Lecture 11

Hypothesis testing continued

Last lecture we discussed conducting one-sample hypothesis testing with the rejection region method. In this lecture we are going to discuss the P-value method. This is the method we will mostly rely on for our hypothesis testing going forward, although you will notice that output from some statistical calculators will provide both the P-value and the critical value so that you can use either method.

In the last lecture, we discussed some procedural steps for conducting a test by the rejection region method. The P-value method differs from this process in two of the middle steps.

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 test statistic into a probability value (P-value).**
5. **Compare the P-value to the significance level.**
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.

The two steps that have been modified are in bold. Instead of calculating a critical value and comparing our test statistic to that value, we will convert our test statistic into a probability, and then compare that probability to the significance level. In some research papers, you may see only a P-value stated and not a specific significance level, allowing readers to apply their own.

Let's look at two of our examples from the last lecture in light of this new method.

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).

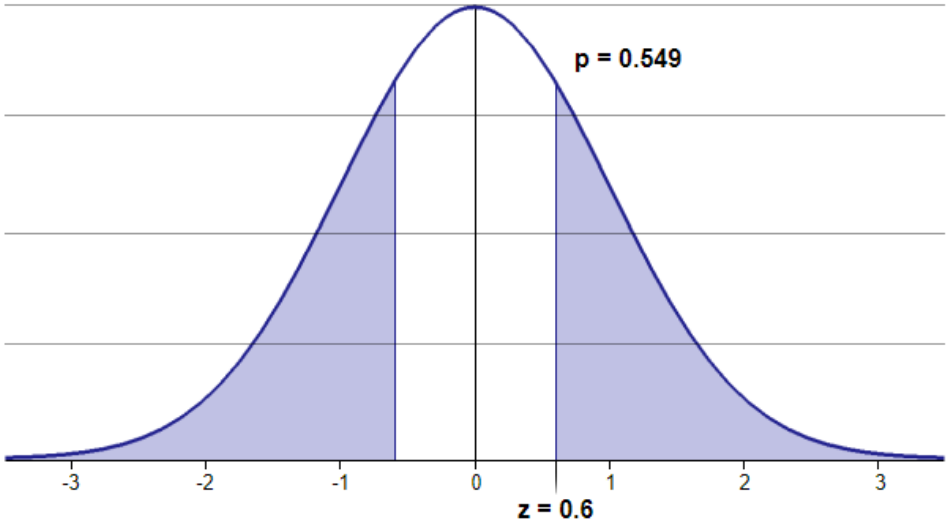
The test statistic for the proportion case is

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

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 $z = \frac{0.53-0.50}{\sqrt{\frac{0.5(0.5)}{100}}} = 0.6$.

So, far, our procedure is exactly the same. But now that we have the test statistic, we want to know the probability that we could obtain a test value that is equal to or more extreme than this value. If the probability is low, that's good evidence against the null hypothesis. If the probability is high, then the result is more likely to be due to random chance.

Our test statistic is from a standard normal distribution. Moreover, since this is a two-tailed test, we need values that are greater than 0.6 and less than -0.6.



Normally, what we will do mathematically is compute $P(Z \geq 0.6)$ or $P(Z \leq -0.6)$ and then multiply the result by 2 since the distribution is symmetric. In this case, our P-value turns out to 0.549. That means, conceptually, that there is more than a 50% chance that this result could have been obtained from the assumptions of the null hypothesis, or that 55% of the time, a result this extreme or more extreme could occur when the null hypothesis is true. That's not good evidence against the null hypothesis.

Procedurally, we compare this P-value to the significance level of 0.05. If the P-value > significance level, then we fail to reject the null hypothesis. If the P-value < significance level, then we reject the null hypothesis.

As we did with the rejection region method, we fail to reject the null hypothesis in this case.

Let's look at the one-tailed case:

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.

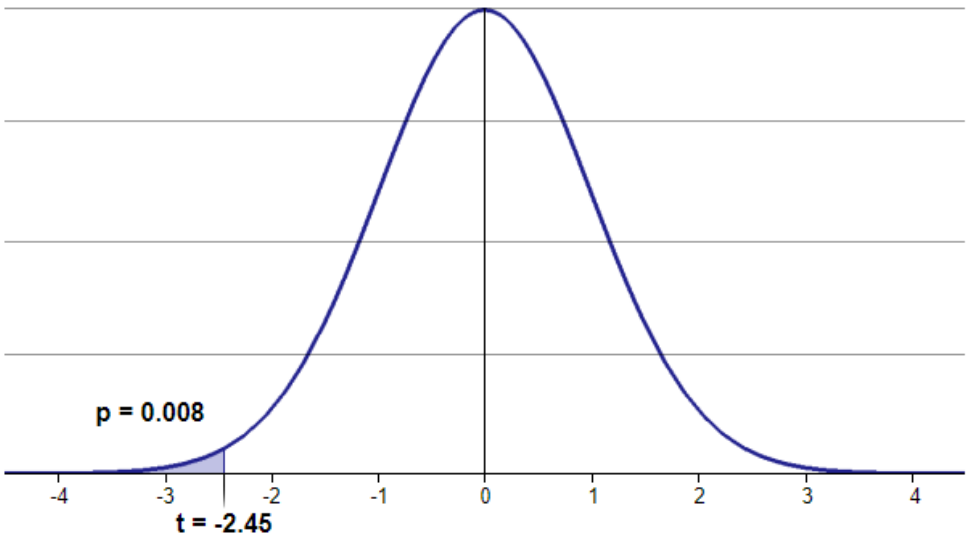
$$H_0: \mu = 1000$$
$$H_a: \mu < 1000$$

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 procedure is the same as before, but now we convert this test statistic to a probability. We only need the one tail, since this is a one-tailed test, so we want $P(T \leq -2.449)$.



The P-value is approximately 0.008 < 0.05, and so we can reject the null hypothesis. We can only expect to obtain this result from the assumptions of the null hypothesis 0.8% of the time. That is less likely than we require, so this is strong evidence for the alternative hypothesis.

Now that we have some greater experience with conducting hypothesis tests, we want to return to a discussion of significance and power.

Recall that α is the variable we use for significance (and its relationship to confidence $= 1 - \alpha$). Sometimes we might want to be more or less careful when creating confidence intervals or analyzing hypotheses. For instance, if the consequences of being in error are high, then we might want to be more certain. This would be equivalent to creating a confidence interval with more confidence (say 99% confidence) or with a lower significance threshold (such as setting $\alpha = 0.01$ or lower). Sometimes, we might want a narrower confidence interval. We might be willing to be wrong more often and set a lower confidence level, which is the same as letting the significance level be higher (say $\alpha = 0.10$). The significance level is the probability of making a Type I error, so we set this value, something we can control, to adjust for these consequences.

For instance, if the consequences of rejecting the null when it is true is high (such as lead in the water), we may want to lower the significance level to be sure that the evidence must be even more compelling in order to reject the null. If the consequences are not serious, such as if we are comparing two learning strategies, both of which are reasonably effective, then the consequences of making a small error and trying the slightly less effective one may not be serious. In such a case, we can raise the significance level so that we can accept somewhat weaker evidence to reject the null.

These considerations are some of the reasons why we should carefully think about where we are setting our significance level before conducting the test. What we do not want to do is look at our P-value and then set the significance level to make it seem significant. We should make an objective assessment of our situation and consider our risk tolerance first. Otherwise, we make a serious statistical error that tends to lead to bad research methods that will come back and bite you later.

The power of a test is related to the probability of a Type II error. Calculating power is somewhat complex since it depends on the significance level, and the specific data we have including the sample size. Before we look at what power is and how we calculate it, let's review the meaning of our Type I and Type II errors, and describe them in context.

Type I and Type II Error

Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = 1 - β
Not rejected	Correct decision True negative Probability = 1 - α	Type II error False negative Probability = β

The table above is similar to one from a couple of lectures ago, but it is deliberately organized differently. You want to understand what the errors mean, not just memorize their locations in a table.

When describing an error, we need to say both what the true state of the world is, and the conclusion we made that contradicts it.

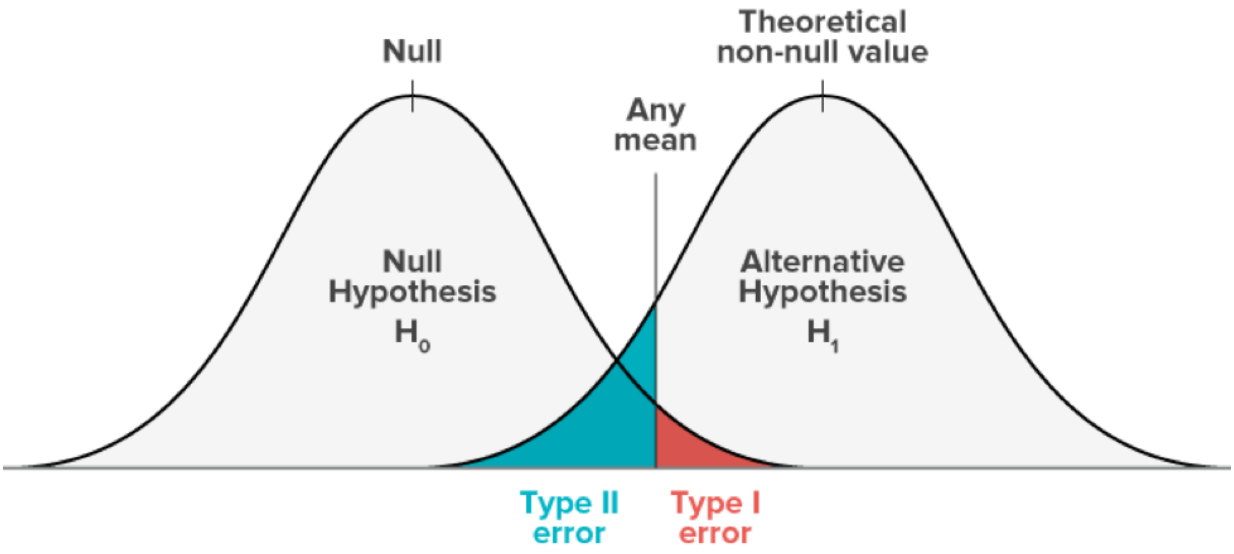
A Type I error is when the true state of the world agrees with the null hypothesis (the null hypothesis is actually true), but based on our data, we rejected the null.

A Type II error is when the true state of the world is that the null hypothesis is false, but based on the data we have collected, we are unable to reject the null.

Let's think about the hypothesis test we conducted on the light bulbs. In context, a Type I error would be saying that the light bulbs do not last for 1000 hours when they actually do. A Type II error would be being unable to say that the light bulbs do not last for 1000 hours (they last less than 1000 hours), when they in fact do not last that long. (The double negatives can be confusing, but the careful language to avoid making overly strong claims is necessary.)

When an experiment is designed, we want to also be able to determine the power of a test, which in some sense is our ability to correctly rule out the null hypothesis given some hypothetical true value for the alternative hypothesis. One of the difficulties with calculating β is that there is a different value for every theoretical value of the alternative that is different from the null. If the alternative is very similar to the value of the null hypothesis, the probability of making a Type II error is large (the power is small). If the hypothetical alternative value is very different from the null, then the power of the test will be high (and the probability of a Type II error will be low).

Conceptually, we are thinking of the following graph. If there is more overlap between the null hypothesis and our hypothetical alternative hypothesis, then the probability of a Type II error will be higher. As the means of the distributions separate, the overlap and the probability of a Type II error will decrease.



While we can compute the power of a test for multiple values of a hypothetical alternative, usually, we calculate it for a specific threshold value that is of interest to the researchers.

For instance, if we think about the example we did in a prior lecture about dropout rates, we might set a threshold of 30% as the point where we want to calculate the power of the test. Perhaps we think that values that are higher than 25% but not as high as 30% don't affect policy as much, but if the true value is as much as 5% higher, then maybe special additional steps will need to be taken and we want to be sure that we can detect it.

This difference is referred to as the effect size.

Recall from our example that we had set α equal to the standard value of 0.05, for a one-tailed test, this set our critical value to 1.645. We can convert this to a region of acceptance (the opposite of the rejection region), but in terms of proportions (or means, depending on the problem). We calculate this as essentially a one-sided confidence interval.

$$0.25 + 1.645\sqrt{\frac{0.25(0.75)}{603}} \approx 0.279$$

Any proportion less than 27.9% will fail to reject the null. Values above it will reject the null. How much of the distribution centered at 30% falls into this reject region?

$$z = \frac{0.279 - 0.30}{\sqrt{\frac{0.3(0.7)}{603}}} \approx -1.1249$$

The numerator is the difference between the critical value (0.279 that we calculated above) and the hypothetical proportion (or mean) that we are calculating the power for. We assume the same sample size as the sample here. And then we convert this score into a probability in the standard normal distribution.

$$P(Z \leq -1.1249) \approx 0.13 = \beta$$

The probability of a Type II error is thus about 13% when the sample proportion is about 30%. Thus the power is $1 - \beta = 87\%$.

The probability distribution for a proportion changes as the proportion changes, but in the case of the means, we generally assume the standard deviation remains fixed in both distributions.

When designing an experiment, we can adjust the sample size to increase the power if needed.

Lower power can cause systematic errors to overwhelm a hypothesis test and lead to missing the desired effect size. If you can't detect the desired difference, then you are kinda wasting your time. Typically, statisticians will try to set the power for their desired effect size to be about 80%. You also don't want the power to be too high. This can make small differences more significant that is meaningful, and it will usually cost more money to get the larger sample sizes. As an example, our light bulb example likely had power that was too large.

One-tailed tests also have higher power (because all of α is on one side). So it can be helpful to test a hypothesis that is clearly either larger or smaller than a null hypothesis and not just "different".

Factors that can impact power:

- effect size: larger effect size = more power
- sample size: larger sample = more power
- significance level: increasing the significance level = more power
- measurement error/variability: smaller variability = more power
- one-tailed vs. two-tailed tests: one-tailed tests = more power

Next time, we'll talk a bit more about hypothesis testing and wrap up our discussion of one-sample tests.

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.scribbr.com/statistics/type-i-and-type-ii-errors/
4. https://www.abtasty.com/blog/type-1-and-type-2-errors/
5. http://www.statdistributions.com/t/
6. https://stattrek.com/hypothesis-test/statistical-power.aspx
7. https://www.scribbr.com/statistics/statistical-power/