Lab 6: Chi-Square Tests

PSC 103B

Marwin Carmo

- This data was 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.

	CLAS	**CA&ES**	**CBS**	**COE**	***Total***
Freshmen	276	147	173	111	707
Transfer	141	76	43	33	293
Total	417	223	216	144	1000

	CLAS	CA&ES	CBS	COE	*Total*
Freshmen	276	147	173	111	707
Transfer	141	76	43	33	293
Total	417	223	216	144	1000

- 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

- Are these differences extreme enough to say that the colleges are *not* represented equally?
- The chi-square goodness-of-fit test compares what we expected to what we observed and tries to see whether the differences are extreme enough to say our expectations were wrong.
- \(H_0\): The 4 colleges are equally represented.

- The chi-square GoF test normally writes the null hypothesis in terms of expected proportions.
- \(H_0\): 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?
- \(H_0\): P = (.25, .25, .25, .25)

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

- And what's our alternative hypothesis?
- Similar to an ANOVA, our alternative would be that *at least* one of these probabilities is not .25
- \(H_A\): P \(\neq\) (.25, .25, .25, .25)
- Let's also make a vector for the observed frequencies

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

- We want to make a vector of frequencies that we would have expected if the null hypothesis was true in this case.
- First, we need to write our total sample size:

```
1 N <- 1000
```

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

```
1 expected_freq <- N * null_prob
2 expected_freq
```

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

- We need to compute the difference between what we expected and what we observed.
- Then, we 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 \((O - E)^2 / E\)

Exercise

Compute the formula below using `obs_freq` and `expected_freq`, and save the results to an object named `diffs`.

We want to find \((O - E)^2 / E\).

```
obs_freq <- c(417, 223, 216, 144)
N <- 1000
expected_freq <- N * null_prob
# diffs <- ?
```

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

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

- Bigger values of these “error” scores represent bigger discrepancies.
- To get our test statistic, we need to add them up.

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

```
[1] 164.04
```

- 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.

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

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

- 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.
- There are post-hoc tests you can do to formally test it (not covered in this class).

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

Exercise

Try to run the `chisq.test()` function yourself first.

```
1 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
```

```r
chisq.test(obs_freq, p = c(.40, .35, .15, .10))
```

- Another type of \(\chi^2\) test 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.
- 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.
- **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.
- **Independence** means knowing the value of one variable doesn't tell you anything about the value of the other variable.

- H_0: Entry status and college are independent of each other.
- H_A: Entry status and college 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.

- To do the test in R, we now need to give R either 2 vectors of data for each category or a table / matrix of the observed frequencies.

Exercise

With the data from the table below create a matrix in R with 2 rows and 4 columns and name it `obs_matrix`:

	CLAS	CA&ES	CBS	COE
Freshmen	276	147	173	111
Transfer	141	76	43	33

```r
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
```

```r
chisq.test(x = 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?
- Entry status and college are dependent on each other.

- You will use a data set built-in in `R` named `UCBAdmissions` .
- It refers to aggregate data on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and sex.
- It is a 3-dimensional array resulting from cross-tabulating 4526 observations on 3 variables. Look at it by calling `UCBAdmissions` on your console.

Aggregate data over departments running the following code:

```
1 agg_UCBAdmissions <- apply(UCBAdmissions, c(1, 2), sum)
```

Reflect on the table you generated for a moment and think about what you expect to find.

What would be your null and alternative hypothesis?

With this new data set, run a chi-squared test to check if the data show evidence of sex bias in admission practices.