ANOVA Part 2: Dealing with Intersectional Groups: Crash Course Statistics #17

Crash Course: Statistics
https://youtube.com/watch?v=wo1xlefg5KI
https://nerdfighteria.info/v/wo1xlefg5KI

Hi, I'm Adriene Hill, and welcome back to Crash Course Statistics.

When comparing groups, there isn't always one single box that we can put someone into. You might be someone's child, but also a parent, and a partner.

You have an ethnicity or maybe a job title, and maybe you're a competitive dog groomer And it's not just people that belong in multiple groups. Your watch might be a smart watch, but also an Apple product, and something that's rose gold. Things and people belong to multiple groups.

And those groups can overlap or interact. So today, we're going to take a look at ANOVAs that include more than one grouping variable. INTRO We want to look at sedan prices to figure out how they're affected by manufacturer and color.

For now, we'll assume that those two factors are independent of each other -- they don't interact. And for this, we use a Factorial ANOVA, which can have just two grouping variables--like car manufacturer and car color--up to hundreds of grouping variables. In this case we're going to look at Toyotas, Hondas, Chevrolets, and Lamborghinis.

And include the colors blue, red, silver, and white. A Factorial ANOVA does almost exactly what a regular ANOVA does: it takes the overall variation--or Sums of Squares--and portions it out into different categories. If we're interested in how car manufacturer and color affect price, we first calculate the overall variation in the dataset called the Sums of Squares Total.

We do this by summing up all the squared distances between each car price and the mean overall car price. Then once we know the total variation in the data set, we set out to use manufacturer and color to explain why these sedans have different prices. Our proposed model looks something like this: Which tells us that we think the price of a car is some baseline cost plus an adjustment for who made the car and what color it is.

And like before, we know that we won't always be exactly spot on. So to complete the General Linear Model form we add an error term which represents how "off" our guess was from the actual price of each car. We're going to use our model and the error to create F statistics for each part of our model, as well as the model as a whole.

The F-statistic is a ratio between the scaled Sums of Squares for a variable and the scaled Sums of Squares for the Error. We call these scaled versions of the Sums of Squares, Mean Squares. When we create these models using statistical software like R, or Python, or even Excel, we'll usually get what we call an ANOVA table as an output.

And the table will give us all the information we need to answer our questions. We can see in this table that the p-value for color is way bigger than our alpha cutoff of 0.05. So we did not find evidence that color has a significant effect on car price.

On the other hand, we did find evidence that manufacturer has a significant effect on car price. And I guess we knew that. But just like with our t-tests, we know that a significant F-test only means this result is statistically significant.

It doesn't always mean it's practically significant to you. If there's a statistically significant effect of manufacturer on car price but it turns out it's only about a $20 difference well that might not have a huge impact on whether or not you decide to buy a particular car. So we need another measure of effect size.

Something that helps us understand how big the effect really is in more practical terms. There are many different measurements of effect size for ANOVAs, but they all share similar ideas, so we'll show you just one: eta squared. Effect sizes try to tell us how large an effect is compared to how much variation we generally expect.

In a t-test, we recognize that a new negotiating technique that only increases salaries by about $2 a year is not that exciting because people's salaries generally vary way more than $2 a year. Eta squared does the same thing for us. To calculate eta squared, you take the Sums of Squares for your particular effect--in this case, car manufacturer--and divide it by the Total Sums of Squares for your entire data set.

Eta squared is always between 0 and 1. And its interpretation is like the interpretation of R-squared. Eta squared tells you the proportion of total Variation that's accounted for by your specific variable.

So here, in our made up data, we see that 46% of the variation in car price is accounted for by manufacturer. Sounds like a lot. But effect size is something that the person analyzing the data will have to interpret for themselves.

It can be pretty subjective. We might also be interested in how well our entire model--with both manufacturer AND color--can predict sedan prices. Say we were designing this model for a car selling website so that they can tell customers what they should expect to pay for their dream car.

They might ask us to calculate eta squared--which is here equivalent to R-squared--for our entire model. And we can do that the formula is exactly the same. So, now we know that our entire model with both factors accounts for about 48% of the variation in the data.

If we could explain 100% variation, we could perfectly predict car price. So 48% means we can predict about half the variation while the rest is explained by other factors we did not include in our model, like size of car and style of car, as well as error. We predicted car price using manufacturer and color with a model assuming that these two factors are independent.

But maybe color has very little effect on the price of cars from less expensive brands like Toyota, Honda, or Chevrolet, whereas if you're getting a fancy Lamborghini, color may have an effect. A lot of people want that bright orange Lambo. If this were the case, then these two factors are not independent.

The effect of color depends on which manufacturer made the car. That's called an interaction because the two factors interact with each other. And these interactions can be really important.

Let's move on from cars and look at how professional and novice olive oil tasters rate olive oil. You're opening an olive oil shop. You've already traveled the world in search of the best olives, you've learned how to extract and process the best oil.

But as you're putting the finishing touches on your storefront and marketing plan, you run into an issue. You're not sure how to bottle your oils. You could shell out a lot of money for very Instagrammable fancy bottles or save some money and go with a simpler bottle (letting your oil speak for itself).

Since you've been watching Crash Course Statistics, you decide to run a small experiment. First, you gather two groups of people: olive oil experts and olive oil novices since those are your two main customer groups. Then, you give them your delicious olive oil from both a fancy and a plain bottle, and ask them to rate their overall impressions.

Once you collect your data, you conduct a TWO-WAY ANOVA, just like the one we did earlier. This time, our TWO factors are expertise and bottle style. Two, hence two-way ANOVA.

But you're curious to see whether expertise and bottle style interact. So you add one more thing to your model, the interaction Term. We won't dwell on the math here, but this new interaction term is calculated similarly to all our other terms.

Since there are 4 different combinations of our two factors--expert with fancy bottle, expert with plain bottle, novice with fancy bottle, novice with plain bottle-- we calculate the sum of the squared distance between the mean of each of these 4 groups, and the overall mean for each point. This is sometimes called the Sums of Squares Between Groups. Also, SSB - Sums of Squares Between Groups.

Then from the Sums of Squares Between Groups, we subtract the sums of squares for each factor in the interaction: expertise and bottle. Sums of Squares Between Groups tell us how much variation is explained by coming from one of the four possible combinations of olive oil expertise and bottle type. When we subtracted the main effects of Expertise and Bottle Type, we were left with the amount of variation explained by how these two factors influence each other.

Here we can seen the means of all four combinations of Expertise and Bottle Type. This type of plot is called an interaction plot, because it shows how these two factors interact. The blue line represents Experts, and the red line, Novices.

You can see that experts rated both bottles of olive oil similarly, they weren't swayed by the fancy bottle. But novices rated olive oil in the fancier bottles higher than olive oil in the simple ones. It seems like the effect of bottle style on olive oil ratings is different depending on whether you're an expert or a novice.

This indicates that there's an interaction between these two factors. If there were NO interaction between Expertise and Bottle Type, we'd expect the red and the blue line to be approximately parallel. This would tell us that regardless of expertise, bottle type has a similar effect. (In this case, both prefer the fancy bottle.) But, we've only looked at graphs so far.

Let's pull up the ANOVA table for this model. Based on our table, we can see that neither Expertise alone, nor Bottle Type alone are significant but their interaction is. When we look at how Experts rate both bottle types, and Novices rate both bottle types, we can see a clear difference, represented by the different slopes of our red and blue lines.

And just like before, we can take our significant effects and calculate an effect size for them, so that we can see how practically significant it is. In this case, the amount of variation in our data due to the interaction between expertise and bottle type. To get effect size, we just divide by the total variation.

In our last example, we talked about eta squared, which is one way to calculate effect sizes for ANOVAs, and is analogous to the R^2 formula we talked about for regression. To calculate eta squared, you just take the Sums of Squares for your desired effect, and divide by the total sums of squares. In this case, the interaction effect of bottle type and expertise accounts for about 9.14% of the total variation in the data.

Effect sizes tell you something about the magnitude of an effect, but it's up to you--or whoever is analyzing the data--to decide whether an effect of that size is useful. In our model, we only had one significant effect: the interaction. But occasionally we'll see other significant effects.

Single variables, like Bottle Type and Expertise, are called main effects. Statistically significant main effects are important, but when you interpret them, you need to do so with caution. For example, if we looked at a study of an allergy medication, we might observe a significant main effect of medication on allergy symptoms.

Which means that on average, people who took the medication had less severe symptoms than those who didn't take it. But, we also recorded whether or not the subjects had a certain allergy related gene, gene Y. It turns out that there's a significant interaction between allergy medication and whether or not you have gene Y.

If you DO have gene Y, the medication doesn't work that well. In fact, you'll feel about the same. But if you DON'T have gene Y, it works incredibly well all of a sudden your sneezes are gone!

If you told everyone that this allergy medication worked….it wouldn't quite be the whole truth. That significant interaction told us that while on average people displayed fewer symptoms on the medication, that allergy relief is different depending on whether you have gene Y. The different slopes for each of our lines in this interaction plot demonstrate how the two groups respond differently.

Back to your olive oil shop. Looking at the data you have--seems like you should go with the fancy bottles. The experts won't be swayed but the rest of your customers will like all the embellishment.

And there's only a couple olive oil professionals in your town. People, cells, animals, and pretty much anything we might be interested in measuring, are parts of multiple groups. So it's important to have the tools to consider multiple groups together with a statistical model.

They allow us to have a richer understanding of how certain things might interact. Like your gender and your ethnicity and your pay. Or your age and generation and favorite Slurpee flavor.

Thanks for watching, I'll see you next time.