Data_606_Lab_5_Foundations for statistical inference - Confidence intervals

Enid Roman

2022-10-10

```
set.seed(500)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

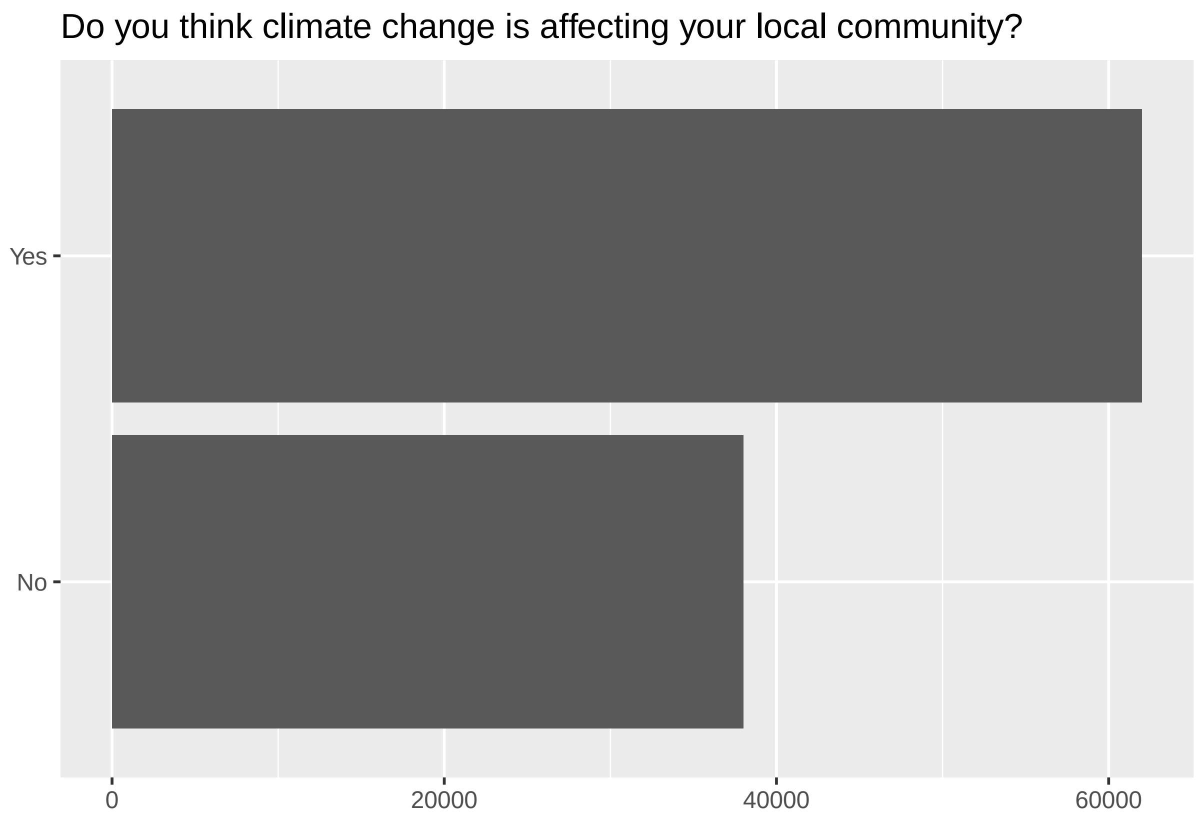
```
library(openintro)
```

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

```
library(infer)
```

```
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
```

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

Exercise 1

```

us_adults %>%
  count(climate_change_affects) %>%
  mutate(p = n /sum(n))
```

62% of the adults in my sample think climate change affects their local community.

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

Exercise 2

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

I would expect another student's sample proportion to be similar but not identical to mine because sample proportions can vary from sample to sample by taking smaller samples from the population. The sample is randomly selected so the sample result would always come out a little different everytime you run the sample code.

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.517    0.767
```

Exercise 3

95% confidence mean 95% confident that the population mean lies within the interval between a lower bound and an upper bound. A confidence interval only provides a plausible range of values.

Exercise 4

Yes confidence interval capture the true population proportion of US adults who think climate change affects their local community. If I was working on this lab in a classroom, my neighbor's interval would have gotten a slightly different confidence interval. The confidence interval is thus a statement about the estimation procedure and not about the specific interval generated in the sample

Exercise 5

The confidence interval ranging from .55 to .783 would be expected to cover the true population proportion 95% of the time because a confidence interval only provides a plausible range of values. While we might say other values are implausible based on the data, this does not mean they are impossible.

LINK TO GITHUB FOR PICTURES OF RESULTS FROM APP

https://github.com/enidroman/Data_606_Satistics_and_Probability_for_Data_Analytics/blob/main/Data_606_Lab_5_Confidence%20Intervals_Images.pdf

Exercise 6 (See pics at github link)

Given a sample size of 60, 1000 bootstrap samples for each interval, and 50 confidence intervals constructed (the default values for the above app), the proportion of my confidence intervals include the true population proportion is 24/25 = .96. It is safe to assume 95% of the confidence intervals demonstrated would include true populationis proportion exactly equal to the confidence level.

Exercise 7 (See pics at github link)

I chose a different confidence level other than 95%. Given a sample size of 60, 1000 bootstrap samples for each interval, and 50 confidene intervals, I chose 90% confidence level. I expect a confidence interval at this level to be narrower (plus or minus 2.5 percent, for example) than the confidence interval I calculated at the 95% confidence level. The reason for this is , if we want an interval with lower confidence, such as 90%, we could use a slightly narrower interval than our original 95% interval. You have that 5% range of incorrectness in this case. When the precision of the confidence interval increases we could assume true population proportion interval decreases.

Exercise 8

```
samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.99)
```

Using the code from the infer package and data from the one sample I have (samp), I chose a confidence level of 99% for the proportion of US Adults who think climate change is affecting their local community. We are 99% level of confident that the proportion of US adults who think climate change affects their local community is between .48 adn .81.

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.483    0.817
```

Exercise 9 (See pics at github link)

Using the app given a sample size of 60, 1000 bootstrap samples for each interval, and 50 confidene intervals, at the confidence level that I chose in the previous question which was 99%, the proportion of my confidence intervals include the true population proportion is 48/50 = .96. Using the code from infer the percentage is lower then using the app.

Exercise 10 (See pics at github link)

```
samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.80)
```

Using the code from the infer package and data from the one sample I have (samp), I chose a confidence level of 80% for the proportion of US Adults who think climate change is affecting their local community. I expect a confidence interval at this level to be more narrower than all the confidence interval I calculated before. The reason for this is , if we want an interval with lower confidence, such as 80%, we could use a slightly narrower interval than our original like the previous interval. We are 80% level of confident that the proportion of US adults who think climate change affects their local community is between .57 adn .73. Using the app the proportion of my confidence intervals include the true population proportion is 38/50 = .76. At a confidence level of 80% this low.

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.567    0.733
```

Exercise 11 (See pics at github link)

Using the app, given a sample size of 100, 1000 bootstrap samples for each interval, and 50 confidence intervals constructed the proportion of my confidence intervals include the true population proportion is 46/50 = .92. Increasing the the sample size to 100 it increased the confidence intervals to the true population porpotion to 92. The width of intervals has gotten much wider then the previous.

Exercise 12 (See pics at github link)

Using the app given a sample size of 60, 10000 bootstrap samples for each interval, and 50 confidence intervals constructed the proportion of my confidence intervals include the true population proportion is 41/50 = .82. Increasing the bootstrap to 10000 it increased the confidence intervals to the true population porpotion to 82. The width of intervals has gotten a little wider then the previous.