# Foundations for statistical inference - Confidence intervals

```
set.seed(1994)
```

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.

```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.

```r
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.

```r
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.

```r
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.

**Insert your answer here**

```r
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
```

From my sample, we can see that **48.3%** of adults think climate change affects their local community.

**End of your answer**

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?

**Insert your answer here**

I expect another student's sample proportion to be a bit higher than what mine was specifically although I expect most samples to be relatively similar. I expect so because these random samples are estimates of the true population proportion and will converge to the 62% true proportion. Because mine is much lower (48.3%), I expect that most other students will have a proportion greater than mine, which is also closer to the true proportion.

**End of your answer**

## 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    0.617
```

- 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?

**Insert your answer here**

The 95% confidence means that if we were to resample the dataset, then the true proportion would have a 95% chance of being within the lower_ci and upper_ci calculated from before.

**End of your answer**

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?

**Insert your answer here**

My confidence interval doesn't contain the true population proportion. I expect my neighbors to capture this value as the chances that both of us do not is 0.25%.

**End of your answer**

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?

**Insert your answer here**

I would expect 95% of these intervals to capture the true population mean. This is because the method of creating the confidence intervals is to create a bound where random samples are 95% likely to contain the population mean.

It's a bit like a self-fulfilling prophecy.

**End of your answer**

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.

**Insert your answer here**

From the chart above, we can see that 2 of the 50 confidence intervals do not contain the population mean. This is equal to 96% which is not exactly equal to the 95% confidence interval which is expected. As the number of confidence intervals created increase, we can expect it to converge closer and closer to the 95%.

**End of your answer**

---

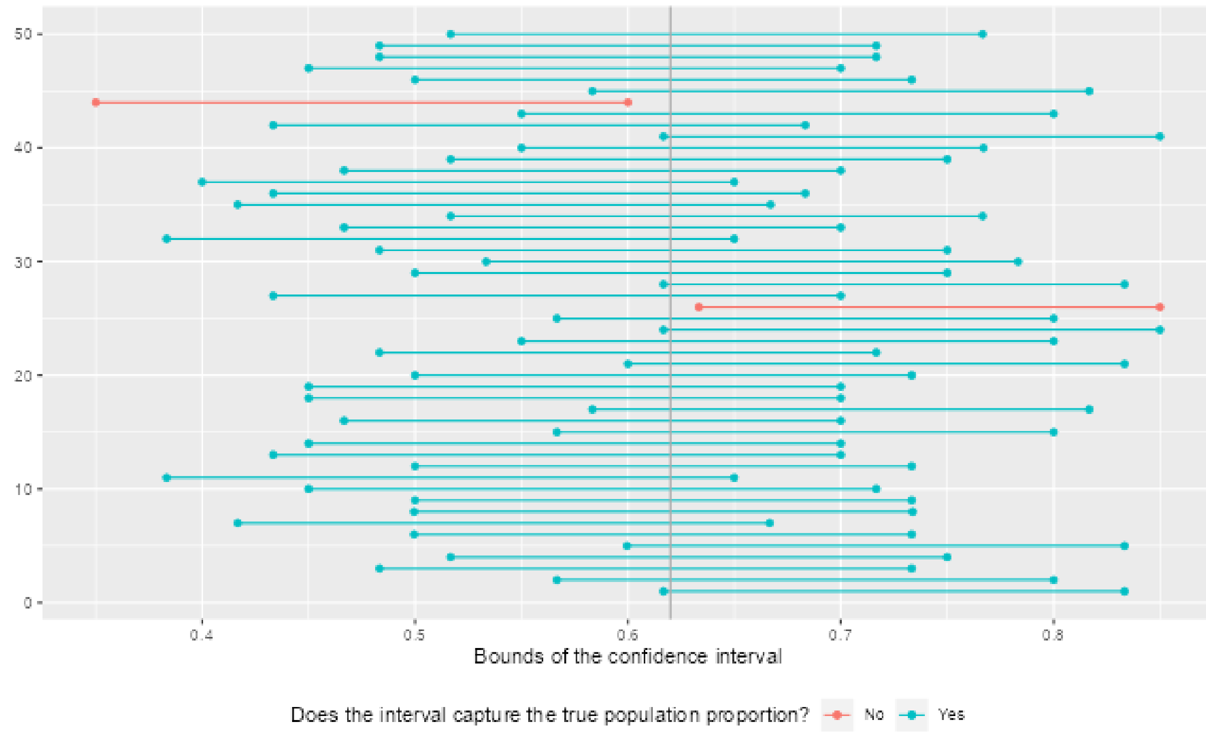


Figure 1: sample size of 60, 1000 bootstrap samples for each interval, and 50 confidence intervals

## 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.

**Insert your answer here**

As the confidence interval decreases, we will see that the range (range = upper_ci - lower_ci) decreases. If the confidence interval were to increase, then the range will also increase. This is because as our confidence changes, so does our expectation that new samples will contain the population mean. Thus, to be more confident we must include more values, widening our range and to be less confident we will exclude more values which will narrow our range.

**End of your answer**

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.

**Insert your answer here**

```
samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.80)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1      0.4    0.567
```

By decreasing the confidence interval to 80%, we can see that our range decreased as our lower_ci = 40% and our upper_ci = 56.7%.

Although we know that this confidence interval doesn't include the population mean, it does demonstrate how a lower confidence interval will increase the lower_ci and decrease the upper_ci.

**End of your answer**

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?

**Insert your answer here**

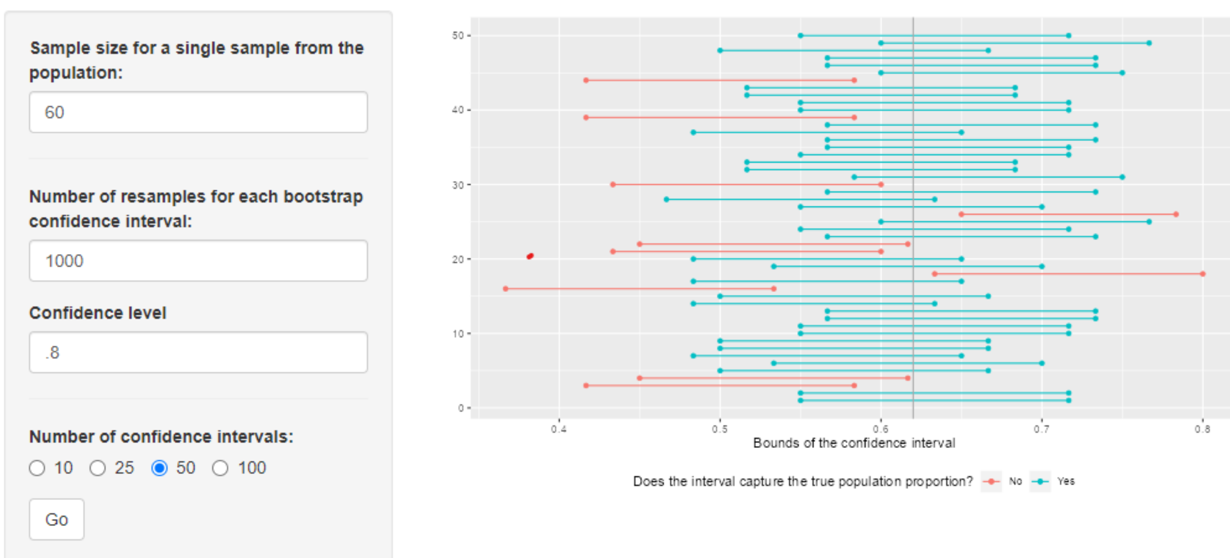


Figure 2: 80% confidence interval

In this example, 10 of the 50 confidence intervals do not include the true population proportion. This is lower than the last one and coincidentally is exactly equal to the confidence interval.

**End of your answer**

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.

**Insert your answer here**

I'm going to use a confidence interval of 10% and here I can expect a much tighter range and that most confidence intervals generated will not contain the population mean.

```r
samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.10)
```

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

Here we can see some proof that the range became even narrower, reducing to 48.3% and 50% for the lower_ci and upper_ci respectively.

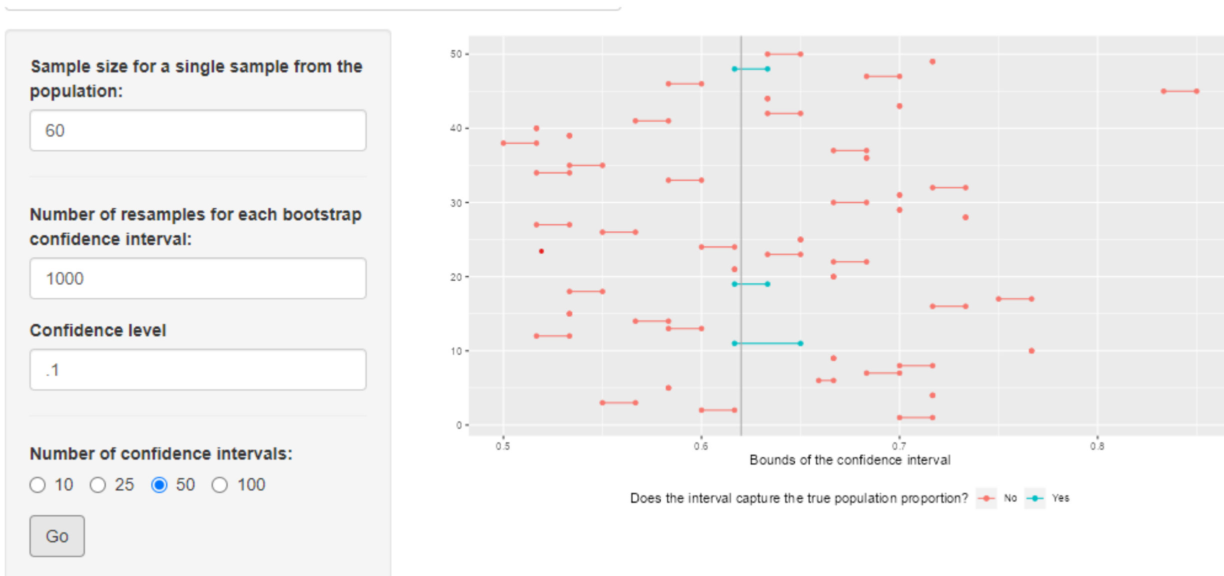


Figure 3: 10% confidence interval

And in this photo, we can see the same. Only 3 of the 50 confidence intervals contain the popluation mean and we can also notice that these ranges are much narrower than the other two calculated above.

**End of your answer**

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).

**Insert your answer here**

As a function of sample size, I've tested 10, 100, and 1000 and found that the confidence intervals do not vary greatly with changes in smaple size.

**End of your answer**

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?

**Insert your answer here**

As we increase the number of bootstrap samples, the range narrows as the standard error decreases when the number of bootstrap samples increases.

**End of your answer**