Lab_5_Data_606

Mathew Katz

2022-10-08

```
knitr::opts_chunk$set(eval = TRUE, message = FALSE, warning = FALSE)
set.seed(1234)
```

In this lab, you will investigate the ways in which the statistics from a random sample of data can serve as point estimates for population parameters. We're interested in formulating a *sampling distribution* of our estimate in order to learn about the properties of the estimate, such as its distribution.

Setting a seed: We will take some random samples and build sampling distributions in this lab, which means you should set a seed at the start 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. We will also use the **infer** package for resampling.

Let's load the packages.

```
library(tidyverse)
library(openintro)
library(infer)
library(shiny)
```

The data

A 2019 Gallup report states the following:

> The premise that scientific progress benefits people has been embodied in discoveries throughout the ages – from the development of vaccinations to the explosion of technology in the past few decades, resulting in billions of supercomputers now resting in the hands and pockets of people worldwide. Still, not everyone around the world feels science benefits them personally.
>
> **Source:** World Science Day: Is Knowledge Power?

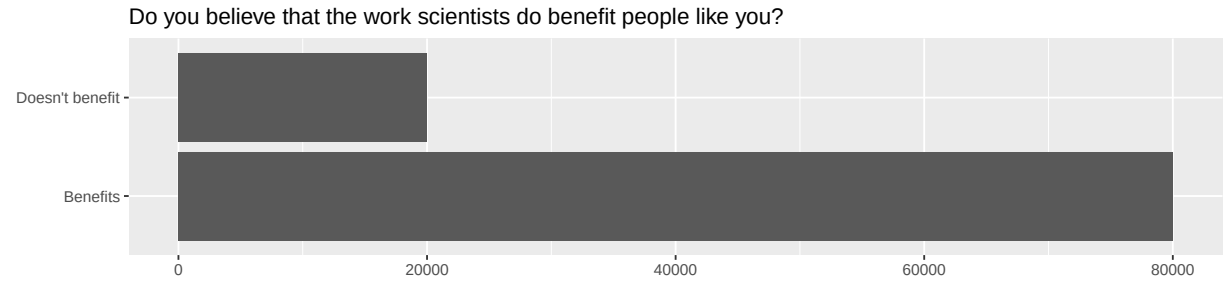
The Wellcome Global Monitor finds that 20% of people globally do not believe that the work scientists do benefits people like them. In this lab, you will assume this 20% is a true population proportion and learn about how sample proportions can vary from sample to sample by taking smaller samples from the population. We will first create our population assuming a population size of 100,000. This means 20,000 (20%) of the population think the work scientists do does not benefit them personally and the remaining 80,000 think it does.

```
global_monitor <- tibble(
  scientist_work = c(rep("Benefits", 80000), rep("Doesn't benefit", 20000))
)
```

The name of the data frame is `global_monitor` and the name of the variable that contains responses to the question *"Do you believe that the work scientists do benefit people like you?"* is `scientist_work`.

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

```
ggplot(global_monitor, aes(x = scientist_work)) +
  geom_bar() +
  labs(
    x = "", y = "",
    title = "Do you believe that the work scientists do benefit people like you?"
  ) +
  coord_flip()
```



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

```
global_monitor %>%
  count(scientist_work) %>%
  mutate(p = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n     p
##   <chr>           <int> <dbl>
## 1 Benefits        80000   0.8
## 2 Doesn't benefit 20000   0.2
```

The unknown sampling distribution

In this lab, you have access to the entire population, but this is rarely the case in real life. Gathering information on an entire population is often extremely costly or impossible. Because of this, we often take a sample of the population and use that to understand the properties of the population.

If you are interested in estimating the proportion of people who don't think the work scientists do benefits them, you can use the `sample_n` command to survey the population.

```
samp1 <- global_monitor %>%
  sample_n(50)
```

This command collects a simple random sample of size 50 from the `global_monitor` dataset, and assigns the result to `samp1`. This is similar to randomly drawing names from a hat that contains the names of all in the population. Working with these 50 names is considerably simpler than working with all 100,000 people in the population.

1. Describe the distribution of responses in this sample. How does it compare to the distribution of responses in the population. **Hint:** Although the `sample_n` function takes a random sample of observations (i.e. rows) from the dataset, you can still refer to the variables in the dataset with the same names. Code you presented earlier for visualizing and summarizing the population data will still be useful for the sample, however be careful to not label your proportion `p` since you're now calculating a sample statistic, not a population parameters. You can customize the label of the statistics to indicate that it comes from the sample.

The distribution of responses in this sample are very similar to the original 100,000 people's responses. Both responses are about 80% 'Benefits' them and 20% 'Doesn't benefit' them.

If you're interested in estimating the proportion of all people who do not believe that the work scientists do benefits them, but you do not have access to the population data, your best single guess is the sample mean.

```
samp1 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits           37  0.74
## 2 Doesn't benefit    13  0.26
```

Depending on which 50 people you selected, your estimate could be a bit above or a bit below the true population proportion of 0.26. In general, though, the sample proportion turns out to be a pretty good estimate of the true population proportion, and you were able to get it by sampling less than 1% of the population.

2. Would you expect the sample proportion to match the sample proportion of another student's sample? Why, or why not? If the answer is no, would you expect the proportions to be somewhat different or very different? Ask a student team to confirm your answer.

I would not expect the sample proportion to match the sample proportion of another student's sample but I wouldn't expect them to be drastically different. Both samples should be relatively close and should be around 60/40.

3. Take a second sample, also of size 50, and call it `samp2`. How does the sample proportion of `samp2` compare with that of `samp1`? Suppose we took two more samples, one of size 100 and one of size 1000. Which would you think would provide a more accurate estimate of the population proportion?

```
samp2 <- global_monitor %>%
  sample_n(50)
```

```
samp1 %>%
  count(scientist_work) %>%
  mutate(p = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n     p
##   <chr>           <int> <dbl>
## 1 Benefits           37  0.74
## 2 Doesn't benefit    13  0.26
```

Samp1 has a 74/26 ratio and Samp2 has an 84/16 ratio. That makes perfect sense as I would expect both to be close to 80/20 but not exactly 80/20. The closer you get to the 100,000 population, the closer you should get to the most accurate estimate of the population proportion.

Not surprisingly, every time you take another random sample, you might get a different sample proportion. It's useful to get a sense of just how much variability you should expect when estimating the population mean this way. The distribution of sample proportions, called the *sampling distribution (of the proportion)*, can help you understand this variability. In this lab, because you have access to the population, you can build up the sampling distribution for the sample proportion by repeating the above steps many times. Here, we use R to take 15,000 different samples of size 50 from the population, calculate the proportion of responses in each sample, filter for only the *Doesn't benefit* responses, and store each result in a vector called `sample_props50`. Note that we specify that `replace = TRUE` since sampling distributions are constructed by sampling with replacement.

```
sample_props50 <- global_monitor %>%
                    rep_sample_n(size = 50, reps = 15000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")
```

And we can visualize the distribution of these proportions with a histogram.

```
ggplot(data = sample_props50, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02) +
  labs(
    x = "p_hat (Doesn't benefit)",
    title = "Sampling distribution of p_hat",
    subtitle = "Sample size = 50, Number of samples = 15000"
  )
```

Next, you will review how this set of code works.

4. How many elements are there in `sample_props50`? Describe the sampling distribution, and be sure to specifically note its center. Make sure to include a plot of the distribution in your answer.

```
sample_props50
```

```
## # A tibble: 15,000 x 4
## # Groups:   replicate [15,000]
##    replicate scientist_work      n p_hat
##        <int> <chr>           <int> <dbl>
##  1         1 Doesn't benefit    11  0.22
##  2         2 Doesn't benefit     8  0.16
##  3         3 Doesn't benefit     8  0.16
##  4         4 Doesn't benefit     6  0.12
```

```
##  5         5 Doesn't benefit    10  0.2
##  6         6 Doesn't benefit    11  0.22
##  7         7 Doesn't benefit    10  0.2
##  8         8 Doesn't benefit    14  0.28
##  9         9 Doesn't benefit    11  0.22
## 10        10 Doesn't benefit    11  0.22
## # ... with 14,990 more rows
```

There are 15,000 elements in 'sample_props50.' The sampling distribution is approximately normal and the spread is centered toward the mean.

Interlude: Sampling distributions

The idea behind the `rep_sample_n` function is *repetition*. Earlier, you took a single sample of size `n` (50) from the population of all people in the population. With this new function, you can repeat this sampling procedure `rep` times in order to build a distribution of a series of sample statistics, which is called the **sampling distribution**.

Note that in practice one rarely gets to build true sampling distributions, because one rarely has access to data from the entire population.

Without the `rep_sample_n` function, this would be painful. We would have to manually run the following code 15,000 times

```
global_monitor %>%
  sample_n(size = 50, replace = TRUE) %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n)) %>%
  filter(scientist_work == "Doesn't benefit")
```

```
## # A tibble: 1 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Doesn't benefit     6  0.12
```

as well as store the resulting sample proportions each time in a separate vector.

Note that for each of the 15,000 times we computed a proportion, we did so from a **different** sample!

5. To make sure you understand how sampling distributions are built, and exactly what the `rep_sample_n` function does, try modifying the code to create a sampling distribution of **25 sample proportions** from **samples of size 10**, and put them in a data frame named `sample_props_small`. Print the output. How many observations are there in this object called `sample_props_small`? What does each observation represent?

```
sample_props_small <- global_monitor %>%
                      rep_sample_n(size = 10, reps = 25, replace = TRUE) %>%
                      count(scientist_work) %>%
                      mutate(p_hat_s = n /sum(n)) %>%
                      filter(scientist_work == "Doesn't benefit")

sample_props_small
```

```
## # A tibble: 23 x 4
## # Groups:   replicate [23]
##    replicate scientist_work      n p_hat_s
##        <int> <chr>           <int>   <dbl>
##  1         1 Doesn't benefit     2     0.2
##  2         2 Doesn't benefit     2     0.2
##  3         3 Doesn't benefit     1     0.1
##  4         4 Doesn't benefit     3     0.3
##  5         5 Doesn't benefit     1     0.1
##  6         6 Doesn't benefit     1     0.1
##  7         8 Doesn't benefit     2     0.2
##  8         9 Doesn't benefit     2     0.2
##  9        10 Doesn't benefit     1     0.1
## 10        11 Doesn't benefit     1     0.1
## # ... with 13 more rows
```

There are 25 observations and each observation represents a person who doesn't believe that the work scientists do benefits people like them.

Sample size and the sampling distribution

Mechanics aside, let's return to the reason we used the `rep_sample_n` function: to compute a sampling distribution, specifically, the sampling distribution of the proportions from samples of 50 people.

```
ggplot(data = sample_props50, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02)
```

The sampling distribution that you computed tells you much about estimating the true proportion of people who think that the work scientists do doesn't benefit them. Because the sample proportion is an unbiased estimator, the sampling distribution is centered at the true population proportion, and the spread of the distribution indicates how much variability is incurred by sampling only 50 people at a time from the population.

In the remainder of this section, you will work on getting a sense of the effect that sample size has on your sampling distribution.

6. Use the app below to create sampling distributions of proportions of *Doesn't benefit* from samples of size 10, 50, and 100. Use 5,000 simulations. What does each observation in the sampling distribution represent? How does the mean, standard error, and shape of the sampling distribution change as the sample size increases? How (if at all) do these values change if you increase the number of simulations? (You do not need to include plots in your answer.)

7. Use the app below to create sampling distributions of proportions of *Doesn't benefit* from samples of size 10, 50, and 100. Use 5,000 simulations. What does each observation in the sampling distribution represent? How does the mean, standard error, and shape of the sampling distribution change as the sample size increases? How (if at all) do these values change if you increase the number of simulations? (You do not need to include plots in your answer.)

Each observation in the sampling distribution represents a person who doesn't believe that the work scientists do benefits people like them. Sample size 10 has a mean of sampling distribution of 0.22 and an SE of sampling distribution of 0.11. Sample size 50 has a mean of sampling distribution of 0.2 and an SE of sampling distribution of 0.06. Sample size 100 has a mean of sampling distribution of 0.2 and an SE of sampling distribution of 0.04. As the

sample size increases, the mean stays constant (0.2) and the SE starts decreasing since there is less spread, and the shape becomes more and more normal with the spread centered toward the mean.

More Practice

So far, you have only focused on estimating the proportion of those you think the work scientists doesn't benefit them. Now, you'll try to estimate the proportion of those who think it does.

Note that while you might be able to answer some of these questions using the app, you are expected to write the required code and produce the necessary plots and summary statistics. You are welcome to use the app for exploration.

7. Take a sample of size 15 from the population and calculate the proportion of people in this sample who think the work scientists do enhances their lives. Using this sample, what is your best point estimate of the population proportion of people who think the work scientists do enchances their lives?

```
set.seed(1234)
samp15 <- global_monitor %>%
  sample_n(15)

samp15 %>%
  count(scientist_work) %>%
  mutate(p_hat5 = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat5
##   <chr>           <int>  <dbl>
## 1 Benefits           11  0.733
## 2 Doesn't benefit     4  0.267
```

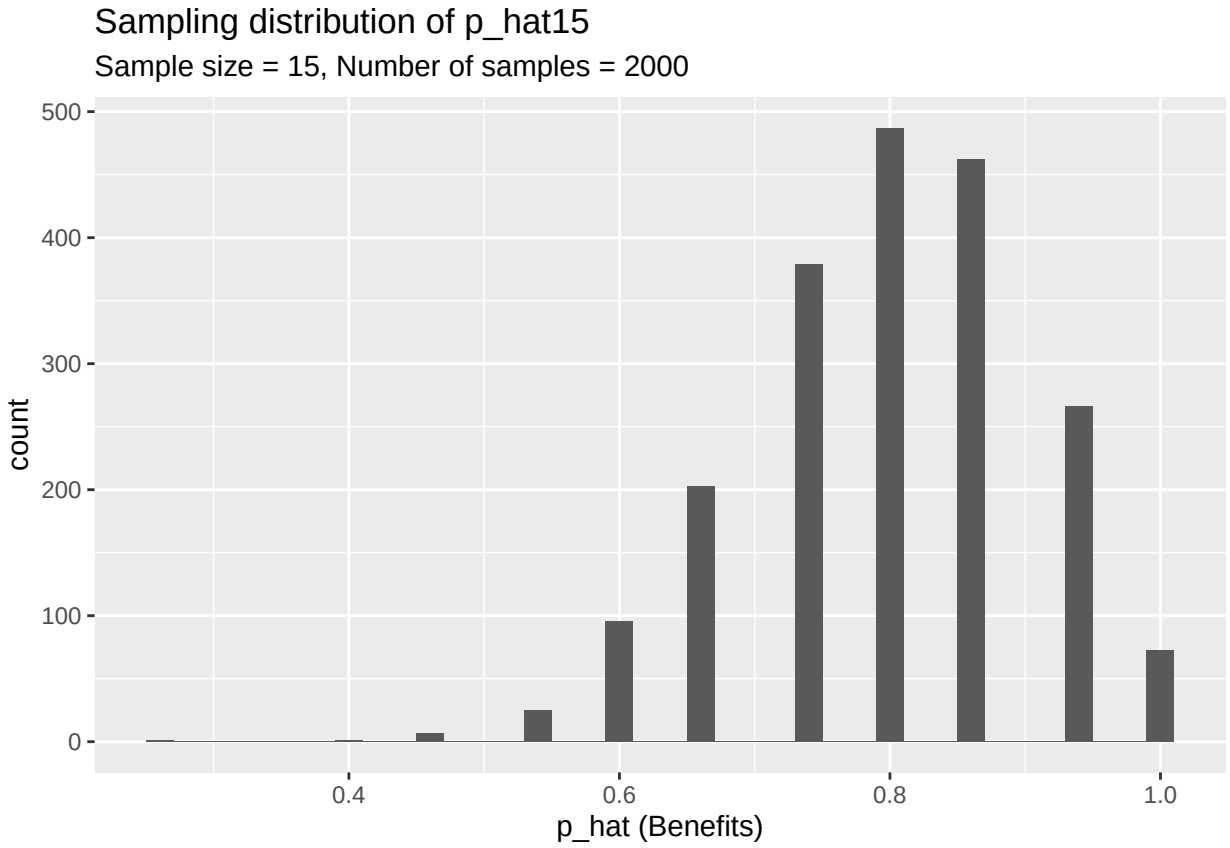
About 73% of the people who think that scientist enhances their lives is most likely the best point estimate of the population.

8. Since you have access to the population, simulate the sampling distribution of proportion of those who think the work scientists do enchances their lives for samples of size 15 by taking 2000 samples from the population of size 15 and computing 2000 sample proportions. Store these proportions in as `sample_props15`. Plot the data, then describe the shape of this sampling distribution. Based on this sampling distribution, what would you guess the true proportion of those who think the work scientists do enchances their lives to be? Finally, calculate and report the population proportion.

```
set.seed(1234)
sample_props15 <- global_monitor %>%
  rep_sample_n(size = 15, reps = 2000, replace = TRUE) %>%
  count(scientist_work) %>%
  mutate(p_hat15 = n /sum(n)) %>%
  filter(scientist_work == "Benefits")

ggplot(data = sample_props15, aes(x = p_hat15)) +
```

```
  geom_histogram(binwidth = 0.02) +
  labs(
    x = "p_hat (Benefits)",
    title = "Sampling distribution of p_hat15",
    subtitle = "Sample size = 15, Number of samples = 2000"
  )
```



The graph looks relatively normal distributed but skewed a bit to the left. Based on this sampling distribution, I would guess the true proportion of those who think the work scientists do enhances their lives to be closer to 80 %

```
summary(sample_props15$p_hat15)
```

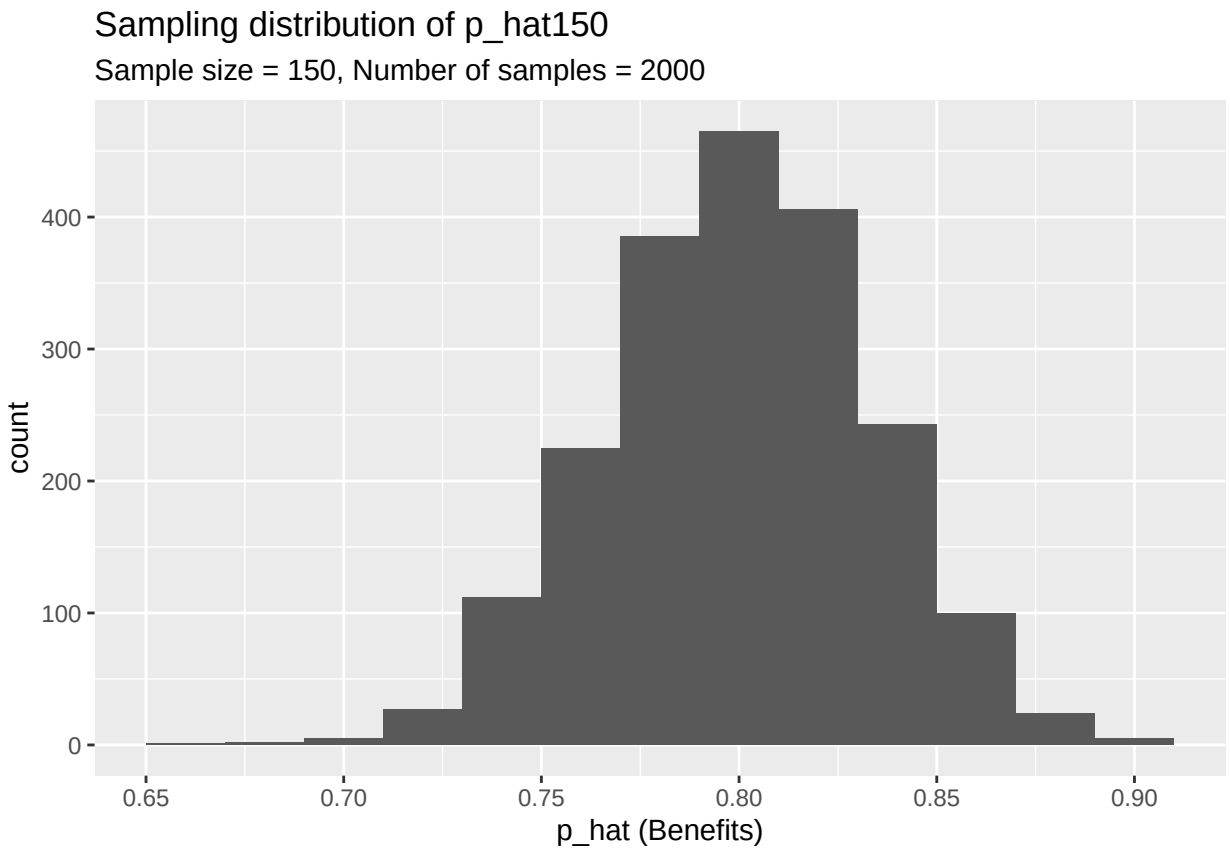
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2667  0.7333  0.8000  0.7997  0.8667  1.0000
```

Calculation also says 80%.

9. Change your sample size from 15 to 150, then compute the sampling distribution using the same method as above, and store these proportions in a new object called `sample_props150`. Describe the shape of this sampling distribution and compare it to the sampling distribution for a sample size of 15. Based on this sampling distribution, what would you guess to be the true proportion of those who think the work scientists do enchances their lives?

```r
set.seed(1234)
sample_props150 <- global_monitor %>%
  rep_sample_n(size = 150, reps = 2000, replace = TRUE) %>%
  count(scientist_work) %>%
  mutate(p_hat150 = n /sum(n)) %>%
  filter(scientist_work == "Benefits")

ggplot(data = sample_props150, aes(x = p_hat150)) +
  geom_histogram(binwidth = 0.02) +
  labs(
    x = "p_hat (Benefits)",
    title = "Sampling distribution of p_hat150",
    subtitle = "Sample size = 150, Number of samples = 2000"
  )
```



The shape of this sampling distribution is normal. Compared to the sampling distribution of size = 15, this sampling distribution is more normal and less spread. I still would guess that the true proportion of those who think the work scientists do, does enhance their lives is closer to 80 %

10. Of the sampling distributions from 2 and 3, which has a smaller spread? If you're concerned with making estimates that are more often close to the true value, would you prefer a sampling distribution with a large or small spread? **3 has a smaller spread. I would prefer the sampling with small spread.**

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**.

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.

```
set.seed(1234)
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                        21  0.35
## 2 Yes                       39  0.65
```

60% of adults in my sample think 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? **I would not expect the sample proportion to match the sample proportion of another student's sample but I wouldn't expect them to be drastically different. Both samples should be relatively close and should be around 80/20.**

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(10)
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    0.533    0.767
```

- 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?

'95% confident' means that we are 95% certain. 95% confidence interval is a range of values that we can be 95% certain contains the true proportion of the population.

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? **Yes; my confidence interval, ci = (0.467, 0.717) captures the true population proportion of US adults who believe climate change affects their local community (0.62)**

2. 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?

Each student should have gotten a slightly different confidence interval due to different samples of US adults that each one will select, but one would expect at least 95% of students to capture the true population mean. This is because there is just a slight difference in confidence interval each one will get, and as we are all working in 95% level, we are all 95% confident that the true population proportion is contained in our confidence 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.

98% of my confidence intervals include the true population. The proportion is not exactly equal to the confidence level but at least 95% of my confidence intervals would include the true population so that could be more than 95%. * * *

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 chose a 90% confidence level and I would expect a confidence interval at this level to be narrower than the confidence interval I calculated at the 95% confidence level. With a 95% confidence level, you have a 5% of being wrong. With a 90% confidence level, you have 10% chance of being wrong. As the precision of the confidence interval increases, the reliability of an interval containing the true population proportion decreases.

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.

```
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
```

I am 90% confident that the true proportion of US adults who think climate change affects their local community is contained in this interval (50%, 70%)

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?

Running the app with 90% confidence level, the percentage of intervals that include the true population proportion is lower (92%) compared to the confidence intervals with 95% confidence level (98%).

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.

```
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    0.583    0.733
```

I expect a confidence interval (80%) at this level to be even narrower. With 80% certainty, the true proportion of US adults who think climate change affects their local community is contained in this interval ci(51.7%, 68.3%) Running the app with 80% confidence level, the percentage of intervals that include the true population proportion is 84%

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 increases, the width of confidence intervals decreases, and when the sample size decreases, the width of confidence intervals increases.

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 will decrease the standard error and the sampling distributions will narrow. The larger the number of bootstrap samples will lead to more precise estimates around the true population.
