Lab 3: Bootstrapping and Confidence Intervals

Javier E. Flores

February 15, 2019

Introduction

Over the past few lectures, we've defined and discussed properties of sampling distributions as well how these distributions might be used in order to obtain interval estimates (i.e. confidence intervals). All of what we've discussed thus far has assumed that we have access to the sampling distribution. This assumption is almost never true. Remember that the sampling distribution is obtained by computing a statistic for each of several different samples. In most practical scenarios, we have access to only a single sample.

Today's lab will introduce **bootstrapping**, which is a technique that allows us to approximate the sampling distribution using a single sample. With this approximate sampling distribution, we can then obtain interval estimates using techniques similar to what was previously discussed in class. Both Minitab and StatKey will be used throughout as we explore these techniques and ideas.

As you work through the lab with your group, you will be asked to answer several questions. Please submit your responses (as a group) in a single, separate document. Include the original questions as well as your group's response in the final submission.

Data

The data we'll be using today are from the 2018 Human Freedom Index (HFI). The HFI quantifies the state of human freedom in the world using over 70 indicators of personal and economic freedom. The HFI contains data on the majority of the world's countries across several years, but we will focus on the 162 available countries for the year 2016. For the sake of this lab, we'll assume that these data are of all world countries in 2016 (i.e. the population). Additionally, rather than explore all 70 indicators contained in the HFI, we will focus on only three summary measures of freedom:

- **Personal Freedom Score (pf_score)**: On a scale ranging from 0 to 10, this variable represents the amount of personal freedom within a country. Higher values are indicative of greater freedom.
- **Economic Freedom Score (ef_score)**: On a scale ranging from 0 to 10, this variable represents the amount of economic freedom within a country. Higher values are indicative of greater freedom.
- **Human Freedom Score (hf_score)**: On a scale ranging from 0 to 10, this variable represents a composite measure of human freedom within a country. Higher values are indicative of greater freedom.

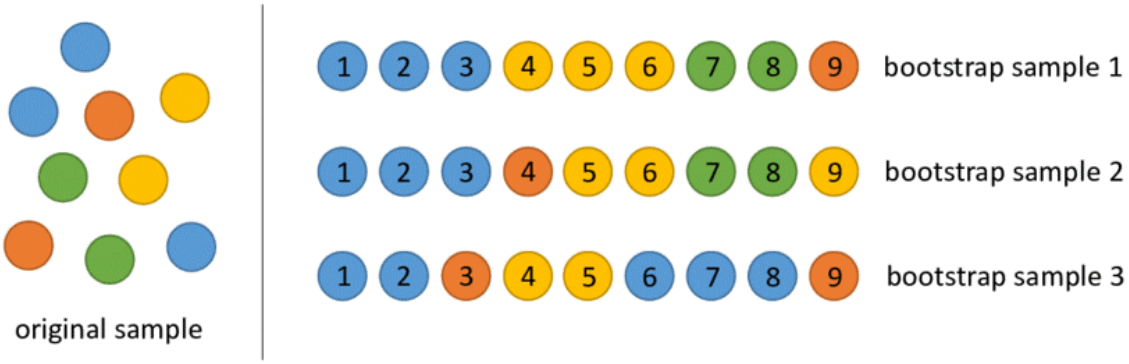
Bootstrapping

The term "bootstrapping" is derived from the old idiom, "Pull yourself up by your bootstraps". This phrase is usually used after one improves a "bad" situation through their own efforts. This is very much what we do, as statisticians, when we bootstrap. Our "bad" situation is having only a single sample to determine a sampling distribution, and we are able to overcome this challenge using the bootstrap!

The key idea behind the bootstrap is to treat our original sample as if it were the population. Then we can simulate a sampling distribution by drawing repeated samples of the same size from this "population"

with replacement. We call each of these repeated samples **bootstrap samples**. Since we are sampling with replacement, cases from our original sample may appear once, more than once, or not at all within each bootstrap sample.

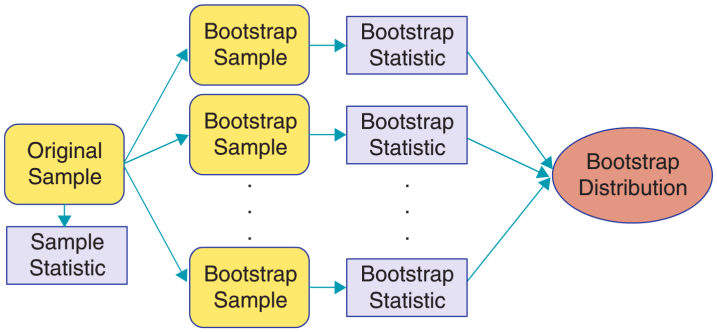
The diagram below provides a visual conceptualization of this process:



In the figure above, each bootstrap sample is obtained by sampling with replacement from the original sample. In the first and third bootstrap samples, you'll notice that there are either more yellow cases or blue cases than in the original sample. In the second bootstrap sample, you'll notice that there is one fewer orange case than in the original sample. This is all due to having been sampled with replacement. Also notice that the bootstrap samples are of the same size (i.e. $n = 9$) as the original sample.

Q1) Why do bootstrap samples need to be drawn with replacement? What would happen if the bootstrap samples were drawn without replacement?

When we perform the bootstrap, there is no limit to the number of bootstrap samples that we can draw. The greater the amount of bootstrap samples we obtain, the more "datapoints" we can use to simulate (approximate) a sampling distribution. In order to generate this approximate sampling distribution, which we refer to as the **bootstrap distribution**, we compute the statistic of interest for each of our bootstrap samples and use the resulting collection to form the bootstrap distribution. If, for example, we were interested in the mean, we would compute the sample mean for each of our bootstrap samples to form the bootstrap distribution of the mean. Shown below is a generalized representation of this process.



Bootstrap Confidence Intervals

Our interest in approximating the sampling distribution really stems from a desire to create an interval estimate. We learned in previous lectures that, if we had the actual sampling distribution, we could generate an interval estimate by using the following formula (provided the distribution is bell shaped and symmetric):

Sample Statistic $\pm$ 2*SE,

where SE denotes the standard error of the sampling distribution. We might also recall that the standard error of the sampling distribution is the same as the standard deviation of the sampling distribution. All this considered, when we approximate the sampling distribution using the bootstrap distribution, the standard deviation of our bootstrap distribution turns out to be a good estimate of our sampling distribution's standard error!

Real World	**Bootstrap World**
• Start with the population	• Start with a sample from the population
• Take several random samples from the population	• Take samples with replacement from the original sample
• Calculate the standard error as the standard deviation of the sampling distribution	• Calculate the standard error as the standard deviation of the bootstrap distribution

Using this (good) estimate of the true sampling distribution standard error, we can construct bootstrap confidence intervals. Fortunately, we don't have to do any of this by hand and can instead rely on the StatKey software. Using StatKey allows us to construct bootstrap confidence intervals for a variety of statistics, but we are required to either use one of its pre-loaded datasets or input our sample. Because we don't want to manually enter the hundred or so observations of our HFI "population" data, we'll learn how to obtain a random sample in Minitab that we can then enter into StatKey to use for bootstrap procedures.

As an example, we'll go back to day one and use the student alcohol consumption data. For this example, our question of interest will be "What percentage of students are in a romantic relationship?" Using Minitab, we'll draw a random sample of 20 students by clicking **Calc -> Random Data -> Sample from Columns**. Enter 20 for the number of rows to sample, and select the variable of interest (romantic) for the column to sample from. Type in an empty column name to store the generated sample.

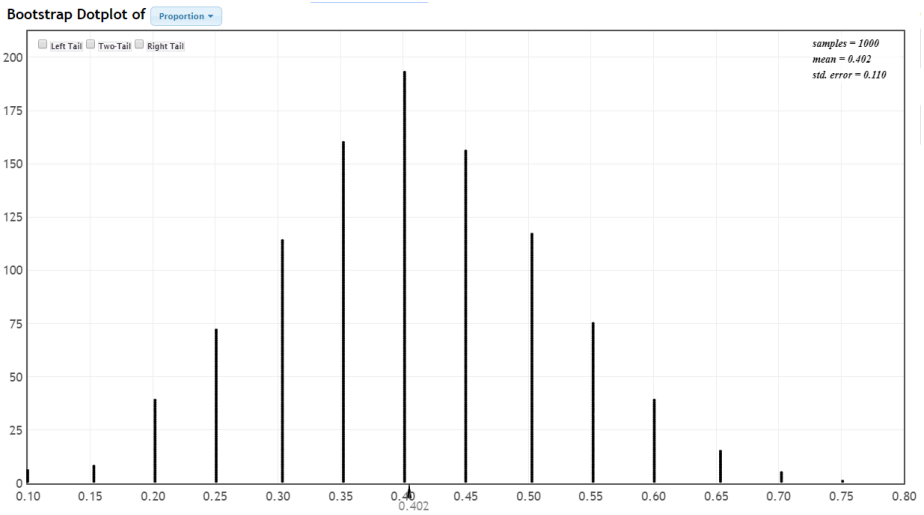
With our random sample now generated, we can move over to StatKey. Since we are interested in a percentage (proportion) for this example, we would select "CI for Single Proportion". After doing so, we can input our Minitab random sample by clicking "Edit Data". All that you are required to enter are a count and sample size. In this example, the count would be the number in our Minitab random sample who are in a romantic relationship. The sample size would be 20. The data entry will vary if you are computing another type of bootstrap confidence interval (e.g. bootstrap for the mean).

Q2) Find the true population mean human freedom score and the correlation between human freedom score and personal freedom score. In your writeup, express these population parameters using the proper notation (i.e. the greek symbols we referenced in class).

Q3) Using Minitab, obtain a random sample of size 40 from the HFI Data. Paste the names of the first 10 countries in your sample into your writeup.

Q4) Find the best estimates of the mean human freedom score and the correlation between human freedom score and personal freedom score.

Once our data are loaded into StatKey, we can obtain our bootstrap distribution by clicking "Generate 1000 Samples". After doing so, you should see output similar to what is shown on the next page:



We can then use the standard error shown to construct our bootstrap confidence interval. We will also need the statistic of interest from our original sample. Using these to pieces of information, our interval would be:

$$\text{OSStat} \pm 2{*}\text{BSE},$$

where "OSStat" is the statistic computed from the original sample and "BSE" is the bootstrap standard error. In this example, my original sample yielded a proportion of 0.667. From the graph, you'll see that the bootstrap distribution has a standard error of 0.110. Therefore the 95% bootstrap confidence interval would be:

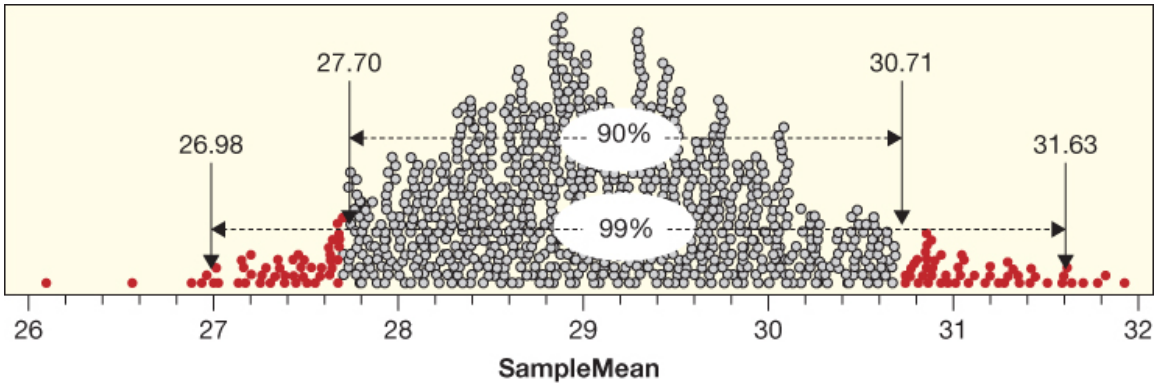
$$0.667 \pm 2 * 0.110 = (0.447, 0.887)$$

Q5) Use StatKey to construct a 95% bootstrap confidence interval for the mean human freedom score using your sample of size 40. Include a copy of the bootstrap distribution you generated in StatKey along with any other work done.

Q6) Repeat this process in StatKey but with a sample of size 80. Include a copy of the bootstrap distribution you generated in StatKey along with any other work done.

Q7) Compare the widths of the intervals generated in the previous two questions. Which is wider? Why? Explain in no more than three sentences.

It isn't always the case that our bootstrap distribution is bell shaped and symmetric, nor is it the case that we'll always want a 95% confidence interval. The approach we've just gone through for bootstrap intervals requires both of these things to be true. An alternative, more general, bootstrap approach to interval estimation exists such that neither of these conditions need be true. This method, called the **percentile bootstrap**, uses percentiles of the bootstrap distribution rather than the standard error. If, for example, we wanted to create a 90% confidence interval, the percentile bootstrap would form an interval by shaving off the lowest and highest 5% of the bootstrap distribution. The figure below illustrates this (as well as how a 99% interval would be obtained).



In StatKey, you can obtain percentile bootstrap confidence intervals by checking the "Two Tail" box. Then, clicking on the box near the center of the distribution allows us to change the confidence level.

Q8) Fill out the following table relating confidence level and interval width using the percentile bootstrap approach on your most recent sample ($n = 80$) to construct interval estimates for the mean human freedom score.

Confidence Level:	**Interval: (A,B)**	**Length: B-A**
50%		
70%		
80%		
90%		
95%		
99%		

Q9) Create a scatterplot relating the confidence level and length from the completed table in the previous question. Is the relationship linear or non-linear? Describe the relationship in no more than three sentences. Include the scatterplot in your response.

Q10) Based upon the various bootstrap intervals we've investigated in this lab, fill out the following table summarizing the impact of changing various factors on confidence interval width:

Change:	**Impact: wider/narrower/negligible?**
Increase n	
Increase number of bootstrap samples	
Increase confidence level	
Increase standard error	
Decrease confidence level	

Q11) Given what you've seen, do you think that all values in a 95% confidence interval are equally plausible? Explain your answer in 1-2 sentences.

Q12) Our investigation thus far has largely centered around the mean. Create either a percentile or standard error 95% bootstrap confidence interval of the correlation between personal freedom score and human freedom score. Choose one or other other depending on what you observe about the bootstrap distribution. Include a plot of your bootstrap distribution in your writeup.

Q13) When introducing the HFI dataset, I mentioned the data contained a vast number of indicators of personal and economic freedom. Visit this site for the codebook containing all variables included in the HFI dataset (scroll through the "Columns" section). Using any two of these variables, formulate a research question requiring you to use either correlation/regression or a difference in means. In your writeup, include 3-5 sentences addressing:

a) Your choices of variables and the research question

b) Your sample size

c) Your bootstrap confidence interval

d) Your interpretation of the interval in the context of your research question

Challenge (Optional)

In this lab, we learned how to generate random samples in Minitab and use those samples to obtain bootstrap distributions in StatKey. Both of these processes may be done in R. For this challenge, load the HFI data into R and obtain a random sample of size 80. Use the sample to generate a bootstrap distribution of the mean human freedom score. In creating the bootstrap distribution, generate at least 100 bootstrap samples. Include either a histogram or dotplot of your generated bootstrap distribution. Provide 95% bootstrap confidence intervals using both the percentile and standard error bootstrap approaches.

You may work on this challenge as a group, but I expect individuals to turn in separate copies of their own code and generated statistics/figures.

Turning in this challenge may prompt a brief assessment in which you will be asked to state the code used to generate any one of the results produced.

This challenge may be turned in and redeemed for credit at any point prior to the first exam, which will tentatively be scheduled one week after finishing chapter 3.