Chi-Square Tests: Crash Course Statistics #29

Crash Course: Statistics
https://youtube.com/watch?v=7_cs1YlZoug
https://nerdfighteria.info/v/7_cs1YlZoug

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

When you're buying a new car, a new house, or a new pair of jeans, you want to make sure you find a good fit. Statistics are the same.

You want to make sure your models or preconceptions are a good fit for the data you have. One way to do that is by comparing our observations to our expectations, just like we've done in statistical tests over the last couple of episodes. Today, we're going to take a break from looking at continuous variables like height and IQ and see how we can measure the fit of categorical variables like hair color, or academic degree, or tax bracket.

Turns out a chi-square model fits like that perfect pair of jeans. INTRO Back in one of our data visualization episodes, we talked about frequency tables, which tell you the counts--or frequencies--of different categories. Maybe you remember pasta madness?

When these tables look at two different categorical variables, like favorite pasta and favorite pasta sauce, we often call them contingency tables or cross tabulations. These tables can help us understand the discrete distribution of one variable, or the relationship between two variables. Just by looking at the data we can see major differences like most people prefer red sauce....

Or the distribution of favorite pastas seems to be different for people who like red sauce vs. white sauce. But sometimes it's hard for us to see smaller differences, or to be sure whether differences we see are statistically meaningful. Just like when we were comparing means, we need a test.

A statistical test that can help us extract the "signal from the noise". Statistical tests usually take this form. And the test we'll use here -- Chi Square -- is only slightly different from the others we've used so far.

The idea in the numerator--looking at the difference between what we observed and what we'd expect if the null is true--is exactly the same. The denominator--average variation--is a little different. Let's figure out why...in an example.

A new video game has come out, League of Lemurs, and it has taken the world by storm. League of Lemurs has hundreds of unique characters you can play as with four main types: Healers, Tanks, Assassins, and Fighters. The official League of Lemurs development team says that on average they see 15% of players choosing Healers, 20% choosing Tanks, 20% choosing Assassins, and 45% choosing Fighters - because who doesn't want to be a fighting lemur.

But you wonder whether that distribution holds in the top ranks of LOL players. The null hypothesis is that the percentages that LOL gave you are correct: And the alternative hypothesis is that at least one of these percentages or proportions are incorrect. So you record 20 games with 10 players each and count the number of Healers, Tanks, Assassins and Fighters.

The data you collected looks like this: According to the numbers the LOL developers gave us, we'd have expected 30 Healers, 40 Tanks, 40 Assassins, and 90 Fighters. So These numbers aren't EXACTLY what we'd expect. But we have to ask whether they're different enough for us to consider it statistically significant.

We need a test statistic. Our general formula applies, with the numerator being the count we observed for each category minus the count we'd expect. But If you tried to add up all these differences: You'll always get zero.

So we need a better way to measure. Using a chi-square, we square them before adding them all up. For the denominator instead of a standard error, we just use the expected counts again.

In this case, the amount that a count deviates from its expected frequency should be scaled by the expected frequency. A deviation of 1 isn't as big of a deal if the expected count is 2000, but if it's 10...that deviation of 1 matters more. You might not think it's worth it to go back to the store to demand a refund if you're overcharged $1 for a $2000 laptop….but it might feel more worth it if you were overcharged $1 for your morning coffee.

When we add up all these calculations We get a chi-square statistic. In this case 3.958. Which, like our other test statistics, helps us quantify how well our sample data fits a certain distribution.

Usually the null distribution. Like a t-statistic, a chi-square statistic has a distribution that we can use to find a p-value. And like t-distributions, chi-square distributions change their shape as degrees of freedom change.

To find our degrees of freedom we have to think about what kinds of independent information we have. A frequency table, like the one we just used for our League of Lemurs example, has a certain number of cells. In this case we had four, one for each character type: That means we have four independent pieces of information: each of the four counts.

But as soon as we know the total counts--in this case the 200 players you recorded--the four values aren't ALL independent anymore. Because if you know 3 of those values and the total, then with some basic math you can figure out the fourth. So in this case, our degrees of freedom is the number of categories we have, 4 minus 1.

In this case, 3. Using our chi-square distribution of 3 degrees of freedom, we can now find our p-value! Our p-value here is 0.3, so if our cutoff was 0.05 we'd fail to reject the null.

The sample we took didn't give us any statistically significant evidence that the game developers' percentages were wrong. All Chi-square tests follow the same formula we just worked through. And there are three main ways that we use chi-square tests.

The one we just did is called a chi-square Goodness of Fit test, because we tested how well certain proportions fit our sample. One way to know that you're looking at a Goodness Of Fit Chi-square test is if it only has one row. We can have many categories, but we're only looking at one variable.

Like in our case, character class. And note: one thing we should always check when doing a Chi-square test is whether the expected frequency for every cell is greater than 5. If the expected frequency is lower than 5, the results of the chi-square test can be off. 5 is arbitrary, just like many of the cutoffs in Statistics, but it's widely accepted.

But chi-square tests aren't limited to analyzing just ONE categorical variable. They can even handle TWO. Like with the Test of Independence, the second type of chi-square test.

Tests of independence look to see whether being a member of one category is independent of the other. For example, let's look at the annual Nerdfighteria Survey -- a survey that Hank and John Green request their audience of nerdfighters take every year. We wanted to take a look at the two most important questions they asked last year: what Hogwarts houses nerdfighters were in AND if they liked pineapple on pizza.

What we want to know is whether Pineapple on Pizza Preference is Independent of Hogwarts House. In other words, does liking Pineapple on Pizza affect the probabilities of you identifying with each of the houses? Luckily, our writer, Chelsea has that data, and she took a random sample of 1000 Nerdfighter's responses so we could answer our questions.

She's a pineapple-loving Ravenclaw, for the record. But let's take a look at the data. Looks like there's a lot of Ravenclaw Nerdfighters.

Unlike our Chi Square goodness of fit test, we're not specifying an exact distribution for Hogwarts houses and comparing our two groups: Yes pineapple and No pineapple. In this situation, we're not too concerned with the exact distribution. We just want to know whether it's different for people who like and don't like pineapple on pizza.

A Chi-square test of Independence, can test whether or not one Variable--pineapple preference--is independent of another--Hogwarts House. And you'll soon see that the calculations we do here are the exact same for the third Chi-Square test: the Chi-Square test of Homogeneity. Test of homogeneity are looking at whether it's likely that different samples come from the same population.

For example you might want to know whether two samples of water are likely from the same lake based on the counts of fish, algae, and bacteria found in them. In essence they're testing similar things, and the calculations we're about to do are the same for both tests. Back to the Nerdfighters.

To calculate the Chi-Square statistic, we need our observed frequencies which we already have, and our expected frequencies, which we need to calculate. But it's not quite as straightforward as in the Goodness of fit test. First we'll need to calculate some row and column totals: We already know that there's 1000 total people, and we can count up all the people who don't like pineapple on their pizza to find that there's 479 of them, which means there must be 521 people who do like pineapple on their Pizza.

We have 3 independent pieces of information here, or 3 degrees of freedom. In general the formula for degrees of freedom for these chi-square tests is rows minus 1 times columns minus 1. Finally, we can calculate those expected frequencies.

Remember, the expected counts are what we would expect if the null hypothesis is true. In this test, the null hypothesis is that the distribution of Hogwarts House is the same for both pineapple lovers and haters. But our null hypothesis says nothing about WHAT that distribution is.

So we can calculate our expected frequencies by taking the total number of Gryffindors and dividing it by the total number of people: This gives us the expected percentage of people who are Gryffindors. Since there's 479 People in our sample who don't like pineapple, we expect 16.1% of them or about 77 of them to also be Gryffindors. Using this same math, we can calculate the expected frequency for all of our cells.

Once we have our expected frequency, we just need to use our Chi-square formula on each cell, and add them all up to get our Chi-Square Statistic: And with our chi-square distribution with 3 degrees of freedom, we can see that our p-value of 0.6 is very large compared to our alpha level of 0.05, so we fail to reject the null hypothesis that the distribution of Hogwarts Houses is the same regardless of pizza preference. If the null were true, we'd expect to see numbers as or more different than ours 60% of the time. So we don't have evidence that Hogwarts House is dependent on Pineapples on Pizza preference.

It's often useful to check our assumptions and to see if they're a good fit. Whether that's testing whether a population is distributed the way we think it is. Are there really the same proportion of Skittles colors in a bag?

Or whether two variables affect each other, like political party preference and cat and dog ownership. Since we, as humans, tend to categorize many things, from dog breed to hair color, it can be useful to check what we think about how and if those categories interact. Thanks for watching.

I'll see you next time.