Foundations for statistical inference - Confidence intervals

Souleymane Doumbia

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)
set.seed(777)
```

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                        27  0.45
## 2 Yes                       33  0.55
```

55 % of adults in sample of 60 think climate change affects their local community.

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 wouldn't anticipate that another student's sample proportion would exactly match mine. This is because each time we use the sample_n() function, it gives us a distinct random sample. Due to this randomness, we might end up with different subsets of the population. However, since both of us are sampling 60 individuals—a rather small number compared to the entire population, it is likely that our proportions might be somewhat similar. This is because such a small sample size doesn't drastically affect the overall proportions of the subgroups within the whole 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
set.seed(777)
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    0.433    0.667
```

- 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 `proportion`.

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" suggests that there's a 95% likelihood that the given range include the actual value. However, there remains a slight 5% possibility that the value might fall outside this range, either due to error or random fluctuations.

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?

Since population proportion of 0.62 falls within my confidence interval [0.45-0.72], my interval covers the true population proportion of US adults who think climate change affects their local community.

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?

Given that each student has constructed a 95% confidence interval for the population proportion, it can be anticipated that, on average, approximately 95% of these individual intervals will encompass the true population proportion. This expectation is rooted in the fact that a confidence level of 95% implies that, if numerous samples were repeatedly drawn from the same population, and confidence intervals were computed using the same method each time, roughly 95% of those intervals would be expected to contain the true population proportion, with the remaining 5% not containing it.

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.

Of my confidence intervals, 94% (or 47 out of 50) capture the true population proportion. This figure doesn't precisely match the expected confidence level of 95%. The reason for this discrepancy is that bootstrapping offers a practical approximation of the sampling distribution, and the inherent randomness in the resampling method can introduce some variability.

More Practice

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.

```r
set.seed(777)
x <- samp %>%
      specify(response = climate_change_affects, success = "Yes") %>%
      generate(reps = 1000, type = "bootstrap") %>%
      calculate(stat = "prop") %>%
      get_ci(level = 0.85)
y <- samp %>%
      specify(response = climate_change_affects, success = "Yes") %>%
      generate(reps = 1000, type = "bootstrap") %>%
      calculate(stat = "prop") %>%
      get_ci(level = 0.95)
dfci <- data.frame(rbind(round(x,2),round(y,2))) %>%
        mutate(CL = c(': ci_85', ': ci_95'))
dfci
```

```
##   lower_ci upper_ci      CL
## 1     0.45     0.63 : ci_85
## 2     0.43     0.67 : ci_95
```

I chose an 85% confidence level. Given the intervals calculated:

CL 85%, CI: 0.45 - 0.63;

CL 95%, CI: 0.43 - 0.67.

I would expect the confidence interval at the 85% confidence level to be narrower than the one at the 95% confidence level. This is consistent with our results, as the 85% CI is indeed narrower than the 95% CI.

The reasoning is that a lower confidence level means there's a greater chance (higher risk) that the true population parameter falls outside the confidence interval. Since we are accepting a higher level of uncertainty with the 85% confidence level, the interval can be narrower. Conversely, at a 95% confidence level, we want to be more certain that our interval captures the true population parameter, so we need a wider interval to account for more variability or uncertainty.

8. 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.

```r
set.seed(777)
samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.98)
```

```
## # A tibble: 1 x 2
```

```
##    lower_ci upper_ci
##       <dbl>    <dbl>
## 1       0.4    0.683
```

With 98% confidence, we estimate that the proportion of US adults who believe climate change is affecting their local community lies between 0.433 and 0.733. This wider interval reflects our increased confidence. Compared to a 95% confidence level, this higher confidence level of 98% means we are even more certain that the true population proportion falls within this interval.

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?

In the previous question, I chose 98% confidence level. Out of 50 confidence intervals, only 2 do not include the true proportion which i expected since it wider interval than 95%. So, 48/50 = 96% of interval that include the true population proportion. This gives a wider interval increasing the chance of covering the true parameter opposing to a 95% level.

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.

```
samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.75)
```

```
## # A tibble: 1 x 2
##    lower_ci upper_ci
##       <dbl>    <dbl>
## 1     0.483    0.633
```

I expect that the proportion of interval to be less than the proportion of my previous question confidence level because 98% confidence will provide a wider interval than 75% confidence level. With 75% confidence level, out of 50 confidence intervals, only 37 intervals include the true proportion. So, 37/50 = 74% of interval that include the true population proportion. This gives a narrow interval reducing the chance of covering the true parameter.

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).

When the sample size decrease, the confidence intervals become wider. A wider interval implies that, to maintain the same level of confidence, we need to consider a broader range of possible values for the true population parameter because there's more uncertainty. In opposite, when the sample size increases, the intervals width become smaller. A narrower interval implies that we can be more confident about where the true population parameter lies without compromising the level of confidence, thus estimates become more precise. Also, The standard

error (SE) of the statistic increases as the sample size decreases. Smaller samples provide less information about the population, leading to less precision in estimates. Conversely, the standard error (SE) of the statistic decreases as the sample size increases; this is because larger samples provide more information about the population.

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 bootstrap samples affect the standard error?

Increasing the number of bootstrap samples helps stabilize the estimate of the standard error, which in turn stabilizes the width of the bootstrap confidence interval. However, after a certain point, further increasing the number of bootstrap samples has a negligible effect on the interval's width. The effect on the standard error diminishes after a sufficiently large number of samples.