Lab 5: The one-sample bootstrap

Prof. Adam Loy

Monday, October 17, 2016

0. Intro

Welcome to Lab 5!

Today's lab will begin to explore the idea of building confidence intervals using the bootstrap.

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.

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

1. The data

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.

```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
```

Question 1. Is this an observational study or an experiment? Briefly justify your answer.

Question 2. Why did Brad randomly select the 30 half hour segments rather than simply choosing his favorites?

Question 3. Calculate the mean length of commercials per half-hour for the basic and extended channels, as well as the difference in means. Does this difference appear to support Brad's hypothesis?

2. The one-sample bootstrap

As we already know, it is unrealistic for us to repeat this sampling process—randomly selecting 10 half hour segments of TV from both the basic and extended cable channels—a large number of times. Instead, we will allow the sample to "pull itself up by its bootstraps" and **use a bootstrap distribution to approximate the sampling distribution**. To utilize the bootstrap *we must assume that the sample is representative of the population so that the population can be "recreated" by "pasting together" many copies of the original sample*; that is, we must be able to *simulate* the population from the sample.

Given that the above assumption holds, we can approximate the sampling distribution by constructing a bootstrap distribution. For simplicity, let's start by considering the bootstrap distribution for the commercial length per half hour on basic cable. Since we are focusing on only one sample, we use the **one-sample bootstrap procedure**:

i) Draw a random sample of n observations, with replacement, from the original sample. This creates one bootstrap sample (also known as a resample).

ii) Calculate the statistic of interest from this bootstrap sample. This statistic is called a bootstrap statistic.

iii) Repeat steps i. and ii. many times, say 10,000.

iv) 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 mean commercial duration per 30-minute spot on basic cable.

To begin, we need to `filter` the data so that we only have basic cable segments

```
basic <- filter(tv, Cable == "Basic")
```

Next we draw a random sample, with replacement, from the rows of the `basic` data frame to obtain our first bootstrap sample.

```
bootstrap_sample1 <- sample_n(basic, size = nrow(basic), replace = TRUE)
bootstrap_sample1
```

```
##     ID Times Cable
## 7    7   8.2 Basic
## 6    6   7.6 Basic
## 7.1  7   8.2 Basic
## 10  10   8.5 Basic
## 2    2  10.0 Basic
## 9    9  11.0 Basic
## 4    4  10.2 Basic
## 9.1  9  11.0 Basic
## 7.2  7   8.2 Basic
## 4.1  4  10.2 Basic
```

Question 4. Calculate the mean length of commercials from your first bootstrap sample. How does it compare to the original sample mean that you calculated in Question 3?

To obtain a second bootstrap sample we repeat the process:

```
bootstrap_sample2 <- sample_n(basic, size = nrow(basic), replace = TRUE)
bootstrap_sample2
```

```
##     ID Times Cable
## 8    8  10.4 Basic
## 3    3  10.6 Basic
## 2    2  10.0 Basic
## 6    6   7.6 Basic
## 1    1   7.0 Basic
## 3.1  3  10.6 Basic
## 4    4  10.2 Basic
## 2.1  2  10.0 Basic
## 6.1  6   7.6 Basic
## 3.2  3  10.6 Basic
```

Question 5. Calculate the mean length of commercials from your second bootstrap sample. How does it compare to your first bootstrap sample mean that you calculated in Question 4? How does it compare to the original sample mean that you calculated in Question 3?

Recall that a 95% confidence interval can be calculated from a unimodal and symmetric sampling distribution using the following equation:

$$\text{statistic} \pm 2 \cdot \text{SE}$$

Since we already have the value of the original statistic in hand, we must determine two things: (i) whether the sampling distribution is approximately unimodal and symmetric, and (ii) the value of the standard error (SE). Both questions can be answered by considering **The Golden Rule of Bootstrapping: The bootstrap statistics are to the original sample statistic as the original sample statistic is to the population parameter.**

This rule tells us a few important things:

- If the center of the bootstrap distribution is approximately at the observed statistic, then the sampling distribution is centered approximately at the parameter.
- The shape of the bootstrap distribution will resemble the shape of the sampling distribution.
- The spread of the bootstrap distribution is approximately the same as the spread of the sampling distribution.

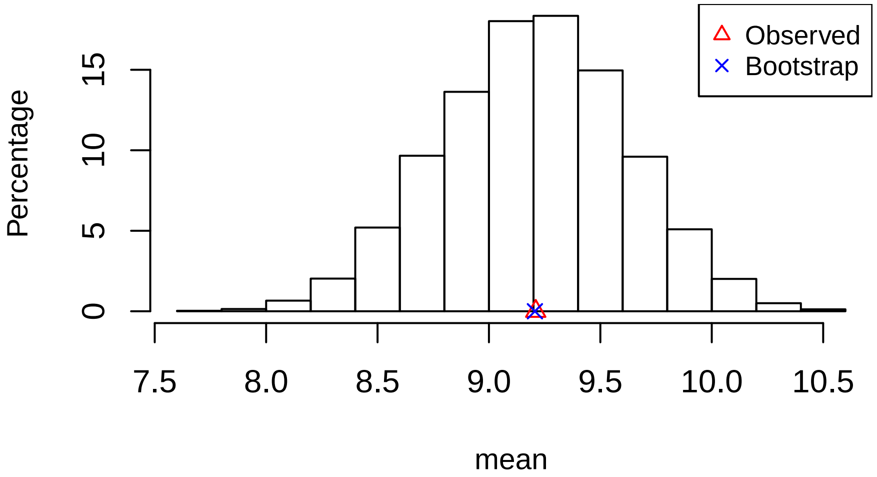
With this knowledge we can calculate a **plug-in bootstrap confidence interval** if the bootstrap distribution is approximately bell-shaped by using the standard deviation of the bootstrap distribution to approximate the standard error—that is, we simply plug in an estimate of what is unknown.

We could manually program a sampling procedure that would repeat the one-sample bootstrap procedure that we carried out above a large number of times, say 10,000, but this is already done for us in the `boot` function found in the `CarletonStats` R package.

To create a bootstrap distribution for the mean commercial length in half-hour segments on basic channels, we use the following command

```
boot(basic$Times, B = 10000)
```

```
##
##  ** Bootstrap interval for mean  **
##
##  Observed  basic$Times : 9.21
##  Mean of bootstrap distribution: 9.20628
##  Standard error of bootstrap distribution: 0.42136
##
##  Bootstrap percentile interval
##  2.5% 97.5%
##  8.38 10.02
##
##      *--------------*
```



Question 6. Describe the shape of the bootstrap distribution.

Question 7. Will the sampling distribution be approximately centered at the value of the population parameter?

Question 8. What is the bootstrap standard error?

Question 9. Calculate a 95% plug-in confidence interval for the mean commercial duration per 30-minute spot on basic cable.

Question 10. Give a one sentence interpretation of the confidence interval you just found in the context of the problem.

Question 11. Create a new data frame by `filter`ing the data so that we only have extended cable segments and run the one-sample bootstrap to create a bootstrap distribution of $B = 10000$ bootstrap statistics.

Question 12. Describe the shape of the bootstrap distribution.

Question 13. Will the sampling distribution be approximately centered at the value of the population parameter?

Question 14. What is the bootstrap standard error?

Question 15. Calculate a 95% plug-in confidence interval for the mean commercial duration per 30-minute spot on extended cable.

Question 16. Give a one sentence interpretation of the confidence interval you just found in the context of the problem.

Question 17. Brad's original question regarded whether there were more commercials on basic cable channels. What do the two confidence intervals we just calculated reveal about this?