Mithilesh singh

Likelihood vs. Probability: What’s the Difference?

likelihood and **probability**.

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.

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.

In the example above, a coin landing on heads only 17 out of 100 times makes us highly suspicious that the truly probability of the coin landing on heads on a given toss is actually p = 0.5.

Hypothesis: -

Hypothesis Tests -→statistical analysis.

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.

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,

we can write the null hypothesis as saying that the probability of getting tails is a half.

Then there's a second hypothesis called the alternative hypothesis.

H_1

It says that there's a different chance process at work.

So in our example,
the alternative hypothesis would be that the probability of getting tails is not equal to one-half.
The exact form of the alternative hypothesis “Ha” will depend on the specific test you are carrying out.

Statistical hypothesis tests are based a statement called the null hypothesis that assumes nothing interesting is going on between whatever variables you are testing. .

The purpose of a hypothesis test is to determine whether the null hypothesis is likely to be true given sample data. If there is little evidence against the null hypothesis given the data, you accept the null hypothesis. If the null hypothesis is unlikely given the data, you might reject the null in favor of the alternative hypothesis: that something interesting is going on. The exact form of the alternative hypothesis will depend on the specific test you are carrying out.

Significance level: -

Once you have the null and alternative hypothesis in hand, you choose a significance level (often denoted by the Greek letter α.). The significance level is a probability threshold that determines when you 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, you reject the null hypothesis in favor of the alternative. This probability of seeing a result as extreme or more extreme than the one observed is known as the p-value.

Let's look at a different 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, in this case,

becomes that there's no change in the blood pressure of the patients,

whereas the alternative hypothesis is, that there's a blood pressure drop.

So, if you put yourselves in the shoes of the company, you see that in this case,

your goal actually is to reject the null Hypothesis.

So oftentimes, the logic behind testing is indirect.

One assumes that nothing extraordinary is happening, and then hopes to reject this assumption.

Making the Decision

In statistics, there are two ways to determine whether the evidence is likely or unlikely given the initial assumption:

- We could take the "**critical value approach**" (favored in many of the older textbooks).

- Or, we could take the **"*P*-value approach"** (what is used most often in research, journal articles, statistical Machine learning software).

PROBABILITY VALUE (p-Value) approach: -

The p-value is the probability of obtaining a result at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. The fact that p-values are based on this assumption is crucial to their correct interpretation.

Usually p-values – an outcome that extreme, 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.

Continue the coin tossing experiment where the null hypothesis is that the coin is fair (p=0.5) and the alternative hypothesis is that the coin is biased in favor of heads (p>0.5).

Suppose the coin is tossed 10 times and 8 heads are observed. Since the alternative hypothesis is p>0.5, more extreme values are numbers of heads closer to 10.

So, to compute the p-value in this situation, you need only compute the probability of 8 or more heads in 10 tosses assuming the coin is fair.

But, the number of heads in 10 tosses of a coin assuming that the coin is fair has a binomial distribution with n=10 and p=0.5.

The p-value is P[8 heads] + P[9 heads] + P[10 heads].

From the binomial probability distribution, P[8 heads]=0.044, P[9 heads]=0.01, and P[10 heads]=0.001.

Thus the p-value is 0.044+0.010+0.001=0.055.

In the coin tosses above, the p-value is 0.055, and if alpha-value is 0.05, you would fail to reject the null hypothesis, that is, you would say 8 heads in 10 tosses is not enough evidence to conclude that the coin is not fair.

For example, you might reject the null only if you observe 9 or 10 heads in the 10 tosses.
The Alpha value or significance level of a test is a value that should be decided upon by the researcher interpreting the data before the data are viewed, and is compared against the p-value or any other statistic calculated after the test has been performed.

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" or "false hit". 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" or "miss". The higher your confidence level, the more likely you are to make a type II error.

The process of testing hypotheses can be compared to court trials.

A person comes into court charged with a crime. A jury must decide whether the person is innocent (null hypothesis) or guilty (alternative hypothesis).

Even though the person is charged with the crime, at the beginning of the trial (and until the jury declares otherwise) the accused is assumed to be innocent. Only if overwhelming evidence of the person's guilt can be shown is the jury expected to declare the person guilty--otherwise the person is considered innocent.

Possible Error: -

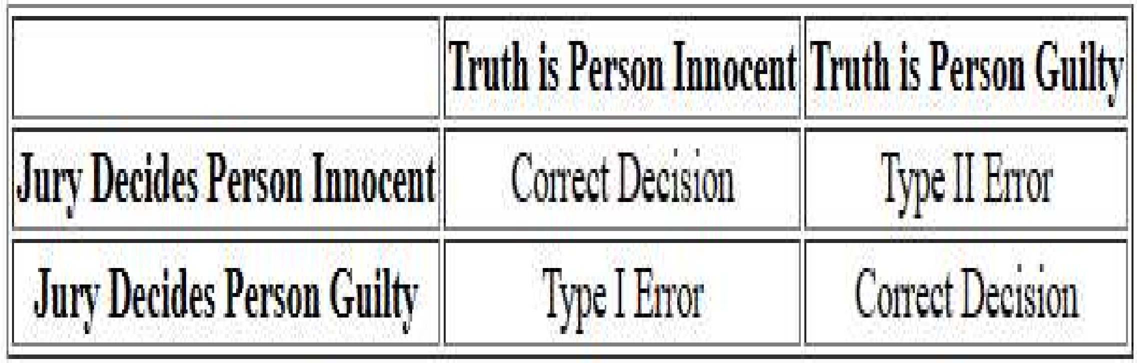
In the jury trial there are two types of errors:

(1) the person is innocent but the jury finds the person guilty, and

(2) the person is guilty but the jury declares the person to be innocent.

In our system of justice, the first error is considered more serious than the second error.

These two errors along with the correct decisions are Possible: -



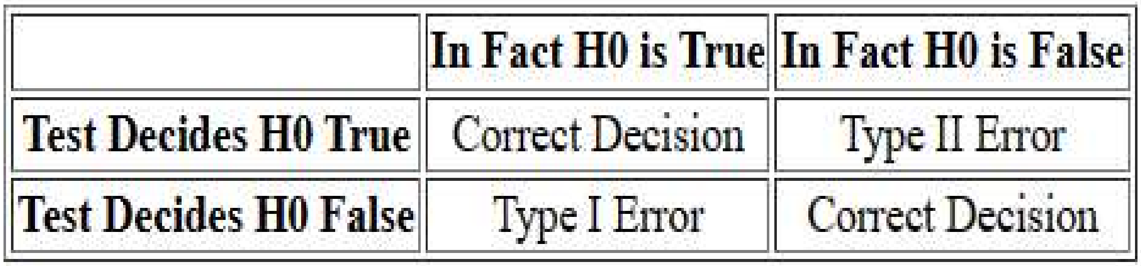
	Truth is Person Innocent	Truth is Person Guilty
Jury Decides Person Innocent	Correct Decision	Type II Error
Jury Decides Person Guilty	Type I Error	Correct Decision

With respect to hypothesis testing the two errors that can occur are:

(1) the null hypothesis is true but the decision based on the testing process is that the null hypothesis should be rejected, and

(2) the null hypothesis is false but the testing process concludes that it should be accepted.

These two errors are called Type I and Type II errors. As in the jury trial situation, a Type I error is usually considered more serious than a Type II error



	In Fact H0 is True	In Fact H0 is False
Test Decides H0 True	Correct Decision	Type II Error
Test Decides H0 False	Type I Error	Correct Decision

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.

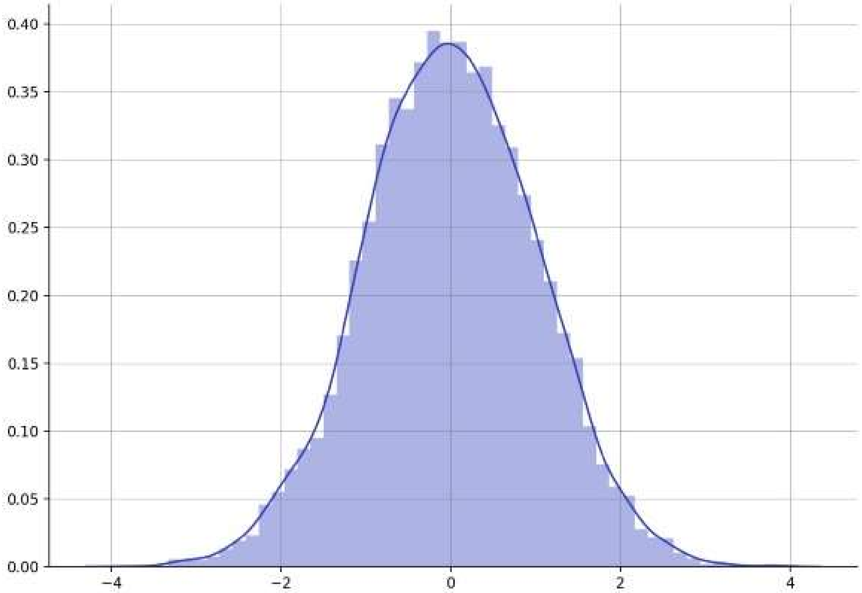
Hypothesis Testing to compare two datasets

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.

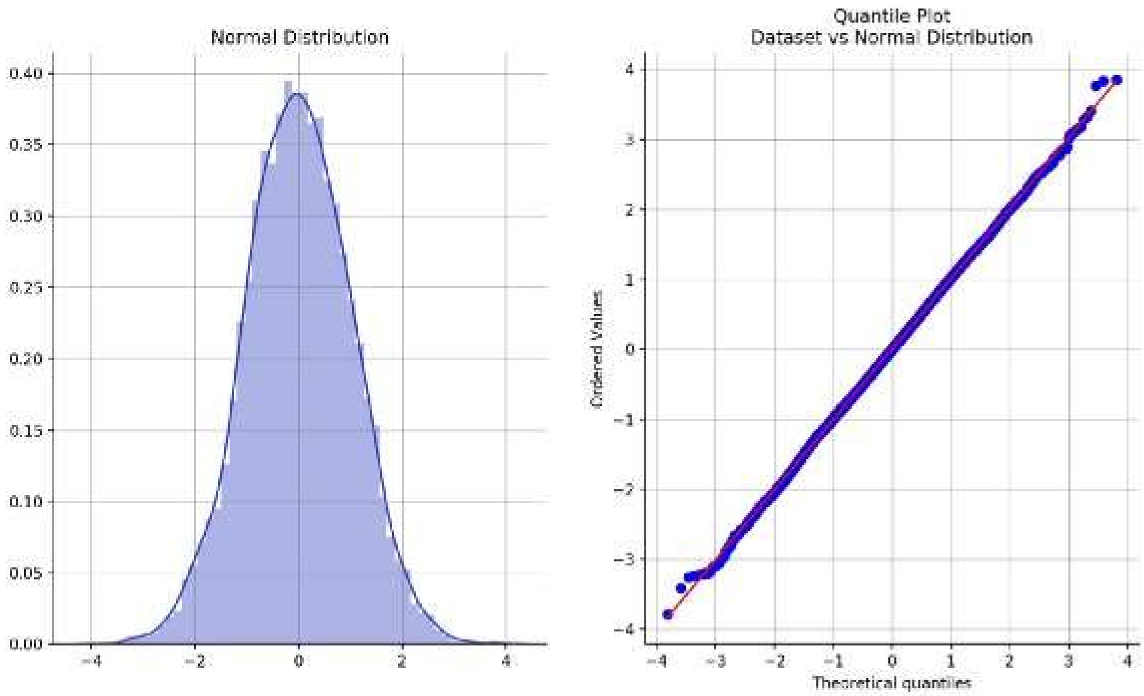
One of the most famous distributions is the so called *Bell Curve*, the Normal Distribution.



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

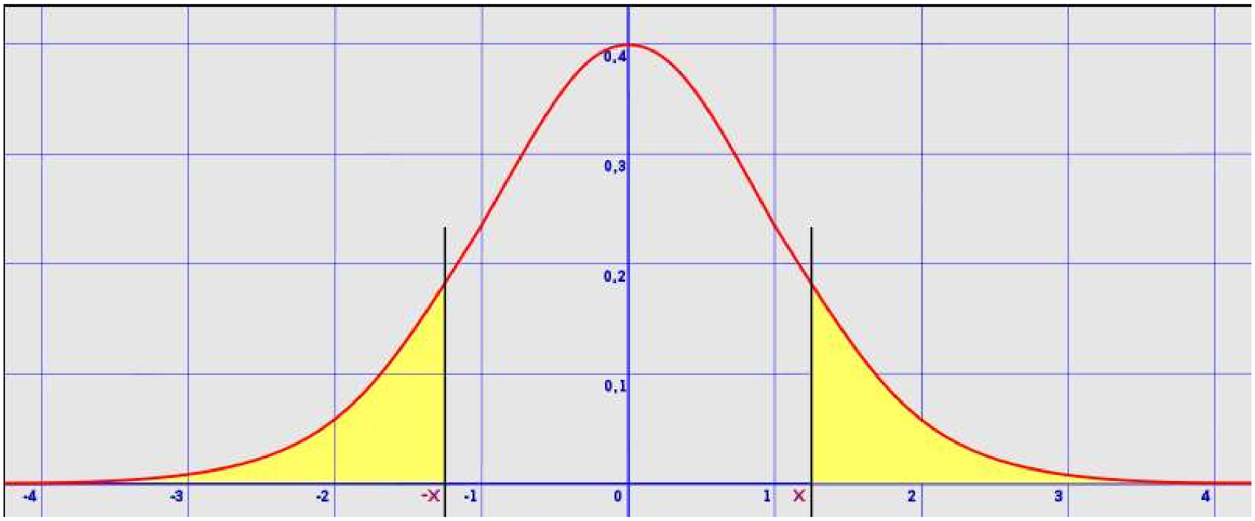
Q-Q plots helps visualize the quantiles of two probability distributions against one another.

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.

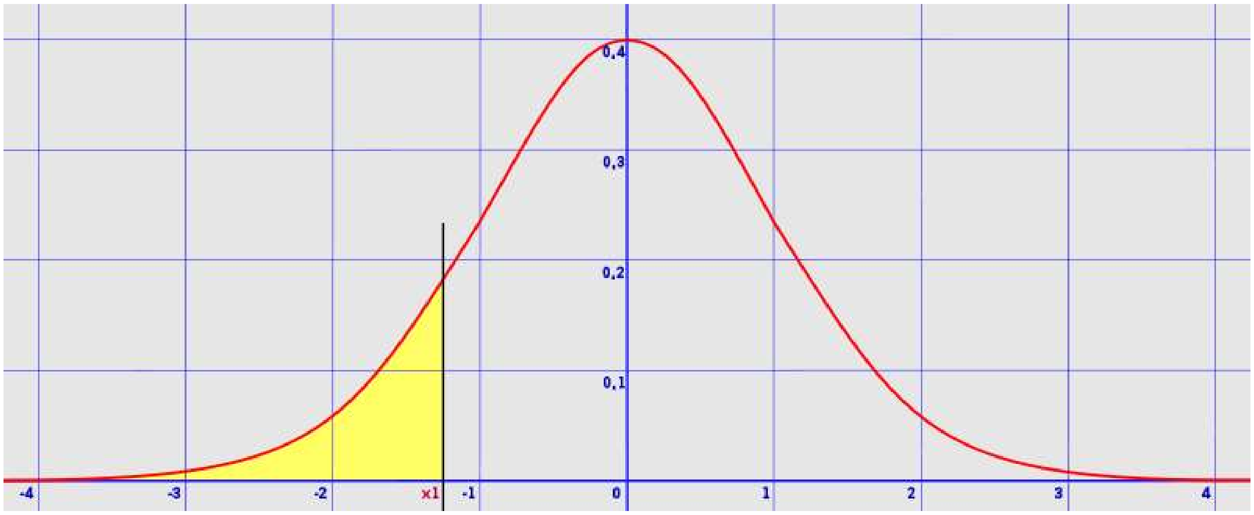
One tail test and Two tail test

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.

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.

When Should You Use a One-Tailed Test?

In the above examples, you were given specific wording like "greater than" or "less than."

Sometimes, the researcher, do not have this information and we have to choose the test.

For example, you develop a drug which you think is just as effective as a drug already on the market (it also happens to be cheaper).

You could run a two-tailed test (to test that it is more effective and to also check that it is less effective).

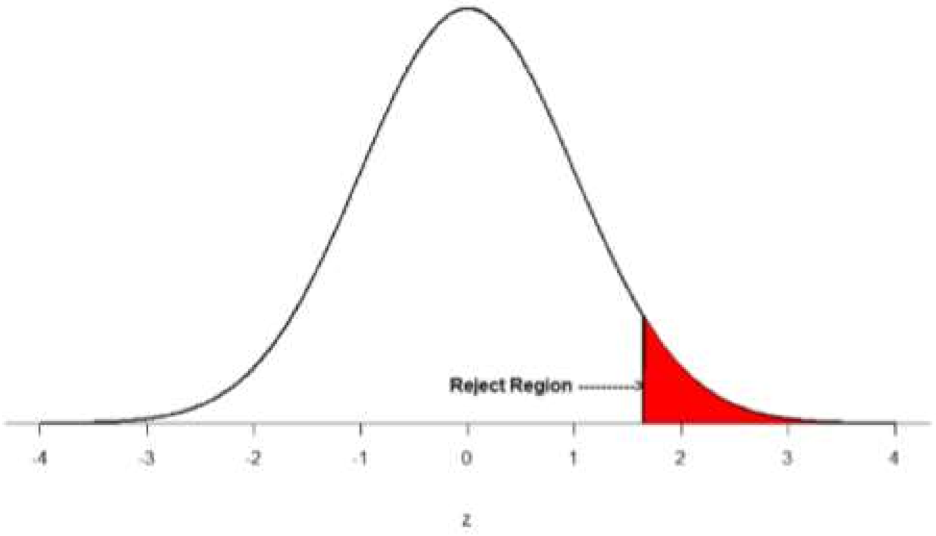
But suppose you don't really care about it being more effective, but you also don't want that it is any less effective. You can run a one-tailed test to check that your drug is at least as effective as the existing drug.

Critical values, p-values, and significance level

The significance level is a threshold we set before collecting data in order to determine whether or not we should reject the null hypothesis.

We set this value beforehand to avoid biasing ourselves by viewing our results and then determining what criteria we should use.

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

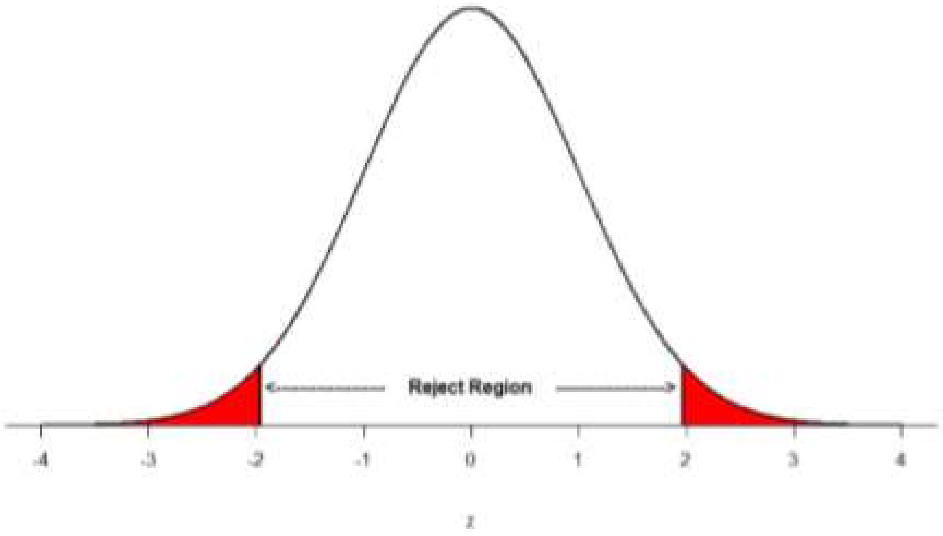
The shaded rejection region takes us 5% of the area under the curve. Any result which falls in that region is sufficient evidence to reject the null hypothesis.

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").

Finding the critical value works exactly the same as finding the z-score corresponding to any area under the curve like we did in Unit 1. If we go to the normal table, we will find that the z-score corresponding to 5% of the area under the curve is equal to 1.645 (z = 1.64 corresponds to 0.0405 and z = 1.65 corresponds to 0.0495, so .05 is exactly in between them) if we go to the right and -1.645 if we go to the left. The direction must be determined by your alternative hypothesis, and drawing then shading the distribution is helpful for keeping directionality straight.

Suppose, however, that we want to do a non-directional 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 (for reasons we will see later). 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.

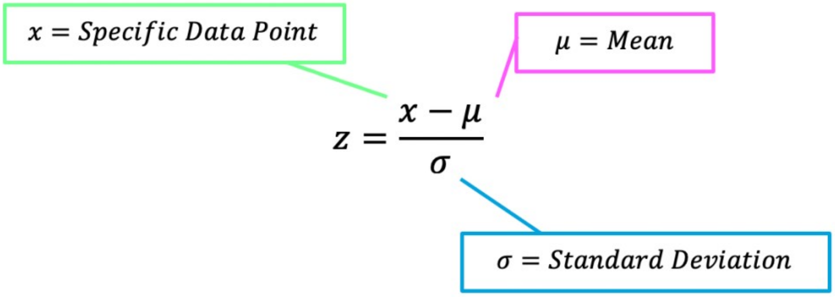
For α = .05, this means 2.5% of the area is in each tail, which, based on the z-table, corresponds to critical values of z_{crit} ("z-critical") = ±1.96.



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 :-


$x = Specific\ Data\ Point$
$\mu = Mean$
$z = \frac{x - \mu}{\sigma}$
$\sigma = Standard\ Deviation$

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 H_0.

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 computer 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. We can directly compare this p-value to α to test our null hypothesis: if $p<\alpha$, we reject Null hypothesis H_0, 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