Lecture 10

One-sample tests

Today we are going to look at conducting the most common (and simple) one-sample tests for means and proportions. We will apply an older strategy (when technology was less available) that uses the rejection region method. We will start with the two-tailed tests as they are the most similar to the confidence interval case, and then we'll look at the one-tailed test case.

Two of our examples from last time with enough information to proceed. Let's look at the means case first.

Example.
Blood glucose levels for obese patients have a mean of 100 with a standard deviation of 15. A researcher thinks that a diet high in raw cornstarch will **have an effect** on blood glucose levels. A sample of 30 patients who have tried the raw cornstarch diet have a mean glucose level of 140. Test the hypothesis that the raw cornstarch had an effect.

Last time we determined what hypotheses we were going to test.

$$H_0: \mu = 100$$
$$H_a: \mu \neq 100$$

The question asks about an effect and does not suggest a direction, so we use not equal to in the alternative hypothesis statement. This is what makes this a two-tailed test. The problem talks about a mean, so this is clearly a test of means.

The next thing, as with confidence intervals, that we have to determine is whether this is a t-test or a z-test. Recall that the z-test is actually the special case. For that, we have to have three things to be true:

- The population standard deviation (not the sample standard deviation) must be available
- The sample size needs to be sufficiently large
- The population must be normally distributed

The problem does provide the population standard deviation in the opening sentence. The sample size is border line (it meets the rule of thumb in some textbooks, but not others). The problem does not state that glucose levels are normally distributed. Many natural processes are, but it would be better if we knew this for sure. Given these uncertainties, I suggest using the t-test. I suspect it won't make a difference in our outcome here, but the t-test is a little more conservative and covers our uncertainties a little more safely.

In order to proceed, we calculate a test statistic. The formulas for the t-test and z-test are similar, so I'll provide them both here.

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \qquad t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Essentially, these formulas are the same. The main difference is in using the sample standard deviation rather than the population standard deviation. μ_0 is the mean of the null hypothesis and $\bar{x}$ is the mean of the sample. The denominator is the standard error of the sampling distribution.

The test statistic here is

$$t = \frac{140 - 100}{\frac{15}{\sqrt{30}}} = 14.6059 \ldots$$

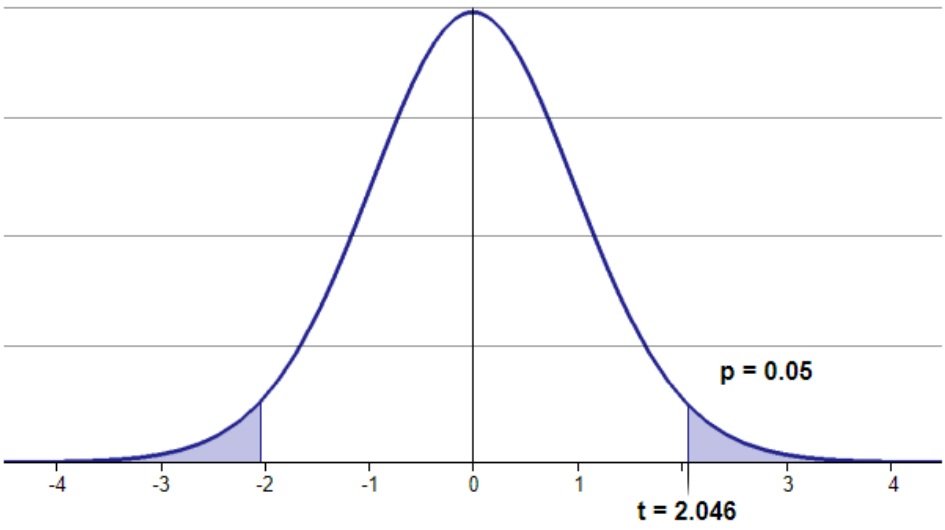
How do we evaluate this test statistic relative to the rejection region? We need to come up with a critical value to compare it to.

By default, the significance level is assumed to be $\alpha = 0.05$ or 5% unless the problem states otherwise. So we will use that to find the critical value.

Since this is a two-tailed test, as in the confidence intervals, that means we need to leave half this amount on either side in the tails. We need to t-value (with 29 degrees of freedom) that have 0.025 less than that value, or 0.975 less than that value.

We obtain the critical value $t_{0.025,29} = 2.045 \ldots$, which is expectedly just a bit more than the 1.96 we'd get from the standard normal distribution.

Let's graph this.



A test statistic that falls into the unshaded portion of this graph in the middle then the probability that his outcome came from the assumptions of the null hypothesis are high enough that we can "keep" it. Any test statistic that falls in the shaded regions have a low probability of coming from the assumptions of the null hypothesis. What this means mathematically is that any test statistic whose absolute value is greater than 2.046, i.e. $|t| > 2.046$ falls into the rejection region. In such a situation we reject the null hypothesis.

Our test statistic is WAY bigger than 2.046.

The terminology around rejecting the null hypothesis has changed somewhat over the decades and so it's important to focus on this for a minute.

If the test statistic is in the rejection region, we say we reject the null hypothesis. That means that there is strong evidence to think that our data was unlikely to have been produced under the assumptions of the null hypothesis. Rejecting the null hypothesis means that we can accept the alternative hypothesis, that there is good evidence for this.

If the test statistic does not fall in the rejection region, we cannot reject the null—the evidence for the alternative hypothesis is too weak. In this case we say that we "**fail to reject**" the null. In the decades before I was born, it was common to say that we "accepted" the null, but this is no longer standard practice. Accepting the null is too strong a statement. We don't have evidence that the null is true, only that it's not strong enough to overturn it. This is akin to saying that a defendant in a criminal trial is "not guilty". We do not say that they are "innocent". Sometimes it can feel like we are using a double negative, or unnecessarily convoluted language when we do this, but we are taking care not to make too strong a claim, one that is not warranted. The precision takes some getting used to, but one gets used to it with practice and experience.

We need to be able to interpret our result in context. We can say here that there is strong evidence to conclude that eating raw cornstarch does affect blood glucose levels.

We can think of conducting a hypothesis test as a step-by-step process.

1. State the null and alternative hypotheses in appropriate notation.
2. Determine the type of test to be conducted.
3. Calculate the test statistic.
4. Convert the significance level to a critical value for the boundaries of the rejection region.
5. Compare the test statistic to the rejection region.
6. State the conclusion of the test (reject or fail to reject the null).
7. Translate the result into a plain English sentence in context.

Let's look at our two-tailed test of proportions.

Example.
The mortgage department of a large bank is interested in the nature of loans of first-time borrowers. This information will be used to tailor their marketing strategy. They believe that 50% of first-time borrowers take out smaller loans than other borrowers. They perform a hypothesis test to determine if the percentage is **the same or different** from 50%. They sample 100 first-time borrowers and find 53 of these loans are smaller than the other borrowers. For the hypothesis test, they choose a 5% level of significance.

State the null and alternative hypotheses.

$$H_0: p = 0.50$$
$$H_a: p \neq 0.50$$

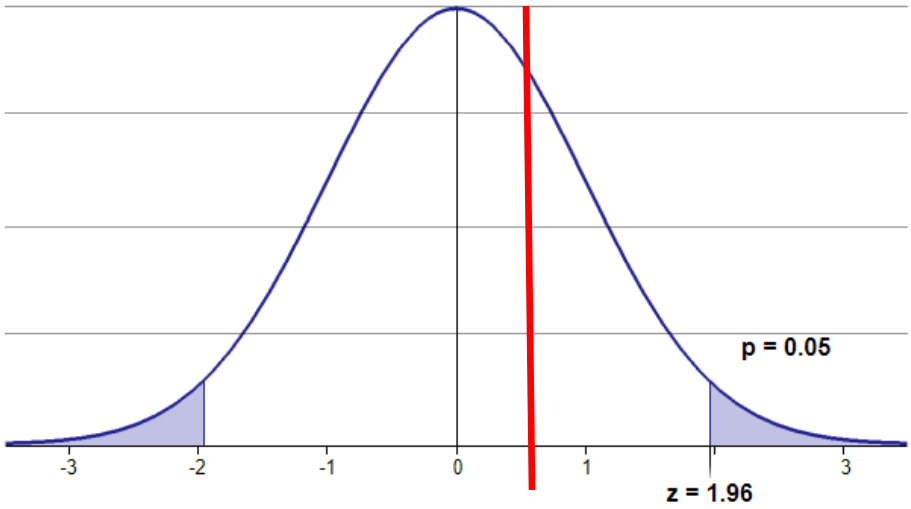
For the test of proportions, we don't have to concern ourselves with choosing the z-test or the t-test. But we should pause a moment to make sure that our situation meets the standards we need for approximating the binomial distribution as a normal distribution. In this case $np(1-p) = 100(0.5)(0.5) = 25$. So we are good. If the test fails, we would have to use the binomial distribution directly (similar to what we did in the simulation lab).

The test statistic for the proportion case is

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

Here $\hat{p}$ is the estimate of the proportion from the sample ($\frac{53}{100} = 0.53$). The p_0 value is the proportion from the null hypothesis. We find the test statistic to be $z = \frac{0.53-0.50}{\sqrt{\frac{0.5(0.5)}{100}}} = 0.6$.

Because this is the standard normal distribution with the standard two-tailed significance level, we already know the critical value is 1.96.



I've added a red line to the graph where our test statistic falls. You can see that it clearly falls into the unshaded region, which means it does not fall in the rejection region.

We therefore conclude that there is insufficient evidence for the alternative, and we fail to reject the null hypothesis.

We conclude there is not enough evidence to claim that first-time borrowers take out smaller loans at a different rate than expected.

Example.
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?

We want to consider now the one-sided case. The scenario here is one such situation. The question asks whether the evidence for the dropout rate being above 25% is strong. First, it's worth noting that if the question asked if it was less, we should first check the sample proportion, because it obviously can't be strong evidence for being less if the proportion is actually greater than the null hypothesis. We could do the math, but this is obviously contradictory, so there isn't much point. That's not the case here, so let's state the null and alternative hypotheses.

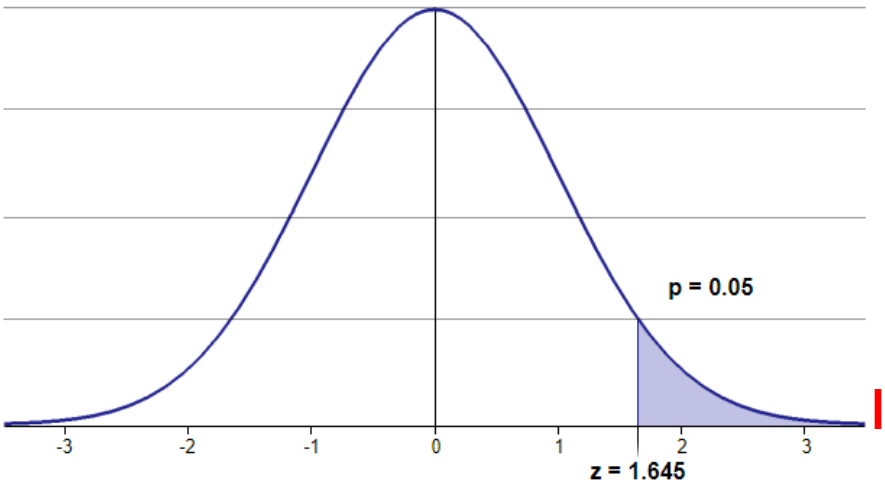
$$H_0: p = 0.25$$
$$H_a: p > 0.25$$

We want to determine if the proportion is greater than 25%... but crucially, enough larger that we could not have obtained the value by chance. It's not enough to simply note that $\frac{190}{603}$ is simply bigger. We need to show that we are unlikely to obtain this result simply because sampling is an estimate.

Calculating the test statistic does not change. We use the one-proportion z-test as we did before.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\frac{190}{603} - 0.25}{\sqrt{\frac{0.25(0.75)}{603}}} \approx 3.691\ldots$$

Even if we use the same significance level we did last time, we can't use the same critical value. That's because we aren't using a two-tailed test, so we don't have to divide the probability up between the two tails. All 5% can go in one tail.



In this case, we use a right-tailed test, since the alternative is greater than. The critical value becomes 1.645. Any test statistic greater than this is in the rejection region. Our test statistic falls just off the graph to the right and is definitely in the rejection region. So we reject the null hypothesis.

This means that there is sufficient evidence to think that the dropout rate is higher than 25%.

If our alternative hypothesis used the less than sign, we'd want a negative critical value for the left-tailed test. We would want our test statistic to be smaller (i.e. more negative) than the critical value to be in the rejection region.

Example.
A light bulb manufacturer claims that its light bulbs will last for 1000 hours. A consumer protection agency questions this claim, so they test 150 light bulbs and obtain an average lifespan of 997 hours with a standard deviation of 15 hours. Is this enough evidence to claim the manufacturer is lying?

The consumer protection agency doesn't care in this case if the lifespan is longer. That's good for consumers. Instead, they naturally care if the manufacturer is exaggerating the life of the bulbs, so they want to know if the bulb life is actually shorter.

Let's set up our null and alternative hypotheses.

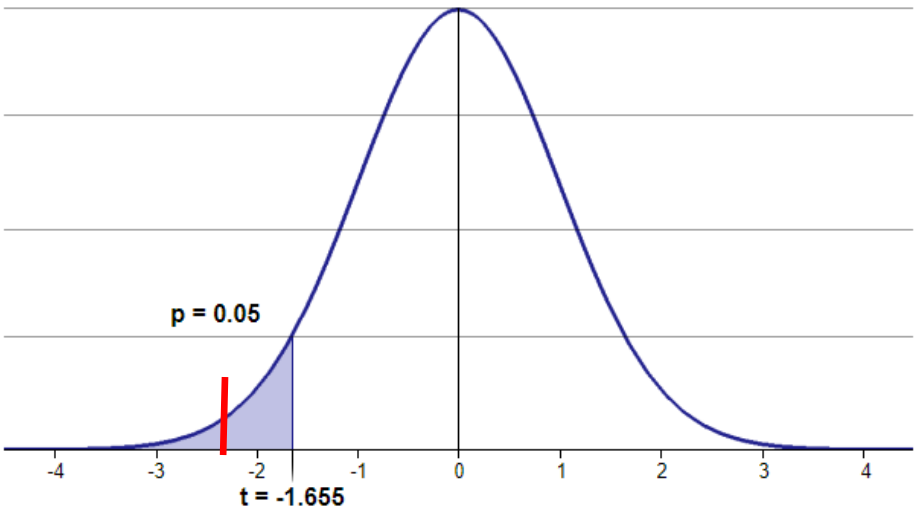
$$H_0: \mu = 1000$$
$$H_a: \mu < 1000$$

This is a one-tailed test, or a left-tailed test since the alternative hypothesis is using a less than inequality.

Let's calculate our test statistic. Since we have only sample data, we definitely want to use the t-test even though we do have a larger sample size.

$$t = \frac{997 - 1000}{\frac{15}{\sqrt{150}}} \approx -2.449\ldots$$

Our rejection region will be similar to the last problem, but on the other side of the distribution.



Our test statistic is less than the boundary of the rejection region and therefore, we can reject the null hypothesis.

There is sufficient evidence that the manufacturer is exaggerating their claims about the life of their lightbulbs.

But, this is a good place to discuss the difference between statistical significance and practical significance. Is a consumer likely to care that the manufacturer rounded up to 1000 if the bulb lifespan is really only 997 hours? Are they likely to care about a 0.3% difference in their claim vs. reality? If we use a large enough sample size, we can make even very small differences statistically significant (and sometimes that is important), but that isn't the same as those differences being meaningful or practical from the perspective of everyday life under normal usage.

In the next lecture, we'll look at reasons to adjust the significance level, and calculate power. We'll also look at the p-value method, which is a more modern (and more technologically dependent) method of evaluating hypotheses. The method is equivalent to our rejection region method but depends on converting the test statistic into a probability instead of converting the significance level into a critical value.

References:

1. https://faculty.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf
2. https://assets.openstax.org/oscms-prodcms/media/documents/IntroductoryStatistics-OP_i6tAI7e.pdf
3. https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/
4. https://opentextbc.ca/introbusinessstatopenstax/chapter/full-hypothesis-test-examples/
5. http://betsymccall.net/prof/courses/spring15/cscc/2470tests.pdf
6. http://www.statdistributions.com/t/
7. https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/one-tailed-test-or-two/