Lab Objective

The purpose of the lab is to help you understand the basic concept of the Bootstrap and learn how to construct confidence intervals by the Bootstrap, without using the Central Limit Theorem.

Lab Procedures

Before coming to lab, read the paper by Efron and Diaconis (1983) “Computer Intensive Methods in Statistics,” *Scientific American*, pp.116–130. 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 the paper, and your goal is to reproduce the analyses of the data.

There are two data sets. The first data set contains information on a population of 82 American law schools participating in a large study of admission practices. The second data set is a random sample of size $n = 15$, drawn from the population of 82 law schools. 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.

You can download the datasets here:

- Sample data (15 schools): *law.dta*
- Full data (82 schools): *law82.dta*
- Bootstrap samples (15 schools): *theta-star.dta*
- Bootstrap samples (82 schools): *theta-hat.dta*

Questions:

1. Calculate the sample correlation coefficient for average LSAT scores and average GPA, using the sample of 15 schools. To calculate the correlation coefficient in Stata, you simply use `correlate LSAT GPA`. 
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 calculated correlation coefficients for each sample. They are found in theta-star.dta (see above).

Use Stata to draw a histogram of the correlation coefficients (drawn from the sample of n = 15). Describe the distribution (e.g. shape, modes, skew). Does it look normal?

histogram varname(s) will show the histogram of the varname(s) you choose. Remember that a true histogram has the area of all the bars sum up to one. If you want the heights to sum up to one, use the fraction option.

3. Use the bootstrap procedure to construct a 95% confidence interval by hand. What is the percentile confidence interval? What is the pivotal confidence interval? In one concise sentence, interpret the pivotal confidence interval.

(a) The percentile confidence interval:

Some percentiles are available with summarize varname(s), detail (specifically, 1, 5, 10, 25, 50, 75, 90, 95, and 99). If you want to find a specific percentile of a variable, Stata has a useful command: pctile varname, percentiles(numlist). This will store all the percentile(s) that are found in numlist, which is just a list of numbers separated by spaces. Note that this will store the percentile(s) in r() and can be seen by typing return list.

Data Analysis Tip: return list allows you to access the results of the most recent command in Stata. With a little bit of programming you can save the results to a more permanent place using scalar and matrix, which we’ll explore in future labs.

(b) The pivotal confidence interval:

You will need to calculate this by hand. Show your work where indicated.

The equation for the pivotal confidence interval is:

\[ L = \hat{\theta} - \hat{\theta}^*_{0.025} \]
\[ U = \hat{\theta} - \hat{\theta}^*_{0.975} \]

Which is \( \hat{\theta} \) and which is \( \hat{\theta}^* \)? (Examine your class notes. If in doubt, consult a T.A.)

Read the formula carefully. Notice that the lower pivotal confidence interval bound uses the upper percentile confidence interval bound.

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.
4. How many different ways can we choose a sample of size 15 from a population of size 82 with replacement? (Note: you do not use Stata to answer this question.)

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 theta-hat.dta (see above). Use Stata to draw a histogram of the correlation coefficients (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% confidence interval by hand. What is the percentile confidence interval? What is the pivotal confidence interval?

   This is essentially the same as the procedure from Question 3. The only difference is that you will obtain both \( \hat{\theta} \) and \( \hat{\theta}^* \) from this single data set.

   Make use of the commands previously introduced in this lab.

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?

8. Explain the utility of the bootstrap procedure in no more than two sentences. (Be concise.)