Data606 - Lab 5b - Confidence intervals

Avery Davidowitz

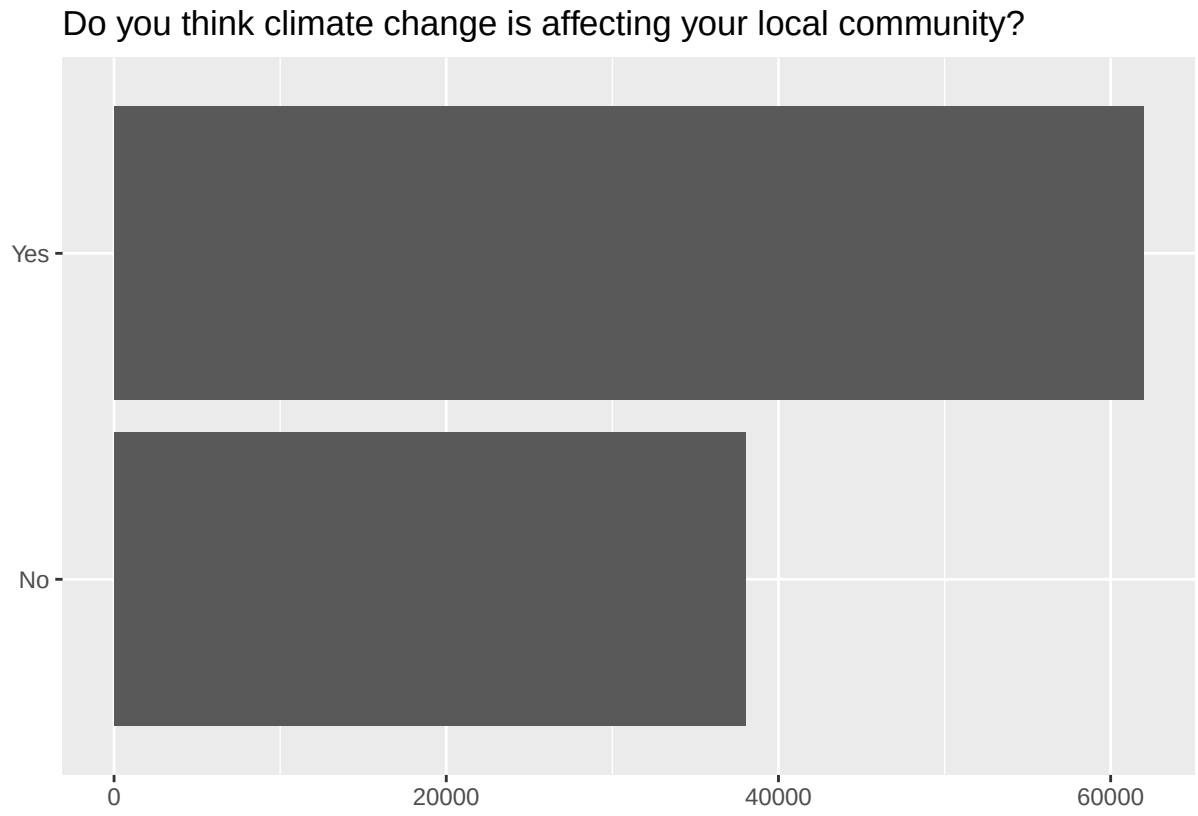
2022-10-16

Load Packages and Set Seed

```
library(tidyverse)
library(openintro)
library(infer)
library(png)
set.seed(21)
```

Generate Data

```
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()
```




```
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
```

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.

```
n <- 60
samp <- us_adults %>%
  sample_n(size = n)
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                        22 0.367
## 2 Yes                       38 0.633
```

63% of sampled adults believe that climate change will affect their community.

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?

I would expect it to be similar but not exactly the same. Since the sample is random and np and n*(1-p) are both greater than 10 the central limit theorem would apply to the samples. Therefore, the random samples should be both centered around the same population proportion.

Exercise 3

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

```
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      0.5     0.75
```

The 95% confidence interval means that due to the sample data we are 95% sure that the true population proportion is between the lower and upper bounds of the confidence interval.

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?

Yes. The true population proportion of 62% falls between the interval of .5 and .75.

Exercise 5

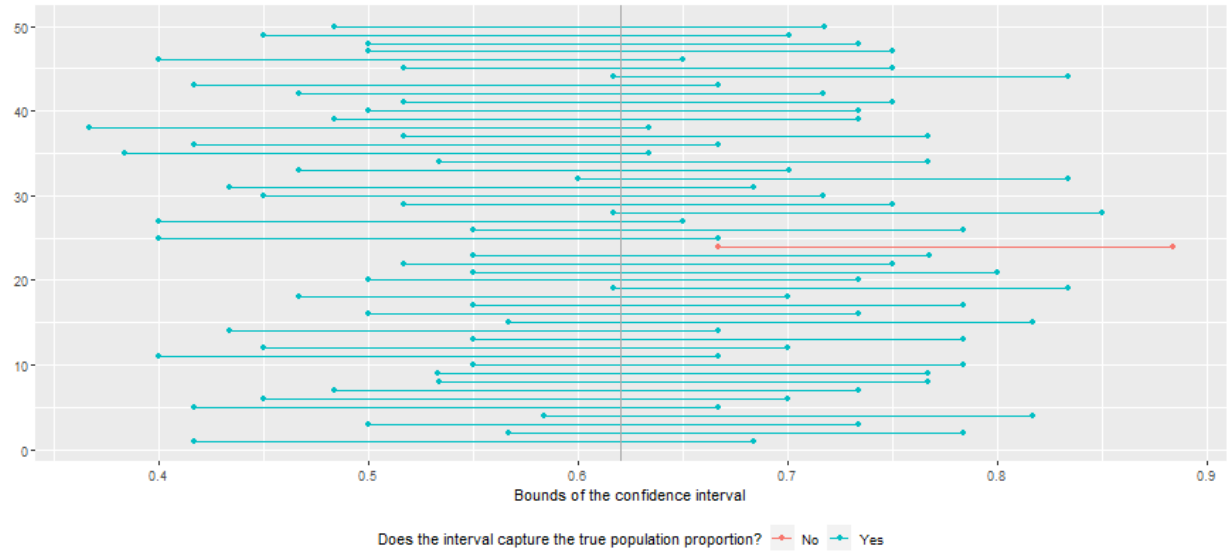
Each student should have gotten a slightly different confidence interval. What proportion of those intervals would you expect to capture the true population mean? Why?

I would expect 95% of the class to have the confidence interval contain the true population proportion of 62%. 95% confidence means that if we repeated that sample 100 times(possibly with a different student each time) we would expect to get at least 95 in the range.

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.

```
img <- readPNG("C:/Users/ADavidowitz/Documents/GitHub/DATA606/Lab5/n601000b50ci.PNG")
grid::grid.raster(img)
```



I would expect that up to 5% of the bootstrap simulated confidence intervals to not include the population proportion. My simulation yielded 1/50 outside of the interval .02 or 2% which is less than 5%.

Exercise 7

Choose a different confidence level than 95%. Would you expect a confidence interval at this level to me wider or narrower than the confidence interval you calculated at the 95% confidence level? Explain your reasoning.

I will use a confidence interval of 98%. Therefore, I expect the range to be wider to allow more certainty.

Exercise 8

Using code from the infer package and data from the one sample you have (samp), find a confidence interval for the proportion of US Adults who think climate change is affecting their local community with a confidence level of your choosing (other than 95%) and interpret it.

```
samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.98)
```

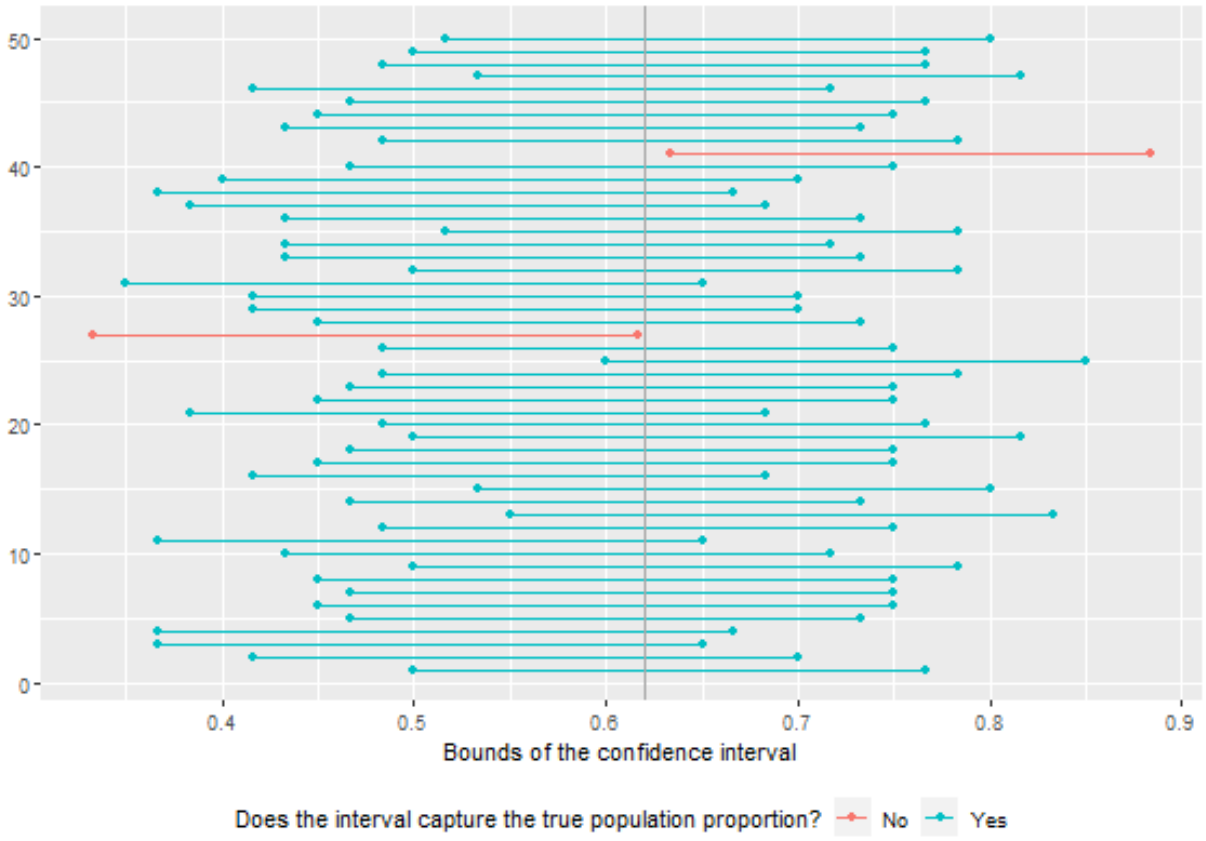
```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.483    0.767
```

As expected the confidence interval for 98% is wider than 95%. We need to allow more of a range to encompass the bigger Z score used in the formula of p hat +/- Z X SE.

Exercise 9

Using the app, calculate 50 confidence intervals at the confidence level you chose in the previous question, and plot all intervals on one plot, and calculate the proportion of intervals that include the true population proportion. How does this percentage compare to the confidence level selected for the intervals?

```
img <- readPNG("C:/Users/ADavidowitz/Documents/GitHub/DATA606/Lab5/n601000b50ci98cl.PNG")
grid::grid.raster(img)
```



I had only 48/50 simulated confidence intervals fall in the correct range despite specifying a 98% confidence level. This does not seem possible because that would be an error of 4% which is higher than the 2% expected error.

Exercise 10

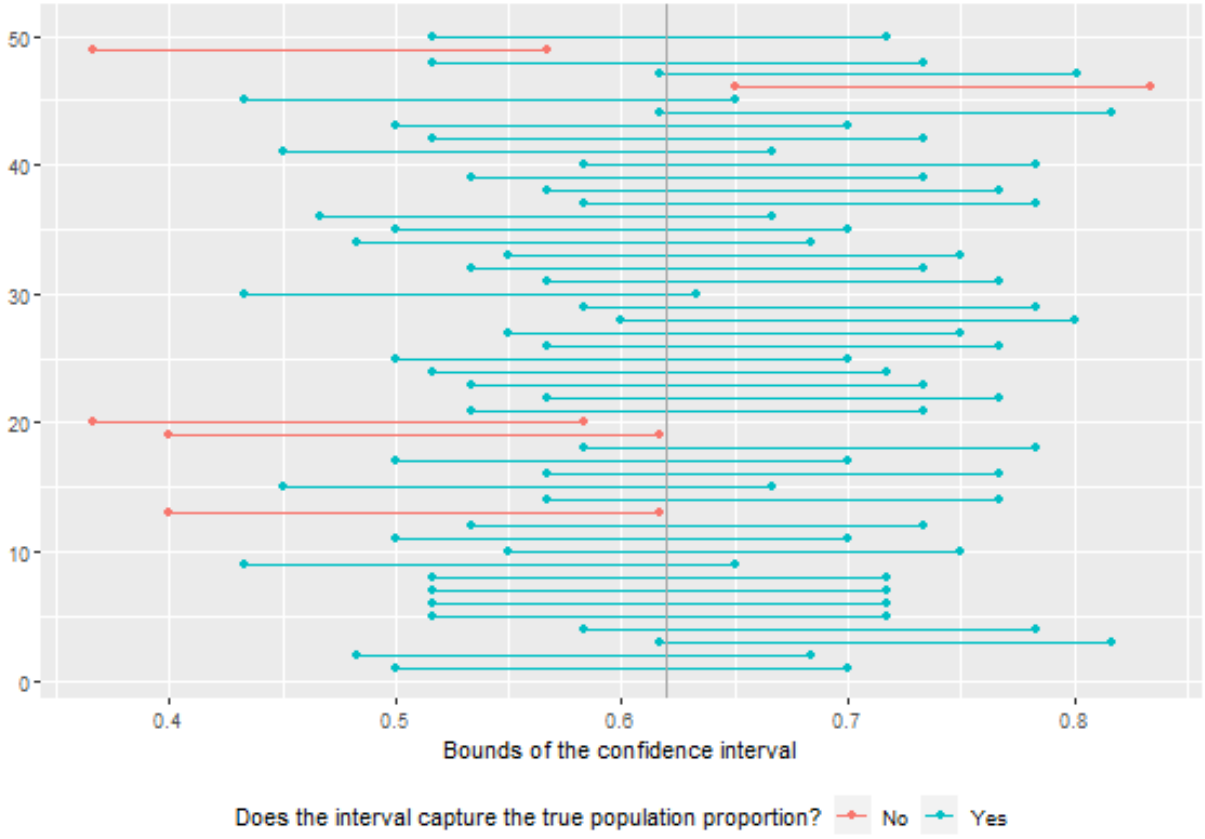
Lastly, try one more (different) confidence level. First, state how you expect the width of this interval to compare to previous ones you calculated. Then, calculate the bounds of the interval using the infer package and data from samp and interpret it. Finally, use the app to generate many intervals and calculate the proportion of intervals that are capture the true population proportion.

I will choose a CL of 90%. I expect the range of the confidence interval to be much smaller.

```
samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.90)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.533    0.733
```

```
img <- readPNG("C:/Users/ADavidowitz/Documents/GitHub/DATA606/Lab5/n601000b50ci90cl.PNG")
grid::grid.raster(img)
```



As expected the CI range was smaller from the infer package and there were an appropriate number of 5/50 (10%) confidence intervals generated by the shiny app that did not contain the desired population proportion. We had more errors because we specified a lower confidence.

Exercise 11

Using the app, experiment with different sample sizes and comment on how the widths of intervals change as sample size changes (increases and decreases).

```
n <- 120
samp2 <- us_adults %>%
  sample_n(size = n)
samp2 %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.90)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.492    0.642
```

```
samp2 %>%
  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    0.475     0.65
```

```
samp2 %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.98)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.467    0.658
```

It appears that the CI for a given level is tighter in spread if the sample size is increased.

Exercise 12

Finally, given a sample size (say, 60), how does the width of the interval change as you increase the number of bootstrap samples. Hint: Does changing the number of bootstrap samples affect the standard error?

The width of the errors does not seem to be closely correlated with the number of bootstrap samples. The width was the same for 500 or 5000 bootstraps.