Getting Started

The data

Confidence intervals

Confidence levels

More Practice

Foundations for statistical inference - Confidence intervals

Your reproducible lab report: Before you get started, download the R Markdown template for this lab. Remember all of your code and answers go in this document:

```
download.file("https://dyurovsky.github.io/85309/post/rmd/lab6.Rmd",
              destfile = "lab6.Rmd")
```

If you have access to data on an entire population, say the size of every house in Ames, Iowa, it's straight forward to answer questions like, "How big is the typical house in Ames?" and "How much variation is there in sizes of houses?". 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 the typical size if you only know the sizes of several dozen houses? This sort of situation requires that you use your sample to make inference on what your population looks like.

> **Setting a seed:** We will take some random samples and build sampling distributions in this lab, which means you should set a seed on top of your lab. For a reminder of what this does, look back to Lab 4 (https://dyurovsky.github.io/85309/post/labs/sampling_distributions.html).

Getting Started

Load packages and data

As usual, we're going to load the `tidyverse` package for data manipulation. We'll also be reading in the same Ames, Iowa dataset we used in Lab 4.

Let's load the packages.

```
library(tidyverse)
ames <- read_csv("https://dyurovsky.github.io/85309/data/lab6/ames.csv")
```

The data

We consider real estate data from the city of Ames, Iowa. This is the same dataset used in the previous lab. The details of every real estate transaction in Ames is recorded by the City Assessor's office. Our particular focus for this lab will be all residential home sales in Ames between 2006 and 2010. This collection represents our population of interest. In this lab we would like to learn about these home sales by taking smaller samples from the full population. Let's load the data.

In this lab we'll start with a simple random sample of size 60 from the population.

```
n <- 60

samp <- ames %>%
  sample_n(n)
```

Note that the data set has information on many housing variables, but for the first portion of the lab we'll focus on the size of the house, represented by the variable `area` .

Exercise 1

Describe the distribution of homes in your sample. What would you say is the "typical" size within your sample? Also state precisely what you interpreted "typical" to mean.

Exercise 2

Would you expect another class member's distribution to be identical to yours? Would you expect it to be similar? Why or why not?

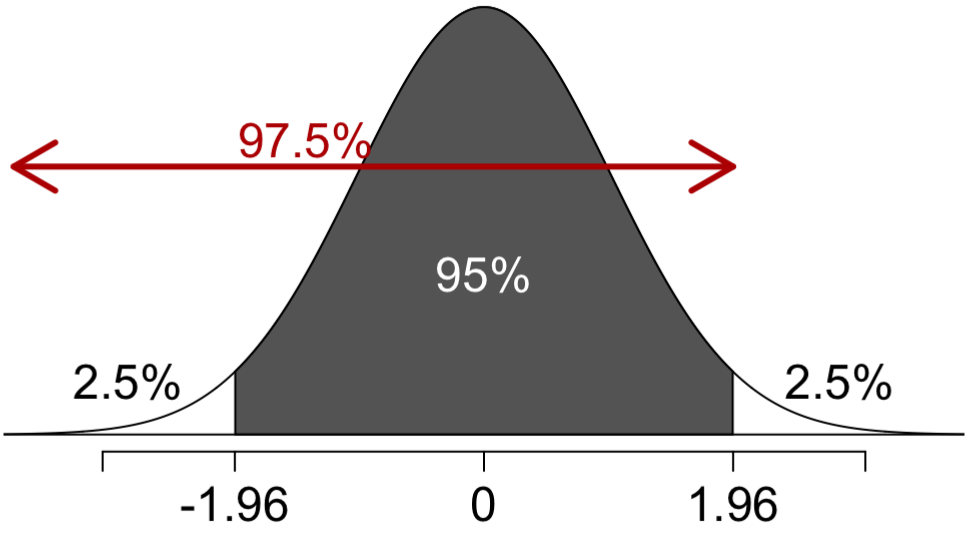
Confidence intervals

Return for a moment to the question that first motivated this lab: based on this sample, what can we infer about the population? Based only on this single sample, the best estimate of the average living area of houses sold in Ames would be the sample mean, usually denoted as $\bar{x}$ (here we're calling it `x_bar`). That serves as a good **point estimate** but it would be useful to also communicate how uncertain we are of that estimate. This uncertainty can be quantified using a **confidence interval**.

A confidence interval for a population mean is of the following form

$$\bar{x} + z^{\star} \frac{s}{\sqrt{n}}$$

You should by now be comfortable with calculating the mean and standard deviation of a sample in R. And we know that the sample size is 60. So the only remaining building block is finding the appropriate critical value for a given confidence level. We can use the `qnorm` function for this task, which will give the critical value associated with a given percentile under the normal distribution. Remember that confidence levels and percentiles are not equivalent. For example, a 95% confidence level refers to the middle 95% of the distribution, and the critical value associated with this area will correspond to the 97.5th percentile.



We can find the critical value for a 95% confidence interal using

```
z_star_95 <- qnorm(0.975)
z_star_95
```

which is roughly equal to the value critical value 1.96 that you're likely familiar with by now.

Let's finally calculate the confidence interval:

```
samp %>%
  summarise(lower = mean(area) - z_star_95 * (sd(area) / sqrt(n)),
            upper = mean(area) + z_star_95 * (sd(area) / sqrt(n)))
```

To recap: even though we don't know what the full population looks like, we're 95% confident that the true average size of houses in Ames lies between the values `lower` and `upper` .

Exercise 3 For the confidence interval to be valid, the sample mean must be normally distributed and have standard error $\frac{s}{\sqrt{n}}$. What conditions must be met for this to be true?

Confidence levels

Exercise 4 What does "95% confidence" mean?

In this case we have the rare luxury of knowing the true population mean since we have data on the entire population. Let's calculate this value so thatwe can determine if our confidence intervals actually capture it. We'll store it in a variable called `mu`.

```
mu <- ames %>%
  summarise(mu = mean(area)) %>%
  pull()
```

Exercise 5 Does your confidence interval capture the true average size of houses in Ames? Does your neighbor's interval capture this value?

Exercise 6 Everyone should have gotten a slightly different confidence interval. What proportion of those intervals would you expect to capture the true population mean? Why?

Using R, we're going to collect many samples to learn more about how sample means and confidence intervals vary from one sample to another.

Here is the rough outline:

- Obtain a random sample.
- Calculate the sample's mean and standard deviation, and use these to calculate and store the lower and upper bounds of the confidence intervals.
- Repeat these steps 50 times.

We can accomplish this using the `replicate` function. By default, `replicate` returns a kind of data structure called an `array` that is a headache to work with in a tidy way. If we specify that instead `simplify = FALSE`, it will return a list of lists which we can `bind_rows` together.

The following lines of code takes 50 random samples of size `n` from population (and remember we defined $n = 60$ earlier), and computes the upper and lower bounds of the confidence intervals based on these samples.

```
one_sample <- function() {
  ames %>%
    sample_n(n) %>%
    summarise(x_bar = mean(area),
              se = sd(area) / sqrt(n),
              me = z_star_95 * se,
              lower = x_bar - me,
              upper = x_bar + me)
}

ci <- replicate(50, one_sample(), simplify = FALSE) %>%
  bind_rows() %>%
  mutate(replicate = 1:n()) # number each replication
```

Let’s view the first five intervals:

```
ci %>%
  slice(1:5)
```

Next we’ll create a plot similar to the you saw in my lecture slides from Lesson 2.5. The First step will be to create a new variable in the `ci` tibble that indicates whether the interval does or does not capture the true population mean. Note that capturing this value would mean the lower bound of the confidence interval is below the value and upper bound of the confidence interval is above the value. Remember that we create new variables using the `mutate` function.

```
ci_captured <- ci %>%
  mutate(capture_mu = if_else(lower < mu & upper > mu, "yes", "no"))
```

The `if_else` function is new. It takes three arguments: first is a logical statement, second is the value we want if the logical statement yields a true result, and the third is the value we want if the logical statement yields a false result.

We now have all the information we need to create the plot. Note that the `geom_errorbar()` function only understands y values, and thus we have used the `coord_flip()` function to flip the coordinates of the entire plot back to the more familiar vertical orientation.

We can accomplish this using the following:

```
ggplot(ci_captured, aes(x = replicate, y = x_bar, color = capture_mu)) +
  geom_errorbar(aes(ymin = lower, ymax = upper)) +
  geom_hline(aes(yintercept = mu), color = "darkgray") + # draw vertical line
  coord_flip()
```

Exercise 7 What proportion of your confidence intervals include the true population mean? Is this proportion exactly equal to the confidence level? If not, explain why. Make sure to include your plot in your answer.

More Practice

Exercise 8 Pick a confidence level of your choosing, provided it is not 95%. What is the appropriate critical value?

Exercise 9 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 mean. How does this percentage compare to the confidence level selected for the intervals? Make sure to include your plot in your answer.

This lab is created and released under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (https://creativecommons.org/licenses/by-nc-sa/4.0/). This lab is adapted from a lab created for OpenIntro by Andrew Bray and Mine Çetinkaya-Rundel.