Agenda

1. Difference of Two Proportions (cont.)
2. Goodness of fit

Difference of two proportions Previously, we considered Ted Williams' batting average of .406 in 1941, which is unmatched in 75 years and counting. In 1994, Tony Gwynn of the San Diego Padres hit .394, but a strike by the player's union shortened the season after only 116 games. Thus, Gwynn accumulated 165 hits in 419 at-bats, whereas Williams had 185 hits in 456 at-bats. Let's assume that Gwynn had an unknown, fixed true batting average of p in 1994.

Just as with the one-proportion test, using the formulas for the expected value and variance of differences between random variables, we can prove that the standard error of $\widehat{p_1 - p_2}$ is $SE_{\widehat{p_1 - p_2}} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2}$. Once again, we'll typically use the normal approximation to the null distribution.

1. Using the normal approximation again, test the hypothesis that Williams and Gwynn had the same true batting averages in 1941 and 1994, respectively.

2. Since we are testing the hyopthesis that $p_1 = p_2$, it is more appropriate to use the *pooled* estimate of the standard error (see page 133). Perform this test.

3. Discuss the extent to which you think the performances of Williams and Gwynn were importantly different.

Practice: Difference of two proportions

1. The University of Michigan survey of consumers asked a sample of Americans Do you rate the governments economic policy as poor? Participants responded Yes or No. They also collected information on whether the participants only had high school diplomas (n = 140), or if they had college degrees (n = 87). Of those with high school diplomas, 70 said yes to the policy question. Of those with college degrees 30 said yes to the question. Test an appropriate hypothesis and state your conclusion.

2. A presidential candidate fears he has a problem with women voters. His campaign staff runs a poll to assess the situation. They randomly sample 300 men and 300 women, asking if they have a favorable impression of the candidate. They find that 59% of the men and 53% of the women have positive images of the candidate. Is there is difference in impression between all men and all women voters? Test an appropriate hypothesis and state your conclusion.

3. Researchers at the National Cancer Institute released the results of a study that investigated the effects of 827 dogs from homes where an herbicide was used on a regular basis, diagnosing malignant lymphoma in 473 of them. Of the 130 dogs from homes where no herbicides were used, only 19 were found to have lymphoma. Is there a different rate of cancer in dogs between homes that use the herbicide and homes that do not? Test an appropriate hypothesis and state your conclusion.

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, we can construct the sampling distribution of $\hat{X^2}$ with simulation, or statitsical theory:

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. 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
```