Lab 4: 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.

Template for lab report

Write your report, or at least run the code and create the plots, as you go so that if you get errors you can ask your TA to help on the spot. Knit often to more easily determine the source of the error.

```
download.file("http://stat.duke.edu/courses/Spring13/sta101.001/labs/lab4.Rmd", destfile = "lab4.Rmd")
```

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 a simple random sample of size 60 from the population. Specifically, this is a simple random sample of size 60. 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 `Gr.Liv.Area`.

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

Before we do the sampling we'll once again set a seed. Choose another team member's birthday and enter use that in the `set.seed` function below. For example, if the birthday is October 4, you would enter `104`.

```
set.seed(104)                # make sure to change the seed
samp <- 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 ot your team mates'. 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 of the:

```
sample_mean <- mean(samp)
```

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 `sample_mean`). That serves as a good *point estimate* but it would be useful to also communicate how uncertain we are of 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 the point estimate.[†]

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

lower <- sample_mean - 1.96 * se

upper <- sample_mean + 1.96 * se

c(lower, upper)
```

> **Exercise 3** Interpret this interval in context of the data. Make sure to include all relevant code for calculating the interval in your report.

This is an important 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`. There are a few conditions that must be met for this interval to be valid.

> **Exercise 4** For the confidence interval to be valid, the sample mean must be normally distributed and have standard error $s/\sqrt{n}$. What conditions must be met for this to be true? Explain why you believe your sample does/does not meet each of these conditions.

Confidence levels

> **Exercise 5** What does "95% confidence" mean? Be specific. 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 6** Does your confidence interval capture the true average size of houses in Ames? Do your team mates' intervals capture this value?

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

Exercise 7 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.

Using R, we're going to recreate many samples to learn more about how sample means and confidence intervals vary from one sample to another. *Loops* come in handy here.[§]

Here is the rough outline:

(1) Obtain a random sample.

(2) Calculate the sample's mean 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 from each sample. 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 calculate the means and standard deviations of 50 random samples.

```
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`.

Exercise 8 View the first interval using the code below. Does this interval capture the true population mean? Note that you will need to include earlier code calculating `lower` and `upper` in your report for the below command to work.

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

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

Exercise 9 Using the following function (which was downloaded with the data set), 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))
```

Exercise 10 Pick a confidence level of your choosing, provided it is not 95%. What is the appropriate critical value ($z^\star$)?

Exercise 11 Calculate 50 confidence intervals at the confidence level you chose in Exercise 10. You do not need to obtain new samples, simply calculate new intervals based on the sample means and standard deviations 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 level selected for the intervals?

Exercise 12 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.