Data 606 Lab 6 - Inference for categorical data

Avery Davidowitz

2022-11-05

Load Packages

```
library(tidyverse)
library(openintro)
library(infer)
```

Data set

```
data('yrbss', package='openintro')
head(yrbss)
```

```
## # A tibble: 6 x 13
##     age gender grade hispa~1 race  height weight helme~2 text_~3 physi~4 hours~5
##   <int> <chr>  <chr> <chr>   <chr>  <dbl>  <dbl> <chr>   <chr>     <int> <chr>
## 1    14 female 9     not     Blac~  NA      NA   never   0             4 5+
## 2    14 female 9     not     Blac~  NA      NA   never   <NA>          2 5+
## 3    15 female 9     hispan~ Nati~   1.73   84.4 never   30            7 5+
## 4    15 female 9     not     Blac~   1.6    55.8 never   0             0 2
## 5    15 female 9     not     Blac~   1.5    46.7 did no~ did no~       2 3
## 6    15 female 9     not     Blac~   1.57   67.1 did no~ did no~       1 5+
## # ... with 2 more variables: strength_training_7d <int>,
## #   school_night_hours_sleep <chr>, and abbreviated variable names 1: hispanic,
## #   2: helmet_12m, 3: text_while_driving_30d, 4: physically_active_7d,
## #   5: hours_tv_per_school_day
```

Exercise 1

The counts within each category for the amount of days these students have texted while driving within the past 30 days.

```
texting_counts <- yrbss |> dplyr::filter(!is.na(text_while_driving_30d)) |>
                count(text_while_driving_30d)

texting_counts
```

```
## # A tibble: 8 x 2
##   text_while_driving_30d     n
##   <chr>                  <int>
```

```
## 1 0                      4792
## 2 1-2                     925
## 3 10-19                   373
## 4 20-29                   298
## 5 3-5                     493
## 6 30                      827
## 7 6-9                     311
## 8 did not drive          4646
```

Exercise 2

The proportion of people who have texted while driving every day in the past 30 days and never wear helmets.

```
never_helm_always_text <- yrbss |> dplyr::filter(text_while_driving_30d == "30" & helmet_12m == "never")
never_helm_always_text_prop <- never_helm_always_text / nrow(yrbss)
never_helm_always_text_prop
```

```
## [1] 0.03408673
```

```
no_helmet <- yrbss %>%
  dplyr::filter(helmet_12m == "never")
no_helmet <- no_helmet %>%
  dplyr::mutate(text_ind = ifelse(text_while_driving_30d == "30", "yes", "no"))
no_helmet <- no_helmet |> filter(!is.na(text_ind)) # I was getting errors without removing NAs
no_helmet %>%
  infer::specify(response = text_ind, success = "yes") %>%
  infer::generate(reps = 1000, type = "bootstrap") %>%
  infer::calculate(stat = "prop") %>%
  infer::get_ci(level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   0.0650   0.0778
```

Exercise 3

The margin of error for the estimate of the proportion of non-helmet wearers that have texted while driving each day for the past 30 days based on this survey.

```
margin_error <- 1.96 * (never_helm_always_text_prop * (1 - never_helm_always_text_prop) / nrow(yrbss))^
margin_error
```

```
## [1] 0.003051546
```

Exercise 4

Calculated confidence intervals for two other categorical variables.

```r
hispanic_pop_ci <- yrbss |> dplyr::filter(!is.na(hispanic)) %>%
  infer::specify(response = hispanic, success = "hispanic") %>%
  infer::generate(reps = 1000, type = "bootstrap") %>%
  infer::calculate(stat = "prop") %>%
  infer::get_ci(level = 0.95)
hispanic_pop_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.249    0.264
```

I can claim with 95% confidence that the true Hispanic proportion of the general population of teens is between 24.9% and 26.4%.

```r
gender_ci <- yrbss |> dplyr::filter(!is.na(gender)) |>
  infer::specify(response = gender, success = "female") %>%
  infer::generate(reps = 1000, type = "bootstrap") %>%
  infer::calculate(stat = "prop") %>%
  infer::get_ci(level = 0.95)
gender_ci
```

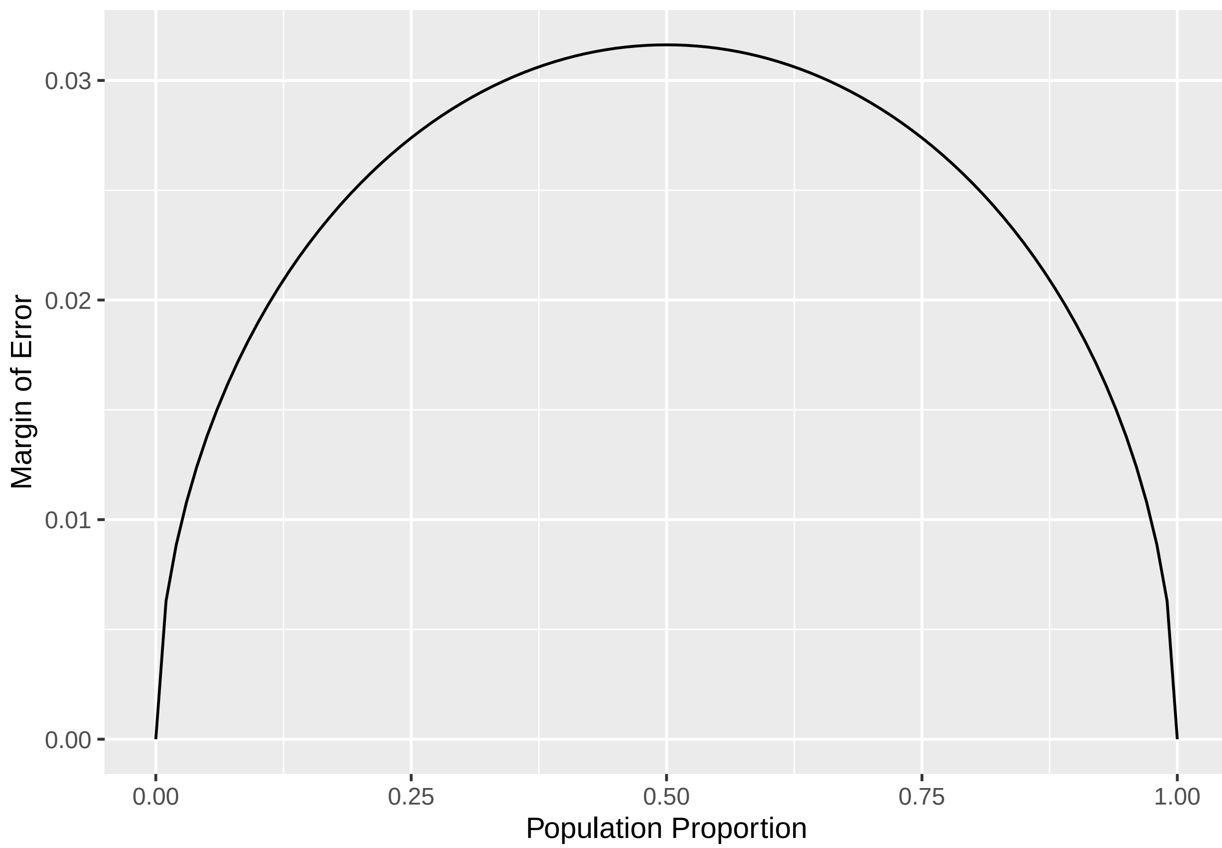
```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.480    0.496
```

I can claim with 95% confidence that the true female proportion of the general population of teens is between 47.9% and 49.6%.

Exercise 5

The margin of error is the largest when p = .5 and the smallest when p approaches 0 and 1. However, when p is very small or very large it may be difficult to obtain a large enough sample to satisfy the success-failure conditions needed to use a normal distribution.

```r
n <- 1000
p <- seq(from = 0, to = 1, by = 0.01)
me <- 2 * sqrt(p * (1 - p)/n)
dd <- data.frame(p = p, me = me)
ggplot(data = dd, aes(x = p, y = me)) +
  geom_line() +
  labs(x = "Population Proportion", y = "Margin of Error")
```



Exercise 6

p hat is narrowly distributed around .1 with a nearly normal shape.

Exercise 7

The distribution of p hats seem very similar for most ranges of p. p hat's distribution does seem to approximate the normal distribution the closer p is to .5. Many other ranges of values of p generate multi-modal distributions.

Exercise 8

The higher the sample size the tighter the spread and the more normal shape for p hat. The distribution of p hat seems more sensitive to p than to sample size.

Exercise 9

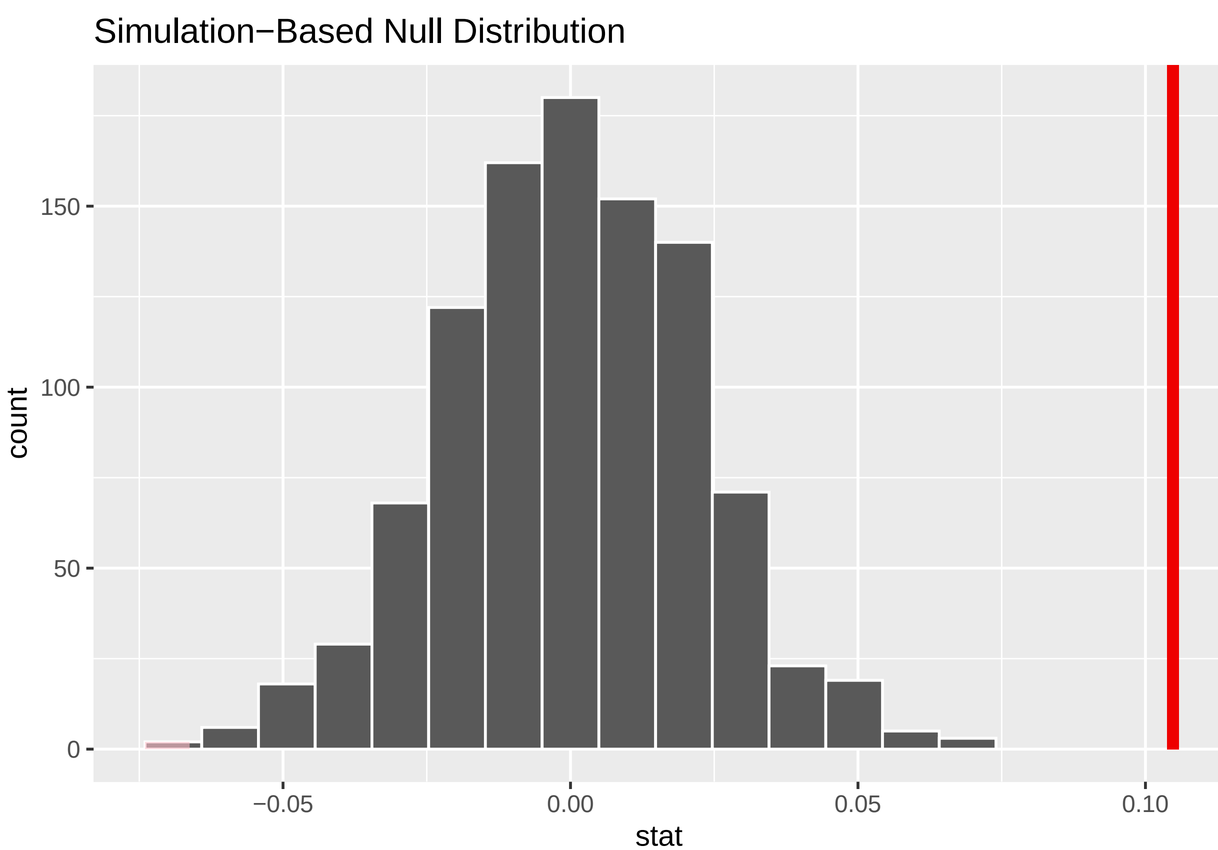
Is there convincing evidence that those who sleep 10+ hours per day are more likely to strength train every day of the week? I would test a null hypothesis of p1 - p0 = 0. Where p1 is the proportion of those who sleep 10+ and strength train everyday and p0 is the proportion of those who strength train everyday but do not sleep 10+. The observes p_hat of the difference between those prortions in .10 which has a p value of 0 which is less than .05 so we reject the null hypothesis. We are 95% sure that the difference should fall between +/- .04 from 0. Therefore they are not independent and we would expect a 10% increase in people who strength train everyday if they sleep over 10 hours.

```
sleep_10_str_every <- yrbss |>
  dplyr::filter(!is.na(strength_training_7d)) |>
  dplyr::filter(!is.na(school_night_hours_sleep)) |>
  dplyr::mutate(str_every = ifelse(strength_training_7d == "7", "yes", "no")) |>
  dplyr::mutate(sleep_10 = ifelse(school_night_hours_sleep =="10+", "yes", "no"))

p_hat <- sleep_10_str_every %>%
  specify(str_every ~ sleep_10, success = "yes") %>%
  calculate(stat = "diff in props", order = c("yes", "no"))

null_dist <- sleep_10_str_every %>%
  infer::specify(str_every ~ sleep_10, success = "yes") |>
  infer::hypothesize(null = "independence") |>
  infer::generate(reps = 1000) |>
  infer::calculate(stat = "diff in props", order = c("yes", "no"))

visualize(null_dist) +
  shade_p_value(obs_stat = p_hat, direction = "two-sided")
```




```
ci <- null_dist |> infer::get_ci(level = 0.95)
ci
```

```
## # A tibble: 1 x 2
```

```
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1  -0.0465   0.0456
```

```
p_val <- null_dist %>%
  get_p_value(obs_stat = p_hat, direction = "two-sided")
p_val
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

Exercise 10

Alpha represents the probability that you reject the null hypothesis incorrectly or a type 1 error. Therefore there would be a 5% chance that your statistical analysis could falsely determine a difference in the likelihood to strength train everyday given sleeping over 10+ hours.

Exercise 11

You can establish an upper bound on the margin of error if you solve for n using a p = q = .5 which produces the highest margin of error. ME = Z $\times$ SE -> n = :

```
n <- (.5*1.96/.01)^2
n
```

```
## [1] 9604
```