Lab 6: Chi-Square Tests

PSC 103B - Winter 2024

Today, we will be learning about chi-square tests which can be used in a few different scenarios We will be going over each of the different scenarios, and the code to conduct these tests in R, today.

But first, let's take a look at the data that we'll be using today. This is data that I adapted from AggieData, which provides statistics on the university, including the student population Here is the website in case you're interested: https://aggiedata.ucdavis.edu/#student. I've changed the total number to 1000 just to make the math a little bit easier, rather than working with tens of thousands but the percentages are the same as reported.

	CLAS	CA&ES	CBS	COE	Total
Freshmen	276	147	173	111	707
Total	417	223	216	144	

Goodness of Fit Test

So this is the data that we have from UC Davis and say that we were interested in whether the 4 colleges were equally represented in our entry-level population. What would it look like if all colleges were represented equally?

There would be 250 in each college - does that roughly look like what we have here? Well we have 223 people in CA&ES and 216 in CBS, so that's not too far off, but we have only 144 people in COE, and 417 in CLAS. Our question is whether these differences are extreme enough to say that the colleges are *not* represented equally. That's what the chi-square goodness-of-fit test does: compares what we expected to what we observed and tries to see whether the differences are extreme enough to say our expectations were wrong.

Let's conduct the chi-square test. First, we need to write up our statement "the 4 colleges are equally represented" into a null hypothesis.

The chi-square GoF test normally writes the null hypothesis in terms of expected proportions H0: P = (P1, P2, P3, P4) where P is a vector or set of probabilities. If the colleges are equally represented, what proportions do we expect? We'd expect 1/4 of the students in CLAS, 1/4 in CA&ES, 1/4 in CBS, and 1/4 in COE

$$H_0 : P = (.25, .25, .25, .25)$$

What do you notice about what these numbers add up to? We need all possible categories to be represented, so all the probabilities need to add up to 1!

Let's save these probabilities in a vector to use later:

```
null_prob <- c(.25, .25, .25, .25)
```

And then what's our alternative hypothesis? Well, similar to an ANOVA, our alternative would be that *at least* one of these probabilities is not .25 so one of them could be different, or all of them could, we're not being specific.

$$H_A : P \neq (.25, .25, .25, .25)$$

Let's also make a vector for the observed frequencies of how many people are in each college.

```
obs_freq <- c(417, 223, 216, 144)
```

And we want to make a vector of frequencies that we would have expected if the null hypothesis was true in this case, we could just write out `c(250, 250, 250, 250)` because we have 1000 people and the math is pretty easy. But sometimes we won't have nice round numbers, so we can do it another way.

First, we need to write our total sample size:

```
N <- 1000
```

And we can multiply this by our expected probabilities to get the expected frequencies:

```
expected_freq <- N * null_prob

expected_freq
```

```
[1] 250 250 250 250
```

To conduct the GoF test, we need to compute some sort of error score to measure the difference between what we expected and what we observed, and then compare this to a distribution to see how big that difference is, and if it's big enough to reject the null.

The formula for this is $(O-E)^2/E$

```
diffs <- (obs_freq - expected_freq)^2 / expected_freq
diffs
```

```
[1] 111.556   2.916   4.624  44.944
```

So we have these "error" scores now, where bigger values represent bigger discrepancies.

To get our test statistic, we need to add up these error scores.

```
test_stat <- sum(diffs)
```

Does a larger test statistic make us more or less likely to reject the null?

More likely! Because a bigger test stat means the discrepancies were bigger. But to see if this test statistic is large enough we need to compare it to the chi-square distribution to make a proper judgement.

The chi-square distribution needs a df, where df = number of categories - 1.

So now we can compute a p-value by seeing the probability of our test statistic or something larger.

```
pchisq(test_stat, df = 3, lower.tail = FALSE)
```

```
[1] 2.461531e-35
```

Do we reject or fail to reject the null? Reject the null! The 4 colleges are not being equally represented in the 2022 class.

But do we know which college is not as expected? No - like the ANOVA, the chi-square test is an omnibus test. We know that one of the proportions is not as expected, but we don't know which one. Of course, we can look at the frequencies and try to make a guess (e.g., the CLAS seems to have way more people than expected), and there are post-hoc tests you can do to formally test it.

As always, we can do this very easily in R using the `chisq.test()` function The arguments are: `chisq.test(observed frequencies, p = expected probabilities)`.

```r
chisq.test(obs_freq, p = c(.25, .25, .25, .25))
```

```
    Chi-squared test for given probabilities

data:  obs_freq
X-squared = 164.04, df = 3, p-value < 2.2e-16
```

Notice that we assumed equal probabilities for this test, but we didn't have to; we could have expected whatever probabilities we wanted and then we would just need to specify them in `p`. We also need to make sure that the order of observed frequencies and p are the same.

Chi-Square Test of Independence

The other chi-square test we can do is the one when we have two categorical variables, and we're interested in testing whether or not they're related or dependent on each other.

In this case, let's say we were interested in testing whether which college a new student was a part of was related to whether they entered as a freshman or transfer student.

Remember that dependence means that knowing the value of one variable gives you an idea of what the value on the second variable is going to be. And independence means knowing the value of one variable doesn't tell you anything about the value of the other variable.

So in this case, independence would mean knowing whether someone entered Davis as a freshman or transfer student tells you nothing about what college they joined.

Here, we write out our hypotheses in words:

- H_0: Entry status and college are independent of each other
- H_A: Entry status and college status are not independent

Just like in the GoF test, the chi-square test of independence computes the difference between what we would expect if the variables were independent, and what we observed.

This is a bit more of a pain to do by hand, so for today we're just going to show how to do it in R.

To do the test in R, we now need to give R either 2 vectors of data for each category (e.g., a vector of each student's college, and a vector of each student's entry type) or a table / matrix of the observed frequencies. Since we don't have access to the raw data here, we're going to make a matrix of the observed frequencies.

Do you recall how to make a matrix? Use the matrix function, and give it your data, number of rows, and the number of columns.

```
obs_matrix <- matrix(c(276, 147, 173, 111,
                       141, 76, 43, 33),
                     nrow = 2, ncol = 4,
                     byrow = TRUE)

obs_matrix
```

```
     [,1] [,2] [,3] [,4]
[1,]  276  147  173  111
[2,]  141   76   43   33
```

I entered my data in one row at a time, so I specified `byrow = TRUE`; if you entered it as the columns, you can say `byrow = FALSE`.

```
chisq.test(obs_matrix)
```

```
    Pearson's Chi-squared test

data:  obs_matrix
X-squared = 18.592, df = 3, p-value = 0.000332
```

What is our p-value? And what do we conclude?

Our p-value is .0003, and since that's less than .05 we reject H_0. Therefore, entry status and college are dependent on each other.