Foundations for statistical inference - Confidence intervals

Alice Ding

Exercise 1

```
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()
```

What percent of the adults in your sample think climate change affects their local community?


```
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)
```

```
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                        26 0.433
## 2 Yes                       34 0.567
```

In my sample, 34/60 (56%) of adults think climage change affects their local 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? If we had the same seed, then yes it'd be identical. Otherwise, our proportions would likely not be identical, however it would be similar. This is due to randomness; our samples should be somewhat representative of the population, so we would both be close to 38%, however they would likely not be exactly the same.

Exercise 3

In the interpretation above, we used the phrase "95% confident". What does "95% confidence" mean? "95% confidence" means that we are 95% certain that the true proportion we are looking for is in between those bounds. It could inversely mean that we have a 5% chance of being incorrect with our boundaries.

Exercise 4

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

Does your confidence interval capture the true population proportion of US adults who think climate change affects their local community?

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.433      0.7
```

My confidence interval of 0.433 to 0.7 does capture the true population proportion of US adults who think climate change affects their local community as the actual number is 62%.

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? 95%. This is because this is a 95% confidence interval – therefore, we are 95% sure that our confidence interval has the true population mean. Extrapolating that to the entire class, that means that 95% of us have a correct range of values and 5% do not.

Exercise 6

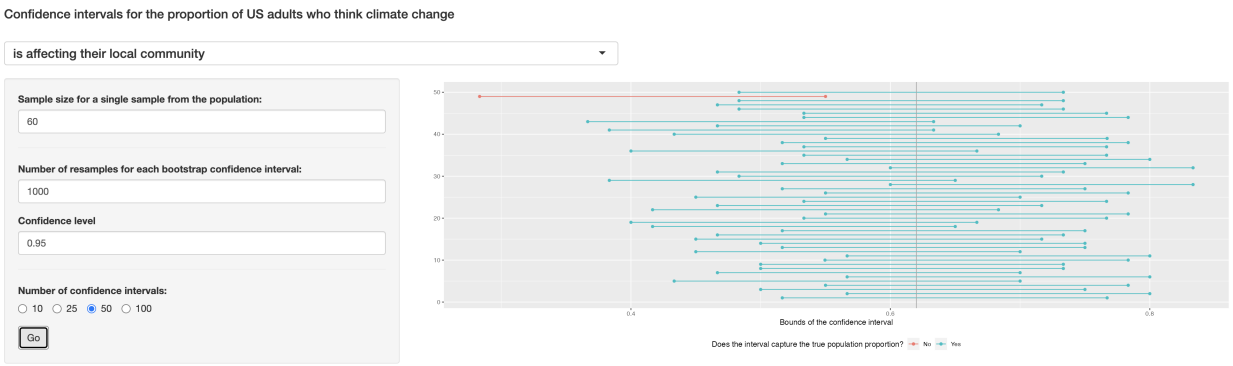

Figure 1: Screenshot of my Confidence Interval Chart

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. It looks like 49/50 of my confidence intervals constructed include the true population proportion. Since we're 95% confident, it's not exact – I'm at 98%. This is likely due to just the randomness of our samples and I managed to get a little luckier where my samples actually matched.

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 choose 90%. I would expect a confidence interval to be narrower than 95% because in this scenario, I would only be 90% confident. The bounds would then be more restrictive since this time, 10% of my intervals wouldn't fit the true proportion rate.

Exercise 8

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

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.

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.417    0.733
```

Here, I chose a 99% confidence level and the bounds are 0.417 to 0.733. This is wider than the 95% confidence interval example and it makes sense as this is a higher level of confidence so it'd be a bit wider. This means we are 99% confident (or sure) that the population proportion is within these bounds and we are correct since 0.62 is between those two numbers.

Exercise 9

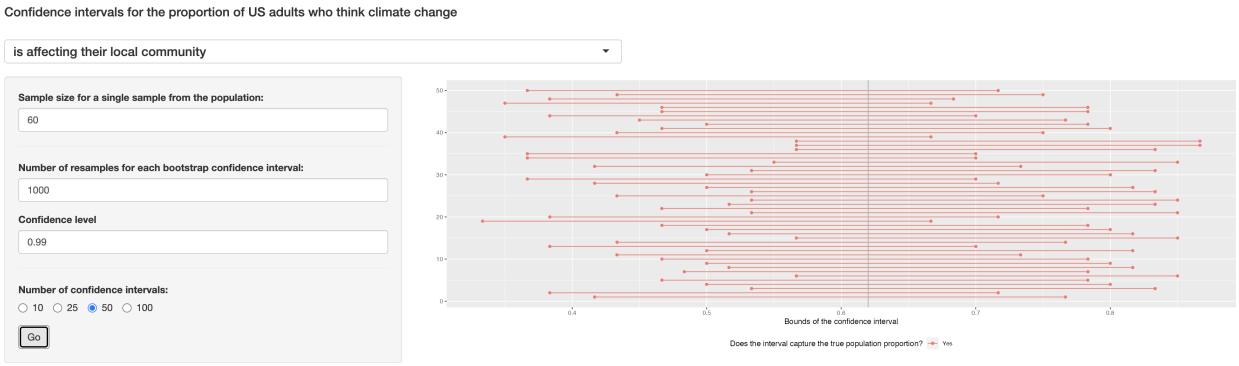

Figure 2: Screenshot of my Confidence Interval Chart at 99%

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? Here, all of my confidence intervals have the true population proportion. This is 100% and the confidence interval is 99% – I just got a bit lucky that it didn't end up at 98% (1 not having it), but this is very likely due to the randomness of sampling.

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'm choosing a confidence interval of 85 – I expect this to be much narrower than the other two confidence levels due to it being much lower.

```
samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.85)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.467    0.667
```

The interval is from 0.467 to 0.667 – much smaller than the other ranges of numbers. Now what does the app show?

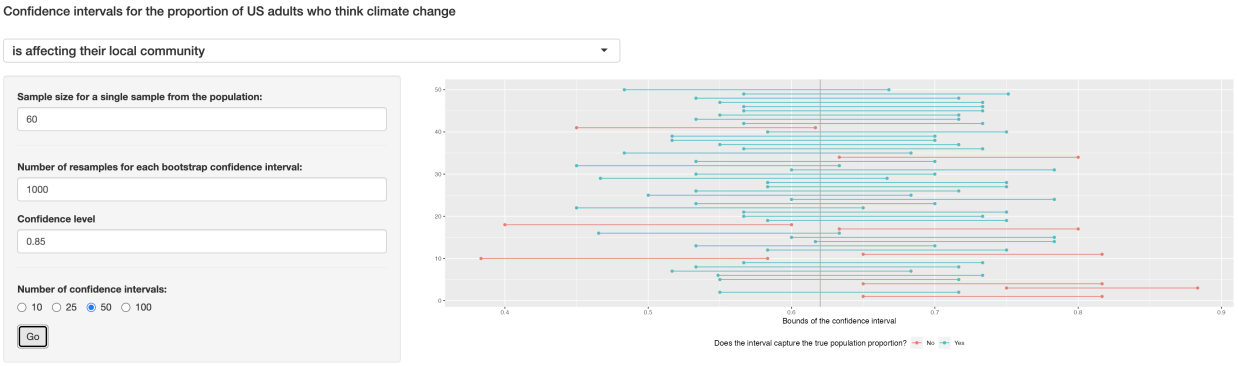

Figure 3: Screenshot of my Confidence Interval Chart at 85%

This time, 9 of the 50 confidence intervals don't contain the true population proportion – this means 82% of them did. Compared to 85%, that's pretty close!

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). The larger the sample size gets, the narrower the interval grows.

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 bootstap samples affect the standard error? The width of the interval grows narrower as we increase the bootstrap samples because the standard error seems to grow smaller when we increase the number of bootstrap samples.