CUNY SPS DATA606 Lab5b: Foundations for statistical inference - Confidence intervals

Chinedu Onyeka

September 30th, 2021

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.

```
set.seed(101)
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.

Solution 1:

```
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                        19 0.317
## 2 Yes                       41 0.683
```

68.3% of US adults in the sample think that climate change affects 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?

Solution 2:

No I would not expect another student's sample proportion to be identical to mine. Our samples will be different since the sampling is random.

Confidence intervals

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

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

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    0.567      0.8
```

Confidence levels

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

Solution 3:

The confidence interval represents the range of plausible values where we are likely to find the population values. A 95% confidence level mean that we are 95% confident that the population parameter will lie within the given interval. Also, a 95% confidence interval corresponds to an interval that contains all values within 2 standard deviations from the mean. Hence, it means that we are 95% confident that the population parameter will fall within 2 standard deviations from the mean.

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?

Solution 4:

Yes the confidence interval captures the true population proportion. Since confidence interval is (0.567, 0.8) and the population proportion is 0.62, the interval captures the population proportion.

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?

Solution 5:

Yes each confidence interval would have been slightly different. However, I would expect atleast 95% of confidence intervals to capture the true population value because each interval was constructed with a 95% level of confidence to ensure that the interval captures the true population value.

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.

Solution 6:

About 98% of confidence intervals include the true population proportion. This is not exactly equal to the confidence interval, but it is within the interpretation of the confidence interval.

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.

Solution 7:

I choose 99% confidence level. This level of confidence is wider to me because it is higher than 95%. Increasing the confidence level basically means widening the interval to be sure that it contains the population parameter. It is more like spreading your net or using a wider net to ensure you catch a fish.

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.

Solution 8:

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

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.517    0.833
```

The confidence interval at 99% level of confidence for the `samp` data set is (0.533, 0.817). This means that we are 99% confident that the population proportion will fall within that interval.

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?

Solution 9:

When I use the app at 99% confidence level, 50 confidence intervals, and 100 bootstraps, the proportion of intervals that include the true population proportion is 98%. It is slightly lower than the confidence level.

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.

Solution 10:

```
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    0.583    0.783
```

Trying another confidence level of 90%, the confidence interval is (0.583, 0.783) which means that we are 90% confident that the population mean will lie within the interval. Using the app for 90% confidence level, the proportion of intervals that capture the true population proportion is 92%.

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).

Solution 11:

Using the app, we can infer that as the sample size increases, the width of the interval decreases and vice versa. Basically, the more we increase the sample size, the less the spread in the data.

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 bootstrap samples affect the standard error?

Solution 12:

It will make the width of the interval narrower(i.e decrease the width). Also, Increasing the bootstrap will affect the sampling distribution and make the standard error to decrease, and the distribution will tend more towards a unimodal symmetric normal distribution.
