Chapter 8 Notes and elaborations for Math 1125-Introductory Statistics

Assignment:

First read my additional notes below. It will tell you when to go read section 8.1. Then do the 8.1 exercises.

8.1: 1 - 13 odd.
8.2: 1 - 9 odd. 15, 17, 21, 23.
8.3: 1, 3, 7, 13, 17, 19.
8.4: 3- 13 odd. (15 -19 odd are good, but overkill for most of you.)

There will be a quiz on the material in Chapter 8.

Chapter 8 is all about hypothesis testing. Before we begin discussing a few specific types of hypothesis tests in detail, I'd like to say a few words about datasnooping. It is embarrassingly easy to claim anything with data once the data is collected. It is one of the primary reasons why the general public is skeptical of statistics. Valid hypothesis testing is done with a procedure in place before any data is collected, detailing virtually every step of the process.

Always, always, always: decide how to do an experiment then collect data. Anything else is called a pilot or exploratory study.

8.0 Hypothesis Testing

As a procedure, hypothesis testing is not very hard. It will involve looking up a number in table, computing a test statistic (subtract and divide, usually - your book calls this the test value), comparing the numbers, and stating a conclusion to reject or not to reject.

Understanding why isn't too hard either. But there are a lot of terms that are easy to get mired in. Expect this to take some time to sink in.

In a nutshell, hypothesis testing is statistical decision making. Actually, it is the only way to make scientific decisions. It is done quite like a criminal trial is done. There are some fundamental ideas behind it that are hard to understand in a vacuum, so I will start this out with an example to ponder.

Example 8.0.0

We have to decide if a coin if fair. We are allowed to flip it 8 times. How will we do it? (Stop reading and think about that for a minute. What would you do?)

Here's how a statistician would tackle it.

First, I'll define a couple of symbols. Let Y be the count of the number of heads in 8 flips. Let p be the true probability of getting a heads that nobody really knows. (If we knew the value of p then we wouldn't be testing if the coin was fair because we'd already know the answer!)

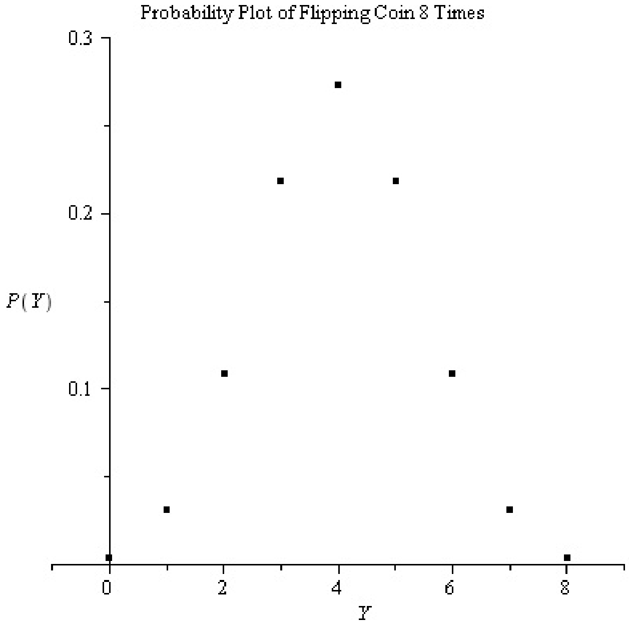
The question “is the coin fair” can be rephrased in math terms: Is $p = 0.5$ or is it the case that $p \neq 0.5$?

In hypothesis testing, there will be two opposing hypotheses about a parameter. I mentioned that this is like a court case. We assume the one with the equals sign in it (in this example it is “p=0.5") is true and see how well that assumption holds up to reality.

If the coin is fair then the distribution of Y is that of a binomial random variable with $p = 0.5$ and $n = 8$. Here’s the distribution table, again, assuming that the coin is fair.

Y	0	1	2	3	4	5	6	7	8
P(*Y*)	0.00391	0.03125	0.10938	0.21875	0.27344	0.21875	0.10938	0.03125	0.00391

Here’s a picture of this distribution:



A reasonable way to think about the problem is like this: let’s assume the coin is fair, collect any relevant data, and decide if the data shows otherwise. This is sort of how a criminal trial operates. Indeed, this is the way hypothesis tests are performed.

Look at the distribution table for 8 flips of a fair coin above. Most of the time a fair coin is flipped 8 times we will see between 1.5 and 6.5 heads. So, my procedure for testing the coin will be like this:

> Flip it 8 times, count the number of heads. Call this number Y. If Y is between 1.5 and 6.5, I’d say “we don’t have any evidence that the coin is not fair.” If it isn’t between 1.5 and 6.5, I’d say that “the data suggests that the coin is unfair.”

I just created a hypothesis test.

All of the hypothesis tests we will perform behave very similar to this coin example, so we will use this example to try and illustrate all of the new terms in this chapter. So, to that end, recall that we had to test if the coin was fair. In the very beginning of the test, we rephrased the question “is the coin fair” into a binary choice: is $p = 0.5$ or is it the case that $p \neq 0.5$? Both of the possibilities have names: the null hypothesis and the alternative hypothesis.

Definition 8.0.0 The **null hypothesis** is denoted as H_0 and always contains a statement about a parameter with equality, i.e., the null hypothesis will always be a statement that some parameter of interest is equal to some conjectured value; we assume the null hypothesis is true while performing the test.

Definition 8.0.1 The **alternative hypothesis** is denoted as H_A (or H_1), and is a statement that opposes the null hypothesis in that it is a statement that some parameter of interest is not equal to (or greater than or less than) some conjectured value.

In the above example, I would have written this:

Let *p* be the true probability of getting a heads in a single trial.

$$H_0: p = 0.5$$
$$H_A: p \neq 0.5$$

We must always state the null and alternative hypotheses before performing the test. We will ultimately conclude one of two things: to reject the null hypothesis or to not reject the null hypothesis.

So, how can we make a mistake? Well, there are two major mistakes that can happen. I could have a really truly fair coin and by chance I get 7 heads. It happens 3.125% of the time from the table in Example 8.0.0. According to my hypothesis test, I would say "the data suggests that the coin is unfair," but my coin is actually fair! This is called a "type I" error.

Definition 8.0.2 A **type I error** occurs when the null hypothesis is true but we decide after performing the test to reject the null hypothesis.

What is the probability of making a type I error in the hypothesis test from Example 8.0.0? Well, making a type I error in this case means that the coin is fair, but we get 0, 1, 7, or 8 heads. If we add the respective probabilities, we get a total of 7.032% chance of rolling a 0, 1, 7, or 8 if the coin is fair. Loosely speaking, if I were to repeat my hypothesis test with a fair coin 100 times, we'd expect to make a type I error about 7 times. This number, 7.032% is called the level of significance, and is denoted as α. It is the probability of making a type I error (calculated before collecting data).

Definition 8.0.3 The **level of significance** of a hypothesis test is the probability of making a type I error and is denoted as α, i.e., P(type I error) = α.

I liken a type I error to convicting an innocent defendant.

There is another type of error we could make. We could have a biased coin (maybe the probability of heads is 60% on one flip) and get exactly 6 heads in the experiment. According to the rule (between 1.5 and 6.5) we are forced to state "we don't have any evidence that the coin is not fair," but the coin is actually not fair! This is called a "type II" error.

Definition 8.0.4 A **type II error** occurs when the null hypothesis is not true but we decide after performing the test to not reject the null hypothesis.

Calculating the probability of this type of error, depends on knowing the actual value of *p*. Notice that we don't know the value of *p* (this is why we're performing the test!). There are some pretty cool things we can do with this mathematically, but it's all way over the level of an introductory stats class. Things you'll need to understand about it go something like this: we call the probability of making a type II error beta. That is, β = P(type II error). I liken this error to letting a guilty defendant back on the streets. (Or these days, back onto Wall St. :~))

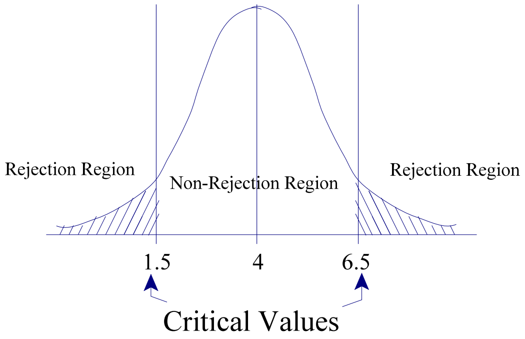
Following is an example of how diagrams are used to illustrate the two possible errors. It is important to notice that the diagram can change depending on the set up. If you do not understand the set up or how to interpret such diagrams, get help immediately.

Example 8.0.1

The smiley face indicates a correct decision.

	Reality	
Decision	null is true	null is false
reject null	type I	☺
do not reject null	☺	type II

The numbers we chose as the boundaries, 1.5 and 6.5, are called **critical values**. The space between them is the **non-rejection region**. All the other places, that is, smaller than 1.5 and larger than 6.5, is called the **rejection region**. The critical values separate the rejection region from the non-rejection region. Consider the following picture:



There is no free lunch. If you want a smaller α, you can have it. Move the critical numbers out (the area in the tails is α). But when you do that β will get bigger! Similarly, if you make the rejection region larger β will go down, but α will go up. There is a way to make them both get smaller. Flip the coin more times. In this case it's pretty easy. In the real world, you might be testing a cancer therapy. Bigger samples means more money and time, and if the therapy works, more lives could be lost before it gets deployed.

The area in the rejection region (the shaded area) is equal to α, and so the area in the non-rejection region is equal to 1 - α. Here's the basic idea of what we will be doing: We will set up a test, take a sample/run an experiment, use data from the sample to compute a test statistic (more on this next - your book calls it a test value), and if the test statistic falls in the rejection region, we reject the null hypothesis. If the test statistic falls between the critical values, i.e., in the non-rejection region, we will not reject the null hypothesis. Here's the application of this process to Example 8.0.0:

Step 1	Set up the test:	H_0: $p = 0.5$ H_A: $p \neq 0.5$
Step 2	Run the experiment :	flip the coin 8 times
Step 3	Compute our test statistic:	count the number of heads
Step 4	Compare the test statistic to cv's:	is test statistic in rejection region or not
Step 5	State the appropriate conclusion:	either the coin is fair or the coin is not fair

This is just a simple example of a hypothesis test. We will now elaborate on more (simple) tests, and the rest of this chapter will be devoted to tests for the means and tests for proportions. Note that the book mentions the power of a test. Technically, the power of a test is 1 - β. Recall β is the probability of a type II error. In both theory and practice, finding this probability is exceptionally difficult. But, if you follow the guidelines laid out in the rest of this section (and others to come), you will always be using the uniformly most powerful test (UMP). That is, the tests described below are the best tests for the respective situations.

Here's an extra side note for those of you who are a bit bored: the α here is the same α as in the construction of confidence intervals.

8.1 and 8.2 of the book will consider hypothesis tests that start with a normal distribution. My example here is binomial. I think here is a good time to start reading chapter 8. Check back when you're done with 8.1.

8.1 One Sample Hypothesis Tests for the Mean

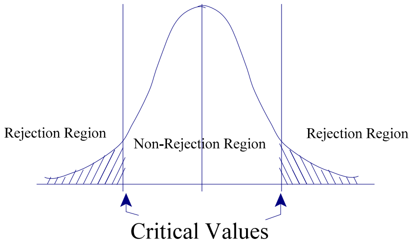
Just as with the construction of confidence intervals, we will have different types of tests depending on the parent distribution, sample size, etc. Before we investigate a few different cases, let's expound in general on hypothesis tests for the mean.

In these types of hypothesis tests, we compare some hypothesized mean (I call it μ_0, you're book calls this *k*) to the true population mean (μ).

Tests for the mean, μ, against some conjectured value μ_0 fall into one of the following three categories: 1) two-tailed test, 2) left-tailed test, and 3) right-tailed test.

1) two-tailed

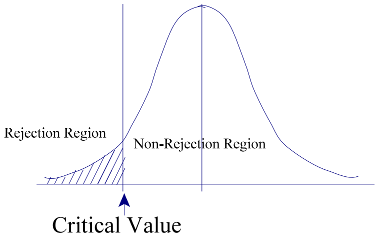
H_0: $\mu = \mu_0$
H_A: $\mu \neq \mu_0$



This is a two-tailed test. We assume that some particular value μ_0 is the same as the population mean (the null) and then hypothesize that μ_0 is *not equal to* μ (the alternative), state the significance level we wish to achieve (this is α), and then we calculate our test statistic (based on sample data). If our test statistic falls in either rejection region, we reject the null hypothesis. That is, if our test statistic is smaller than the lesser critical value or larger than the greater critical value, we reject. If the test statistic falls between the two critical values, we do not reject the null.

2) left-tailed

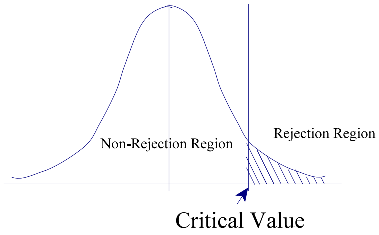
H_0: $\mu = \mu_0$
H_A: $\mu < \mu_0$



This is a one-tailed test. In particular, this is a left-tailed test. We hypothesize that some particular value μ_0 is *greater than* μ, state the significance level we wish to achieve, and then we calculate our test statistic. If our test statistic falls in the rejection region, we reject the null hypothesis. That is, if our test statistic is smaller than the critical value, we reject. If the test statistic is greater than the critical value, we do not reject the null.

3) right-tailed

$H_0: \mu = \mu_0$
$H_A: \mu > \mu_0$



This is also one-tailed test. We hypothesize that some particular value μ_0 is *less than* μ, state the significance level we wish to achieve, and then we calculate our test statistic. If our test statistic falls in the rejection region, we reject the null hypothesis. That is, if our test statistic is greater than the critical value, we reject. If the test statistic is less than the critical value, we do not reject the null.

Recall that the first step in performing a hypothesis test is to set up the test. So, figuring out which category your test falls into is the first step. Note that you should always, always draw a corresponding picture.

There are two types of tests for the mean that we will be performing, the *z*-test and the *t*-test. Have your *z*-table and *t*-table handy. In fact, don't put them away for the rest of the course. As with confidence intervals, the following four cases are considered:

Case 1. Our sample is from a normal population, and it's variance is *known*.
Case 2. Our sample is from a normal distribution, it's variance is *unknown*, and $n \geq 30$.
Case 3. Our sample is from a "well-enough behaved" distribution and $n \geq 30$.
Case 4. Our sample is from a normal population, it's variance is *unknown*, and $n < 30$.

Again, we will, in general, assume that when the sample size is larger than 30, that the sample average is distributed like a normal random variable via the CLT. It is simply for convenience. In many situations it isn't a safe assumption, and in some others 30 is complete overkill. We assume this for the sake of learning the associated concepts.

Before going into each case with examples, let us give the formulas for two different test statistics that we will be using to test for means and discuss them a bit in the context of testing for means. Here they are:

Formula 8.1.0

$$T = \frac{\bar{X} - \mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)},$$

and

Formula 8.1.1

$$T = \frac{\bar{X} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)},$$

where $\bar{X}$ is the sample mean, μ_0 is the assumed mean in the null hypothesis, σ is the standard deviation for the population, s is the sample standard deviation, and n is the sample size. Let's talk for a second about these formulas.

Recall the CLT. It tells us that the average of a bunch of samples is essentially distributed normal with mean μ and standard deviation $\sigma/\sqrt{n}$. Under the null hypothesis, we are assuming that $\mu = \mu_0$. So, via this assumption, our sample average is distributed $N(\mu_{0,} \sigma/\sqrt{n})$. These formulas are just *z*-scores of the sample average, where the first is with the population standard deviation and the second is with the sample standard deviation (remember how to standardize: take a value, subtract the mean, divide by the standard deviation).

In order to use a *z*-test, i.e., in order to use the *z*-table, the underlying test statistic, T (note: using T to denote the test statistic is standard), must be a standard normal random variable under the assumptions of the null hypothesis. So, when is T normally distributed?

The test statistic given in Formula 8.1.0 is normal only when the sample comes from a normal distribution to begin with. However, if the sample is "large enough" and it comes from a distribution with a finite variance, i.e., the distribution is "well-enough behaved", then we may assume it is approximately normal from the central limit theorem. The test statistic given in Formula 8.1.1 isn't normal even when the sample comes from a normal distribution because there is a random variable in the denominator (this situation gets messy pretty quickly)! However, with a large enough sample it becomes approximately normal as well.

We always prefer the test statistic in Formula 8.1.0, although it's usually the case that we must use the other. Let's now explicitly review each case. Notice we now add a Step 0 to the process we outlined earlier; step 0 is to decide the appropriate test.

Case 1. Our sample is from a normal population, and it's variance is *known*.

In this case, we know σ and the test statistic given in Formula 8.1.0 is distributed $N(0, 1)$ (because our sample comes from a normal population), so we use a *z*-test and Formula 8.1.0. Here's an example of how it goes.

Example 8.1.0

A saw at a sawmill is set to cut boards into 4.25" strips. It is known that the width of the strips is normally distributed and the variance is given by $\sigma^2 = 0.0025$. However, as the blade wears over time, it is suspected that the mean width of the boards changes. The foreman decides to test this hypothesis one week after installing a new blade. So, he takes a sample of 22 boards and finds an average width of 4.23". Are the suspicions correct? Does the mean width of the boards change as the blade wears? Test this with level of significance of a) 0.10 and b) 0.05.

a) **Step 0**: Decide the appropriate test.

Our sample is from a normal and we know the population variance. So, we use a *z*-test with Formula 8.1.0.

Step 1: State the hypotheses and significance level.

Let μ be the true mean width of a board.

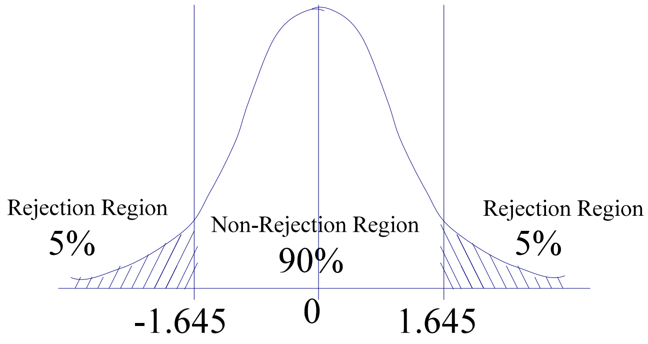
$\alpha = 10\%$

H_0: $\mu = 4.25"$

H_A: $\mu \neq 4.25"$

Step 2: Find the critical value(s).

This is a two-sided *z*-test, so it will have two critical numbers From the *z*-table we see that the critical numbers are $z = \pm 1.645$. The rejection region will contain any number smaller than -1.645 and any number bigger than 1.645. Here's a picture of the situation:

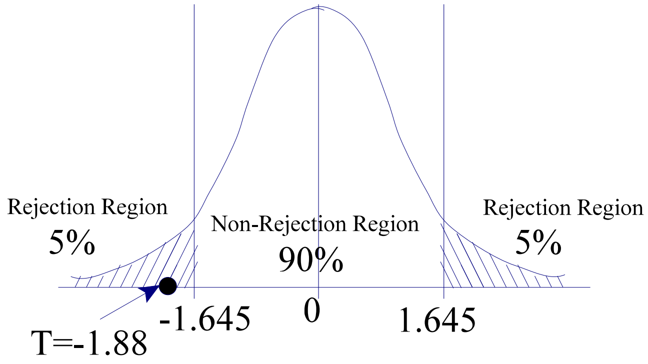


Step 3: Calculate the test statistic.

We use $T = \frac{\overline{X} - \mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)}$. Let's identify all the pieces we need:

$\overline{X} = 4.23''$, $\mu_0 = 4.25''$, $\sigma = 0.05$, and $n = 22$.

Computing the numerator we have -0.02". The denominator is about 0.0107". Thus, T = -1.88 rounded to the nearest thousandth.



Step 4: Compare the test statistic to the critical value and make a decision (to reject or to not reject).

We see that T is in the rejection region. We reject H_0.

Step 5: State the conclusion.

We have evidence suggesting that the mean width of a board is no longer 4.25".

b) **Step 0**: Decide the appropriate test.

Our sample is from a normal and we know the population variance. So, we use a *z*-test with Formula 8.1.0.

Step 1: State the hypotheses and significance level.

Let μ be the true mean width of a board.

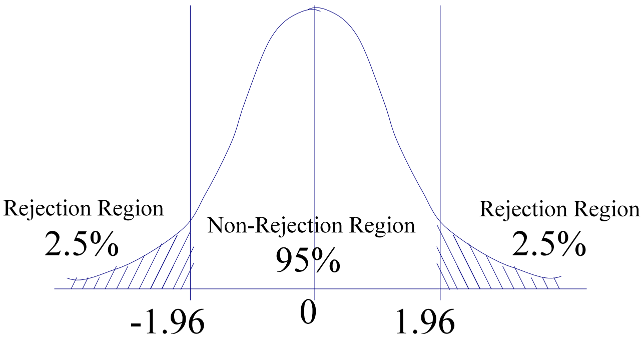
$\alpha = 0.05$

H_0: $\mu = 4.25''$

H_A: $\mu \neq 4.25''$

Step 2: Find the critical value(s).

This is a two-sided *z*-test, so it will have two critical numbers From the *z*-table we see that the critical numbers are $z = \pm 1.96$. The rejection region will contain any number smaller than -1.96 and any number bigger than 1.96. Here's a picture of the situation:

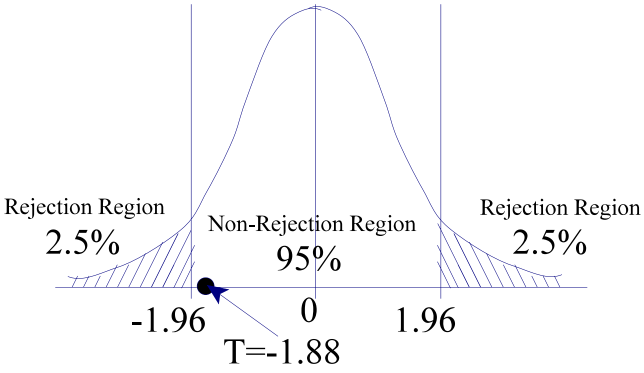


Step 3: Calculate the test statistic.

We use $T = \dfrac{\overline{X} - \mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)}$. Let's identify all the pieces we need:

$\overline{X} = 4.23''$, $\mu_0 = 4.25''$, $\sigma = 0.05$, and $n = 22$.

Computing the numerator we have -0.02". The denominator is about 0.0107". Thus, T = -1.88 rounded to the nearest thousandth.



Step 4: Compare the test statistic to the critical value and make a decision (to reject or to not reject).

We see that T is not in the rejection region. We do not reject H_0.

Step 5: State the conclusion.

We have evidence suggesting that the mean width of a board is still 4.25".

Notice that the level of significance matters. We rejected the null hypothesis at $\alpha = 0.10$ and we did not reject the null at $\alpha = 0.05$. Which level of significance we would use in such a case would really depend on a cost-benefit analysis, e.g., how much waste is tolerable, cost of new blades, downtime of machines while replacing new blades, etc.

Case 2. Our sample is from a normal distribution, it's variance is *unknown*, and $n \geq 30$.

In this case, we don't know σ, but our sample size is large enough and so the test statistic given in Formula 8.1.1 is approximately distributed $N(0, 1)$ (because our sample comes from a normal population), so we use a *z*-test and Formula 8.1.1. Here's an example of how it goes.

Example 8.1.1

A student suspects the cost of a date is no longer \$20.00. Assume that the average cost of a date is normally distributed. They take a random sample of 160 people from their dormitory and found an average cost of \$21.17 with a sample standard deviation of \$5.51. Test at a) $\alpha = 0.05$ and b) $\alpha = 0.01$ to check their claim.

a) **Step 0**: Decide the appropriate test.

Our sample is from a normal distribution, we don't know the population variance, but the sample size is large enough. So, we use a *z*-test with Formula 8.1.1.

Step 1: State the hypotheses and significance level.

Let μ be the true mean cost of a date.

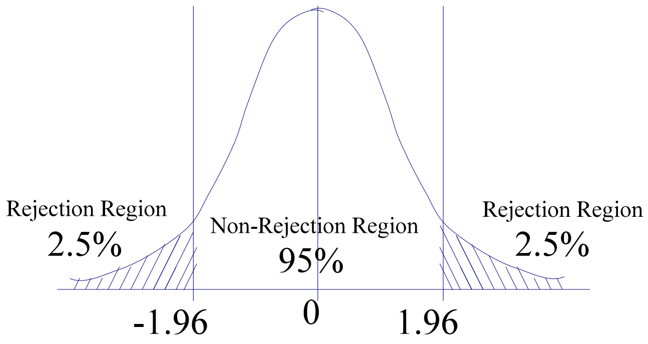
$\alpha = 5\%$

H_0: $\mu = \$20$

H_A: $\mu \neq \$20$

Step 2: Find the critical value(s).

This is a two-sided *z*-test, so it will have two critical numbers From the *z*-table we see that the critical numbers are $z = \pm 1.96$. The rejection region will contain any number smaller than -1.96 and any number bigger than 1.96.

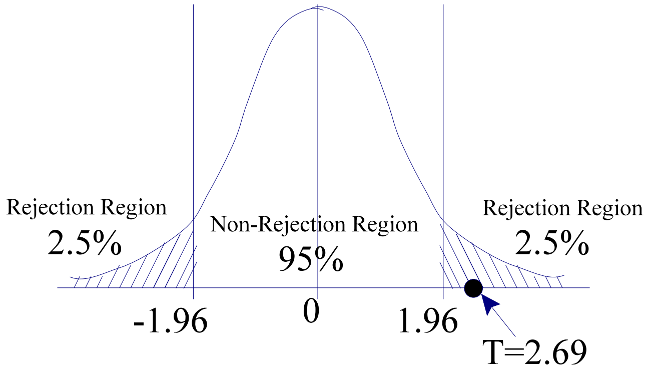


Step 3: Calculate the test statistic.

We use T = $\frac{\bar{X} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$. Let's identify all the pieces we need:

$\bar{X}$ = \$21.17, μ_0 = \$20, s = \$5.51, and n = 160.

Computing the numerator we have \$1.17. The denominator is about \$0.44. Thus, T = 2.69 rounded to the nearest thousandth.



Step 4: Compare the test statistic to the critical value and make a decision (to reject or to not reject).

We see that T is in the rejection region. We reject H_0.

Step 5: State the conclusion.

We have evidence suggesting that the cost of a date is no longer \$20.00.

b) **Step 0**: Decide the appropriate test.

Our sample is from a normal distribution, we don't know the population variance, but the sample size is large enough. So, we use a *z*-test with Formula 8.1.1.

Step 1: State the hypotheses and significance level.

Let μ be the true mean cost of a date.

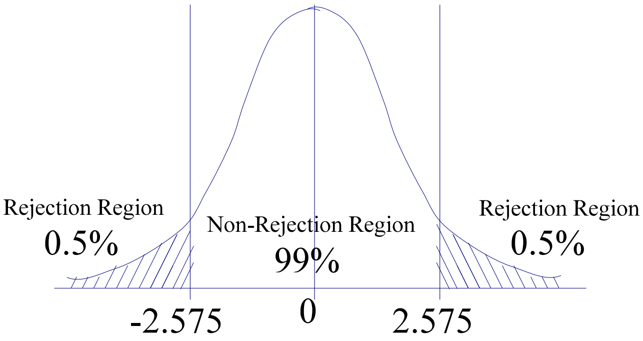
$\alpha = 1\%$

H_0: μ = \$20

H_A: μ ≠ \$20

Step 2: Find the critical value(s).

This is a two-sided *z*-test, so it will have two critical numbers From the *z*-table we see that the critical numbers are $z = \pm 2.575$. The rejection region will contain any number smaller than -2.575 and any number bigger than 2.575.

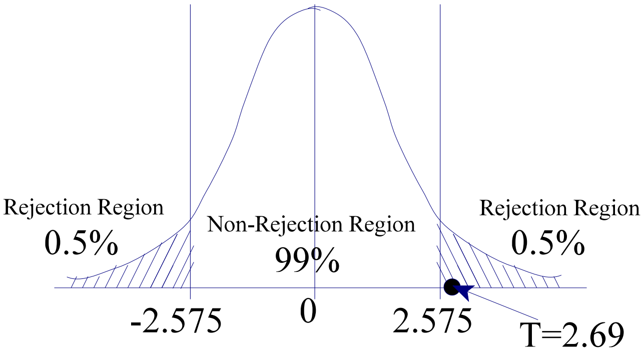


Step 3: Calculate the test statistic.

We use $T = \dfrac{\bar{X} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$. Let's identify all the pieces we need:

$\bar{X}$ = \$21.17, μ_0 = \$20, s = \$5.51, and n = 160.

Computing the numerator we have \$1.17. The denominator is about \$0.44. Thus, T = 2.69 rounded to the nearest thousandth.



Step 4: Compare the test statistic to the critical value and make a decision (to reject or to not reject).

We see that T is still in the rejection region. We reject H_0.

Step 5: State the conclusion.

We have evidence suggesting that the cost of a date is no longer $20.00.

In this example, the change in significance level did not change the conclusion. What do you think the conclusion would be if we made the significance level even smaller than 1%? You're correct if you think we'd reject.

Case 3. Our sample is from a "well-enough behaved" distribution and $n \geq 30$.

In this case, our parent distribution is nice enough and our sample is large enough, so we use a *z*-test with Formula 8.1.0 if we know σ and we use a *z*-test with Formula 8.1.1 if we don't know σ.

Example 8.1.2

An irate consumer group decides to study the claim that Kangaroo Chili will raise your math test score. They wish to test the claim at the 10% level of significance, i.e, $\alpha = 0.1$. It is known that the mean score on the YAMS Test (Yet Another Math Standardized test) is 100. A randomly selected group of 49 students is chosen and given some chili right before the exam. The mean score of the group is 103 with a sample standard deviation of 15. What is the result of the hypothesis test?

Step 0: Decide the appropriate test.

We know that math test scores have a nice enough distribution (test scores have finite variance), and since the sample size, 49, is sufficiently large (that is, bigger than 30), we can use the *z*-test (although we might want to check that the histogram of the data set looks reasonably normal first). We don't know the population variance, so we use Formula 8.1.1.

Step 1: State the hypotheses and significance level.

Let μ be the true mean score for people under the influence of Kangaroo Chili on the YAMS.

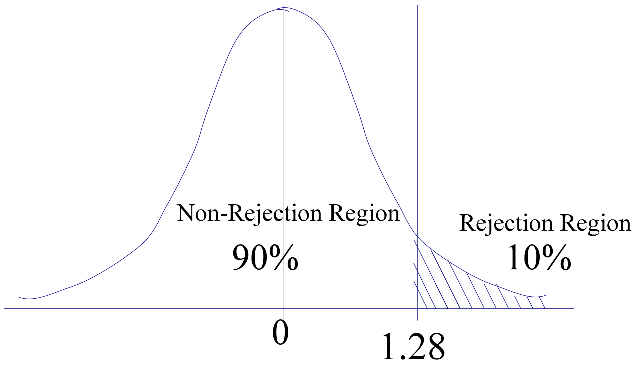
$\alpha = 0.10$

$$H_0: \mu = 100$$
$$H_A: \mu > 100$$

This is because the claim is that the score is greater than before chili. This claim must go in the alternative hypothesis because it doesn't have the equal sign included in the statement.

Step 2: Find the critical value(s).

This is a one-sided test, so it will be one critical value. It is a right-tailed test because the alternative favors larger numbers; it will be the positive *z*-score corresponding to α = 0.10. Here's a picture to illustrate the situation:



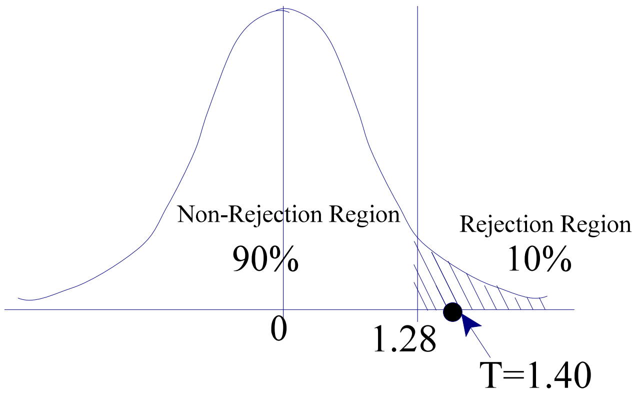
From the *z*-table (Table E) we see that the critical number is $z = 1.28$ (it is the *z*-value corresponding to 90% or 0.90). We will reject the null hypothesis if our test statistic is any number bigger than 1.28.

Step 3: Calculate the test statistic.

We use $T = \frac{\bar{X} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$ since we do not know the population standard deviation. Let's identify all the pieces we need:

$\bar{X} = 103$, $\mu_0 = 100$, $s = 15$, and $n = 49$.

Computing the numerator 103-100 = 3.0. The denominator is (15 / 7) = 2.1429. Thus, T = 1.40, rounded to the nearest hundredth. Now we have this picture:



Step 4: Compare the test statistic to the critical value and make a decision (to reject or to not reject).

We see that T falls in the rejection region. We reject H_0.

Step 5: State the conclusion.

There is evidence that eating Kangaroo Chili right before the YAMS appears to be improving math scores. (Maybe it's just eating? That would require another test.)

If we performed this test with α = 0.05, our critical value would be 1.645 and we would not reject.

Case 4. Our sample is from a normal population, it's variance is *unknown*, and $n < 30$.

This situation requires a *t*-test, which was first developed for testing small samples of beer (Guinness) by William Gosset. The situation is as follows: we are sampling from a distribution that is known to be normal (or really close to normal), we don't know the population standard deviation, and $n < 30$.

Because the sample is small, we can't assume the test statistic is normally distributed. It is actually a student-*t* distribution, or a *t*-distribution with n - 1 degrees of freedom (note that *t*-distributions are **not** normals, they are *t*'s; we assume the population from which we are sampling is normal). Other than the table value lookup, the test is run in an identical fashion as the *z*-test. That is, we use the *t*-table and Formula 8.1.1.

Example 8.1.3

A large study of 30 states yield that the average one-way commute time to work is 29.4 minutes. The mayor of a small town believes that his town has a lower average one-way commute. She takes a random sample of 20 commuters and finds a mean commute time of 26.2 with a standard deviation of 7. With $\alpha = 0.10$, is her assertion plausible?

Step 0: First, we decide that we can't do this hypothesis test with the tools that we have available. Notice the requirements for a small sample test requires the data to be normal or close to normal. This requirement is not met.

Now I'll change the example to assume that the data comes from a normal distribution. As the sample is small and we don't know the population standard deviation, the hypothesis test will be a *t*-test. (Note that this was just to make a point, you can't just add assumptions in reality.)

Step 1: State the hypotheses and significance level.

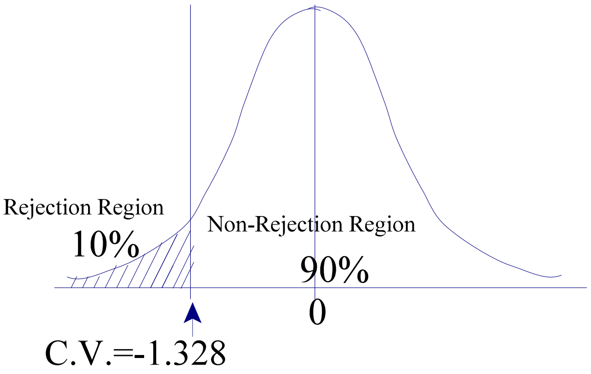
Let μ be the true mean commute time for people of the small town.

$\alpha = 0.10$

H_0: $\mu = 29.4$
H_A: $\mu < 29.4$

Step 2: Find the critical number(s).

This is a one-sided *t*-test, so it will be one critical number, and because the alternative favors smaller numbers, it is left-tailed, i.e., it will be the negative *t*-score. The degrees of freedom is 20-1=19. From the table we see that the critical number is $t = -1.328$. The rejection region will contain any number smaller than -1.328. Here's a picture:



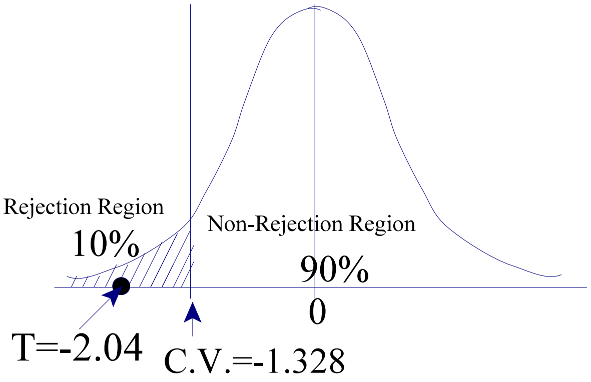
Step 3: Calculate the test statistic.

We use $T = \frac{\bar{X} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$. Let's identify all the pieces we need:

$\bar{X} = 26.2$, $\mu_0 = 29.4$, $s = 7$, and $n = 20$.

Computing the numerator 26.2-29.4 = -3.2. The denominator is about 1.56525. Thus, $T \approx -2.0444$.

Now our picture is:



Step 4: Compare the test statistic to the critical value.

We see that T is in the rejection region. We reject H_0.

Step 5: State the conclusion.

Our evidence suggests that the mayor is quite correct.

If we were to use $\alpha = 0.05$, our critical value would be -1.729, and we would still reject. However, if $\alpha = 0.025$, our critical value would be -2.093, and we would not reject the null. Again, the level of significance matters.

Here's another example.

Example 8.1.4

Let's revisit Example 8.1.1 with a different sample size.

A student suspects the cost of a date is no longer \$20.00. Assume that the average cost of a date is normally distributed. They take a random sample of 16 people from their dormitory and found an average cost of \$21.17 with a sample standard deviation of \$5.51. Test at $\alpha = 0.05$ to check their claim.

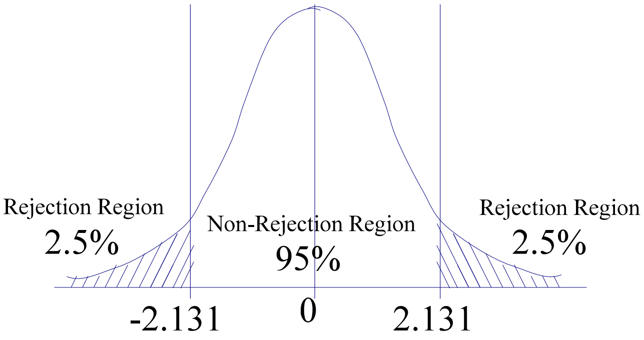
Step 0: Our sample is from a normal distribution, we don't know the population variance, and the sample size is small. So, we use a *t*-test.

Step 1: Let μ be the true mean cost of a date.

$\alpha = 5\%$

H_0: $\mu = \$20$
H_A: $\mu \neq \$20$

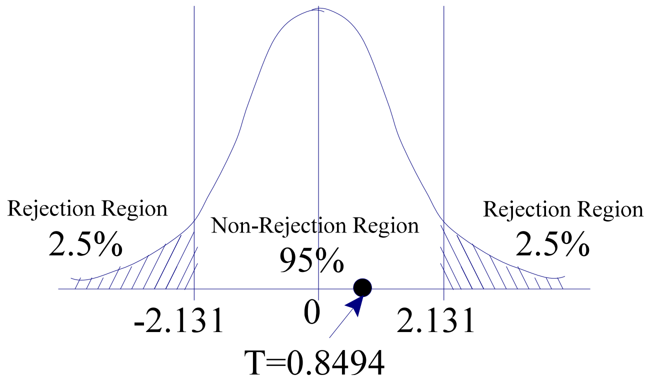
Step 2: This is a two-sided *t*-test, so it will have two critical numbers The degrees of freedom is 15. From the table we see that the critical numbers are t = ±2.131. The rejection region will contain any number smaller than -2.131 and any number bigger than 2.131.



Step 3: We use T = $\dfrac{\bar{X} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$. Let's identify all the pieces we need:

$\bar{X} = \$21.17$, $\mu_0 = \$20$, $s = \$5.51$, and $n = 16$.

Computing the numerator we have \$1.17. The denominator is \$1.3775. Thus, T = 0.8494, rounded to the nearest ten-thousandth.



Step 4: We see that T is in the non-rejection region. We fail to reject the H_0.

Step 5: We have no evidence suggesting that the cost of a date is no longer $20.00.

Notice that here, we reject the null at α = 5%, where we didn't in Example 8.1.1. The sample size dramatically affects the result in this case.

Now we try to summarize in a nice to use format (for quiz/test purposes, etc.). Even if you don't print up all the notes, I'd at least print this summary and keep it handy while taking assessments (hint, hint).

Summary (one sample) - So when do we use which test when testing for means?

***z*-test appropriate situation:**

1. If the sample came from a normal distribution and the population standard deviation is known, use Formula 8.1.0 and the *z*-table.

2. If the sample came from a normal distribution and the sample size is bigger than 29, then we may substitute the sample standard deviation for the population standard deviation if it is unknown (which is usually the case). That is, use Formula 8.1.1 and the *z*-table.

3. If the sample is "large enough" that $\overline{X}$ is close to being normally distributed, i.e., our sample is from a "well-enough behaved" distribution and $n \geq 30$, we can use Formula 8.1.1 and the *z*-table.

***t*-test appropriate situations:**

4. If the sample came from a normal distribution, the population SD is unknown, and the sample size is less than 30, use Formula 8.1.1 with the *t*-table.

Consult a statistician or start doing some research if:

$\overline{X}$ isn't almost normally distributed for your given sample size. If you can't answer that, you probably want to check with someone.

8.2 One Sample Tests for Proportion

We use a test for proportions when individuals in a population can be classified into one of two categories, denoted with 0's and 1's, where 1 indicates ‘success’ (which happens with probability *p*) and 0 indicates ‘failure’ (which happens with probability 1-*p*). Formally, these types of trials are called Bernoulli trials (written Bernoulli(*p*)). Our intent is to make inferences regarding the proportions that belong to each category.

Example 8.2.0

Drug X is given to a sample of 150 patients who have a particular disease. We find that 50% of the subjects in the sample recover from the disease. We want to test whether Drug X is better than Drug Y (at some level of significance), which is known to produce a recovery rate of 45%.

Before going any further here, we would like to comment that, properly done, we test via a proper binomial test in this situation (computers do this for us very easily). We show here the theoretical normal approximation to the binomial. The central limit theorem applies pretty quickly provided $np \geq 5$, and $n(1-p) \geq 5$ (some people prefer to use 10 instead of 5).

So, when testing hypotheses, we compare some hypothesized proportion (p_0) to the population proportion (p). And we use the same types of set-ups and processes as in testing for means, i.e., our test will fall into one of the following three categories.

1)	$H_0: p = p_0$ $H_A: p \neq p_0$	2)	$H_0: p = p_0$ $H_A: p < p_0$	3)	$H_0: p = p_0$ $H_A: p > p_0$

But our test statistic is now

Formula 8.2.0

$$T = \frac{(\hat{p} - p_0)}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

where $\hat{p}$ is the sample proportion, p_0 is the hypothesized proportion, and *n* is the sample size.

We compare the test statistic given in Formula 8.2.0 to appropriate critical values obtained from the *z*-table.

Example 8.2.1

Drug X is given to a sample of 170 patients who have a particular disease. We find that 50% of the subjects in the sample recover from the disease. We want to test whether Drug X produces a higher recovery rate than Drug Y, which is known to produce a recovery rate of 45%. Test this claim at a) $\alpha = 5\%$ and b) $\alpha = 10\%$.

The two categories are recover from the disease (1) and does not recover from the disease (0). For Drug Y, we know that p = 0.45. That is, the proportion of subjects in category (1) is 45%. What we wish to find out is if the proportion of people who took Drug X and recovered, 50%, is statistically significantly greater than the proportion of people who took Drug Y and recovered.

a) **Step 0**: We have n = 170 and p = 0.45, so np = 67.5 and $n(1-p)$ = 82.5, so we can use the normal approximation to a binomial and run a test for proportions.

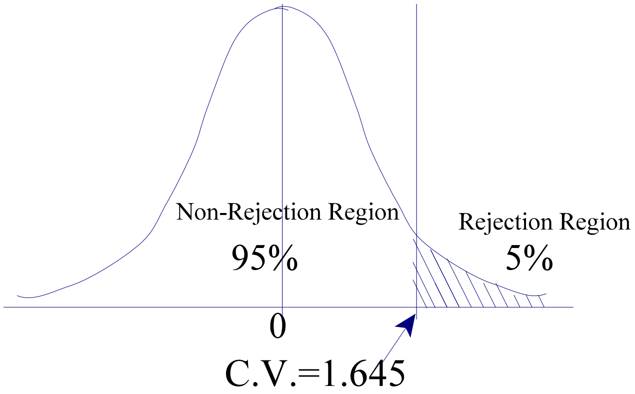
Step 1: Let p be the true proportion of population who recover from Drug X.

$\alpha = 5\%$

H_0: $p = .45$

H_A: $p > .45$

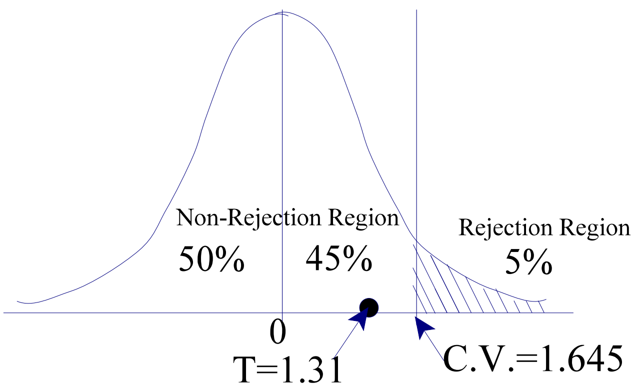
Step 2: This is a one-sided proportion test, so it will have one critical number, 1.645. That is, any test statistic larger than 1.645 will result in a rejection of the null hypothesis. Our picture looks like this:



Step 3: We use $T = \dfrac{(\hat{p} - p_0)}{\sqrt{\frac{p_0(1-p_0)}{n}}}$. Let's identify all the pieces we need:

$\hat{p}$ = 0.50, p_0 = 0.45, and n = 150.

Computing the numerator we have 0.50-0.45=0.05. The denominator is $\sqrt{\frac{.45(1-.45)}{170}}$, which is approximately equal to 0.03816. Thus, T = 0.05/0.03816 ≈ 1.31, rounded to the nearest hundredth. So now our picture is



Step 4: We see that T is in the non-rejection region. We fail to reject H_0.

Step 5: We have no evidence suggesting that Drug X is better than Drug Y.

b) **Step 0**: We have n = 150 and p = 0.45, so np = 67.5 and $n(1-p)$ = 82.5, so we can use the normal approximation to a binomial and run a test for proportions.

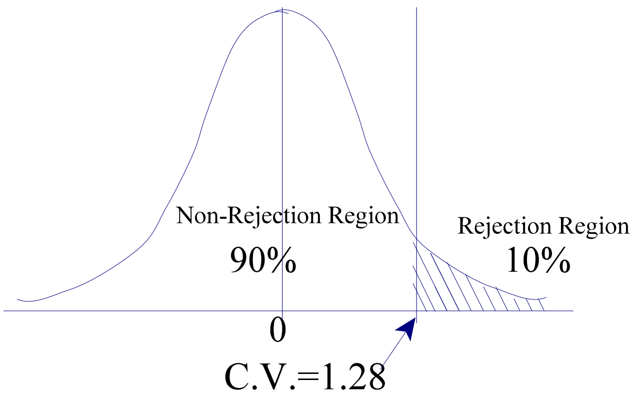
Step 1: Let p be the true proportion of population who recover from Drug X.

$\alpha = 10\%$

H_0: $p = .45$

H_A: $p > .45$

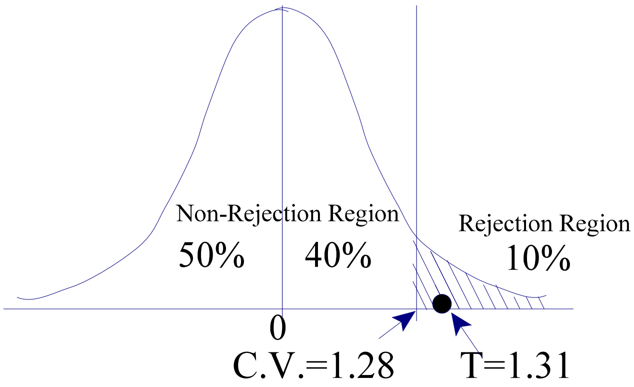
Step 2: This is a one-sided proportion test, so it will have one critical number, 1.28. That is, any test statistic larger than 1.28 will result in a rejection of the null hypothesis. Our picture now looks like this:



Step 3: We use $T = \frac{(\hat{p} - p_0)}{\sqrt{\frac{p_0(1-p_0)}{n}}}$. Let's identify all the pieces we need:

$\hat{p} = 0.50$, $p_0 = 0.45$, and $n = 150$.

Computing the numerator we have 0.50-0.45=0.05. The denominator is $\sqrt{\frac{.45(1-.45)}{170}}$, which is approximately equal to 0.03816. Thus, T = 0.05/0.03816 ≈ 1.31, rounded to the nearest hundredth. So now our picture is



Step 4: We see that T is in the rejection region. We reject H_0.

Step 5: We have evidence suggesting that Drug X is better than Drug Y.

Now recall the motivating coin example at the beginning of these notes.

Example 8.2.2

Someone hands you a coin and tells you to test if it is fair. Recall that in our previous coin example, we were only allowed to flip the coin 8 times. Suppose we beg to flip it 10 times instead, and we are granted this (remember that we need $np \geq 5$, and $n(1\text{-}p) \geq 5$ in order to use the *z*-table). So we flip the coin 10 times and we get 4 heads. Test at a level of significance of 5%.

The two categories are heads (1) and tails (0).

a) **Step 0**: We have $n = 10$ and $p = 0.5$, so $np = 5$ and $n(1\text{-}p) = 5$, so we can use the normal approximation to a binomial and run a test for proportions.

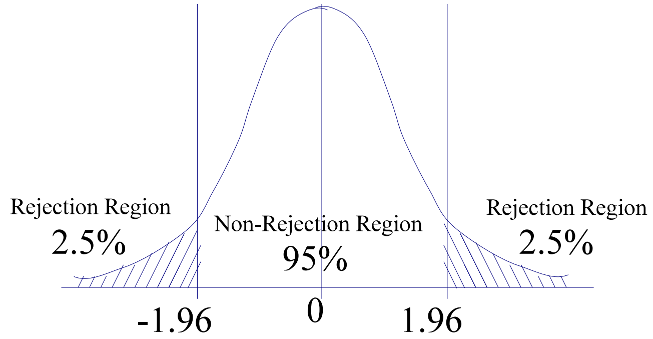
Step 1: Let *p* be the true proportion of heads flipped.

$\alpha = 5\%$

H_0: $p = .5$

H_A: $p \neq .5$

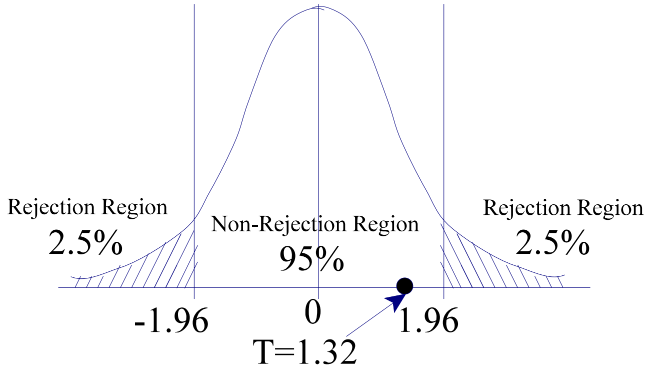
Step 2: This is a two-sided proportion test, so it will have two critical numbers, ±1.96. Our picture looks like this:



Step 3: We use $T = \frac{(\hat{p} - p_0)}{\sqrt{\frac{p_0(1-p_0)}{n}}}$. Let's identify all the pieces we need:

$\hat{p} = 0.50$, $p_0 = 0.4$, and $n = 10$.

Computing the numerator we have 0.50-0.40=0.10. The denominator is $\sqrt{\frac{0.24}{10}}$. Thus, T $\approx$1.32, rounded to the nearest hundredth. So now our picture is



Step 4: We see that T is in the non-rejection region. We fail to reject H_0.

Step 5: We have no evidence suggesting that the coin is not fair.

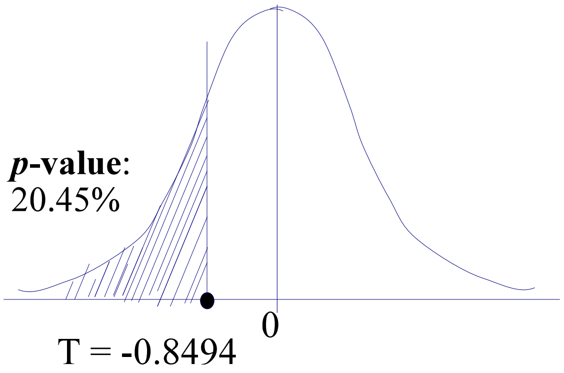
If we had flipped 6 heads, our test statistic would have been -1.31.
If we had flipped 7 heads, our test statistic would have been -3.01.

8.3 Testing Using p-values

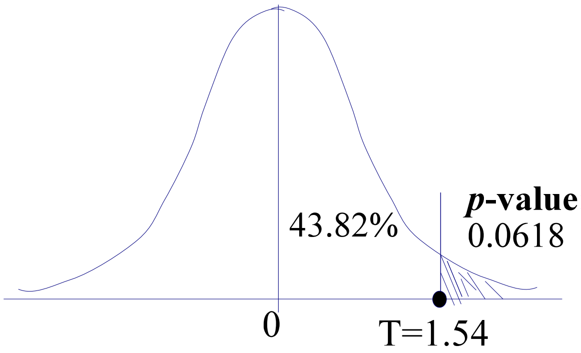
Testing using *p*-values is actually quite common, and not really so different from the testing we've done thus far. In fact, the processes are almost identical. Reporting a *p*-value is just the in thing to do these days.

Instead of choosing a significance level, we report back the level (the value of α) that would "just balance" the test between rejection and a failure to reject. That is, instead of calculating a test statistic and comparing it to critical value(s), we forget about the critical values, calculate the test statistic, find the appropriate corresponding area, and this is what we report. Following are a few of examples of what this can look like.

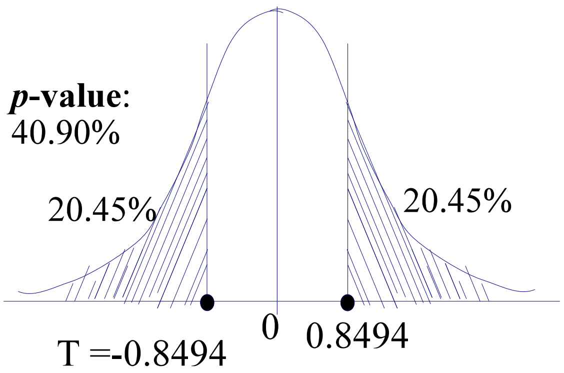
left-tailed *t*-test with test statistic of -0.8494



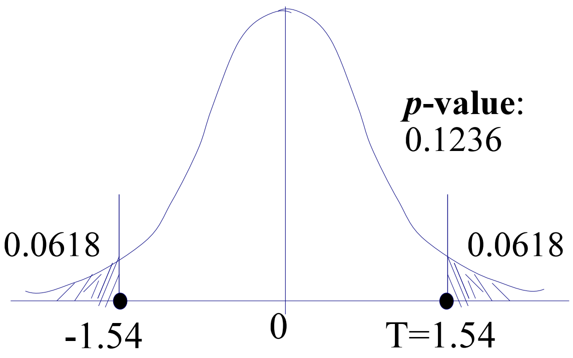
right-tailed *z*-test with test statistic of 1.54



two-tailed t-test with test statistic of -0.8494



two-tailed *z*-test with test statistic of 1.54



Hopefully you see that once we have the *p*-value, we can easily compare this area (probability) to various levels of significance. For any α smaller than our *p*-value, we do not reject the null hypothesis and for any α larger than our *p*-value, we reject the null hypothesis. Do you see this? Think about it.

An interpretation of a *p*-value is as follows: If the null is true and the test is repeated over and over, how often would we see results "as strange" as the result we saw. That is, what is the probability of seeing as extreme a test statistic in a repeated trial, given that the null hypothesis is true. So, larger *p*-values are not that significant, because a large *p*-value means the test statistic is not that strange. The smaller the p-value is, the stranger the test statistic is and the more likely we are to reject the null hypothesis.

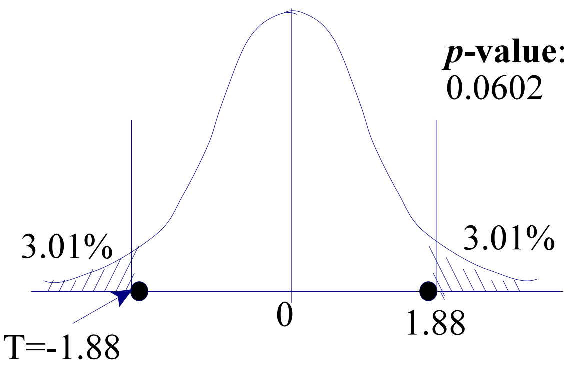
Let's look at a few of the previous examples using *p*-values.

Example 8.1.0
A saw at a sawmill is set to cut boards into 4.25" strips. It is known that the width of the strips is normally distributed and the variance is given by $\sigma^2 = 0.0025$. However, as the blade wears over time, it is suspected that the mean width of the boards changes. The foreman decides to test this hypothesis one week after installing a new blade. So, he takes a sample of 22 boards and finds an average width of 4.23". Are the suspicions correct? Does the mean width of the boards change as the blade wears? Test this with level of significance of a) 0.10 and b) 0.05.

Steps 0-1 are exactly the same. Step 2 can be omitted. Step 3 is exactly the same. At step 4 we can compute the *p*-value.

From Step 3 we have a test statistic of T =-1.88. Exactly how odd is this if the null hypothesis is true? First we look at the alternative hypothesis. It is two sided, so our *p*-value is found by finding the area in the left tail and multiplying by 2. We need the area under the standard normal curve to the left of -1.88. This is now a "simple" table value lookup. I find that the value in the *z*-table for -1.88 is 0.0301. Twice this is the *p*-value. So, the *p*-value here is 6.02%. Here's a picture:

two-tailed *z*-test with test statistic -1.88



What does a *p*-value of 6.02% mean? If we repeat the experiment and the null hypothesis is true, then the probability of getting a test statistic as strange as -1.88 is 6.02%, which is pretty small. Thus, this *p*-value is fairly significant.

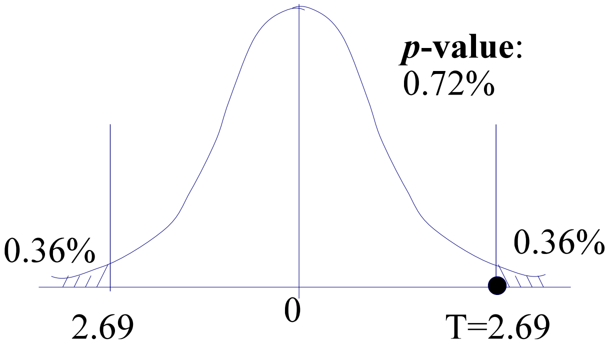
Recall that we did not reject at a 10% level of significance, and we did reject at a 5% level of significance. How does this relate to the *p*-value here? (Think about this!!)

Example 8.1.1
A student suspects the cost of a date is no longer \$20.00. Assume that the average cost of a date is normally distributed. They take a random sample of 160 people from their dormitory and found an average cost of \$21.17 with a sample standard deviation of \$5.51. Test at a) $\alpha = 0.05$ and b) $\alpha = 0.01$ to check their claim.

Steps 0-1 are exactly the same. Step 2 can be omitted. Step 3 is exactly the same. At step 4 we can compute the *p*-value.

From Step 3 we have a test statistic of T = 2.69. Exactly how odd is this if the null hypothesis is true? First we look at the alternative hypothesis. It is two sided, so our *p*-value is found by finding the area in the right-tail and multiplying by 2. We need the area under the standard normal curve to the right of 2.69. This is now a “simple” table value lookup. I find that the value in the *z*-table for 2.69 is 99.64%. We want the right tail, so the answer is 100%-99.64% = 0.36%. Twice this is the *p*-value. So, the p-value here is 0.72%. Here’s a picture:

two-tailed *z*-test with test statistic 2.69

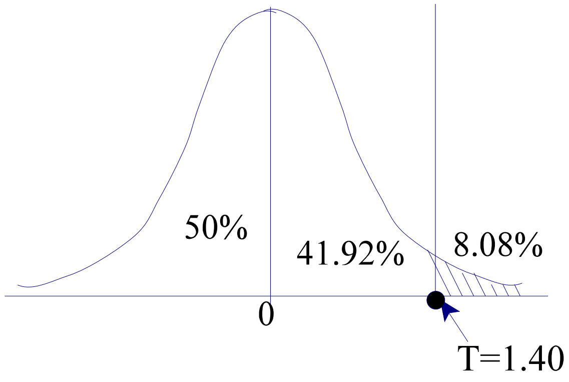


What does a *p*-value of 0.72% mean? If we repeat the experiment and the null hypothesis is true, then the probability of getting a test statistic of 2.69 or bigger or getting a -2.69 or smaller (“seeing something as strange as 2.69") is 0.72%, which is not very likely at all. Thus, this *p*-value is very significant. Recall that we rejected the null even at the 1% level of significance.

Example 8.1.2
Steps 0-1 are exactly the same. Step 2 can be omitted. Step 3 is exactly the same. At step 4 we can compute the *p*-value.

From Step 3 we have a test statistic of T = 1.40. Exactly how odd is this if the null hypothesis is true? First we look at the alternative hypothesis. It is one sided and favors large values of the test statistic. So our *p*-value is found by using the right-tail. We need the area under the standard normal curve to the right of 1.40. This is now a “simple” table value lookup. I find that the value in the *z*-table for 1.40 is 91.92%. We want the right tail, so the answer is 100%-91.92% = 8.08%. This is the *p*-value. Here’s a picture:

Right-tailed z-test with test statistic 1.40

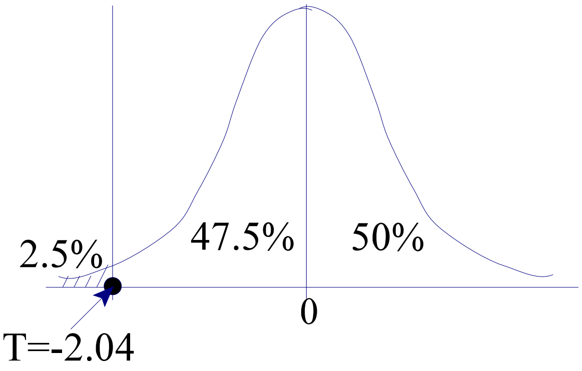


What does a *p*-value of 8.08% mean? If we repeat the experiment and the null hypothesis is true, then the probability of getting a test statistic of 1.40 or bigger (“seeing something as strange as 1.40") is 8.08%, i.e., iIf α =

10%, then we would reject the null hypothesis. If α = 5% then we fail to reject the null hypothesis.

Example 8.1.3
A large study of 30 states yield that the average one-way commute time to work is 29.4 minutes. The mayor of a small town believes that his town has a lower average one-way commute. She takes a random sample of 20 commuters and finds a mean commute time of 26.2 with a standard deviation of 7. With α = 0.10, is her assertion plausible?

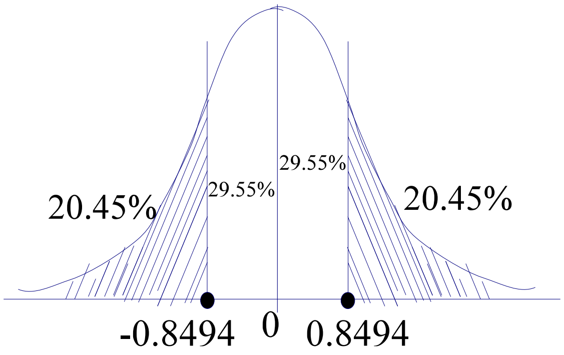
> Steps 0-1 are exactly the same. Step 2 can be omitted. Step 3 is exactly the same. At step 4 we can compute the *p*-value.
>
> From Step 3 we have a test statistic of T ≈ -2.0444. Exactly how odd is this if the null hypothesis is true? First we look at the alternative hypothesis. It is one sided and favors small values of the test statistic. So our *p*-value is found by using the left-tail of the *t*-distribution. This is now a "simple" table value lookup. I find that the value in the *t*-table for -2.0444 is approximately 0.025 (this is the area that corresponds to 2.093, which is very close to 2.0444). This is the *p*-value (actually, the true *p*-value is just a bit larger than 2.5%, but not much - check here http://www.stat.tamu.edu/~west/applets/tdemo.html).



What does a *p*-value of 2.5% mean? If we repeat the experiment and the null hypothesis is true, then the probability of getting a test statistic of -2.0444 or smaller ("seeing something as strange as -2.0444") is about 2.5%. This is pretty significant. We reject for any level of significance larger than 2.5%.

Example 8.1.4
A student suspects the cost of a date is no longer $20.00. Assume that the average cost of a date is normally distributed. They take a random sample of 16 people from their dormitory and found an average cost of $21.17 with a sample standard deviation of $5.51. Test at α = 0.05 to check their claim.

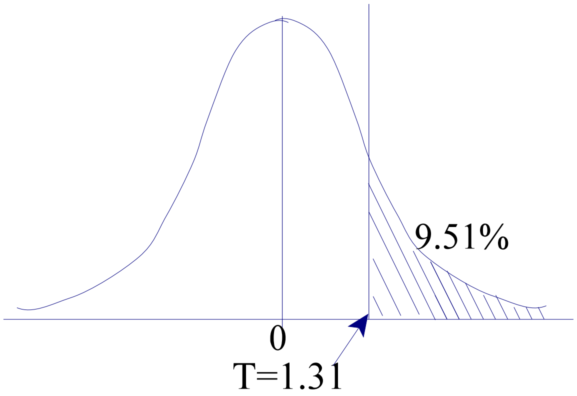
> Steps 0-1 are exactly the same. Step 2 can be omitted. Step 3 is exactly the same. At step 4 we can compute the *p*-value.
>
> From Step 3 we have a test statistic of T = 0.8494. Exactly how odd is this if the null hypothesis is true? First we look at the alternative hypothesis. It is two sided. So our *p*-value is found by using both tails of the *t*-distribution. The area to the right of 0.8494 is about 0.2045. Since we use both tails, we take twice this, and so our *p*-value is about 40.9%. This is a very large *p*-value and so is not significant at all. We would fail to reject the null for any reasonable significance level.



Example 8.2.1
Drug X is given to a sample of 170 patients who have a particular disease. We find that 50% of the subjects in the sample recover from the disease. We want to test whether Drug X produces a higher recovery rate than Drug Y, which is known to produce a recovery rate of 45%. Test this claim at a) α = 5% and b) α = 10%.

The two categories are recover from the disease (1) and does not recover from the disease (0). For Drug Y, we know that p = 0.45. That is, the proportion of subjects in category (1) is 45%. What we wish to find out is if the proportion of people who took Drug X and recovered, 50%, is statistically significantly greater than the proportion of people who took Drug Y and recovered.

> Steps 0-1 are exactly the same. Step 2 can be omitted. Step 3 is exactly the same. At step 4 we can compute the *p*-value.
>
> From Step 3 we have a test statistic of T ≈ 1.31. Exactly how odd is this if the null hypothesis is true? First we look at the alternative hypothesis. It is one sided and favors large values of the test statistic. So our *p*-value is found by using the right-tail of the standard normal distribution. This is now a “simple” table value lookup. I find that the value in the *z*-table for 1.31 is 90.49%. The area to the right of this is 9.51%. So, our *p*-value is 9.51%. Here’s a picture:



> What does a *p*-value of 9.51% mean? If we repeat the experiment and the null hypothesis is true, then the probability of getting a test statistic of 1.31 or larger (“seeing something as strange as 1.31") is about 9.51%. If α = 10%, then we would reject the null hypothesis. If α = 9.5%, then we would fail to reject the null hypothesis.