Section 5
Exploring Hypothesis Testing

Last week, we introduced the main ideas of hypothesis testing, building on what we modeled and simulated in TinkerPlots. We will continue this exploration this week and explore two-sided alternatives, errors, power, and the limitations of hypothesis testing.

5.1 – Two-sided Alternatives

Last week, we explored tests with alternative hypotheses that are one-sided in nature. This is typically used when you have some expectation about whether you expect your data to be greater or less than your null hypothesis even before collecting data. But sometimes, no such expectation may be present.

Example: You have found a coin on the ground that has seen better days and are curious to determine if it is still a fair coin when flipped – that is, heads and tails are still equally likely to occur. You flip the coin 100 times and get 58 heads. Does this result still seem plausible if the coin were still fair? Conduct a hypothesis test to evaluate this question, and use simulation to see what would happen by random chance.

H_0: _____

H_a: _____

What does this mean about our *p*-value, and what results are considered "as extreme or more" than what we observed?

5.2 – Errors and Power

Errors in hypothesis testing

The difficulty in making statistical decisions is that we never have the ability to prove anything. Last section, we saw that 14 out of 16 infants choosing the helper toy was strong evidence toward the idea that infants were making the decision consciously and not randomly. However, it's always possible that it could have happened by random chance. Similarly, if we flip 16 coins, it's unlikely to get 14 heads to come up, but it is definitely possible.

Thus, it's possible that these infants really didn't consciously choose the helper toy 14 times, and this was just statistical improbability. If we assume this to be the case, then going on to conclude that they were making this choice consciously and preferring helpful behaviors would be an error.

The type of error that could have been made in this case would be called a _____, where you reject H_0 when H_0 was actually true. On the flip side, if you fail to reject H_0 and H_a is actually true, this would be called a _____. The table that follows shows the possible outcomes of a hypothesis test:

		Decision	
		Fail to Reject H_0	Reject H_0
The true hypothesis	H_0		
	H_a		

Probabilities of errors

An important measure for statisticians to understand when testing is the probability of errors occurring. The probability that a type I error occurs has already been defined – it's α. Why? If the null is actually true, then our null distribution that we would simulate not only reflects the null, it would reflect actual samples we would obtain in the real world too. Because these match, the *p*-values we get match the actual probabilities of obtaining a result like that. Getting a low *p*-value would result in an error here (since the null is true!) so the percentage of results that produce a *p*-value lower than the significance level would just be the significance level itself.

As a result, we can think about setting α at the beginning of a test as a way to say how often we are comfortable with rejecting the null incorrectly. Hypothesis tests are often used to make real-world decisions, so when setting α, you should think about what the consequences are of being wrong. If your hypothesis tests will make a decision regarding releasing a new prescription to the public, and you were testing for its side effects being potentially deadly, a lower α value might be a good idea. If your consequences are less dire, then a higher value of α is potentially acceptable.

We use the symbol β to represent the probability of a type II error. While α is something that we know and set at the outset of a hypothesis test, β is not easily calculated. One common misconception is that α and β are complementary probabilities, but they are not. The probability of a type I error is based on the assumption that the null is true, where type II error is based on the

assumption that the alternative is true, so they can't be complementary because they are based on different assumptions!

The complement of β does have its own definition though: the quantity $1 - \beta$ is the _____. This value tells you a probability of rejecting the null hypothesis when the alternative is true, a probability of utmost importance to researchers. Hypothesis testing is designed so that the researcher's theory is reflected in the alternative hypothesis. Collecting data can be expensive, so before conducting the test, they will want to know what the likelihood is that that they will get results that match their theories.

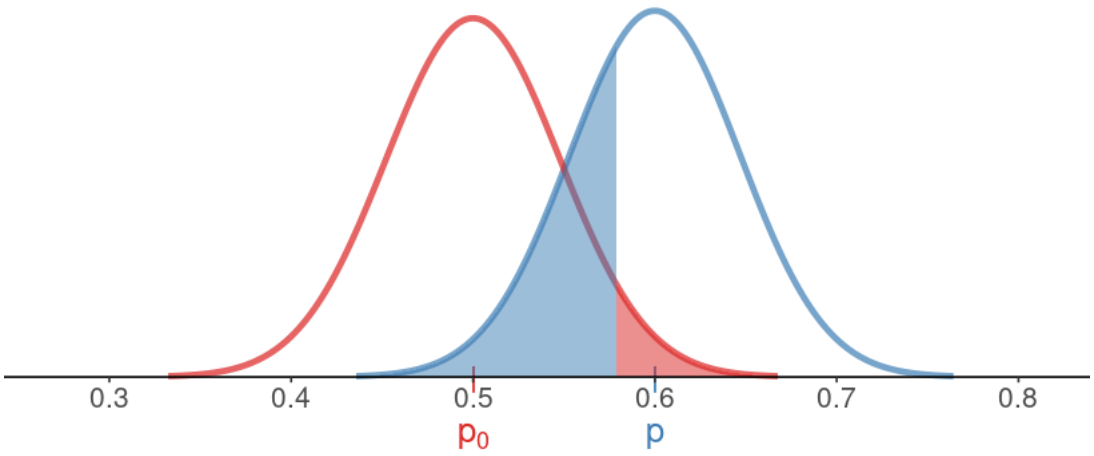
If one is interested in increasing the power of a test, there are two ways to accomplish this:

- _____: This is a somewhat artificial way of increasing power, but by making more p-values able to reject the null hypothesis, you also make it more likely that you do so correctly. The cost here is that you also increase the probability that you reject the null hypothesis when the null is true too.

- _____: By increasing this, you get more information about your population, and thus make better and more informed decisions. This makes it more likely to distinguish your data from the null hypothesis.

To illustrate these relationships, try using the applet at the link below:

https://istats.shinyapps.io/power/



Practical significance

It's important to remember that the purpose of hypothesis testing is to see if your data fits within the null hypothesis or not. We noted earlier that you can increase the power of your test by increasing the sample size, but sometimes this may go too far. Consider the next example:

> *Example*: Trident's advertisements always claim that 4 out of 5 dentists recommend Trident gum to their patients over other brands of chewing gum. Someone thought that they might be overstating how many dentists really recommend their gum, so they took a random sample of 4000 dentists in the US and asked them if they would recommend Trident gum to their patients. 3132 of those dentists said that they would recommend Trident. Evaluate Trident's claim based on this data with $\alpha = 0.01$.

According to this test, we would conclude that Trident's claim is false. But is that fair to say? If we check the percentage of dentists that recommended Trident, we get 3132/4000 = 78.3%. The hypothesis test that we conducted told us that if 80% of all dentists recommend Trident, then seeing the data that we obtained of 78.3% at such a high sample size is very unlikely. But does that make Trident's claims in their commercials invalid? A difference of 1.7% is not terribly meaningful, and for advertising purposes, rounding to an easy fraction is better for delivering their message. It's important to keep in mind that hypothesis testing can only tell us if the null hypothesis is not consistent with our data, and that the actual difference between the data and the null hypothesis might not hold any practical importance. You may be able to have a more powerful test with a larger sample size, but that larger sample size may not end up telling you results that are meaningful, even if they are statistically significant.

To wrap up this section, let's finish with one more example that does all steps of a hypothesis test using R, and interprets the results of the test. So far, our null hypotheses have all been stated in terms of fractions (e.g. 1 in 2, 1 in 4, 4 in 5), but this does not always have to be the case – your null model could just be represented by a spinner with any proportion!

Example: The 2010 census showed that 55.3% of US households do not have any children. A random sample of 500 US households is taken now to assess if this population proportion has changed since 2010. Of the 500 households, 305 had no children. Use a $\alpha = 0.05$ to carry out the test.

Write out hypotheses:

H_0: _____

H_a: _____

Conduct the test and evaluate the evidence:

Interpret the *p*-value:

Make a decision for the test:

Draw conclusions:

5.3 – Additional Practice

Example: Revisit the additional practice question from last section:

In Western Countries, only about 12% of the population identifies as "left-handed." While part of hand preference may be environmentally conditioned, genetics may also play a role. One theory posits that red-headed people are more likely to be left-handed! Based on genetic theory, we'd like to see if red-headed people might be more likely to be left-handed than the general population. Let's say that we take a random sample of 125 red-headed people. We found that 40 of them had a preference for their left-hand. We conducted this test using a significance level of $\alpha = 0.01$.

- In the previous section, you should have found that the null hypothesis was rejected. What kind of error could have been made in this scenario?

- Suppose that the actual data from this study found that 24 out of 75 red-headed people were left-handed. How would the following change if a hypothesis test were conducted on this data?
    - *p*-value
    - Power
    - Probability of a Type II error

- Suppose that we are now considering the study of 125 people again, but this time, the test was conducted with significance level $\alpha = 0.03$. How would the following change?
    - *p*-value
    - Power
    - Probability of a Type II error