Select Reading Question Responses (9/13/2021)

> What does it mean exactly when talking about "levels" of a categorical variable?

The *levels* of a categorical variable are the values that the categorical variable can take. For example, if you have a categorical variable `eyeColor` describing eye colors in a sample of Americans, its levels would be `brown`, `blue`, `green`, `gray`, ...

> I noticed some similarities between the conditions for a chi-squared test and the conditions for a nearly normal distribution. Both necessitate an independent sample, and meeting some parameter that requires a certain number of observations/cases. It seems like these guidelines for investigating data are particularly important, and foundational to a lot of statics concepts we've looked at thus far.

I agree! I would just propose rephrasing your first sentence more precisely as follows: "The conditions required for the sampling distribution of the chi-square test statistic to roughly follow a chi-square distribution are similar to the conditions for the sampling distribution of a proportion (or a difference of proportions) to roughly follow a normal distribution."

> For the sample size condition for chi-square tests, why must the expected counts be greater than 5? Why did they use 5 instead of a different number?

> Why is the sample per scenario for a chi-squared test required to be 5?

Good question! You should in fact find it a little strange that the textbook uses 10 for the success-failure condition but then 5 when talking about chi-square. For example, when you're testing the hypothesis H_0 that the distribution of a single binary variable is equal to something, the calculation you do when you do a p-value test using the method of 5.3–6.1 is actually identical to the calculation you would do when doing a chi-square test using the method of 6.3. It would be better if they used the same bound here, and I'm not sure why they went with different numbers.

> Will chi-square tests replace confidence intervals moving forward, or just be along with them?

When we studied single proportions and differences of proportions, we were able to both construct confidence intervals as well as do p-value tests. For chi-square tests, we'll only learn to do p-value tests. These p-value tests generalize the p-value tests for proportions and differences of proportions, but the confidence intervals for proportions and differences of proportions are still independently useful!

> I was a bit confused by how we can establish independence in chi-square test. Should we just assume that a random sampling method = independence?

> How can you determine if it is appropriate to assume that classmates, spouses, etc. . . do not influence each others decisions in a sample and that independence is satisfied?

If you know that your sample is a simple random sample taken from a very large population, you know definitively that independence is satisfied. Otherwise, you have to use your judgment about whether or not independence is satisfied, and your judgment might defer from other people's judgments.

For example, in the setup of exercise 6.33, the textbook decides that independence might be reasonable. I'm not sure I really agree with this. For example, probably the percentage of students who are willing to read a textbook entirely online (not purchasing a hard copy and not printing it out themselves) is increasing over time, so one class at a given point in time is not a random sample of all textbook uses. If a student sees many of their classmates using the textbook in one format, they might feel more comfortable using the textbook in the same format; in other words, one student's decision might not be completely independent of everyone else's decision. Probably you can come up with other examples of reasons why independence might not be satisfied.

Anyway, the point is that the judgment of whether or not independence is satisfied in a given situation can be subjective. The only situation where you know for sure that there's no reason to worry is if the sample was collected by generating some random numbers and taking a simple random sample from a very large population, but that's not easy to do in practice. In most realistic situations, if you think hard enough you'll probably be able to find some reason to worry about independence not being satisfied. You might still go ahead with the statistical test even if you find a reason to worry, just to see what it says, but the results of the test should be taken with a grain of salt because you know that one of the hypotheses of the test might not have been satisfied.