ANOVA: Crash Course Statistics #33

Crash Course: Statistics
https://youtube.com/watch?v=oOuu8IBd-yo
https://nerdfighteria.info/v/oOuu8IBd-yo

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

In many of our episodes we've looked at t-tests, which among other things, are good for testing the difference between two groups. Like people with or without cats.

Families below the poverty line...and families above it. Petri dishes of cells that are treated with a chemical and those that aren't. But the world isn't always so binary.

We often want to compare measurements of MORE than two groups. Things like ethnicity, medical diagnosis, country of origin, or job title. So today, we're going to apply the General Linear Model Framework we learned in the last episode to test the difference between multiple groups using a new model called the ANOVA.

INTRO The GLM Framework takes all the information that our data contain, and partitions it into two piles: information that can be explained by a model that represents the way we think things work, and error, which is the amount of information that our model fails to explain. So let's apply that to a new model: the ANOVA. ANOVA is an acronym for ANalysis Of VAriance.

It's actually very similar to Regression, except we're using a categorical variable to predict a continuous one. Like using a soccer player's position to predict the number of yards he runs in a game. Or using highest completed degree to predict a person's salary, note that this alone isn't evidence that getting a degree causes a higher salary, just that knowing someone's degree might help estimate how much they get paid.

Like Regression, the ANOVA builds a model of how the world works. For example, my model for how many bunnies I'll see on my walk into work might be that if it's raining I'll see 1 bunny, and if it's sunny, I'll see 5. I walk through a bunny preserve... 1 and 5 are my predictions for how many bunnies I'll see, based on whether or not it's raining.

Yesterday it rained. And I saw two bunnies! My model predicted 1, and my error is 1.

And we can represent this model as a sort of Regression where there are ONLY two possible values that the Variable Weather can have. 0--if it rains--or 1--if it doesn't. In this case, expected number of bunnies on a rainy day is 1 and beta is the difference between the two means, 5-1 = 4. Which means our ANOVA model looks like this: In a Regression we did a statistical test of the slope and that's what this simple ANOVA is doing too.

Since we assigned rainy days to be coded as 0, and sunny days as 1, the change in the X-direction is just one (1-0). So the slope of this line is the difference between mean bunny count on sunny days, five, minus mean bunny count on rainy days, one. This difference of 4 is the change in the Y direction.

We test this difference in the same way that we tested the regression slope. And this slope tells us the difference between the means of the two groups. Usually we'll like to think of this slope as the difference between two group means.

But, knowing that our model treats it like a slope helps us understand how ANOVAs relate to regression. In a regression the slope tells you how much an increase in one unit of X affects Y. Like for example, how much an increase of 1 year increases shoe size in kids.

An ANOVA actually does the same thing. It looks at how much an increase from 0 (rainy days) to 1 (non-rainy days) affects the number of bunnies you'd see. Now...to another example.

Let's look at the ratings of various chocolate bars based on the type of cocoa bean used. We'll use a data-set you can find at Kaggle.com courtesy of Brady Brelinski. Our three groups are chocolate bars made with Criollo beans, Forastero beans, or Trinitario beans.

Chocolate making is complex, so we took a small sample of bars that only contained 1 of these three beans. And the chocolate taster used a scale--with 5 as the highest score --transcending beyond the ordinary limits. 1 was "mostly unpalatable"... But is there really "mostly unpalatable" chocolate out there?

We want to know if the type of bean affects our taster's ratings. To find out, we need the ANOVA model! Like Regression, we can calculate a Sums of Squares Total by adding up the squared differences between each chocolate rating, and the overall mean chocolate rating.

This gives us our Sums of Squares Total, or SST. If that sounds like how we calculated variance, that's because it is! SST is just N times Variance.

This Sum represents the total amount of variation, or information, in the data. Now, we need to partition this variation. When we previously used a simple linear regression model, we partitioned this variation into two parts: Sums of Squares for Regression, and Sums of Squares for Error.

And the ANOVA does the same thing. The first step is to figure out how much of the variation is explained by our model. In an ANOVA--what we're using here--our best guess of a chocolate bar's rating is its group mean.

For bars made with Criollo beans 3.1, Forastero beans 3.25, and Trinitario beans 3.27. So we sum up the squared distances between each point and its group mean. This is called our Model Sums of Squares (or SSM) because it's the variation our model explains.

So now that we have the amount of variation explained by the model. In other words, how much variation is accounted for if we just assumed each rating value were it's group mean rating. We're also going to need the amount of variation that it DOESN'T explain.

In other words, how much ratings vary within each group of Cacao beans. So, we can sum up the squared differences between each data point and its group mean to get our Sums of Squares for

Error: the amount of information that our model doesn't explain. Now that we have that information, we can calculate our F-statistic, just like we did for regression. The F-statistic compares how much variation our model accounts for vs. how much it can't account for. The larger that F is, the more information our model is able to give us about our chocolate bar ratings. Again, SSM is the variation our model explains and SSE is the variation it doesn't explain. We want to compare the two. But we also need to account for the amount of independent information that each one uses. So, we divide each Sums of Squares by its degrees of freedom. Our ANOVA model has 2 degrees of freedom. In general, the formula for degrees of freedom for categorical variables (like cocoa bean types) in an ANOVA is k-1, where k is the number of groups. In our case we have 3 groups. Our Sums of Squares for Error has 787 degrees of freedom because we originally had 790 data points, but we calculated 3 means. The general formula for degrees of freedom for your errors is n minus k where n is the sample size and k is the number of groups. For our test, we got an F-statistic of 7.7619. This F-statistic--sometimes called an F-ratio--has a distribution that looks like this: And we're going to use this distribution to find our p-value. We want to know whether the effect of bean type on chocolate bar

ratings is significant. In this case we have a p-value of 0.000459. Small enough to reject the null. So we've found evidence that beans influenced the chocolate bar ratings. A statistically significant result means that there is SOME statistically significant difference SOMEWHERE in the groups, but it doesn't tell you where that difference is. Maybe Trinitario is significantly different from Criollo but not Forastero beans.. An F-test is an example of an Omnibus test, which means it's a test that contains many items or groups. When we get a significant F-statistic, it means that there's SOME statistically significant difference somewhere between the groups, but we still have to look for it. It's kinda like walking into your kitchen and smelling something reallllllly stinky. You know there's SOMETHING gross, but you have to do more work to find out exactly what is rotting... We already have tools to do this, in statistics at least, because you can follow up a significant F-test in an ANOVA with multiple t-tests, one for every unique pair of categories your variable had. We had 3, which means we only need to do 3 t-tests in order to find the statistically significant difference or differences. To conduct these T-tests, we take just the data in the two categories for that t-test, and calculate the t-statistic and p-value. For our first t-test we just look at the bars with Trinitario and Criollo beans. First, we follow our Test statistic general formula: We take the difference between the mean rating of chocolates made with Trinitario and Criollo beans. And divide by the standard error. And once we do this for all three comparisons, we can see where our statistically significant differences are. It looks--from our graph--like ratings of chocolate bars made with Criollo beans are different...in a statistically significant way... than those made with Trinitario or Forastero beans. And our graph and group means show that Criollo bars have a slightly lower mean rating. But bars made with Trinitario beans are NOT statistically significantly different than those made with Forastero beans. So our ANOVA F-test told us that there WERE some differences, and our follow up t-tests told us WHERE they were. And this is interesting. Criollo beans are generally considered a delicacy and of a much higher quality than Forastero. And Trinitario are hybrid of the two. But we found...in this data set... that Criollo bars had statistically significantly lower ratings. This might be because we excluded bars with combinations of our three bean types...or because the rater has a different preference...or even be caused by some other unknown factor that our model does not include. Like who made the chocolate. Or the country of origin of the beans. We can also use ANOVAs for more than 3 groups. For example, the ANOVA was first created by the statistician R. A. Fisher when he was on a potato farm looking at studies of fertilizer. In one of the first experiments he described, he looked at 12 different species of potato and the effect of various fertilizers. Let's look at a simple version of Fisher's potato study. Here we have 12 different varieties of potato. We'll represent each of them with a letter A through L. There are 21 of each of the potato plants, for a total of 252 potato plants. We give our future french fries about a season to grow, then we dig them up and weigh each one. This graph shows the potato weights that we recorded, as well as the total mean potato weight and each group mean potato weight. Using these numbers, we can calculate our Total Sums of Squares, Model Sums of Squares, and Sums of Squares error. We're going to let a computer do that for us this time. And our computer spit out this: the degrees of freedom, sums of squares, mean squares, F-statistic, and p-value. This is called an ANOVA table and it organizes all the information our ANOVA models give us. Here we can see that our Model had an F-statistic--or F-value--of around 3, and a p-value of 0.000829. So we reject the null hypothesis. We found evidence that the potato varieties don't all have the same mean weight. But since this was an Omnibus test, our statistically significant F-test just means that there is some statistically significant difference somewhere in those 12 potato varieties. We don't know where it is. In that way, ANOVAs can be thought of as a first step. We do an overall test that tells us whether there's a needle in our haystack. If we find out there is a needle, then we go looking for it. However, if our test tells us there's no needle, we're done. No need to look for something that probably doesn't exist. But you can see that this significant F-statistic for potato varieties will require MANY follow up tests. 12 choose 2. Or 66.

We showed a lot of calculations today, but there's two big ANOVA ideas to take away from this. First, a lot of these different statistical models are more similar than they are actually different. ANOVAs and Regressions both use the General Linear Model form to create a story about how the world might work. The ANOVA says that the best guess for a data point--like the rating of a new chocolate bar--is the mean rating of whatever Group it belongs to. Whether that's Criollo, Trinitario , or Forastero. If we don't know anything else, we'd guess that the rating of a Criollo chocolate bar is the mean rating for all Criollo bars. Also, an ANOVA is a great example of filtering. If there's no evidence that bean type has an overall effect on chocolate-bar ratings, we don't want to go chasing more specific effects. Our time is precious...and we want to use it as best as we can. So we have more time out in the world...to look for bunnies. Thanks for watching, I'll see you next time.