# Lab 5 Part II: The two-sample bootstrap

*Prof. Adam Loy*

*Wednesday, October 19, 2016*

## 0. Intro

Welcome back to Lab 5!

Today's lab will continue to explore the idea of building confidence intervals using the bootstrap. Specifically, it will introduce how you can calculated a confidence interval for the difference in means.

### Setup

You can download the .Rmd file for this lab and the data from the course webpage as a zip file. If you are using a Mac, then you will need to use a browser other than Safari for the download.

If you are using the RStudio server, then you can upload the entire zip folder directly onto the server.

We will use the following packages during this lab. Make sure that you have downloaded all of them before running the commands.

```
library(ggplot2)
library(dplyr)
library(CarletonStats)
```

## 1. The data

We will explore the same data set as in the first part of the lab, the description is repeated below for your convenience.

Brad Efron was annoyed by TV commercials. He suspected that there were more commercials on the "basic" TV channels, the channels that come with a cable TV subscription, than in the "extended" channels you pay extra for. To check this, he collected the data on 20 randomly selected half hour segments of TV.

The data are contained in the file `TV.csv`. The rows describe the half hour TV segments, and the columns describe:

- **ID**: an ID number uniquely identifying segment,
- **Times**: minutes of commercials per half-hour,
- **Cable**: whether the channel was a `basic` channel or on `extended` cable.

The below code loads the data in to R.

```
tv <- read.csv("data/TV.csv")
summary(tv)
```

```
##        ID            Times             Cable   
##  Min.   : 1.00   Min.   : 3.40   Basic   :10  
##  1st Qu.: 5.75   1st Qu.: 7.45   Extended:10  
##  Median :10.50   Median : 8.10                
##  Mean   :10.50   Mean   : 8.04                
##  3rd Qu.:15.25   3rd Qu.: 9.70                
##  Max.   :20.00   Max.   :11.00
```

## 2. The two-sample bootstrap

In the first part of this lab we computed two confidence intervals: one for the average commercial duration for basic TV channels, the other for the average commercial duration for extended cable channels. Instead of trying to determine whether there are more commercials, on average, on the basic channels using two intervals, we can create a *confidence interval for the difference in mean commercial time* in half-hour blocks between basic and extended cable channels. The beauty of the bootstrap is that we can easily extend it to many different situations (i.e., the estimation of many different parameters). Since we are considering the difference in means, we must change the process slightly to utilize the two-sample bootstrap:

i) Draw a random sample of $n_1$ observations, with replacement, from the original sample for the first group.

ii) Draw a random sample of $n_2$ observations, with replacement, from the original sample for the second group.

iii) Calculate the statistic of interest from these bootstrap samples (e.g., the difference in means). This statistic is called a bootstrap statistic.

iv) Repeat steps i–iii many times, say 10,000.

v) Combine all of the bootstrap statistics to form the bootstrap distribution.

Now that you are familiar with the process in theory, let's create a bootstrap distribution for the difference in mean commercial duration.

To begin, we can draw a single bootstrap sample from our original samples using commands in the `dplyr` package that we first used in lab 2.

```
grouped_tv <- group_by(tv, Cable)
tsboot1 <- sample_frac(grouped_tv, size = 1, replace = TRUE)
```

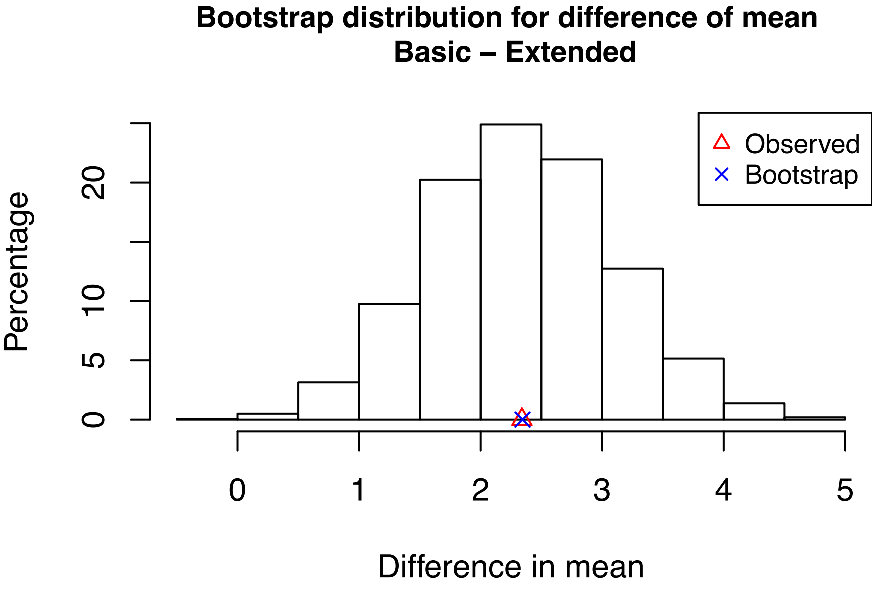
Recall that the `group_by` function allows us to partition our data set so that we can apply functions to each group. Next, we use the `sample_frac` function to sample a fraction of the original data set within each group. Specifying `size=1` indicates that we want data sets of the same size as the originals and specifying `replace=TRUE` tells R to sample with replacement.

**Question 1.** Calculate the mean for each group from your bootstrap sample and then calculate the difference in means. Compare your bootstrap statistics with those of your group members to get a sense for the variability of the bootstrap statistics.

To create a bootstrap distribution for the difference in mean commercial length in half-hour segments on basic channels, we use the following command. Note that the formula specified takes the form: `Response_variable ~ Explanatory_variable`.

```
boot(Times ~ Cable, data = tv, B = 10000)
```

```
##
##  ** Bootstrap interval for difference: mean  **
##
##  Observed difference of  mean :  Basic - Extended =  2.34
##  Mean of bootstrap distribution: 2.34478
##  Standard error of bootstrap distribution: 0.75966
##
##  Bootstrap percentile interval
##    2.5%   97.5%
## 0.87975 3.87000
##
##      *---------------*
```



**Question 2.** Explain what an individual case for the bootstrap distribution.

**Question 3.** Where is the bootstrap distribution centered?

**Question 4.** Describe the shape of the bootstrap distribution.

**Question 5.** Calculate a 95% plug-in confidence interval for the difference in mean commercial time in half hour blocks between basic and extended cable channels.

**Question 6.** Give a one sentence interpretation of the confidence interval you just found in the context of the problem.

**Question 7.** Does the confidence interval you just found provide evidence of a difference in mean commercial time in half hour blocks between basic and extended cable channels? Justify why or why not.

## 3. Wrapping up

Now that you understand the fundamentals of the one- and two-sample bootstrap procedures you can tackle a wide variety of problems. Using the bootstrap you are not restricted to inference for the mean, you can also conduct inference for the median, trimmed mean, and many other statistics. Using the `boot` function in the `CarletonStats` package this is done by adding a `fun` argument, which specifies the statistic to be calculated. For example `fun = median` will create a bootstrap distribution for the median, or difference in medians. Additionally, with some creativity (which we will explore at some point in Math 107), you can create confidence intervals for regression coefficients and the correlation. Next time we will investigate how to build bootstrap confidence intervals for different confidence levels.

At this point I want to reassure you that the bootstrap procedure is not simply a parlor trick, or a procedure that is only used by statistics professors to illustrate the principles of statistical inference. The bootstrap, along with other similar resampling procedures, are important in practice and in some situations provide the only practical way to conduct inference (i.e., build confidence intervals or conduct hypothesis tests). Tim Hesterberg from Google provides this example (adapted from arXiv:1411.5279):

> In Google Search they estimate the average number of words per query, in every country. The dataset is immense, and is "sharded"—stored on tens of thousands of machines. Google can count the number of queries in each country, and the total number of words in queries in each country, by counting on each machine and adding across machines, using the MapReduce algorithm. But they also want to estimate the variance, for users in each country, in words per query per user. The queries for each user are sharded, and it is not feasible to calculate queries and words for

every user. But there is a bootstrap procedure they use to accomplish this. (I will omit the details here since they are rather complex.)