Select Reading Question Responses (9/9/2021)

> Is there a general rule about which sample group is p_1 and which is p_2 in a situation where you are just asked the difference between the two? Or would it not matter which way you stated your answer?

It doesn't matter; setting it up either way is fine!

> If one of two groups fails the success-failure test, is it still possible to calculate a confidence interval?

Yes, but we won't discuss it in this class. When the success-failure condition is not satisfied, all that means is that some binomial distribution that appears in the sampling distribution isn't well approximated by a normal distribution. That means you just have to use the binomial distribution directly to model the sampling distribution, instead of trying to approximate it with a normal distribution. The calculations involved are a little ugly, but it's possible to do.

> In 6.2.5 of the notes, it explains how the equation for standard error for 2 sample proportions is derived by saying "If X and Y are two independent random variables with variances σ_x^2 and σ_y^2, then the variance of $X - Y$ is $\sigma_x^2 + \sigma_y^2$," but I am still confused about why you would *add* these variances if you are finding the variance of the *difference*?

> Why the standard error for the difference two sampling distributions is the square root of the sum of the two standard errors?

An intuitive way of thinking about this is as follows. You might think of the variance of a random variable as some measure of the uncertainty of the value of that random variable. If we have some uncertainty about X, and we have some uncertainty about Y, we have *even more* uncertainty about $X - Y$ than we did about X and Y individually, just because $X - Y$ involves both X and Y! It doesn't make sense to *subtract* the uncertainties, since that would mean that we're *less* uncertain about $X - Y$ than we are about X and Y individually.

More formally, this goes back to that formula on page 123. Notice that $X - Y = 1 \cdot X + (-1) \cdot Y$ is a linear combination of X and Y, so that formula tells us that

$$\mathrm{Var}(X - Y) = 1^2 \cdot \mathrm{Var}(X) + (-1)^2 \cdot \mathrm{Var}(Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Why would you want to check if the difference between p_1 and p_2 is anything other than 0?

I'm a bit confused on when we'd use a null hypothesis H0 where $p_1 - p_2 =$ a non-zero number.

Let's say I'm interested in the difference between the percentage p_1 of current math majors in the US who identify as female to the percentage p_2 of current physics majors in the US who identify as female. I have historical data suggesting that 20 years ago, there were 10% more math majors who identified as female than physics majors. I might then want to know: has this difference in proportions changed since 20 years ago?

In this case, the appropriate null hypothesis is that this proportion has not changed, ie, that $p_1 - p_2 = 0.10$ still (and the alternative is that $p_1 - p_2 \neq 0.10$).

Probably it's most common to consider the case where the null hypothesis is that the difference in proportions is 0, but the method we've learned does work just fine for other situations as well.