Agenda

1. Difference of Two Proportions from last time
2. Goodness of fit

Goodness of Fit Previously, we considered inference for a single proportion. That proportion was the fraction of the outcomes of a binary response variable that had a certain value. For example, respondents could either say that they preferred Coke, or that they preferred Pepsi. But what if the variable can have more than two outcomes? Can we still test the hypothesis that the sample was drawn from a known population?

The US Census Bureau reports that in 2000, among the population 15 years and older:

- 54.3% are married
- 27.1% have never been married
- 9.7% are divorced
- 6.6% are widowed
- 2.2% are separated

We can encode these percentages as a vector in R:

```
us <- c("Divorced" = 0.097, "Married" = 0.543, "Never married/single" = 0.271,
        "Separated" = 0.022, "Widowed" = 0.066)
# normalize to make sure proportions sum to 1
us <- us / sum(us)
```

The `openintro` package contains a sample of 500 Americans collected in the 2000 Census. In this sample, the percentages are different:

```
library(openintro)
library(mosaic)
marital_summary <- census %>%
  mutate(maritalStatus =
           forcats::fct_recode(maritalStatus, Married = "Married/spouse absent",
                               Married = "Married/spouse present")) %>%
  group_by(maritalStatus) %>%
  summarize(status_obs = n()) %>%
  mutate(marital_status_pct = status_obs / nrow(census), marital_status_us = us)
marital_summary$marital_status_pct

## [1] 0.076 0.412 0.444 0.006 0.062
```

Is it reasonable to conclude that the sample from 2000 reflects the overall US population?

In the previous case, the test statistic was the observed sample proportion $\hat{p}$. In this case, we have more than two outcomes, so there is nothing quite analogous to $\hat{p}$. The test statistic that we will use will be labelled X^2, and it's formula is:

$$X^2 = \sum_{i=1}^{k} Z_i^2 = \sum_{i=1}^{k} \left(\frac{observed_i - expected_i}{\sqrt{expected_i}} \right)^2 = \sum_{i=1}^{k} \frac{(observed_i - expected_i)^2}{expected_i},$$

where k is the number of different outcomes (which in this case is 5). As always, our goal is to put X^2 in context by determining where it lies in the null distribution. First, let's compute the test statistic:

```
n <- nrow(census)
k <- nrow(marital_summary)
marital_summary <- marital_summary %>%
  mutate(status_exp = marital_status_us * n)
X2_hat <- marital_summary %>%
  summarize(X2 = sum((status_obs - status_exp)^2 / status_exp)) %>% unlist()
```

1. Write out the full calculation for X^2 using a table

We want to test the null hypothesis that our sample came from the population, whose marital status breakdown is known. Since this implies that the observed counts will match the expected counts exactly, this would result in a test statistic of $\hat{X^2} = 0$. Our observed value of $\hat{X^2}$ is very different from 0, but in order to understand *how* different, we need to know what the null distribution of $\hat{X^2}$ is. In this case, it is *not* normal!

Just as before, there are at least three different ways to construct the sampling distribution of $\hat{X^2}$:

1. Simulation: The procedure is the same it has been: sample from the hypothesized distribution and compute the test statistic many thousands of times.

```
sim <- do(1000) *
  marital_summary %>%
  sample_n(size = n, replace = TRUE, weight = marital_status_us) %>%
  group_by(maritalStatus) %>%
  summarize(status_obs = n(), status_exp = first(status_exp)) %>%
  mutate(X2_i = (status_obs - status_exp)^2 / status_exp) %>%
  summarize(X2 = sum(X2_i))
qplot(data = sim, x = X2)
```

   The p-value can be obtained using the `pdata` function, since the sampling distribution comes from simulated data in our workspace. Note also that since the distribution is non-negative, our test is one-sided.

```
pdata(~X2, X2_hat, data = sim, lower.tail = FALSE)

## X2
##  0
```

2. Probability Theory: Last time, we worked with a *binary* variable, and that led to a *binomial* distribution. This time, we have a categorical variable that can take on more than two values, and that leads to a *multinomial* distribution. For the purposes of this class, you do not need to know what a multinomial distribution is, but it is the multivariate extension of the binomial distribution (i.e. the binomial distribution is the special case of the multinomial distribution when the number of outcomes is 2).

   We will not discuss this approach any further, but based on what you saw last time, hopefully you can believe us that: a) it exists; b) it requires some non-trivial probability theory; and c) it is computationally burdensome.

3. Chi-Squared Test: Since the multinomial distribution is very cumbersome to work with, statisticians have constructed a parametric approximation to the sampling distribution of $\hat{X^2}$. It

follows from probability theory that as long as the expected count of each outcome is at least 5, the test statistic follows a distribution that is closely approximated by a χ^2-distribution on $k-1$ degrees of freedom.

```
plotDist("chisq", params = list(df = k-1), lwd = 3)
```

The p-value can be obtained using the `pchisq` function, since the sampling distribution follows a χ^2-distribution.

```
pchisq(X2_hat, df = k-1, lower.tail = FALSE)

##           X2
## 2.63096e-16
```

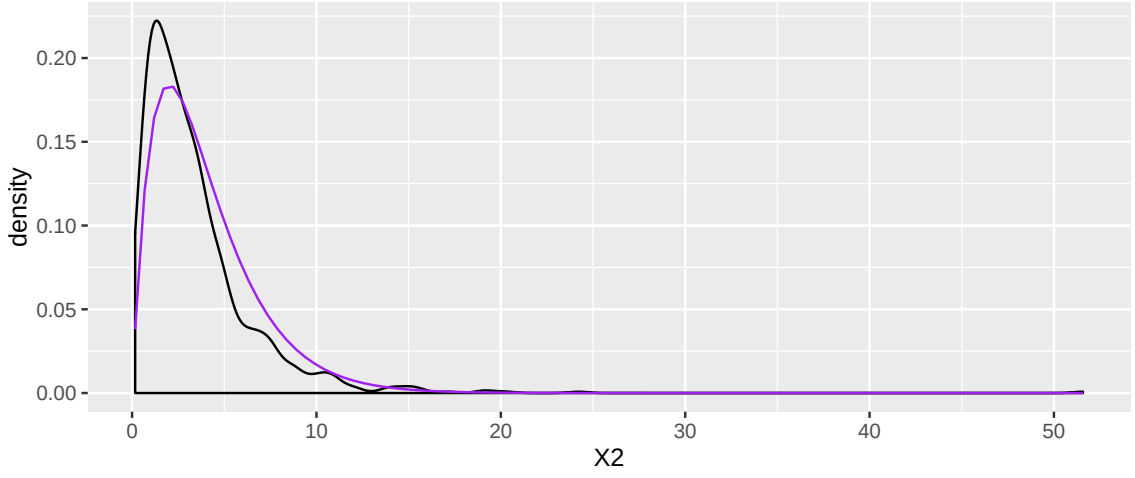
Notice that the p-value is a one-tailed area in this case, since the distribution is non-negative. There is also a built-in function in R that will perform a χ^2-test.

```
with(marital_summary, chisq.test(status_obs, p = marital_status_us))

##
##  Chi-squared test for given probabilities
##
## data:  status_obs
## X-squared = 79.154, df = 4, p-value = 2.631e-16
```

What Can Go Wrong? Once again, the condition that the expected count for each category is at least 5 is important, because if that condition is not met, the χ^2-distribution may not be a sufficiently good approximation. Note that the deviation in each count is approximately normal, so the approximation can fail for any of the outcomes.

```
n <- 35
sim <- do(1000) *
  marital_summary %>%
  mutate(status_exp = marital_status_us * n) %>%
  sample_n(size = n, replace = TRUE, weight = marital_status_us) %>%
  group_by(maritalStatus) %>%
  summarize(status_obs = n(), status_exp = first(status_exp)) %>%
  mutate(X2_i = (status_obs - status_exp)^2 / status_exp) %>%
  summarize(X2 = sum(X2_i))
qplot(data = sim, x = X2, geom = "density") +
  stat_function(fun = dchisq, args = list(df = k-1), color = "purple")
```



In-Class Exercise, OI, 3.40 Evolution vs. creationism A Gallup Poll released in December 2010 asked 1019 adults living in the Continental U.S. about their belief in the origin of humans. These results, along with results from a more comprehensive poll from 2001 (that we will assume to be exactly accurate), are summarized in the table below:

	Year	
Response	2010	2001
Humans evolved, with God guiding (1)	38%	37%
Humans evolved, but God had no part in process (2)	16%	12%
God created humans in present form (3)	40%	45%
Other / No opinion (4)	6%	6%

1. Calculate the actual number of respondents in 2010 that fall in each response category.

2. State hypotheses for the following research question: have beliefs on the origin of human life changed since 2001?

3. Calculate the expected number of respondents in each category under the condition that the null hypothesis is true.

4. Conduct a chi-square test and state your conclusion. (Reminder: verify conditions.)