Lab 4B: Foundations for statistical inference - Confidence levels

Sampling from Ames, Iowa

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.

The data

In the previous lab we looked at the population data of houses from Ames, Iowa. Let's start with loading that data set.

```
download.file("http://www.openintro.org/stat/data/ames.RData", destfile = "ames.RData")

load("ames.RData")
```

In this lab we'll start with just a sample from the population, which is a more realistic situation. Specifically, this is a simple random sample of size 60. Note that the data set has information on many variables on these houses, but for the first portion of the lab we'll focus on the size of the house, represented by the variable `Gr.Liv.Area`.

```
population <- ames$Gr.Liv.Area

sample <- sample(population, 60)
```

> **Exercise 1** Describe the distribution of your sample. What would you say is the "typical" size within your sample? Also state precisely what you interpreted "typical" to mean.

> **Exercise 2** Now compare your distribution to your neighbor's. Do they look similar? Are they identical? Why, or why not?

Confidence intervals

One of the most common ways to describe the typical or central value of a distribution is to use the mean. In this case we can calculate the mean age size our sample:

```
sample_mean <- mean(sample)
```

Return for a moment to the question that first motivated this lab: based on this sample, what can we infer about the population? More specifically, what is our best estimate of the mean size of the houses in

This lab was written for OpenIntro by Andrew Bray and Mine Çetinkaya-Rundel.

Ames? Based only on this single sample, our best estimate would be the sample mean, what we usually denote as $\bar{x}$ (here we're calling it `sample_mean`). That serves as a good *point estimate* but it would be useful to also communicate how uncertain we are in that estimate. This can be captured by using a *confidence interval*.

We can calculate a 95% confidence interval for a sample mean by adding and subtracting 1.96 standard errors to our point estimate.[†]

```
se <- sd(sample)/sqrt(60)

lower <- sample_mean - 1.96 * se

upper <- sample_mean + 1.96 * se

c(lower, upper)
```

This is a big inference that we've just made: 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`. This implies that we have a pretty good idea of what the distribution of $\bar{x}$ looks like. This requires some condition.

Exercise 3 What conditions have to be met for the sample mean to be nearly normally distributed with standard error $s/\sqrt{n}$, and for the above interval to be valid.

Confidence levels

Before we start investigating what confidence levels really mean, let's first revisit its definition.

Exercise 4 What does "95% confidence" mean? If you're not sure, see Section 4.2.2.

In this case we have the luxury of knowing the true population mean since we have data on the entire population. This value can be calculated using the following command:

```
mean(population)
```

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 Each student in your class should have gotten a slightly different confidence interval. What proportion of those intervals would you expect to capture the true population mean? Why? Collect data on the intervals created by other students in the class and calculate the proportion of intervals that capture the true population mean.

So far we used confidence intervals obtained by individual students when exploring the concept of confidence levels. Using computation, each individual student can also obtain many random samples and explore this concept using confidence intervals based on these samples. *Loops* come in handy here.[§]

[†] See Section 4.2.3 if you are unfamiliar with this formula.

[§] If you are unfamiliar with loops, review Lab 4A.

Here is the rough outline:

(1) Obtain a random sample.

(2) Calculate its means and standard deviation.

(3) Use these statistics to calculate a confidence interval.

(4) Repeat steps (1)-(3) 50 times.

But before we do all of this, we need to first create empty vectors where we can save the means and standard deviations that will be calculated in step (2). And while we're at it, let's also store the desired sample size as `n`.

```
samp_mean <- rep(NA, 50)

samp_sd <- rep(NA, 50)

n <- 60
```

Now we're ready for the loop where we obtain 50 random samples and quickly calculate and save their means and standard deviations.

```
for (i in 1:50) {
    samp <- sample(population, n)  # obtain a sample of size n = 60 from the population
    samp_mean[i] <- mean(samp)  # save sample mean in ith element of samp_mean
    samp_sd[i] <- sd(samp)  # save sample sd in ith element of samp_sd
}
```

Lastly, we construct the confidence intervals.

```
lower <- samp_mean - 1.96 * samp_sd/sqrt(n)

upper <- samp_mean + 1.96 * samp_sd/sqrt(n)
```

Lower bounds of these 50 confidence intervals are stored in `lower`, and the upper bounds are in `upper`. Let's view the first interval.

```
c(lower[1], upper[1])
```

But browsing these intervals one-by-one will get tedious...

On your own

1. Using the following custom function, plot all intervals. What proportion of your confidence intervals include the true population mean? Is this proportion exactly equal to the confidence level? If no, explain why.[†]

```
plot_ci(lower, upper, mean(population))
```

2. Pick a confidence level of your choosing. What is the appropriate critical value?

3. Calculate 50 confidence intervals at this confidence level. You do not need to obtain new samples, simply calculate new intervals based on the samples you have already collected. Using the `plot_ci` function plot all intervals and calculate the proportion of intervals that include the true population mean. How does this percentage compare to the confidence level you picked?

4. What concepts from the textbook are covered in this lab? What concepts, if any, are not covered in the textbook? Have you seen these concepts elsewhere, e.g. lecture, discussion section, previous labs, or homework problems? Be specific in your answer.

[†] This figure should look familiar, if not, see Section 4.2.2.