Prelude

We're not getting into the part of the class where we'll get bombarded with many different types of hypothesis tests. Here's a little table that helps me keep these straight; maybe it'll help you too. Note that, in this table, a "binary variable" is a special kind of categorical variable which has only two categories.

Number and types of variables	Hypothesis test	Sections where discussed
1 binary	Proportion	5.3, 6.1
2 binary	Difference of proportions	6.2
1 categorical	χ^2 test	6.3
2 categorical	χ^2 test	6.4
1 numerical	t test	7.1-2
1 numerical, 1 binary	Difference of means	7.3
1 numerical, 1 categorical	ANOVA	7.5
2 numerical	Regressions	8.1–4

Select Reading Question Responses (3/10)

> How do you calculate the p-value in a two-sided hypothesis test? Is it the same method as one side and then multiply that by 2?

Yes! ☺

> How do you find tail area in a normal distribution? I don't understand the jump from z-score to tail area.

You have to use the `pnorm` function in R, which converts z-scores to percentiles. Note that you can't plug a z-score into `pnorm` and just blindly use the output as your p-value. You have to remember that the output of `pnorm` will be a *percentile*, which is not necessarily the same as the p-value that you're looking for. Draw pictures to make sure you're computing the right thing!

> I was confused by when it is considered appropriate to use a sample proportion and when to use a population proportion. If a previous study found that $p = 0.55$, is that now a population proportion if we decided to replicate it?

When the success-failure condition is satisfied, the sampling distribution for a proportion is roughly normal $\mathcal{N}(p, \sqrt{p(1-p)/n})$. If we don't know the population proportion p, then

it makes sense to approximate it using our sample proportion $\hat{p}$. This is what happens, for example, when we compute confidence intervals. On the other hand, when you conduct p-value hypothesis tests, you want to *use your null hypothesis to describe your sampling distribution*. That way, the p-value you compute is really the probability of observing the data assuming the null hypothesis!

For example, if a previous study found tht $p = 0.55$, and the question you're interested in is "Has p changed since that last study?" then you might conduct a hypothesis test where your null hypothesis is something like "$p = 0.55$ still." In other words, you *assume* that the population proportion is still 0.55 like it was for the previous study, and use that assumption to describe the sampling distribution. In this case, you approximate our sampling distribution using $p = 0.55$ as $\mathcal{N}(0.55, \sqrt{0.55 \cdot 0.45/n})$ instead of using the approximation $p \approx \hat{p}$. That way, the p-value you compute in the end will be the probability of observing the data you observed if indeed it is still true that $p = 0.55$. If that probability turns out to be very small, it means that the data you observed would be very unlikely if p is still 0.55, so you can be confident in rejecting the hypothesis that p is still 0.55.

> It makes sense why failing to find strong evidence for the alternate hypothesis is not enough to accept the null hypothesis as true, but are there any cases where we would end up accepting the null hypothesis as true? What would that process look like?

This is very difficult to do because it involves taking larger and larger samples. Remember that null hypotheses are usually *precise*: something like "$p = 0.55$" in the question above. How would you verify that the population proportion is actually equal to 0.55 *on the nose*?

Let's say I'm running tests with significance level $\alpha = 0.05$ and H_0 is the statement that $p = 0.55$. I might collect a sample size that's big enough so that me observing $\hat{p} = 0.6$ would be statistically significant. In other words, the sample size would be big enough that observing $\hat{p} = 0.6$ would be enough to reject H_0. But likely that same sample size would still not be big enough to reject H_0 if what I observe is $\hat{p} = 0.56$. For that, I would need a bigger sample size. But even that bigger sample size might not be enough to reject H_0 if what I observe is $\hat{p} = 0.551$. I could increase my sample size again so that the difference between 0.55 andd $\hat{p} = 0.551$ would be statistically significant, but then maybe my sample size would still not be big enough to distinguish between 0.55 and $\hat{p} = 0.5501$. And so forth.

In the limit, the only way of really verifying that $p = 0.55$ on the nose is actually just "sampling" the *entire population*! In most situations, that's just not going to be feasible.

That being said, depending of what kind of a situation you're in, maybe the difference between 0.55 and 0.551 is too small to matter. In that case, I would choose a sample size that's large enough to detect the difference between 0.55 and 0.551. I might collect a couple of samples of that size and then not find evidence to reject H_0. And then maybe I repeat that process once or twice more, with the same results. In that case, I can probably go about my life pretending like $p = 0.55$ without too much harm... ☺

And *that* being said, it's also worth pointing out that this somewhat peculiar language ("We find evidence to reject the null hypothesis" and "We do not find evidence to reject the null hypothesis") is drilled into statisticians because it reflects a kind of deference to a crucial principle in the philosophy of science: that we never know anything completely definitively, that we have some data now, and later we might have more data, and the later data might or might not support the conclusions we've drawn from our current data. So, even if you do what's described in the previous paragraph, you might still be met with consternation if you told a statistician "I've accepted my null hypothesis that $p = 0.55$," because it would signal to them that you haven't completely understood this very important principle (or, even worse, that you haven't bought into it). If you understand and buy into this principle in the philosophy of science, I would encourage you to use this peculiar language as it's been established.

> I got curious and decided to dig into the Wikipedia article on the Clopper-Pearson interval that the book had mentioned. After skimming the article I was wondering how the Clopper-Pearson interval can somehow simultaneously have a 95% confidence interval rating but yet also be wider than a traditionally calculated 95% confidence interval? Wouldn't that change the percentage of the confidence interval then?

Great question! I like that you're being curious and looking up things outside the book ☺

Based on some calculations I've just done in R, I think it's actually not true that the Clopper-Pearson interval is necessarily wider than the "traditional" confidence interval, if by "traditional" we mean the one that we learned about in our textbook (on the Wikipedia page, the "traditional" one is called the "normal approximation interval"). What Wikipedia gives an example of is the Clopper-Pearson interval being wider than the *Wilson* interval (which is yet another variant of the confidence interval, but is probably not worth dwelling on too much here).

In any case, I think the substance of your question still stands, because it is true is that the "traditional" confidence interval and the Clopper-Pearson interval are not exactly the same width. Even if we're working with a fixed significance level, sometimes the "traditional" one is wider, and sometimes the Clopper-Pearson one is wider. The reason for this difference in width is that, when we make a "traditional" confidence interval, we assume the success-failure condition is satisfied and use the fact that our sampling distribution is *approximately* normal because the success-failure is satisfied. The Clopper-Pearson interval doesn't make this approximation and uses the *actual* sampling distribution (which is a binomial distribution). Sometimes the normal approximation results in a slightly over-wide interval, and sometimes it results in a slightly under-wide interval. But the difference between the widths are guaranteed to be quite small if the success-failure condition is satisfied.

You might ask why we use the "traditional" method at all if it's only an approximation (and since it only works when the success-failure condition is satisfied). I think the answer here is just because doing computations with normal distributions is actually much easier

for computers than doing computations with binomial distributions!

> I'm having a really hard time keeping track of all the different equations. Is there a resource somewhere out there that has an organized index of all the equations and their purpose/situations where we'd use them?

You should make this resource — and it might even be your handwritten sheet of notes that you get to use for your quiz! I've often found that trying to organize a large amount of information into a small condensed form has been very beneficial for my learning (far more beneficial than just using a pre-existing resource of this form). The table at the top of this document is an example of one way that I've tried to condense information in a way that feels manageable to me. You should try to come up with your own ways of condensing information like this.