Select Reading Question Responses (9/15/2021)

> When checking the three conditions required to run a ANOVA test, how do you determine if the variability across groups is equal?

You can eyeball variability by making side-by-side box plots or hollow histograms.

You also have two numerical measures of variability: standard deviation, and IQR. You could compute standard deviation within each group and check to see that all of those standard deviations are similar. Or you could compute IQRs within each group and check to see those are all similar.

> If one or a few of the distributions appear skewed does this mean that the normality condition [for ANOVA] is not met?

Technically, yes, the condition is not met. But you can be a bit flexible about this! If a couple of the distributions are only kind of skewed, maybe it's okay to proceed with ANOVA anyway. But, if there are some distributions that are *really* skewed, then ANOVA won't give reliable results.

> I struggled to understand exercise 7.39(c).

That's okay, don't bother ☺It's good to have tried, but the calculations are somewhat horrible and not worth dwelling on at this stage. I think the important things about ANOVA for now are that you know (a) when it's an appropriate test to use, and (b) how to interpret the results of the test.

> In the book they mention how we shouldn't inspect the data before picking groups to analyze, but how can we decide which groups to analyze without having some familiarity with the data and implications we'd want to inspect? I'm unclear on where the line is between simply looking at datasets to find something to analyze and true data snooping.

The line probably is probably a little fuzzy. But if you have a clear understanding of how to interpret p-values, you'll likely be able to avoid committing any serious statistical no-nos. Remember that p-values represent the probability of seeing "patterns" in sample data when in fact there is no underlying pattern in the population. If you go looking for enough patterns, you'll probably be able to find something in your data just by random chance. Here's an xkcd comic that describes something relevant: even if jelly beans have nothing to

do with acne, if you do enough tests, one of your tests will find a correlation just by random chance.

To be completely on the *safe* side, you should precisely frame the hypothesis test you want to run first, then collect data, and then run the hypothesis test you had planned. The result of that test is probably useful.

Entirely on the *bad* side would be if you have a hypothesis test in mind, you collect data to run that test, you run the test, and you don't find a statistically significant result. Then you start analyzing the data looking for anything that *is* statistically significant in your data, just so that your data collection wasn't in vain. That is *decidedly* data snooping. Don't do that.

Somewhere in the gray area is if you collect data before you're sure what hypothesis test you'll be running. You should recognize that if you do this, you're *at risk* of data snooping. If you start analyzing your data and look for more than 20 patterns, it's completely reasonable to expect at least one of the corresponding hypothesis tests to come back with a p-value of less than 0.05 — just by random chance! It might be an interesting observation, but you shouldn't take any serious action based on that observation alone. Instead, you should probably treat this observation as a call for further data collection: you now have a precise hypothesis test you're interested running, so you can go back, collect new data specifically to run that test, and see what that new data says.

> In the textbook, in section 7.5.6, it states that "it is possible to reject the null hypothesis using ANOVA and then to not subsequently identify differences in the pairwise comparison." This sentence just made me really curious of what type of data or studies would experience this situation.

Actually, something similar happens with χ^2 tests — and it might be easier to understand what's going on in the χ^2 setting.

Let's imagine we have an alien planet where we have four kinds of aliens: red, green, blue, and pink. The four colors are equally distributed in the population. Suppose further that a disease starts running through this alien population. We might be interested in understanding how this disease correlates with color. Suppose we take a sample of 1000 diseased aliens and find the following color distribution in our sample.

Red	Green	Blue	Pink	Total
230	225	275	270	1000

Let's say we first run a chi-square test to test the hypothesis H_0 that all colors are equally represented among diseased aliens. We have the following expected counts:

	Red	Green	Blue	Pink	Total
Observed	230	225	275	270	1000
Expected	250	250	250	250	1000

Then we compute our test-statistic χ^2 using the usual formula of

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

and find $\chi^2 = 8.2$. We have 3 degrees of freedom here, so our p-value is `1 - pchisq(8.2, df=3)`, which evaluates to 0.042. Using the standard significance level of $\alpha = 0.05$, we reject H_0.

In other words, this test leads us to think that certain colors might be over- or under-represented among diseased aliens. So now we ask some follow-up questions. First up, let's ask: is it true that a quarter of all diseased aliens are red? In other words, let p_R be the proportion of diseased aliens who are red. We're testing the null hypothesis H_0 which says $p_R = 0.25$. This test uses the process described in 5.3–6.1. We calculate our test statistic

$$Z = \frac{0.23 - 0.25}{\sqrt{\frac{0.23 \cdot 0.75}{1000}}} \approx -1.46$$

which yields a p-value of `2*pnorm(-1.4)`, which evaluates to 0.144. We do not find evidence to reject H_0.

We repeat this with green now. Our test statistic is $Z \approx -1.83$ and the p-value is 0.068, so we don't reject. Now we do it with blue. Our test statistic is $Z \approx 1.83$ and the p-value is 0.068, so we don't reject. And finally we do it with pink. Our test statistic is $Z \approx 1.46$ and our p-value is 0.144, so again we don't reject.

In other words, even though our χ^2 test suggests that colors are not evenly distributed among diseased aliens, the individual proportion tests all say that the proportions of each color we observed in our sample of aliens who have this disease are not statistically significantly different from 0.25.

Bizarre?! Well, maybe not — if you remember how p-values should be interpretted! Suppose it's true that the colors are evenly distributed among diseased aliens. My individual p-value tests tell me that, if I take a sample of 1000 diseased aliens, there's a:

- 14.4% chance of seeing 230 red aliens in my sample,[1]
- 6.8% chance of seeing 225 green aliens in my sample,
- 6.8% chance of seeing 275 blue aliens in my sample, and
- 14.4% chance of seeing 270 pink aliens in my sample.

[1]This isn't quite right. It's rather a 14.4% chance of seeing a red alien count that's at least as distant from 250 as the count that we actually saw, which is 230. But that's more verbose, so let me simplify.

What is the probability that *all four* of those events happen? It's not just the minimum of those probabilities! Alien color counts are also not independent (eg, if I know that there are 1000 red aliens in my sample, I know for sure that there must be 0 blue aliens in my sample), so it's also not the product of those four probabilities. The probability that all four of those events happen is rather the p-value of my χ^2 test, which is 4.2%.