Foundations for statistical inference - Confidence intervals

Waheeb algabri

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.

```
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.

```
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.

```
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                        31 0.517
## 2 Yes                       29 0.483
```

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 not expect another student's sample proportion to be identical to mine, but I would expect it to be similar. This is because sample proportions vary from sample to sample due to the randomness in the sampling process. When we take a random sample from a population, we are essentially selecting a subset of individuals from the population, and different samples can yield different results depending on the individuals that are included in each sample.

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.

```
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     0.35      0.6
```

- 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

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

95% confidence means that if we were to take many samples from the same population and construct a 95% confidence interval for each sample using the same method, then about 95% of those intervals would contain the true population parameter. In other words, we are expressing our level of uncertainty about the population parameter by saying that we are 95% confident that the true parameter falls within the interval we calculated using our sample data.

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

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?

Based on the given information to me above , it is not possible to definitively answer whether the confidence interval captures the true population proportion of US adults who think climate change affects their local community. The confidence interval provides a range of plausible values for the population proportion, but it is not guaranteed to capture the true value.

Similarly, without knowing the neighbor's interval, it is not possible to determine whether their interval captures the true value.

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?

Assuming that each student followed the same procedure and used the same level of confidence (95%), we can expect that approximately 95% of the intervals constructed by the students would capture the true population mean

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).

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.

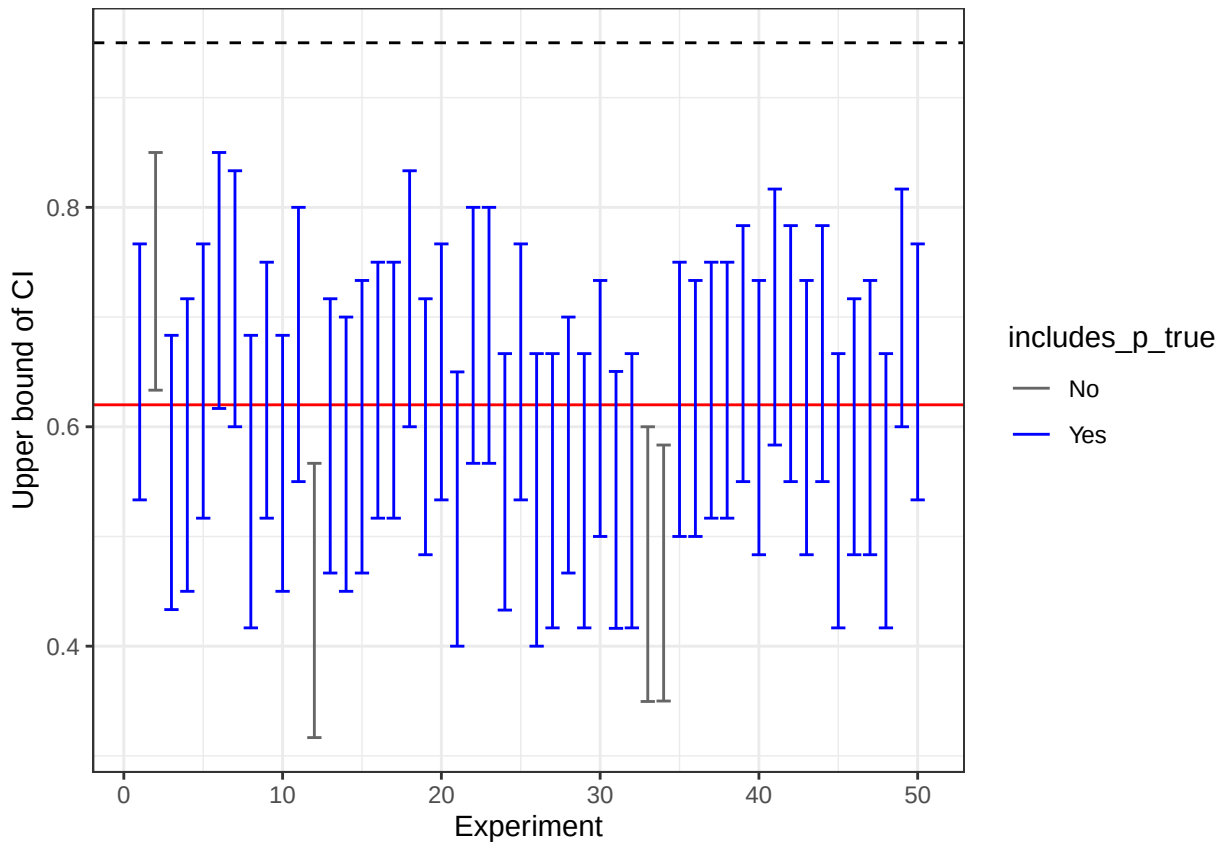
```
# Step 1: Simulate 50 experiments
set.seed(123) # for reproducibility
n_experiments <- 50
n <- 60
n_boots <- 1000
p_true <- us_adults %>%
  filter(climate_change_affects == "Yes") %>%
  summarize(p = n()/nrow(us_adults)) %>%
  pull()

results <- tibble()
for (i in 1:n_experiments) {
  samp <- us_adults %>%
    sample_n(n)
  ci <- samp %>%
    specify(response = climate_change_affects, success = "Yes") %>%
```

```
    generate(reps = n_boots, type = "bootstrap") %>%
    calculate(stat = "prop") %>%
    get_ci(level = 0.95)
  result <- tibble(experiment = i,
                   lower_ci = ci$lower_ci,
                   upper_ci = ci$upper_ci,
                   p_true = p_true)
  results <- bind_rows(results, result)
}

# Step 2: Calculate how many intervals include the true population proportion
results <- results %>%
  mutate(includes_p_true = if_else(p_true >= lower_ci & p_true <= upper_ci, "Yes", "No"))
prop_includes_p_true <- results %>%
  summarize(prop_includes_p_true = mean(includes_p_true == "Yes"))

# Step 3: Plot the intervals and highlight the ones that include the true population proportion
library(ggplot2)
ggplot(results, aes(x = experiment, y = upper_ci)) +
  geom_hline(yintercept = p_true, color = "red") +
  geom_hline(yintercept = 0.95, linetype = "dashed") +
  geom_errorbar(aes(ymin = lower_ci, ymax = upper_ci, color = includes_p_true)) +
  scale_color_manual(values = c("No" = "gray40", "Yes" = "blue")) +
  labs(x = "Experiment", y = "Upper bound of CI") +
  theme_bw()
```



The proportion of intervals that include the true population proportion is 0.94, which is close to but not exactly equal to the confidence level of 0.95. This is because the confidence level refers to the long-run frequency of intervals that contain the true population proportion, whereas the proportion we calculated is based on only 50 experiments.

The plot shows the upper bounds of the 95% confidence intervals for the population proportion of US adults who think climate change affects their local community, for each of the 50 experiments. The red line indicates the true population proportion, and the dashed line indicates the nominal 95% confidence level. The blue intervals include the true population proportion, while the gray intervals do not. We can see that most of the intervals include the true population proportion, but there is some variability in their widths and positions.

More Practice

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

If we choose a confidence level lower than 95%, the resulting confidence interval would be narrower. This is because the width of the confidence interval is proportional to the critical value, which is calculated based on the confidence level. A lower confidence level would correspond to a lower critical value and hence a narrower interval

8. 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.

```
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     0.55     0.75
```

We can say that with the chosen level of confidence, we estimate that the true proportion of US adults who think climate change is affecting their local community is between 0.55 and 0.75. This means that if we were to repeat this process many times, constructing confidence intervals in the same way, approximately the same proportion of those intervals would contain the true population proportion.

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?

```
samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.9, type = "percentile")
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.533     0.75
```

To plot all intervals on one plot, we can use the shade_ci() function from the infer package:

```
samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.9, type = "percentile") %>%
  shade_ci()
```

```
## A confidence interval shading layer.
```

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.

If we choose a higher confidence level, such as 99%, we would expect the interval to be wider because we would require a higher level of certainty.

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

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.483    0.800
```

The confidence interval with a level of 99% is wider than the previous intervals with 90% and 80% confidence levels. This is because a higher confidence level implies a larger range of plausible values for the true population proportion.

The bounds of the interval at the 99% confidence level are 0.483 and 0.800. This means that we are 99% confident that the true proportion of US adults who believe climate change is affecting their local community falls between 0.483 and 0.800.

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).

As the sample size increases, the width of the confidence interval tends to decrease. This is because larger sample sizes lead to more precise estimates of the population proportion, which in turn leads to narrower confidence intervals. Also conversely, as the sample size decreases, the width of the confidence intervals increases. This is because smaller sample sizes tend to produce less precise estimates, which leads to wider intervals.

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?

Increasing the number of bootstrap samples generally does not affect the standard error, but it can improve the precision of the estimate. As the number of bootstrap samples increases, the width of the interval generally becomes more stable and tends to converge to a specific value.
