

17.1 Randomization test for the difference in proportions

Example 1: The cardiopulmonary resuscitation (CPR) experiment consisted of two treatments on patients who underwent CPR for a heart attack and were subsequently admitted to a hospital. Each patient was randomly assigned to either receive a blood thinner (treatment group) or not receive a blood thinner (control group). The outcome variable of interest was whether the patient survived for at least 24 hours. ([Böttiger et al. 2001](#))

The `cpr` data can be found in the `openintro` R package.

Group	Died	Survived	Total
Control	39	11	50
Treatment	26	14	40
Total	65	25	90

In this study, a larger proportion of patients who received blood thinner after $\hat{p}_t = \frac{14}{40} = 0.35$, survived compared to those who did not receive blood thinner, $\hat{p}_c = \frac{11}{50} = 0.22$. However, based on these observed proportions alone, we cannot determine whether the difference ($\hat{p}_t - \hat{p}_c = 0.35 - .22 = .13$) provides *convincing evidence* that blood thinner usage after CPR is effective.

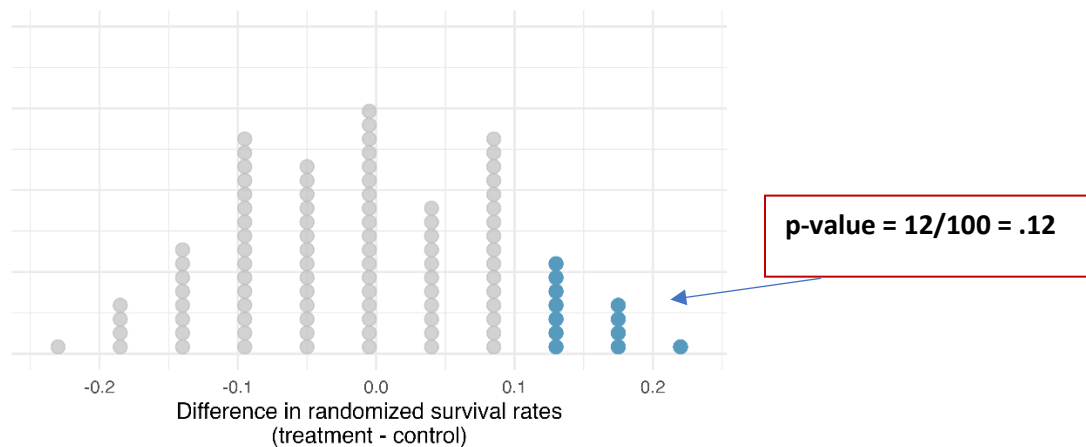
As we saw in Chapter 11, we can re-randomize the responses (survived or died) to the treatment conditions assuming the null hypothesis is true and compute possible differences in proportions.

How often would you observe a difference of at least 13% (0.13) according to the figure? Is this a rare event?

- H_0 : Independence model. Blood thinners after CPR have no effect on survival, and we just happened to observe a difference that would only occur on a rare occasion. $H_0: p_t = p_c$
- H_a : Alternative model. Blood thinners after CPR **increase chance of survival**, and the difference we observed was actually due to the blood thinners after CPR being effective at increasing the chance of survival, which explains the difference of 13%. $H_a: p_t > p_c$

17.1.2 Variability of the statistic

Below is a stacked plot of the differences found from 100 randomization simulations (i.e., repeated iterations as described in Figure 11.8), where each dot represents a simulated difference between the infection rates (control rate minus treatment rate).



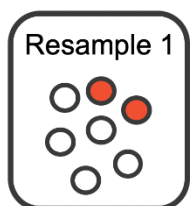
Assuming that the independence model was true, that blood thinners after CPR have no effect on survival. Under the null hypothesis, we expect the difference to be near zero with some random fluctuation, where *near* is pretty generous in this case since the sample sizes are so small in this study.

Since we determined that the outcome is not that rare (12% chance of observing a difference of 13% or more under the assumption that blood thinners after CPR have no effect on survival), we fail to reject H_0 , and conclude that the study results do not provide strong evidence against the independence model. This does not mean that we have proved that blood thinners are not effective, it just means that this study does not provide convincing evidence that they are effective in this setting.

Statistical inference, is built on evaluating how likely such differences are to occur due to chance if in fact the null hypothesis is true. In statistical inference, data scientists evaluate which model is most reasonable given the data. Errors do occur, just like rare events, and we might choose the wrong model. While we do not always choose correctly, statistical inference gives us tools to control and evaluate how often these errors occur.

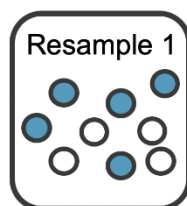
17.2 Bootstrap confidence interval for the difference in proportions

The variability of the statistic (the difference in sample proportions) can be calculated by taking one bootstrap resample from Sample 1 and one bootstrap resample from Sample 2 and calculating the difference in the bootstrap proportions.



Difference 1

$$\hat{p}_{1,boot1} - \hat{p}_{2,boot1}$$



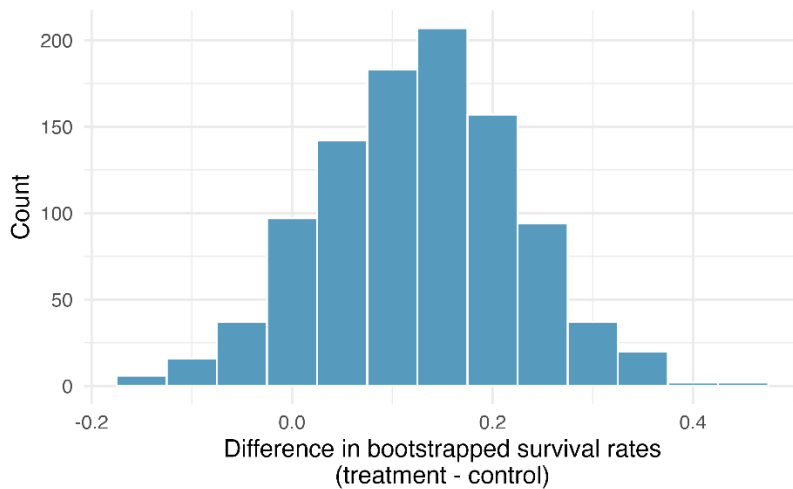
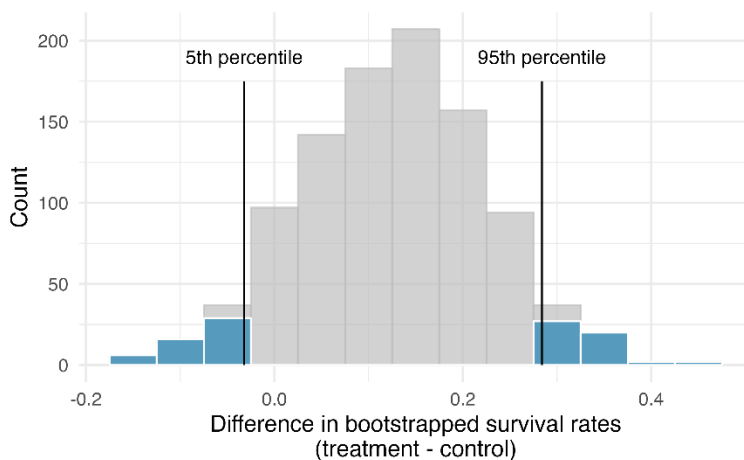


Figure 17.7: A histogram of differences in proportions from 1000 bootstrap simulations of the CPR data. Note that because the CPR data has a larger sample size than the illustrated example, the variability of the difference in proportions is much smaller with the CPR histogram.

The interval created from the percentiles of the distribution is called the **percentile interval**. Note that here we calculate the 90% confidence interval by finding the 5th and 95th percentile values from the bootstrapped differences.

The bootstrap 5 percentile proportion is -0.032 and the 95 percentile is 0.284. The result is: we are 90% confident that, in the population, the true difference in probability of survival for individuals receiving blood thinners after CPR is between -0.032 lower and 0.284 higher than those who did not receive blood thinners.



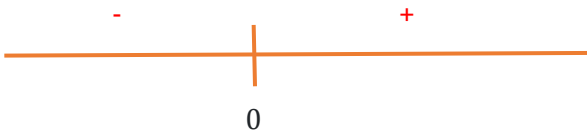
Bootstrap SE confidence interval

Alternatively, we can use the variability in the bootstrapped differences to calculate a standard error of the difference. The resulting interval is called the **SE interval**. Section [17.3](#) details the mathematical model for the

standard error of the difference in sample proportions, but the bootstrap distribution typically does an excellent job of estimating the variability of the sampling distribution of the sample statistic.

$$SE(\widehat{p}_t - \widehat{p}_c) \approx SE(\widehat{p}_{t_{boot}} - \widehat{p}_{c_{boot}}) = 0.098$$

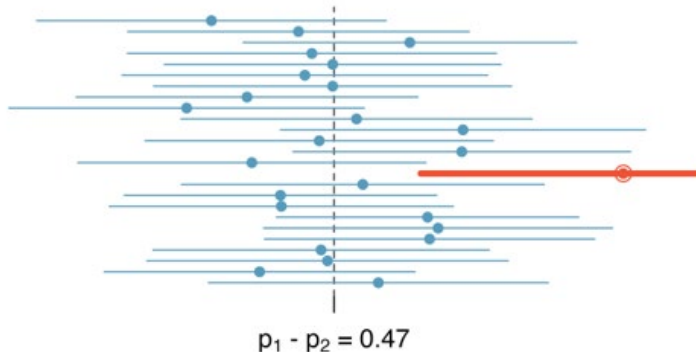
$$(\widehat{p}_t - \widehat{p}_c) \pm z^*SE = \frac{14}{40} - \frac{11}{50} \pm 1.960(0.098) = (-0.067, 0.327)$$



Conclusion: Because zero is included in the interval, the study provides very little evidence about the effectiveness of blood thinners.

17.2.4 What does 95% mean?

If we are making 95% intervals, then about 5% of the intervals we create over our lifetime will *not* capture the parameter of interest (e.g., will be red as in Figure 17.9). What we know is that over our lifetimes as scientists, about 95% of the intervals created and reported on will capture the parameter value of interest: thus the language “95% confident.”



Example 2: A 30-year study was conducted with nearly 90,000 participants who identified as female. During a 5-year screening period, each participant was randomized to one of two groups: in the first group, participants received regular mammograms to screen for breast cancer, and in the second group, participants received regular non-mammogram breast cancer exams. No intervention was made during the following 25 years of the study, and we’ll consider death resulting from breast cancer over the full 30-year period. Results from the study are summarized in table below.

The [mammogram](#) data can be found in the [openintro](#) R package.

If mammograms are much more effective than non-mammogram breast cancer exams, then we would expect to see additional deaths from breast cancer in the control group. On the other hand, if mammograms are not as effective as regular breast cancer exams, we would expect to see an increase in breast cancer deaths in the mammogram group.

Death from breast cancer?			
Treatment	Yes	No	Total
Control	505	44405	44911
Mammogram	500	44425	44925

1. Set up hypotheses to test whether there was a difference in breast cancer deaths in the mammogram and control groups.
2. Be sure to test the success-failure conditions
3. State the null and alternative hypotheses
4. Perform the hypothesis test using coding
5. State your conclusions

Chapter 17 interactive notes

R Saidi

Load the libraries and data

```
library(tidyverse)
library(openintro)
library(tidymodels)
data("gss")
```

A question in two variables

Do women and men join democratic party at different rates?

Let p be the proportion that are democrats.

$$H_o: p_f = p_m \quad H_a: p_f \neq p_m$$

H_o : Women and men join the democratic party at the same rates.

H_a : Women and men join the democratic party at different rates.

We're curious to know: do men and women join the democratic party at different rates?

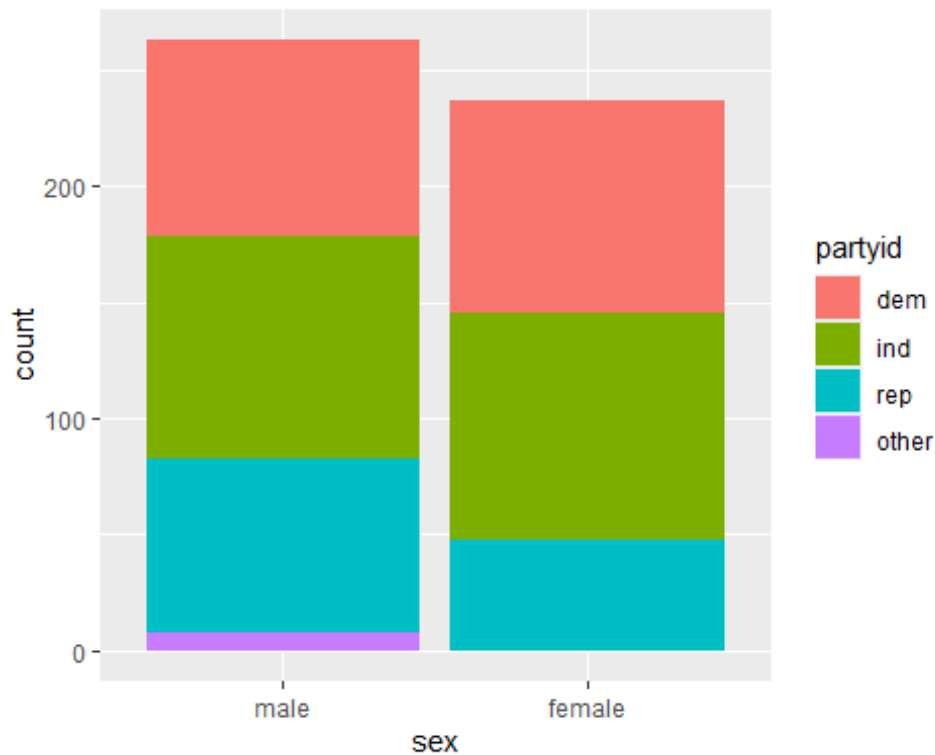
Let's let p be the true proportion that join democrats.

We can then phrase this question as the null hypothesis that the difference in the proportion of men and the proportion of women that join democrats is zero.

The alternative hypothesis would then be that that difference is non-zero.

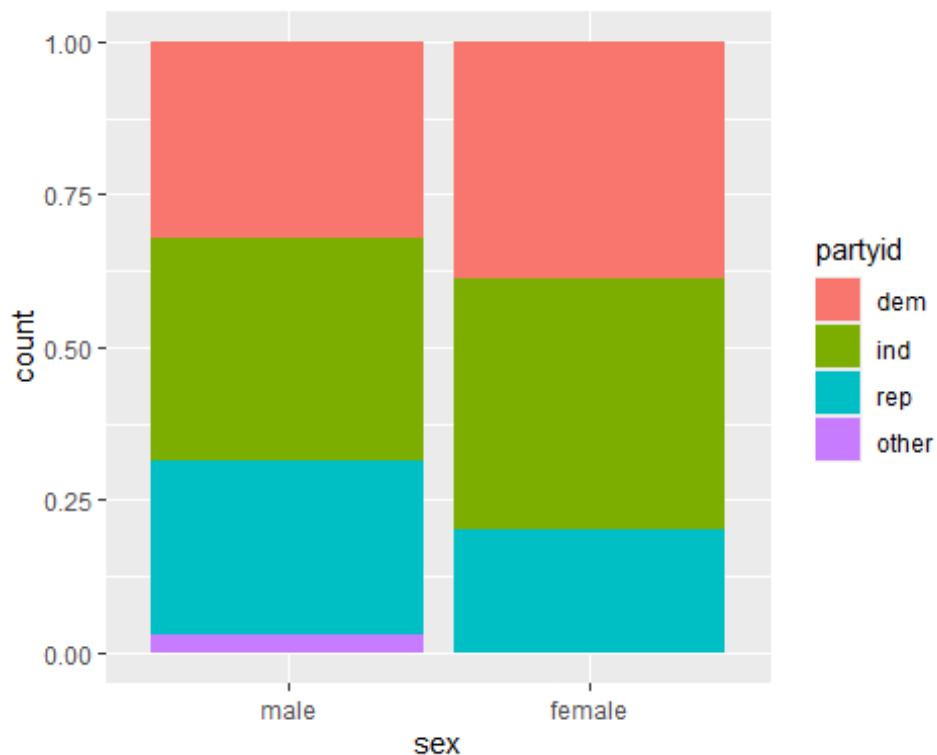
Do women and men have different proportions of affiliation to the democratic party?

```
ggplot(gss, aes(x = sex, fill = partyid)) +  
  geom_bar()
```



Let's take a look at how these proportions compare in the gss data set. The data live in two columns, partyid and sex, so we can map sex to the x-axis and their political party to the color fill of the bars. If we add a geom_bar layer, we get a stacked bar chart that shows us that we have more females in our data set than males and that opinions are split.

```
ggplot(gss, aes(x = sex, fill = partyid)) +  
  geom_bar(position = "fill")
```



We can convert these to proportions by adding the position equals “fill” argument. It looks like the proportion for men is a bit lower than the proportion for women.

Create p-hat values

Sample proportions for males and females to identify with democrats - then calculate the difference of the proportions using d_hat

```
p_hats <- gss |> # phat = sample proportion affiliated with dem party
  group_by(sex) |>
  filter(!is.na(partyid)) |>
  summarize(mean(partyid == "dem")) |>
  pull()

d_hat <- diff(p_hats)
d_hat # observed (sample) difference in proportions

[1] 0.06499174
```

We can calculate the difference in these proportions by using our normal summarize method of calculating a proportion, but add in a group_by line to indicate we want to calculate that proportion for men and women separately. The result is a vector of two proportions. We take their difference with the diff function and save it as d-hats, which we learn is 0.064.

Generating data from H_0

$$H_0: p_f = p_m$$

There is no association between political party affiliation as democrats for sex of a subject.

The variable partyid = democratic is independent from the variable sex.

⇒ Generate data by permutation

Mutate partyid to be a binary variable

```
gss1 <- gss |>
  filter(!is.na(partyid)) |>
  mutate(two_party = ifelse(partyid == "dem", "democrat", "not-democrat"))
```

Do women and men have different political party affiliations?

Check basic assumptions

$np \geq 10$ $n(1-p) \geq 10$

```
gss1 |>
  group_by(sex) |>
  count(two_party) |>
  filter(two_party == "democrat")
```

```
# A tibble: 2 × 3
# Groups:   sex [2]
  sex    two_party    n
<fct> <chr>    <int>
1 male  democrat    85
2 female democrat    92
```

Because np and $n(1-p) \geq 10$, we have met the basic assumptions.

Build up a full null distribution

- specify: syntax for difference of 2 groups is always $y \sim x$
- hypothesize: null = “independence”
- generate: type = “permute” (for hypothesis test for 2 groups)
- calculate: stat = “diff in props”, order = c(“factor1”, “factor2”)

```
null <- gss1 |>
  specify(two_party ~ sex, success = "democrat") |> # y~x
  hypothesize(null = "independence") |>
  generate(reps = 1000, type = "permute") |>
  calculate(stat = "diff in props", order = c("female", "male"))
null
```

Response: two_party (factor)

Explanatory: sex (factor)

Null Hypothesis: indep...

```
# A tibble: 1,000 × 2
  replicate    stat
  <int>    <dbl>
1         1 -0.0473
2         2  0.000818
3         3 -0.0232
4         4 -0.0152
```



```

5      5 -0.0313
6      6 -0.0313
7      7 -0.0473
8      8  0.0249
9      9 -0.0232
10     10  0.000818

```

i 990 more rows

```

obs_diff <- gss1 |>
  specify(two_party ~ sex, success = "democrat") |> # y~x
  hypothesize(null = "independence") |>
  calculate(stat = "diff in props", order = c("female", "male")) |>
  pull()

```

Message: The independence null hypothesis does not inform calculation of the observed statistic (a difference in proportions) and will be ignored.

obs_diff

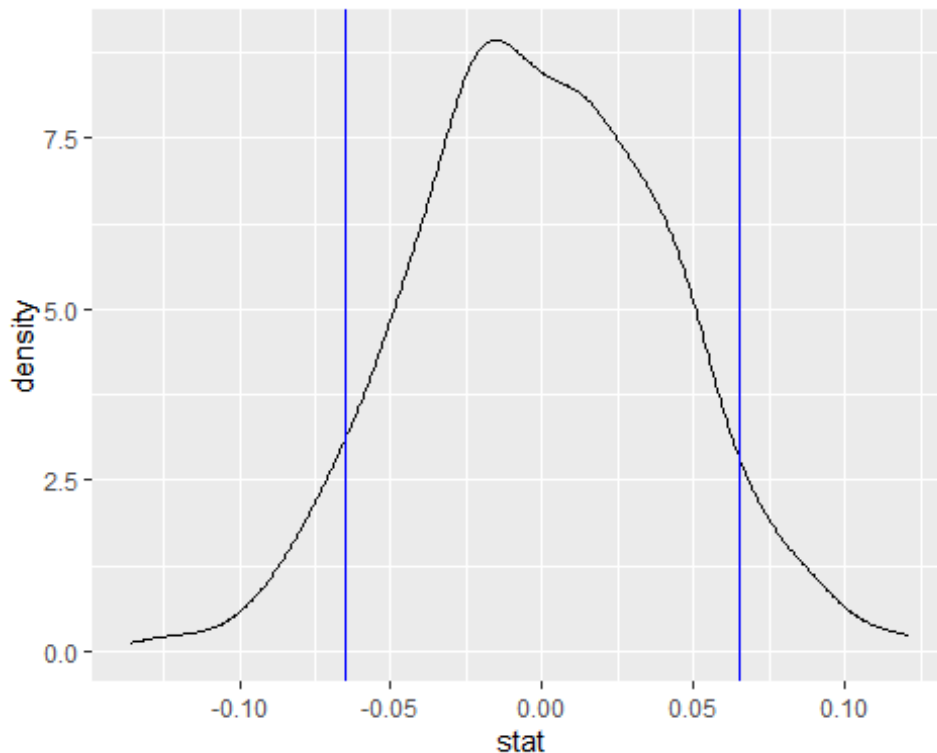
[1] 0.06499174

Plot a density curve

```

ggplot(null, aes(x = stat)) +
  geom_density() +
  geom_vline(xintercept = c(-obs_diff, obs_diff), color = "blue")

```



This distribution of data suggests that there may be a difference between sexes in proportion of affiliation to democrats with females at higher proportions than males..

alternative and efficient p-value calculation

```
p_value <- get_p_value(null, obs_diff, direction = "two-sided")
p_value

# A tibble: 1 × 1
  p_value
  <dbl>
1    0.142
```

The two-tailed p-value is 0.14. We fail to reject the null. There is no compelling evidence that there is a difference in proportions of females and males affiliated to the democratic party.

Create a 95% Bootstrap CI

The center of the bootstrap confidence interval is the observed difference in the proportions of males and females affiliated with the democratic party?

```
# to convert from scientific to standard notation for values
#options(scipen = 0)

# Create the bootstrap distribution
boot <- gss1 |>
  # Specify the variables and success
  specify(two_party ~ sex, success = "democrat") |>
  # Generate 500 bootstrap reps
  generate(reps = 1000, type = "bootstrap") |>
  # Calculate the statistics
  calculate(stat = "diff in props", order = c("female", "male"))
boot

Response: two_party (factor)
Explanatory: sex (factor)
# A tibble: 1,000 × 2
  replicate    stat
  <int>    <dbl>
1         1 0.0881
2         2 0.130
3         3 0.109
4         4 0.0410
5         5 0.107
6         6 0.0895
7         7 0.0773
8         8 0.0757
9         9 0.146
10        10 0.120
# i 990 more rows

# Compute the standard error
SE <- boot |>
  summarize(se = sd(stat)) |>
  pull()

# Form the CI (lower, upper)
c(d_hat - 1.960 * SE, d_hat + 1.960 * SE)
```

```
[1] -0.01999622 0.14997969
```

alternative and efficient ci calculation

```
ci <- get_ci(boot, level = .95)
ci
```

```
# A tibble: 1 × 2
  lower_ci upper_ci
  <dbl>    <dbl>
1 -0.0181 0.151
```

We are 95% confident that the true difference in proportions of male and female affiliation to the democratic party is between -1.19% and 14.79%. Because zero is included there is no compelling evidence that there is a difference in sex with respect to affiliation to the democratic party.

For your textbook homework `prop.test()`

```
# prop.test(x = c(x1, x2), n = c(n1, n2)) # this defaults to 2-tailed test
# this uses the Yates Continuity Correction
prop.test(x = c(505, 500), n = c(44911, 44925))
```

2-sample test for equality of proportions with continuity correction

```
data: c(505, 500) out of c(44911, 44925)
X-squared = 0.017386, df = 1, p-value = 0.8951
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.001282986 0.001512587
sample estimates:
 prop 1      prop 2 
0.01124446 0.01112966
```

Homework Chapter 17

1. Review section 17.4 (the chapter review)
2. **Suggested:** from textbook section 17.5 exercises: 1, 3, 7, 9, 11, 14, 16, 22
3. **Suggested:** tutorials

2 - [Hypothesis tests to compare proportions](#)