Foundations for statistical inference - Confidence intervals

If you have access to data on an entire population, say the opinion of every adult in the United States on whether or not they think climate change is affecting their local community, it's straightforward to answer questions like, "What percent of US adults think climate change is affecting their local community?". Similarly, if you had demographic information on the population you could examine how, if at all, this opinion varies among young and old adults and adults with different leanings. If you have access to only a sample of the population, as is often the case, the task becomes more complicated. What is your best guess for this proportion if you only have data from a small sample of adults? This type of situation requires that you use your sample to make inference on what your population looks like.

Setting a seed: You will take random samples and build sampling distributions in this lab, which means you should set a seed on top of your lab. If this concept is new to you, review the lab on probability.

Getting Started

Load packages

In this lab, we will explore and visualize the data using the **tidyverse** suite of packages, and perform statistical inference using **infer**.

Let's load the packages.

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

The data

A 2019 Pew Research report states the following:

To keep our computation simple, we will assume a total population size of 100,000 (even though that's smaller than the population size of all US adults).

> Roughly six-in-ten U.S. adults (62%) say climate change is currently affecting their local community either a great deal or some, according to a new Pew Research Center survey.
>
> **Source:** Most Americans say climate change impacts their community, but effects vary by region

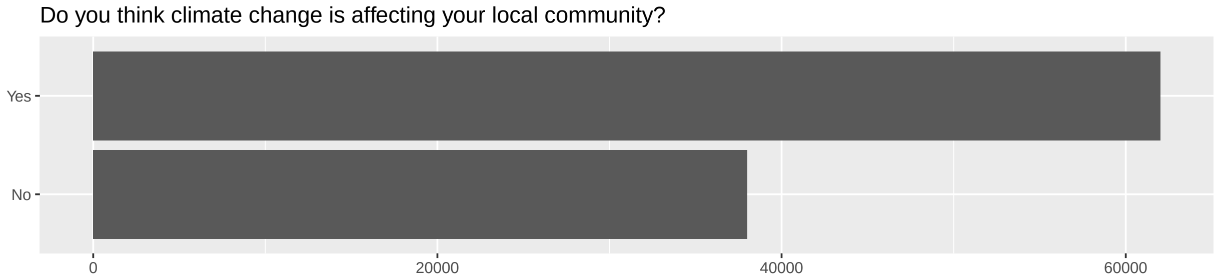
In this lab, you will assume this 62% is a true population proportion and learn about how sample proportions can vary from sample to sample by taking smaller samples from the population. We will first create our population assuming a population size of 100,000. This means 62,000 (62%) of the adult population think climate change impacts their community, and the remaining 38,000 does not think so.

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

The name of the data frame is `us_adults` and the name of the variable that contains responses to the question *"Do you think climate change is affecting your local community?"* is `climate_change_affects`.

We can quickly visualize the distribution of these responses using a bar plot.

```
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()
```



We can also obtain summary statistics to confirm we constructed the data frame correctly.

```
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
```

In this lab, you'll start with a simple random sample of size 60 from the population.

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

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.

```
library(dplyr)

# Creating a tibble with the climate change data
us_adults <- tibble(climate_change_affects = c(rep("Yes", 62000), rep("No", 38000)))

# Setting up the seed for reproducibility
set.seed(1234)
```

```r
# Sample 60 individuals from the above given dataset
samp <- sample_n(us_adults, size = 60)

# Calculating the percentage of respondents who believe climate change affects them
percentage_yes <- mean(samp$climate_change_affects == "Yes") * 100

# Displaying the result
percentage_yes
```

```
## [1] 61.66667
```

Approximately 61.7% of the adults in the sample believe that climate change impacts their local community.

1. 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 anticipate that another student's sample proportion would be comparable but not exactly the same as mine. Since they probably used a different seed, their sample would differ, leading to a different proportion. However, I expect it to be similar because both samples are drawn from the same population.

Confidence intervals

Return for a moment to the question that first motivated this lab: based on this sample, what can you infer about the population? With just one sample, the best estimate of the proportion of US adults who think climate change affects their local community would be the sample proportion, usually denoted as $\hat{p}$ (here we are calling it `p_hat`). That serves as a good **point estimate**, but it would be useful to also communicate how uncertain you are of that estimate. This uncertainty can be quantified using a **confidence interval**.

One way of calculating a confidence interval for a population proportion is based on the Central Limit Theorem, as $\hat{p} \pm z^\star SE_{\hat{p}}$ is, or more precisely, as

$$\hat{p} \pm z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Another way is using simulation, or to be more specific, using **bootstrapping**. The term **bootstrapping** comes from the phrase "pulling oneself up by one's bootstraps", which is a metaphor for accomplishing an impossible task without any outside help. In this case the impossible task is estimating a population parameter (the unknown population proportion), and we'll accomplish it using data from only the given sample. Note that this notion of saying something about a population parameter using only information from an observed sample is the crux of statistical inference, it is not limited to bootstrapping.

In essence, bootstrapping assumes that there are more of observations in the populations like the ones in the observed sample. So we "reconstruct" the population by resampling from our sample, with replacement. The bootstrapping scheme is as follows:

- **Step 1.** Take a bootstrap sample - a random sample taken **with replacement** from the original sample, of the same size as the original sample.
- **Step 2.** Calculate the bootstrap statistic - a statistic such as mean, median, proportion, slope, etc. computed on the bootstrap samples.

- **Step 3.** Repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics.
- **Step 4.** Calculate the bounds of the XX% confidence interval as the middle XX% j knof the bootstrap distribution.

Instead of coding up each of these steps, we will construct confidence intervals using the **infer** package.

Below is an overview of the functions we will use to construct this confidence interval:

Function	Purpose
`specify`	Identify your variable of interest
`generate`	The number of samples you want to generate
`calculate`	The sample statistic you want to do inference with, or you can also think of this as the population parameter you want to do inference for
`get_ci`	Find the confidence interval

This code will find the 95 percent confidence interval for proportion of US adults who think climate change affects their local community.

```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
```

- In `specify` we specify the `response` variable and the level of that variable we are calling a `success`.
- In `generate` we provide the number of resamples we want from the population in the `reps` argument (this should be a reasonably large number) as well as the type of resampling we want to do, which is `"bootstrap"` in the case of constructing a confidence interval.
- Then, we `calculate` the sample statistic of interest for each of these resamples, which is `prop`ortion.

Feel free to test out the rest of the arguments for these functions, since these commands will be used together to calculate confidence intervals and solve inference problems for the rest of the semester. But we will also walk you through more examples in future chapters.

To recap: even though we don't know what the full population looks like, we're 95% confident that the true proportion of US adults who think climate change affects their local community is between the two bounds reported as result of this pipeline.

Confidence levels

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

A 95% confidence level indicates that we are 95% certain the true population proportion falls within the interval we calculated from our sample data. This level of confidence corresponds to a normal distribution where 5% of the area remains unshaded, reflecting the likelihood that the true proportion lies outside this interva

In this case, you have the rare luxury of knowing the true population proportion (62%) since you have data on the entire population.

1. 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?

```r
library(dplyr)
library(infer)

# Set the seed for reproducibility
set.seed(1234)

# Generate bootstrap confidence intervals for the proportion
ci_results <- samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)

# Display the confidence interval results
ci_results
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.500     0.75
```

Yes, the confidence interval provided below accurately encompasses the true population proportion of adults who believe climate change impacts their local community. Since the actual proportion is 62%, it lies within the range of 49.99% to 75.00%. If I were to conduct this analysis using different seeds, I would expect the intervals to capture this true proportion approximately 95% of the time.

1. 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 anticipate that approximately 95% of the intervals would include the true population mean. Given that this is a 95% confidence interval, there is a 95% chance that the population mean lies within this range. This also implies that 95% of the time, the true mean will be contained within the interval.

In the next part of the lab, you will collect many samples to learn more about how sample proportions and confidence intervals constructed based on those samples vary from one sample to another.

- Obtain a random sample.
- Calculate the sample proportion, and use these to calculate and store the lower and upper bounds of the confidence intervals.

- Repeat these steps 50 times.

Doing this would require learning programming concepts like iteration so that you can automate repeating running the code you've developed so far many times to obtain many (50) confidence intervals. In order to keep the programming simpler, we are providing the interactive app below that basically does this for you and created a plot similar to Figure 5.6 on OpenIntro Statistics, 4th Edition (page 182).

1. 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.

In my simulation, 55 out of 60 confidence intervals captured the true population proportion, which is 0.92. This result doesn't perfectly match the confidence level. A 95% confidence interval indicates that, on average, about 95% of the intervals will include the true proportion, but this does not guarantee it will happen every time. In some cases, we might find that 94% or 96% of the intervals contain the true mean. Therefore, it reflects an average percentage rather than an exact value.

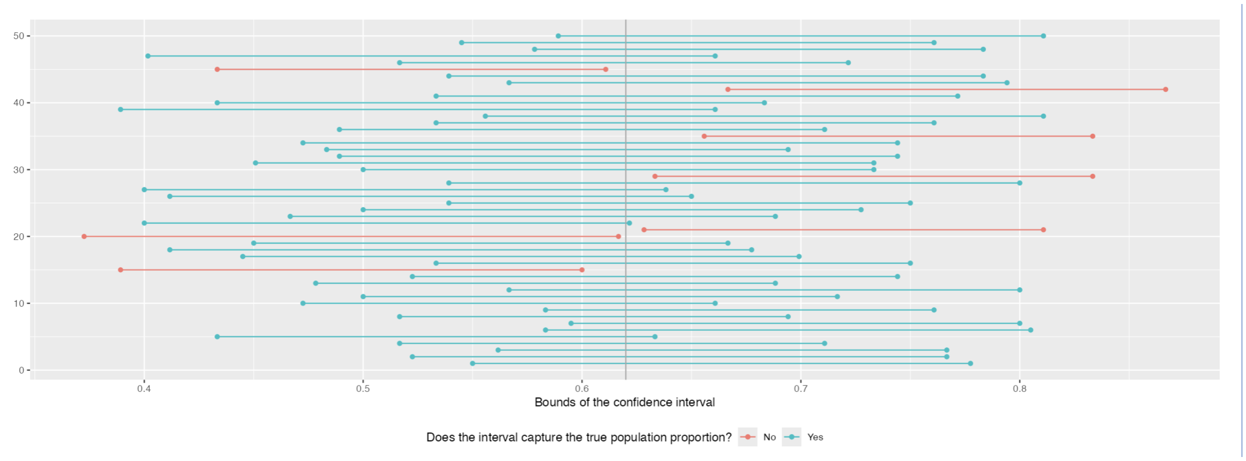


Figure 1: alt text here

More Practice

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.

I selected a confidence level of 90%, which I anticipate will result in a narrower interval compared to the 95% confidence interval. With a lower confidence level, there is a reduced likelihood that the mean falls within this range, so the interval is likely to be narrower. This means it will encompass fewer values, leading to a decreased probability of including the true mean.

1. Using code from the **infer** package and data fromt 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.

```r
library(dplyr)
library(infer)

# Setting the seed for reproducibility
set.seed(1234)

# Generating bootstrap confidence intervals for the proportion at a 90% confidence level
ci_results_90 <- samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  # Generating 1000 bootstrap resamples for the proportion
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.90)

# Displaying the confidence interval results
ci_results_90
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.517    0.717
```

This is a 90% confidence interval, indicating that I am 90% confident the true population proportion of 62% lies within the range of 0.5167 to 0.7167. Since the mean is included in this range, it aligns with the expectation that it is more likely for the mean to fall within this interval than outside of it. If I were to generate 100 confidence intervals without setting a seed, approximately 90 of those intervals would encompass 0.62, while around 10 would not.

1. 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?

I selected a 90% confidence interval, and out of the 60 intervals, 7 did not include the mean, while 53 did. Calculating this gives 53 / 60 = 0.8833, meaning 88.33% of the intervals contained the mean. This result is quite close to the expected 90%.

1. 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.

```r
library(dplyr)
library(infer)

# Setting the seed for reproducibility
set.seed(1234)

# Generating bootstrap confidence intervals for the proportion at a 50% confidence level
ci_results_50 <- samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.50)

# Display the confidence interval results
ci_results_50
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.579    0.667
```

I will attempt a confidence level of 50%, and I anticipate that this interval will be significantly narrower than those I calculated previously. As shown from app/code, the true population proportion of 0.62 did fall within this range, which was somewhat surprising given that there was only a 50-50 chance of this occurring. However, since it was very close to the bounds, this outcome was not entirely unexpected. When I utilized the app, the proportion of intervals that included the true population proportion was around 0.44.

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

As the sample size grew, the widths of the confidence intervals narrowed. Conversely, when the sample size was reduced, the widths of the intervals expanded. Both of these statements hold true as long as the confidence level remains unchanged. However, it was difficult to discern this trend because the widths of the confidence intervals varied in one of the graphs.

1. 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 widths of the intervals appeared to shrink as the number of bootstrap samples increased. However, it was difficult to confirm this observation because the confidence interval widths varied across the graph.
