Data 606 Lab 5B

Joao De Oliveira

2025-10-11

Overview

This report intends to explore the foundations of the statistical inference and confidence intervals.

Load Packages

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(openintro)
```

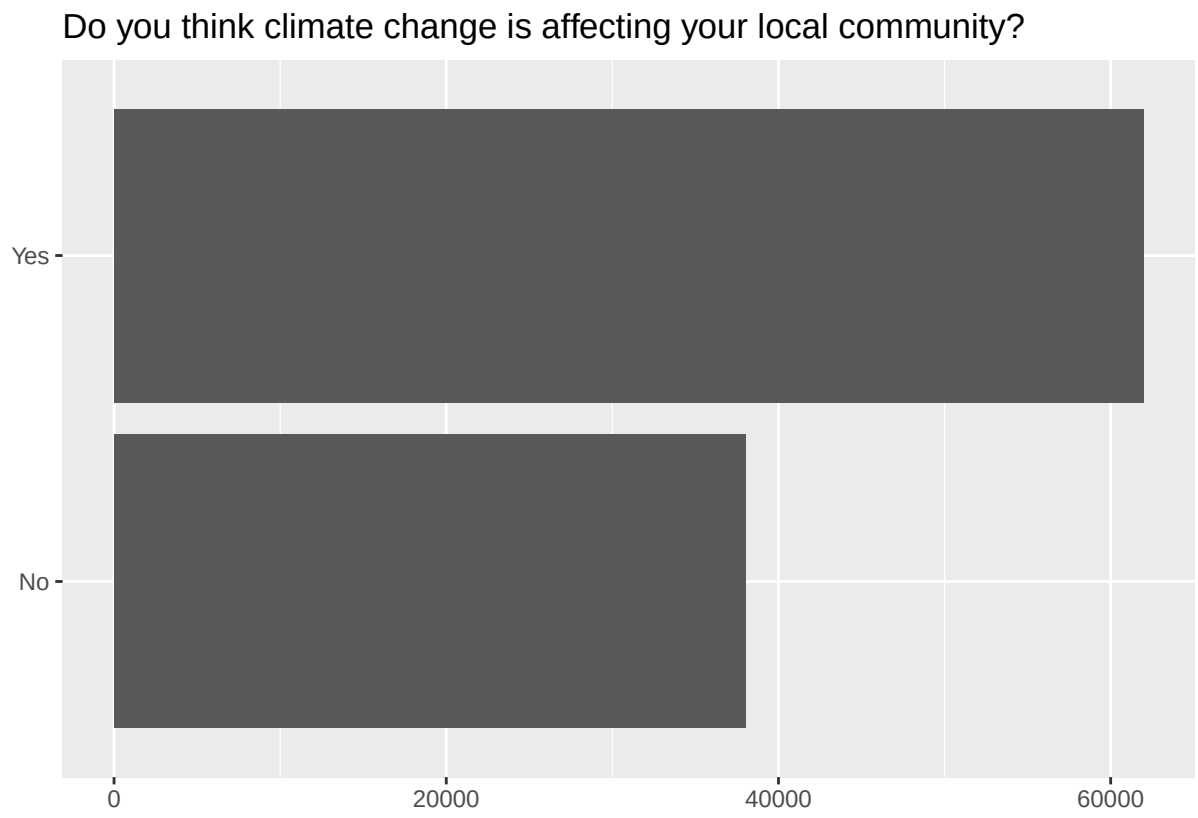
```
## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata
```

```
library(infer)
```

Exercise 1

What percent of the adults in your sample think climate change affects their local community? Hint: Just like we did with the population, we can calculate the proportion of those in this sample who think climate change affects their local community.

```r
set.seed(1234)

us_adults <- tibble(
  climate_change_affects = c(rep("Yes", 62000), rep("No", 38000))
)

ggplot(us_adults, aes(x = climate_change_affects)) +
  geom_bar() +
  labs(
    x = "", y = "",
    title = "Do you think climate change is affecting your local community?"
  ) +
  coord_flip()
```




```r
# summary stats population
us_adults %>%
  count(climate_change_affects) %>%
  mutate(p = n /sum(n))
```

```
## # A tibble: 2 x 3
##   climate_change_affects     n     p
##   <chr>                  <int> <dbl>
## 1 No                     38000  0.38
## 2 Yes                    62000  0.62
```

```r
n <- 60
samp <- us_adults %>%
  sample_n(size = n)

# summary stats sample with size 60
samp %>%
  count(climate_change_affects) %>%
  mutate(p = n / sum(n))
```

```
## # A tibble: 2 x 3
##   climate_change_affects     n     p
##   <chr>                  <int> <dbl>
## 1 No                        23 0.383
## 2 Yes                       37 0.617
```

Answer 1

In the sample with size 60, the percent of adults who think climate change affects their local community is 61.7%, which is a very close value to the population proportion of 62%.

Exercise 2

Would you expect another student's sample proportion to be identical to yours? Would you expect it to be similar? Why or why not?

Answer 2

I would expect another student's sample proportion to be similar to mine since we are using the same population (with proportion 62%) but not the same since the sampling method is random, so my sample of 60 adults is different than another student's sample.

```r
samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.483     0.75
```

Exercise 3

In the interpretation above, we used the phrase "95% confident". What does "95% confidence" mean?

Answer 3

A 95% confidence level means that the method used is able to generate intervals that capture the true proportion population value 95% of the time.

Exercise 4

Does your confidence interval capture the true population proportion of US adults who think climate change affects their local community? If you are working on this lab in a classroom, does your neighbor's interval capture this value?

Answer 4

My 95% confidence interval for the proportion of US adults who think climate change affects their local community was from 0.483 up to 0.75. Since we know that the true population proportion is 0.62, this value falls within my confidence interval, which means my interval successfully captured the true population parameter. I am not in a classroom but I assume that 95% of my colleagues would have a confidence interval that included the true population proportion, while 5% wouldn't due to the randomness of sample selection.

Exercise 5

Each student should have gotten a slightly different confidence interval. What proportion of those intervals would you expect to capture the true population mean? Why?

Answer 5

I would expect that 95% of students' intervals would have the true proportion value within their intervals since the 95% confidence interval means that the method used would have the true parameter in around 95% of those intervals.

Exercise 6

Given a sample size of 60, 1000 bootstrap samples for each interval, and 50 confidence intervals constructed (the default values for the above app), what proportion of your confidence intervals include the true population proportion? Is this proportion exactly equal to the confidence level? If not, explain why. Make sure to include your plot in your answer.

```
p_true <- 0.62

n <- 60
reps_boot <- 1000
n_intervals <- 50

# sample and bootstrap a 95% confidence interval
one_sample_ci <- function() {
  samp <- us_adults %>% sample_n(n)
  p_hat <- mean(samp$climate_change_affects == "Yes")
  ci <- samp %>%
    specify(response = climate_change_affects, success = "Yes") %>%
    generate(reps = reps_boot, type = "bootstrap") %>%
    calculate(stat = "prop") %>%
    get_ci(level = 0.95, type = "percentile")
  tibble(p_hat = p_hat, lower = ci$lower_ci, upper = ci$upper_ci)
}

# 50 confidence intervals
```

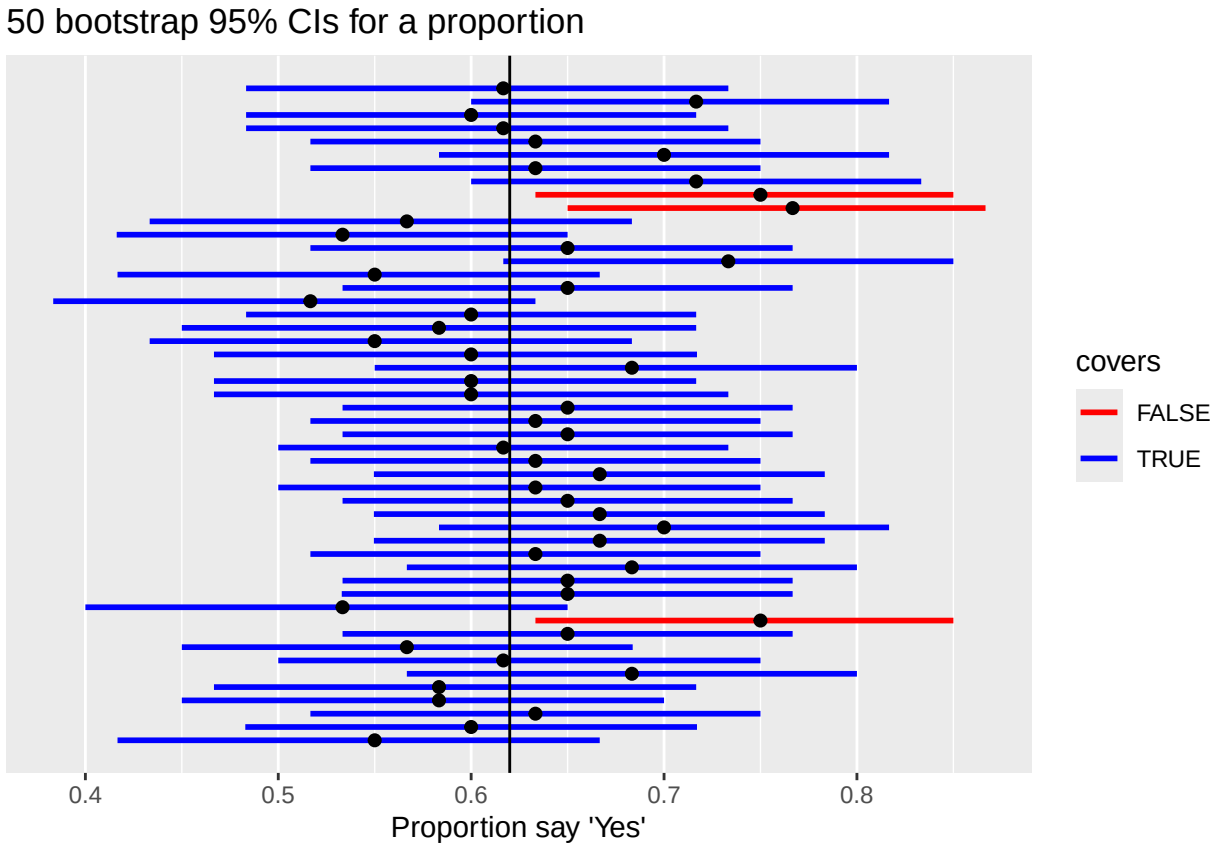
```r
confidence_intervals <- map_dfr(1:n_intervals, ~ one_sample_ci()) %>%
  mutate(id = row_number(),
         covers = (lower <= p_true & upper >= p_true))

coverage <- mean(confidence_intervals$covers)
coverage
```

```
## [1] 0.94
```

```r
# table counts
table(covers = confidence_intervals$covers)
```

```
## covers
## FALSE  TRUE
##     3    47
```

```r
# plot
ggplot(confidence_intervals, aes(y = id)) +
  geom_segment(aes(x = lower, xend = upper, yend = id, color = covers), linewidth = 1) +
  geom_point(aes(x = p_hat), size = 1.8) +
  geom_vline(xintercept = p_true) +
  scale_y_continuous(NULL, breaks = NULL) +
  scale_color_manual(values = c(`TRUE` = "blue", `FALSE` = "red")) +
  labs(
    x = "Proportion say 'Yes'",
    title = "50 bootstrap 95% CIs for a proportion"
  )
```


50 bootstrap 95% CIs for a proportion
0.4
0.5
0.6
0.7
0.8
Proportion say 'Yes'
covers
FALSE
TRUE

Answer 6