

Hypothesis Testing

Hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true.

Hypothesis Testing can be broken into 10 basic steps.

- 1.) Data Testing
- 2.) Assumptions
- 3.) Hypothesis
- 4.) Test Statistic
- 5.) Distribution of Test Statistic
- 6.) Decision Rule
- 7.) Calculation of Test Statistic
- 8.) Statistical Decision
- 9.) Conclusion
- 10.) p-values

It is important to note that the initial steps can be done in different orders or at the same time, specifically, I would suggest you come up with your research question first, before going out to do data testing. Moving along:

Let's now break down these steps:

#####Step 1: Data Testing

This one is pretty simple, to do any sort of statistical testing, we'll need some data from a population.

#####Step 2: Assumptions

We will need to make some assumptions regarding our data, such as assuming the data is normally distributed, or what the Standard Deviation of the data is. Another example would be whether to use a T-Distribution or a Normal Distribution.

#####Step 3: Hypothesis

In our Hypothesis Testing, we will have two Hypothesis: *The Null Hypothesis* (denoted as H_0) and the *Alternative Hypothesis* (denoted as H_A). The Null Hypothesis is the hypothesis we are looking to test against the alternative hypothesis. Let's get an example of what this looks like:

Let's assume we have a data set regarding ages of customers to a restaurant. Let's say for this particular data set we want to try to prove that the mean of that sample population is *not* 30. So we set our null hypothesis as:

$$H_0 : \mu = 30$$

And our Alternative Hypothesis as:

$$H_A : \mu \neq 30$$

We could have also set out Alternative Hypothesis to be that the mean is greater than 30 or another Alternative Hypothesis we could have proposed is the mean is less than 30. We will discuss these other options later when we go into one-tail versus two-tail tests.

Continuing on with our steps:

#####Step 4: Test Statistic

Based on our assumptions, we'll choose a suitable test statistic. For example, if we believe to have a normal distribution with our data we would choose a z-score as our standard error:

$$z = \frac{\bar{x} - \mu_o}{\sigma/\sqrt{n}}$$

Or if we assume to have a t-distribution we would choose a t-score (estimated standard error):

$$t = \frac{\bar{x} - \mu_o}{s/\sqrt{n}}$$

Then our test statistic is defined as:

$$\text{Test Statistic} = \frac{\text{Relevant Statistic} - \text{Hypothesized Parameter}}{\text{Standard Error of Relevant Statistic}}$$

#####Step 5: Distribution of Test Statistic

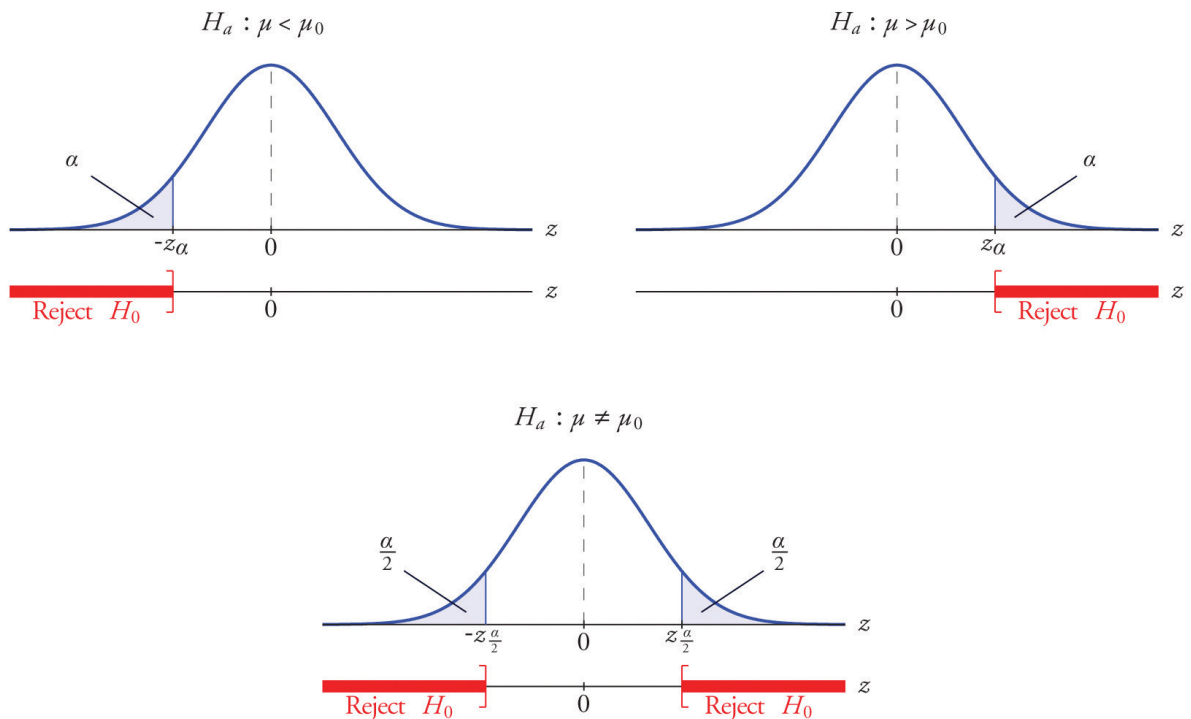
As discussed earlier, verify that whether your data should use a t or z distribution.

#####Step 6: Decision Rule

Considering the distribution, we need to establish a significance level, usually denoted as alpha, α . Alpha is the probability of having a Null Hypothesis that is true, but our data shows is wrong. So alpha is the probability of rejecting a *true* Null Hypothesis. By convention, alpha is usually equal to 0.05 or 5%. This means that 5% of the time, we will falsely reject a true null hypothesis, this is best explained through a picture.

```
In [11]: from IPython.display import Image
url = 'http://images.flatworldknowledge.com/shafer/shafer-fig08_004.jpg'
Image(url,height=400,width=600)
```

Out[11]:



The above image shows three types of tests. It shows 2 types of One-Tail Tests on top, and a Two-Tail Test on the bottom. You apply your significance level to a region depending on your Alternative Hypothesis (shown above each distribution).

Let's look at the bottom two-tail test. With an $\alpha=0.05$, then whenever we take a sample and get our test statistic (the t or z score) then we check where it falls in our distribution. If the Null Hypothesis is True, then 95% of the time it would land in between the $\alpha/2$ markers. So if our test statistic lands in between the alpha markers (for a two tail test in between $\alpha/2$, for a one tail test either below or above the α marker depending on the Alternative Hypothesis) we accept (or don't reject) the Null Hypothesis. If it lands outside of this zone, we reject the Hypothesis.

Now let's say we actually made a mistake, and rejected a Null Hypothesis that unknown to us was True, then we made what is called a Type I error. If we accepted a Null Hypothesis that was actually False, we've made a Type II error. Below is a handy chart depicting the error types.

```
In [13]: url='http://www.personal.ceu.hu/students/08/Olga_Etchevskaia/images/errors.jpg'
         Image(url,height=200,width=300)
```

Out[13]:

| | | Truth about the population | |
|--------------------------|--------------|----------------------------|------------------|
| | | H_0 true | H_a true |
| Decision based on sample | Reject H_0 | Type I error | Correct decision |
| | Accept H_0 | Correct decision | Type II error |

#####Step 7: Calculate the Test Statistic Now that we understand our decision rule we can calculate our t or z score.

#####Step 8: Statistical Decision

We take what we understand from Step 6 and see where our test statistic from Step 7 lies.

#####Step 9: Conclusion

We check our statistical decision and conclude whether or not to reject the Null Hypothesis.

#####Step 10: Calculate a p-value

The P value (<http://en.wikipedia.org/wiki/P-value>) is the estimated probability of rejecting the null hypothesis of a study question when that null hypothesis is actually true. In other words, the P-value may be considered the probability of finding the observed, or more extreme, results when the null hypothesis is true – the definition of ‘extreme’ depends on how the hypothesis is being tested.

It is very important to note: Since the p-value is used in Frequentist inference (and not Bayesian inference), it does not in itself support reasoning about the probabilities of hypotheses, but only as a tool for deciding whether to reject the null hypothesis in favor of the alternative hypothesis.

Let's now go over an example to go through all the steps!

Example Problem 1

There are two main types of single-sample hypothesis tests:

- 1.) When the population standard deviation σ is known or given.
- 2.) When the population standard deviation σ is NOT known, and therefore we have to estimate with s , the sample standard deviation.

For Case 1, we use a normal (Z) distribution, for Case 2, we use the T-distribution.

Let's start with an example. We are hired by a fast food chain to figure out the mean age of customers. The current CEO believes the mean age is 30, however we don't believe that to be True. So we now have our Null and Alternative Hypotheses:

$$H_o : \mu = 30$$

$$H_A : \mu \neq 30$$

It is important to remember that the Null and Alternative Hypotheses must account for all possible outcomes.

So now we will see if we can reject the Null Hypothesis in order to accept our Alternative Hypothesis.

Step 1: Data

So we take a random sample of customers and come up with the following data. From our *sample population*, we get a mean age of 27 years from a sample of 10 customers. Note: This is a sample population of the total population (referred to as just the population).

Step 2: Assumptions

We will begin and get some assumptions. Based on some textbook sources you check, you assume that the population follows a normal distribution (Note: There are specific tests (http://en.wikipedia.org/wiki/Normality_test) to check this). We also see in this textbook that given the population we could assume a variance of 20, so a standard deviation of:

$$\sigma = \sqrt{20}$$

It's important to note that in a real world case it is unlikely to know the population variance, and we will discuss that in the next example. Usually you will only know your sample population's standard deviation (not the *entire* population's standard deviation). Given that we know the standard deviation, we can use a Z Distribution.

Step 3: Hypothesis

We already have our Hypotheses from above:

$$H_o : \mu = 30$$

$$H_A : \mu \neq 30$$

Step 4: Test Statistic

Since we have the standard deviation, we can use the Z-Statistic:

$$z = \frac{\bar{x} - \mu_o}{\sigma/\sqrt{n}}$$

Step 5: Distribution

Again we have chosen the Z-Distribution, having a mean and max at zero.

Step 6: Decision Rule

We will set our significance level (alpha) at 5%. Meaning we are trying to minimize our Type I error. So we are trying to make sure we can reject the Null Hypothesis and minimize the chance that it was actually True. By choosing 5% we will have a certainty of 95% that we are correct if we reject the Null Hypothesis.

Step 7: Calculation of Test Statistic

Let's go ahead and calculate our Test Statistic:

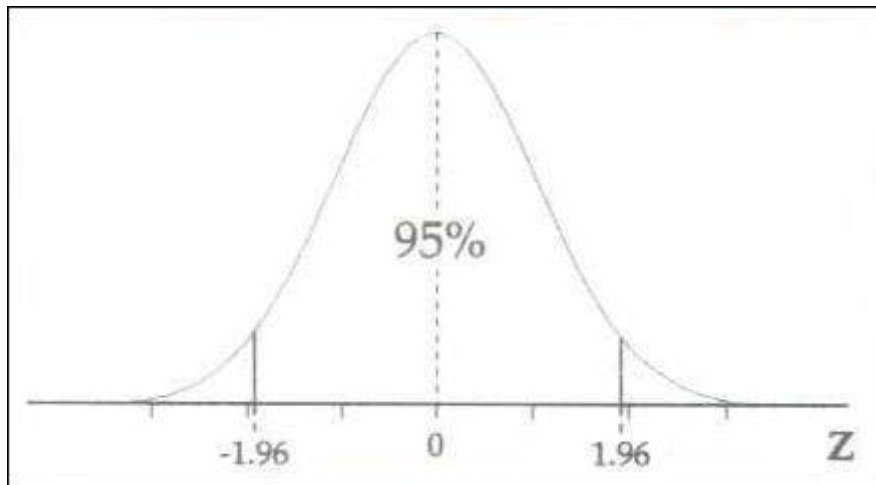
$$z = \frac{\bar{x} - \mu_o}{\sigma/\sqrt{n}} = \frac{27 - 30}{\sqrt{20}/\sqrt{10}} = -2.12$$

Step 8: Conclusion

In order to accept or reject the Null Hypothesis we need to see where our Z score lands. If we take alpha to be 95% then we know from a Z-Score Table (<http://www.stat.ufl.edu/~athienit/Tables/Ztable.pdf>) that we have to lie above 1.96 or below -1.96 as seen in the figure below:

```
In [14]: url='http://www.duncanwil.co.uk/norm_files/image056.jpg'  
Image(url)
```

Out[14]:



Given that our Z-score for our Test Statistic was -2.12, we can see that we can reject the Null Hypothesis! So we reject that the mean of the customer age is 30. Remember, this test does not actually tell us what the mean is, only that it is unlikely to be 30. How unlikely? This is where the p-value comes into play.

Step 9: p-value

From a table, we can see that 98.3% corresponds to the percentage of the area in between our +/- of our Z-Score, 2.12. This means that if the reality is the Null Hypothesis is actually True, we would only find that data 100-98.3= 1.7% of the time. So our p-value is 1.7% (or 0.017). This is pretty unlikely so we can be quite confident in our rejection of the Null Hypothesis.

Confidence Intervals

We could describe the above process using confidence intervals. The equation for a confidence interval is:

$$\bar{x} \pm z\sigma_{\bar{x}}$$

From our work above we can fill in the equation as:

$$27 \pm 1.96 * \left(\frac{\sqrt{20}}{\sqrt{10}} \right)$$

Which (with rounding) is equal to:

$$27 \pm 2.77$$

This gives us a range between 24.23 and 29.77. Since the Null Hypothesis mean of 30 is NOT in that range, we can reject that Null Hypothesis. A simple approach to the above problem!

Now what if we don't know the standard deviation? Let's go ahead and look at an example problem:

Example Problem 2

Assuming the same problem as before, we want to know whether or not we can reject the Null Hypothesis.

Step 1: Data

Let's say after getting our data, we have a mean of 27 from 10 customers. Our *sample population* has a standard deviation of 3.

Step 2: Assumptions

In this second case we don't consult a textbook so we don't know what the entire population's standard deviation is.

Step 3: Hypotheses

Same as before:

$$H_o : \mu = 30$$

$$H_A : \mu \neq 30$$

Step 4: Test Statistic

We now have to use the t-distribution since we don't know the population's standard deviation. So our test statistic will be:

$$t = \frac{\bar{x} - \mu_o}{s/\sqrt{n}}$$

Step 5: Distribution

Since we only have 10 customers in our sample and don't know the entire population's standard deviation we use a t-distribution. Remember as a sample population grows beyond 30, the t distribution begins to resemble and behave like a standard Z distribution. Keep in mind a t-distribution is "flatter" and operates on degrees of freedom (n-1).

We can use a T-table (<http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf>) to find that for a 95% significance level, our region for accepting the Null Hypothesis falls between +/- 2.262

Step 6: Decision Rule

As before: We will set our significance level (alpha) at 5%. Meaning we are trying to minimize our Type I error. So we are trying to make sure we can reject the Null Hypothesis and minimize the chance that it was actually True. By choosing 5% we will have a certainty of 95% that we are correct if we reject the Null Hypothesis.

Step 7: Calculate our Test Statistic

Plugging into:

$$t = \frac{\bar{x} - \mu_o}{s/\sqrt{n}}$$

we get:

$$t = \frac{27 - 30}{3/\sqrt{10}} = -3.16$$

Step 8: Statistical Decision

Given the rejection regions was $t = \pm 2.262$, then our $t = -3.16$ allows us to reject the Null Hypothesis.

Step 9: Conclusion

Just as before we can reject the Null Hypothesis.

Additional Resources

Here are some more great resources for learning about Hypothesis Testing:

1.) Wolfram (<http://mathworld.wolfram.com/HypothesisTesting.html>) 2.) Stat Trek (<http://stattrek.com/hypothesis-test/hypothesis-testing.aspx>) 3.) Wikipedia (http://en.wikipedia.org/wiki/Statistical_hypothesis_testing) 4.) Probability Book Sample (http://www.sagepub.com/upm-data/40007_Chapter8.pdf)