Lab 6: Bootstrap Method

STA 111 (Summer Session II)

Learning Objectives

The purpose of the lab is to help you understand the basic concept of the Bootstrap and learn how to construct confidence intervals by the Bootstrap, without using the Central Limit Theorem or compare to it.

Dataset 1: Mustang Prices

The exercises below pertain to a dataset of a random sample of 25 used Mustangs being offered for sale on a website. The dataset contain information on price (in $1,000s), mileage (in thousands of miles), and age (in years) of these cars.[1]

```
mustang <- read.csv("http://shaobohan.net/sta111/mustang.csv")
```

Exercise 1. What is the average price of a used Mustang in this sample?

The Bootstrap

Using this sample we would like to construct a bootstrap confidence interval for the average price of all used Mustangs sold on this website. Below is a quick reminder of how bootstrapping works:

1. Take a bootstrap sample (a random sample with replacement of size 25) from the original sample.
2. Record the mean of this bootstrap sample.
3. Repeat steps 1 and 2 many times to build a bootstrap distribution.
4. Calculate the XX% interval as teh middle XX% of the bootstrap distribution.

Since we're going to do some random sampling, let's start by setting a seed. You may replace 12345 below with any number.

```
set.seed(12345)
```

Now let's take 100 bootstrap samples, and record their means in a new vector called `boot_means`.

```
boot_means <- rep(NA, 100)
for (i in 1:100){
  boot_sample <- sample(mustang$price, 25, replace = TRUE)
  boot_means[i] <- mean(boot_sample)
}
```

Exercise 2. Make a dot plot of the bootstrap distribution.
(The dot plot function is in a contributed R package, so we'll first need to install and load that.
Use the command `install.packages("BHH2")`) in R console.

```
library(BHH2)
dotPlot(boot_means)
```

[1] Source: Statistics: Unlocking the Power of Data.

Exercise 3. Estimate (by eyeballing) a 90% confidence interval for the average price of Mustangs sold on this website, explain briefly how you estimated the interval, and interpret this interval in context of the data.

The `inference` function

Next we'll introduce a new function that you'll be seeing more in the upcoming labs - a function that allows you to apply many statistical inference method that you'll be learning in this course. Since this is a custom function, we need to first go and download it from the course website.

```
source("http://shaobohan.net/sta111/inference.R")
```

We are going to explore this function more throughout the semester, but for now, we'll just use it to construct a bootstrap interval, without having to write our own for loop. By default this function takes 10,000 bootstrap samples and creates a bootstrap distribution, and calculate the confidence interval.

```
inference(mustang$price, type = "ci", method = "simulation", conflevel = 0.9, est = "mean")
```

We can easily change the confidence level to 95% by changing the `conflevel`:

```
inference(mustang$price, type = "ci", method = "simulation", conflevel = 0.95, est = "mean")
```

Or create an interval for the median instead of the mean:

```
inference(mustang$price, type = "ci", method = "simulation", conflevel = 0.95, est = "median")
```

Exercise 4. Create a 95% confidence interval for the average *mileage* of used Mustangs sold on this website using the inference function.

Lab Questions

Dataset 2: Salaries of College Professors

In the next part of the lab, we will work with a new dataset that contains salary information (in $1,000s) on randomly sampled college professors.[2]

```
prof <- read.csv("http://shaobohan.net/sta111/prof.csv")
```

1. What does each row represent in this dataset? How many cases are there?
2. How many variables are there? Determine if each variable is numerical or categorical.
3. Using the `inference` function, construct a 99% bootstrap confidence interval for the average salary of college professors, and interpret it in context of the data.
4. Now construct a 99% confidence interval using a theoretical method, i.e., a z-interval. Note that you will need to change `method = "theoretical"`. Comment on whether the two approachs yield similar or difference results.

[2]Source: Statistics: Unlocking the Power of Data.

5. Create a 90% confidence interval (`method = "theoretical"`) and a 90% bootstrap confidence interval (`method = "simulation"`) for the difference in mean salary by gender, compare and interpret in context. To do this you will need to add `group = as.factor(prof$gender)` in the arguments of function `inference`. Does one gender appear to be paid more than the other?

This lab is slightly modified from Lab 5 of STA101 taught by Professor Mine Çetinkaya-Rundel in Spring 2013.