Lab 4B: Foundations for statistical inference - Confidence levels

Sampling

If you have access to data on an entire population, say the entire US population, it's straight forward to answer questions like, "What is the average weight of a young adult living in the US?" and "How much variation is there weights in these young adults?". However, 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 average weight if you only know the weights of several dozen individuals? 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 CDC data on 20,000 individuals living in the US in 2002. Let's start by loading that data set.

```
source("http://www.openintro.org/stat/data/cdc.R")
```

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 several variables, but for the first portion of the lab we'll focus on the heights of individuals, represented by the variable **height**.

```
population <- cdc$height

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

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

Exercise 2 Would you expect another student's distribution to be identical to yours? Would you expect it to be similar? 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 using,

```
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 height of people living in the US in 2002 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)
```

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 average height of people living in the US lies between the values **lower** and **upper**. There are a few conditions that must be met for this interval to be valid.

The general formula for calculating a confidence interval is

$$(PointEstimate) \pm (CriticalValue) * (StandardError)$$

where the critical value can be calculated using the **qnorm** function. This function will return the lower quantile (by default) of the normal distribution given a probability (this can be changed to the upper quantile by specifying lower.tail=FALSE). It is important to remember that if we wanted the critical value for say a 95% confidence interval, we would need to find the quantile for probability $0.05/2$, since we are interested in the middle 95% of the normal distribution.

> **Exercise 3** 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?

Confidence levels

> **Exercise 4** What does "95% confidence" mean?

Let's look at the mean of the the larger population we have access to. For the sake of this lab, we will call this the "true mean" (even though it is not based on the entire population):

```
mean(population)
```

> **Exercise 5** Does your confidence interval capture the average height of the 20,000 individuals in our larger sample? If you are working on this lab in a classroom, 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? If you are working in this lab in a classroom, 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:

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

(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_vector <- samp_mean - 1.96 * samp_sd/sqrt(n)

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

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

```
c(lower_vector[1], upper_vector[1])
```

On your own

1. Using the package **plotrix**, you will be able to plot the 50 confidence intervals you just created and stored. First, install the package and then run the code below. How do your confidence levels look? Do approximately 95% of them contain the true population mean?

   ```
   require(plotrix)
   mean50 <- rep(mean(population), 50)  # Vector of 'true' mean repeated 50 times
   plotCI(1:50, mean50, ui = upper_vector, li = lower_vector)
   ```

2. Pick a confidence level of your choosing, provided it is not 95%. What is the appropriate critical value?
3. Calculate 50 confidence intervals at the confidence level you chose in the previous question. 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 **plotCI** 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 selected for the intervals?