Lab Five: Bootstrap

STA 111: Probability & Statistical Inference

Lab Objective

To illustrate how bootstrap works.

Lab Procedures

The Bootstrap was invented by Efron in 1979. It was a revolution in statistics and has had a great impact on scientific communities. In this lab, you will have access to the law school data presented in his paper, and your goal is to understand how bootstrap works.

There are four data sets. The first data set contains information on a population of 82 American law schools participating in a large study of admission practices (*"Law82"*). The second data set is a random sample of size n = 15, drawn from the population of 82 law schools (*"Law"*). Two measurements were made on the entering class of each school in 1973: LSAT, the average score for the class on a national law test, and GPA, the average undergraduate grade point average for the class. For the third data set, we obtained 1,000 bootstrap samples and calculated correlation coefficients between GPA and LSAT for each sample (*"Thetastar"*). Finally, for the fourth data set, we selected 1,000 random samples of size 15 each from the population and calculated the correlation coefficient for each sample (*"Thetahat"*). Read in the all the data sets:

```
Law82 = read.table("http://www2.stat.duke.edu/~oma9/STA111_SummerI_2018/Labs/Law82.txt",header=T)
Law = read.table("http://www2.stat.duke.edu/~oma9/STA111_SummerI_2018/Labs/Law.txt",header=T)
Thetastar = read.table("http://www2.stat.duke.edu/~oma9/STA111_SummerI_2018/Labs/Thetastar.txt",header=T)
Thetahat = read.table("http://www2.stat.duke.edu/~oma9/STA111_SummerI_2018/Labs/Thetahat.txt",header=T)
```

Questions

1. Calculate the sample correlation coefficient for average LSAT scores and average GPA, using the sample of 15 schools (in the *"Law"* data set).

   Hint: recall the "cor" function.

2. When we do not have access to the population data, we can apply the bootstrap procedure to mimic the empirical distribution of the sample correlations coefficient. We drew 1,000 bootstrap samples and calcu-

lated correlation coefficients for each sample. They are found in *"Thetastar"*. Use R to draw a histogram of the correlation coefficients in this data set (drawn from the sample of n = 15). Describe the distribution (e.g. shape, modes, skew). Does it look normal?

Hint: recall the "hist" function.

3. Use the Bootstrap procedure to calculate:

   (a) The 95% percentile confidence interval. This requires finding the 2.5th percentile as the lower bound and the 97.5th percentile as the upper bound. Use the quantile function to find the required percentiles. Type `quantile(Data$Var,probs=c(0.025,0.975))` where Data is the data frame containing your data and Var is the particular variable you want to examine. The probs option specifies the list of percentiles you need.

   (b) The 95% pivotal confidence interval. You can calculate this by hand or use the R console as a calculator. The equation for the pivotal confidence interval is:

$$L = 2\hat{\theta} - \hat{\theta}^{\star}_{0.975}$$

$$U = 2\hat{\theta} - \hat{\theta}^{\star}_{0.025}$$

   where $\hat{\theta}$ is the sample correlation coefficient from the sample of 15 schools from question 1, $\hat{\theta}^{\star}_{0.975}$ is the 97.5th percentile from (a) above and $\hat{\theta}^{\star}_{0.025}$ is the 2.5th percentile from (a) above

4. Calculate the population correlation coefficient for average LSAT scores and average GPA, using the full data of 82 schools (in the *"Law82"* data set).

   The whole point of a bootstrap interval is to cover, with approximate probability equal to the confidence, the true value for the population. In this case, we are fortunate to have the data for average GPA and average LSAT scores of all 82 U.S. law schools in 1973 (see above), so we can see whether the intervals we have set are correct. Does your percentile confidence interval contain the true correlation coefficient? What about your pivotal confidence interval?

5. Empirical histograms of the sample correlation will converge to the probability histogram of the sample correlation. We selected 1,000 random samples from the population and calculated the correlation coefficient for each sample. They are found in *"Thetahat"* (see above). Use R to draw a histogram of the correlation coefficients in this data set (drawn from the sample of n = 15). Describe the distribution (e.g. shape, modes, skew). Does it look normal?

6. Pretend that the population of 82 schools is a sample. Use the bootstrap procedure to construct a 95% percentile confidence interval. What is the 95% pivotal confidence interval?

   This is essentially the same as the procedure from Question 3. The only difference is that you will obtain $\hat{\theta}$ from the population and both $\hat{\theta}^{\star}_{0.975}$ and $\hat{\theta}^{\star}_{0.025}$ from *"Thetahat"*.

7. Briefly compare the bootstrap confidence intervals obtained from the sample of size 15 with those obtained from the population of size 82. What do you find?

This ends the lab. Remember to turn in your lab reports on Sakai.