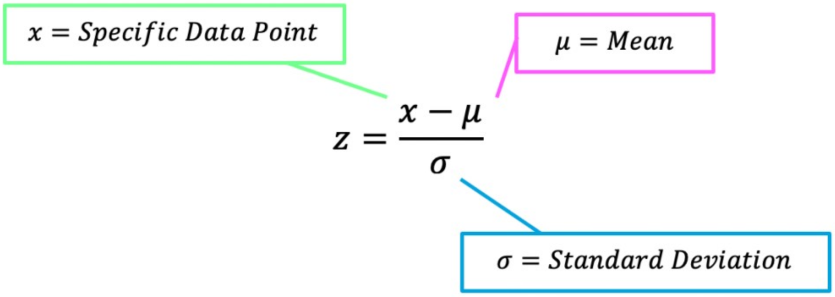
For α = .05, this means 2.5% of the area is in each tail, which, based on the z-table, corresponds to critical values of $zcrit$ ("z-critical") = ±1.96.



Two tail rejection region.

Thus, any z-score falling outside ±1.96 (greater than 1.96 in absolute value) falls in the rejection region.

When we use z-scores in this way, the obtained value of z (sometimes called z-obtained) is something known as a test statistic, which is simply an inferential statistic used to test a null hypothesis. The formula for our z-statistic :-



$$z = \frac{x - \mu}{\sigma}$$

$x = Specific\ Data\ Point$

$\mu = Mean$

$\sigma = Standard\ Deviation$

To formally test our hypothesis, we compare our obtained z-statistic to our critical z-value. If Zobt>Zcrit, that means it falls in the rejection region and so we reject $H0$.

If Zobt<Zcrit, we fail to reject.

Please note: - as z gets larger, the corresponding area under the curve beyond z gets smaller. Thus, the proportion, or p-value, will be smaller than the area for α, and if the area is smaller, the probability gets smaller.

The z-statistic is very useful when we are doing our calculations by hand. However, when we use ML software, it will report to us a p-value, which is simply the proportion of the area under the curve in the tails beyond our obtained z-statistic.

so zobt and p value are inversly prob

We can directly compare this p-value to α to test our null hypothesis: if $p<\alpha$, we reject Null hypothesis $H0$, but if $p>\alpha$, we fail to reject Null Hypothesis.

When the null hypothesis is rejected, the effect is said to be statistically significant. statistically significant signifies that the effect is real and not due to chance.