AE 15: Hypothesis tests for independence

Dav King

03/01/2022

```
library(tidyverse)
library(tidymodels)
```

```
yawn <- read_csv("yawn.csv")
```

Learning goals

Use and understand simulation-based methods to …

- test a claim about a population proportion
- test a claim about independence between two groups

Different options inside generate

- Discussion of bootstrap, draw, and permute here.
  - Bootstrap: with replacement, for confidence intervals or for a single mean.
  - Draw: only for hypothesis testing for one proportion.
  - Permute: "shuffles" the data without replacement- see explanation here. You might use this for a difference in means or to test for independence (diff in props).

If you want to double check, I would recommend checking the vignette here. This is not an exhaustive list, just some common ones we use in this course.

Part 1: Hypothesis test for a single proportion[1]

A large university knows that about 70% of the full-time students are employed at least 5 hours per week. The members of the Statistics Department wonder if the same proportion of their students work at least 5 hours per week. They randomly sample 25 majors and find that 15 of the students work 5 or more hours each week. Use the code below to create a data frame of the results.

```
stats_work <- tibble(work_hours = c(rep("At least 5", 15),
                                    rep("Less than 5", 10)))
```

[1]This question was adapted from *Introduction to Modern Statistics*.

- State the null and alternative hypotheses.

$H_0 : p = 0.7$ $H_A : p \neq 0.7$

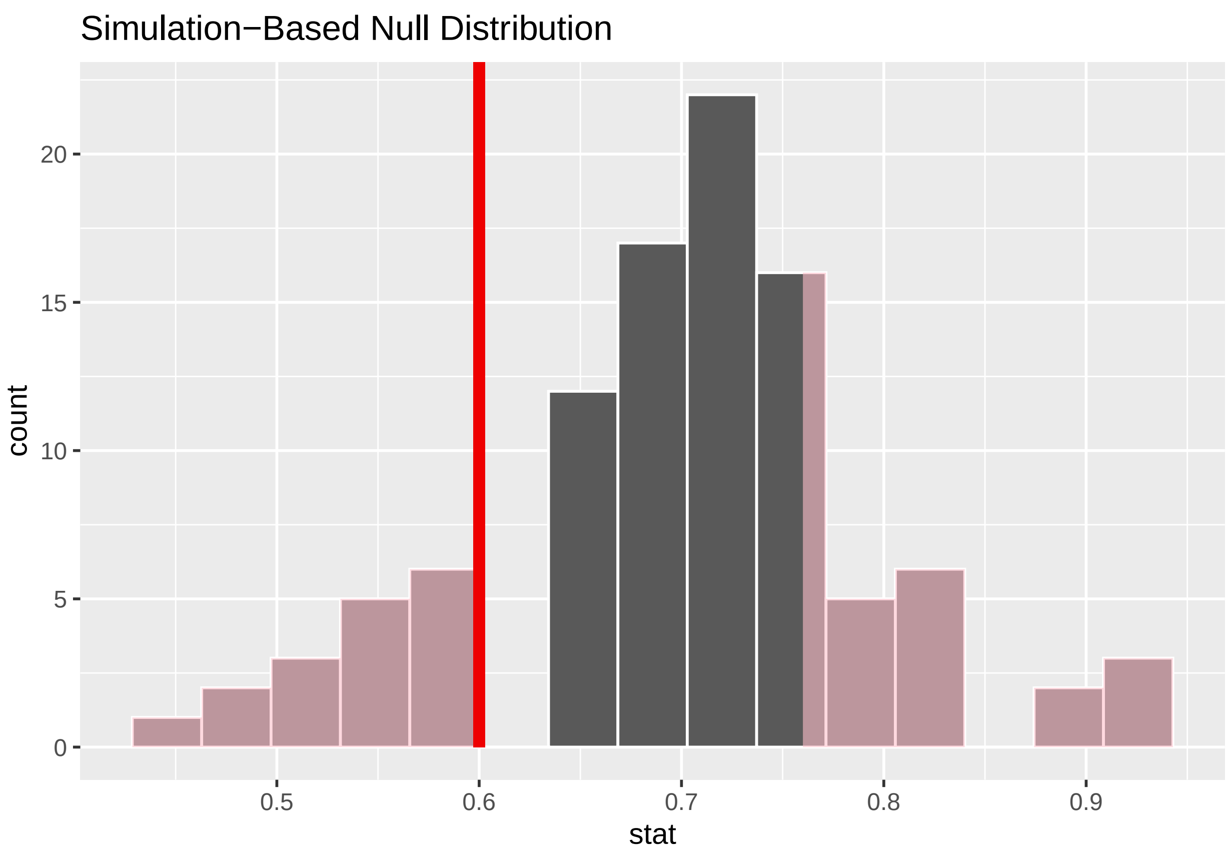
- Fill in the code to generate the null distribution.

```
set.seed(101821)
null_dist <- stats_work %>%
  specify(response = work_hours, success = "At least 5") %>%
  hypothesize(null = "point", p = 0.7) %>%
  generate(reps = 100, type = "draw") %>%
  calculate(stat = "prop")
```

- Visualize the null distribution and shade in the area used to calculate the p-value.

```
p <- stats_work %>%
  count(work_hours) %>%
  summarize(work_hours, p = n/sum(n)) %>%
  filter(work_hours == "At least 5") %>%
  pull()

visualize(null_dist) +
  shade_p_value(obs_stat = p, direction = "both")
```



Note, the odd asymmetry we seem to be getting around 0.7 seems to be a result of the seed we are using.

- Calculate the p-value. Then use the p-value to make your conclusion using a significance level of 0.05.

```
null_dist %>%
  get_p_value(obs_stat = p, direction = "both")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1    0.34
```

At $\alpha = 0.05$, we fail to reject H_0. We do not have significant evidence to say that the proportion of stat majors who work at least 5hr/wk differs from the student body as a whole.

Part 2: Test for independence

- First let's, watch the experiment from *Mythbusters* here [https://www.youtube.com/watch?v=mrr_UjNLbhE].
- Let t be the treatment group who saw a person yawn, c be the control group who did not see anyone yawn, and p be the proportion of people who yawned.

Exercise 1

We want to use simulation-based inference to assess whether or not yawning and seeing someone yawn are independent.

- State the null and alternative hypotheses in words:

H_0: the proportion of people who yawned will not differ according to which treatment group they were a part of. H_A: the proportion of people who yawned will differ according to which treatment group they were a part of.

- Select the appropriate null and alternative hypotheses written in mathematical notation:

a. $H_0 : p_t = p_c$ vs. $H_a : p_t < p_c$
b. $H_0 : p_t = p_c$ vs. $H_a : p_t > p_c$
c. $H_0 : p_t = p_c$ vs. $H_a : p_t \neq p_c$
d. $H_0 : \hat{p}_t = \hat{p}_c$ vs. $H_a : \hat{p}_t < \hat{p}_c$
e. $H_0 : \hat{p}_t = \hat{p}_c$ vs. $H_a : \hat{p}_t > \hat{p}_c$
f. $H_0 : \hat{p}_t = \hat{p}_c$ vs. $H_a : \hat{p}_t \neq \hat{p}_c$

C is correct.

Exercise 2

Let's use the data from the *Mythbusters* episode and simulation-based inference in R to test this claim. Based on their experiment, do yawning and seeing someone yawn appear to be dependent?

Evaluate this question using a simulation based approach. We will use the same null and alternative hypotheses as before. The data from *Mythbusters* is available in the `yawn` data frame.

- Fill in the code below to generate the null distribution. Uncomment the code once it is complete.

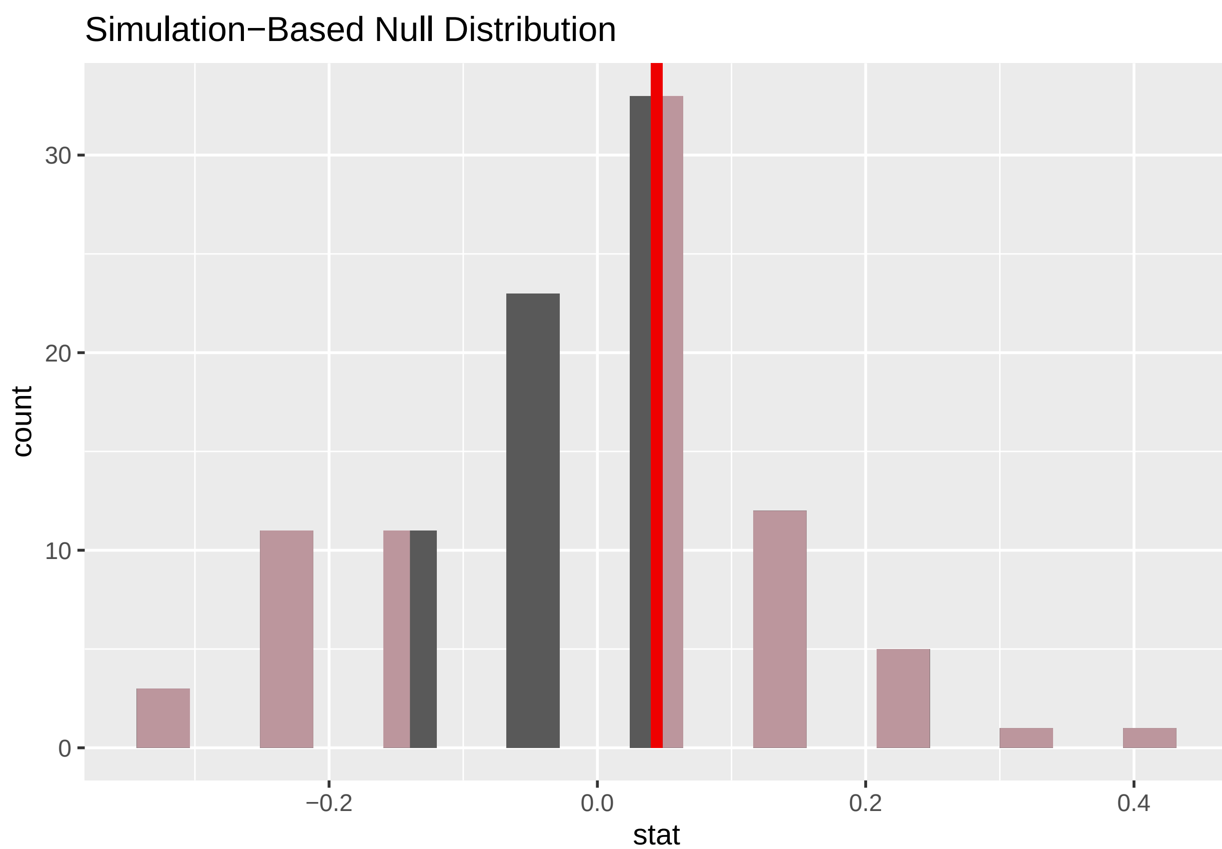
```
set.seed(101821)
null_dist <- yawn %>%
  specify(response = result, explanatory = group, success = "yawn") %>%
  hypothesize(null = "independence") %>%
  generate(100, type = "permute") %>%
  calculate(stat = "diff in props",
            order = c("trmt", "ctrl"))
```

- Visualize the null distribution and shade in the area used to calculate the p-value.

```
p2 <- yawn %>%
  group_by(group) %>%
  count(result) %>%
  summarize(result, p_yawn = n/sum(n)) %>%
  filter(result == "yawn") %>%
  ungroup() %>%
  summarize(p2 = p_yawn[group == "trmt"] - p_yawn[group == "ctrl"]) %>%
  pull()
```

```
## `summarise()` has grouped output by 'group'. You can override using the `.groups` argument.
```

```
visualize(null_dist) +
  shade_p_value(obs_stat = p2, direction = "both")
```



- Calculate p-value. Then use the p-value to make your conclusion using a significance level of 0.1.

```
null_dist %>%
  get_p_value(obs_stat = p2, direction = "both")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       1
```

At $\alpha = 0.1$ (or, to be honest, any alpha level), we fail to reject H_0. We do not have significant evidence to claim that there is a difference in the proportion of people who yawned across treatment groups.

Exercise 3: Confidence interval

Construct a 90% confidence interval for the difference in proportion of yawners between those who see someone else yawn and those who don't.

```
boot_dist <- yawn %>%
  specify(response = result, explanatory = group, success = "yawn") %>%
  generate(100, type = "bootstrap") %>%
  calculate(stat = "diff in props",
            order = c("trmt", "ctrl"))
```

- Why are we using "bootstrap" instead of "permute" here? We don't want to assume that the outcome is independent of treatment group - we are interested in exploring that fact.

```
# calculate the lower and upper bounds for the 90% ci
boot_dist %>%
  summarise(lower = quantile(stat, 0.05),
           upper = quantile(stat, 0.95))
```

```
## # A tibble: 1 x 2
##    lower upper
##    <dbl> <dbl>
## 1 -0.148 0.257
```

- Interpret the interval in the context of the data. We are 90% confident that the true difference in yawning proportions between those who saw someone yawn and those who didn't is between -0.148 and 0.257.

- Suppose you use the confidence interval to evaluate the hypotheses in Exercise 2. Is the conclusion drawn from the confidence interval consistent with the conclusion from the the hypothesis test? Yes, it is consistent. We failed to reject H_0, which suggested that there was no true difference between these groups. Similarly, this interval contains 0 - which would be the value of no difference between groups.

Exercise 4

Do you buy the conclusions from this experiment? Why or why not?

I would be hard pressed to not buy the conclusions from the experiment - this is what they found. However, I'm not sure that it really extends beyond this experiment. While it's good to simulate a very controlled experiment to measure the impact of somebody yawning on others yawning (and while much of our evidence of mirror neurons, the proposed mechanism for how one person's yawn leads to a yawn in another, is fishy at best), the fact is that the real world is not a controlled environment. If people think that other people yawning will make them yawn, then seeing someone else yawn will make them yawn.

Exercise 2 on Sakai, which is not in this Rmd

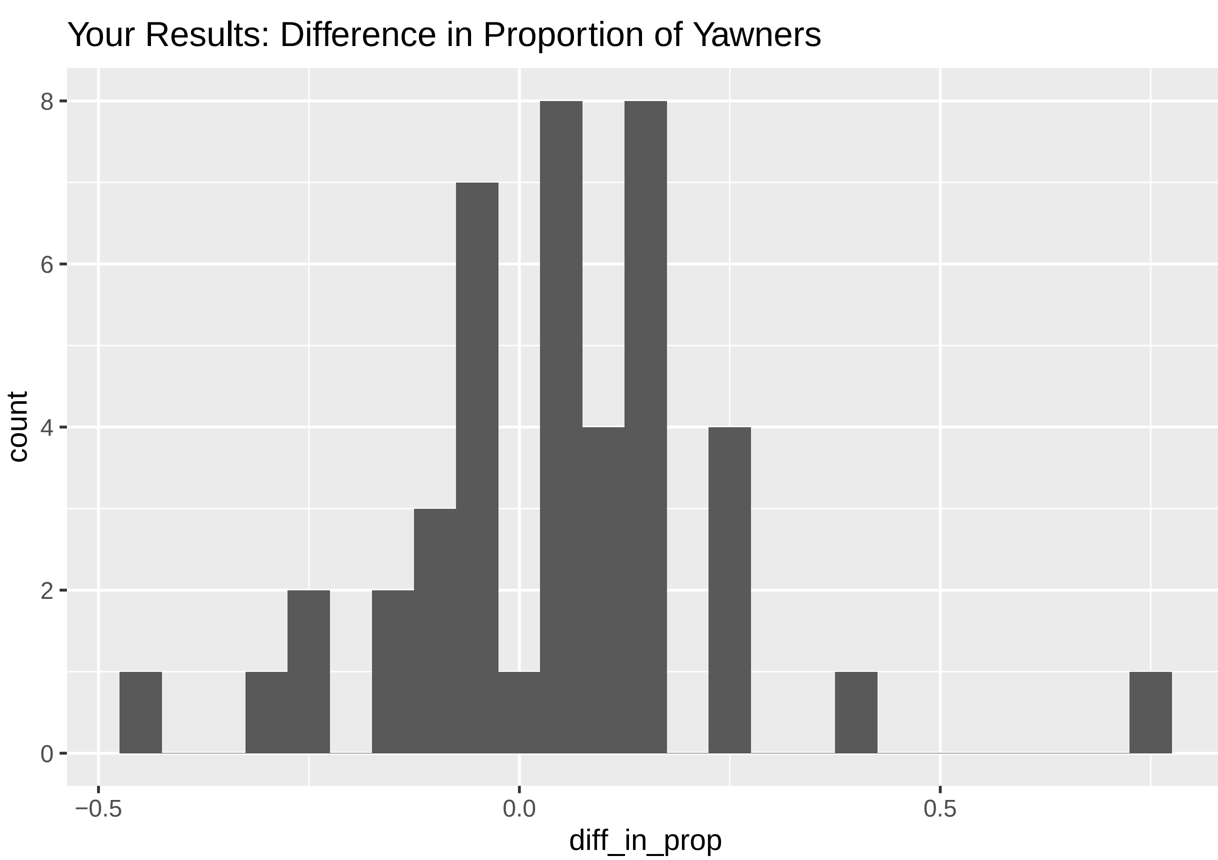
```
sim_data <- read_csv("https://sta199-fa21-003.netlify.app/appex/data/yawn-sim.csv")
```

```
## Rows: 43 Columns: 1

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (1): diff_in_prop

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ggplot(sim_data, aes(x = diff_in_prop)) +
  geom_histogram(binwidth = 0.05) +
  labs(title = "Your Results: Difference in Proportion of Yawners")
```



The approximate center of this distribution is a little higher than 0 - maybe 0.1 or 0.2.

Yawning and seeing someone yawn do not appear to be dependent - these are quite tiny differences.