Lab 4: Bootstrap confidence intervals

Template for lab report

```
download.file("http://stat.duke.edu/courses/Summer13/sta104.01-1/labs/lab4.Rmd",
    destfile = "lab4.Rmd")
```

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.

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 contains information on price (in $1,000s), mileage (in thousands of miles), and age (in years) of these cars.[†]

```
mustang = read.csv("http://stat.duke.edu/courses/Spring13/sta101.001/labs/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 the middle XX% of the bootstrap distribution.

Since we're going to do some random sampling, let's start by setting a seed.

Exercise 2 Choose another team member and use their birthday as the seed for random sampling (replace `xxx` below with their birthday).

```
set.seed(xxx)
```

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)
}
```

[†]Source: *Statistics: Unlocking the Power of Data*.

Exercise 3 Make a dot plot of the bootstrap distribution.

```
dotPlot(boot_means)
```

Exercise 4 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 a lot more of in the upcoming labs - a function that allows you to apply any 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://stat.duke.edu/courses/Spring13/sta104.01-1/resources/R/inference.R")
```

We're going to explore this function 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 calculates 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 5 Create a 95% confidence interval for the average **mileage** of used Mustangs sold on this website using the inference function.

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.[†]

```
prof = read.csv("http://stat.duke.edu/courses/Spring13/sta101.001/labs/prof.csv")
```

Exercise 6 What does each row represent in this dataset? How many cases are there?

Exercise 7 How many variables are there? Determine if each variable is numerical or categorical.

[†]Source: *Statistics: Unlocking the Power of Data*.

Exercise 8 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.

Exercise 9 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 approaches yield similar or different results. Also, use the argument `eda_plot = FALSE` to turn of the exploratory data analysis (EDA) histogram, i.e. histogram of the data.