# Data 606 - Lab 5B

Cameron Smith

2020-10-05

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

## Getting Started

### 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.*

55% of the population within the sample drawn think climate change affects their community.

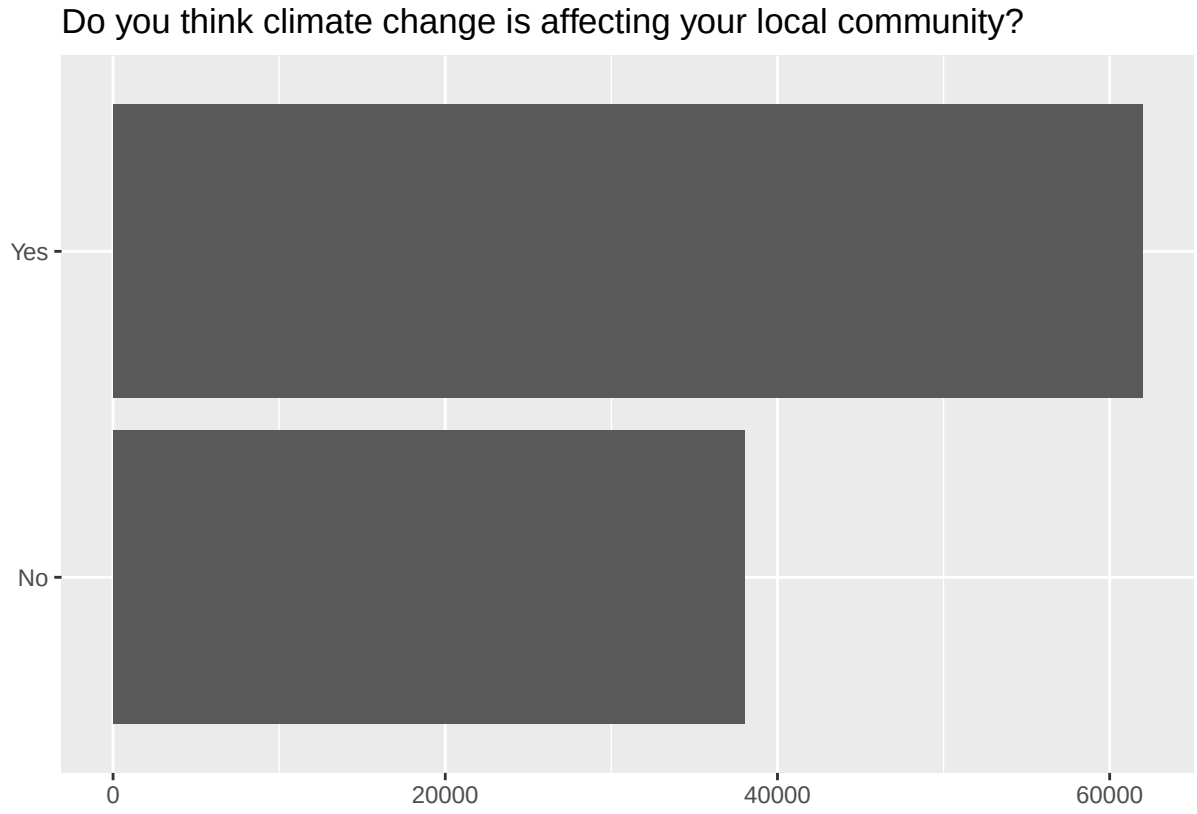
```
# Lab intro code

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

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
```

```
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()
```




```
# Code for exercise 1 starts here

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                        21  0.35
## 2 Yes                       39  0.65
```

### 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 another student's sample proportation to be similar, but not identical because they would have most likely used a different random seed.

## Confidence intervals

### Exercise 1

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

In this context 95% confident means that we are 95% confident that these figures are reflective of the total population.

```
samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = .95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.533    0.767
```

### Exercise 2

*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?*

Not necessarily, but at 95% it is pretty close. My neighbor's interval would not necessarily capture the same value. I simulated this below with a new sample which shows bounds of 53.3% and 76.7% (compared to 41.7 % and 66.7%). The range is similar but the bounds are different.

```
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 = .95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.517     0.75
```

\newline

### Exercise 3

*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 them to capture the true population of the mean as that is the confidence level used.

### Exercise 4

*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.*

47 / 50, or 94% of the confidence intervals include the true population proportion. It is not exactly equal to the confidence level (95%), though it's pretty close. When I ran it again, but based on 100 confidence intervals, it was exactly 95%. The reason for this is that although intervals provide a range of plausible values, it is not necessarily true that values out of that range are impossible - only implausible.

```
samp3 <- us_adults %>%
  sample_n(size = 60)

samp3 %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop")
```

```
## # A tibble: 1,000 x 2
##    replicate  stat
##        <int> <dbl>
##  1         1 0.6
##  2         2 0.567
##  3         3 0.667
##  4         4 0.5
##  5         5 0.533
##  6         6 0.483
##  7         7 0.7
##  8         8 0.667
##  9         9 0.567
## 10        10 0.55
## # ... with 990 more rows
```

## More Practice

### Exercise 1

*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.*

If choosing a level higher than 95% then I would expect an interval to be wider, and if less than 95% I would expect it to be narrower. It goes back to the fishing analogy in the book - if you want to be more confident you'll catch a fish then you need to use a bigger net.

### Exercise 2

*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.*

Based on this sample, with sample size of 60, 1000 resamples for each bootstrap CI, and a confidence level of 90%, the range is between .51 to .71.

```
samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = .90)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     0.55     0.75
```

### Exercise 3

*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?*

Of the 50 confidence intervals, 45 of them, or 90% include the true population proportion. This is spot on with the 90% confidence level that I used.

### Exercise 4

*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.*

For this exercise I chose a 98% confidence level. I expect the width to be much wider than the previous ones I calculated given that a "wider net" will be needed to "catch" the number.

The calculated bounds are between 47% and 77%, which is in line with the above expectation.

I used the app to generate 100 different intervals, and of those 96 (96%) of them captures the true population proportion.

```
samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = .98)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.517    0.783
```

### Exercise 5

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

The below is all based on 1000 resamples and a 95% confidence level used, with 50 confidence intervals. The results are below, and I observed that as the sample size increases the interval width decreases.

Sample size 5: 45/50 capture true population, or 90% Sample size 30: 45/50 capture true population, or 90% Sample size 50: 46/50 capture true population, or 92% Sample size 100: 48/50 capture true population, or 96% Sample size 1000: 46/50 capture true population, or 92%

### Exercise 6

*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?*

It seems that the confidence intervals are narrower based on the higher the number of intervals used in the app. The reason seems to be that yes, the standard error changes based on the number of bootstrap samples. More samples results in lower standard error.