STA 111: Probability & Statistical Inference

Lecture Fifteen – Bootstrap

D.S. Section 12.6

Instructor: Olanrewaju Michael Akande

Department of Statistical Science, Duke University

Outline

- Questions from Last Lecture
- Some History
- How the Bootstrap Works
- Examples
- Recap

Motivation

A lot of theoretical statistics has focused on developing methods for setting confidence intervals and testing hypotheses. A key tool for doing this is the Central Limit Theorem, which says that for large samples, the average is approximately normally distributed.

With some work, the CLT allows confidence intervals on the mean, the proportion, the sum, the difference of means, and the difference of proportions. But what can we do if we want to set confidence intervals on a correlation or an sd or a ratio?

Motivation

For many years, statisticians could not set confidence intervals on many parameters of interest without having to make strong and often unrealistic assumptions about the distribution from which the data were obtained.

For example, there is theory that tells how one can set a confidence interval on the sd, provided the data come from a normal distribution. But if one is interested on the sd of income in the U.S., we know from the histogram that there is a very long right tail. Income is not normally distributed, but economists still need to estimate the sd.

Similarly, there is theory on how to estimate confidence intervals for ratios, provided that both the numerator and the denominator are independent normal random variables. But for many applications, this is untrue—income per hour worked is an example.

Motivation

In 1979, Brad Efron invented the bootstrap. This is a computer-intensive procedure that substitutes fast computation for theoretical math.

The main benefit of the procedure is that it allows statisticians to set confidence intervals on parameters without having to make unreasonable assumptions. It was a revolution.

This was one of the first of many breakthroughs in computational statistics, which is the way that nearly all work is done now.

How the Bootstrap Works

If one samples from a population without replacement and makes a histogram of the results, then as the sample size increases, the histogram of results converges to the probability histogram for that population.

Thus if one draws 10^7 people at random and makes a histogram of their incomes, one can use this to approximate, with pretty good accuracy, the probability that the next draw will be, say, a millionaire.

How the Bootstrap Works

Let n be the sample size, and suppose one observes a random sample $X_1, \ldots, X_n$. One can form the histogram of the data by putting rectangles of size $1/\text{n}$ at each of the X_i values.

As the sample size increases one can let the width of the rectangle go to zero, and in the limit, by the convergence, one gets the probability histogram.

Note that this ensures that the total area under the histogram is 1, as required. If two of the observations are identical, then one gets a bar of twice the height, suggesting that the observation is more common.

Consider the case of determining the number of hours that people work in a week.

How the Bootstrap Works

Caveat: The following is an approximation to the real mathematical notation.

Let F be the real probability histogram for the population. And let $\hat{F}_n$ be the histogram you have built from a sample of n random draws. Both of these can be thought of as special kinds of distributions, similar to the normal or chi-squared or t-distributions.

Suppose you want to estimate some parameter θ of F, such as the sd or the ratio of the third quartile to the first. Denote your estimate of that parameter of interest by $\hat{\theta} = T(\{X_i\}, F)$. (Here $\{X_i\}$ is the set representing the entire sample, $X_1, \ldots, X_n$.)

Note that this estimate depends upon both the true probability histogram F and the particular sample $\{X_i\}$ that you happened to draw.

How the Bootstrap Works

To set a confidence region on the parameter one needs to know how chance variation in drawing the sample affects your estimate. The strategy behind the bootstrap method for doing this reflects the reflexivity in its name.

Then we can draw a new random sample of size n, with replacement, from $\hat{F}_n$. This is like drawing with replacement from a box in which each ticket is labeled with an observation in the initial random sample.

This second sample is called a bootstrap sample. For that bootstrap sample, we can calculate an estimate of the parameter of interest, say sd, for $\hat{F}_n$. Denote this by $\hat{\theta}^*$. Note that the sd of $\hat{F}_n$ will probably be different than the sd for F, but that is irrelevant.

Since we know the box perfectly, we can draw as many bootstrap samples of size n as we want.

How the Bootstrap Works

Suppose we use a computer to draw 1000 bootstrap samples of size n. For each such sample, we can calculate a new estimate of the parameter of interest.

Rank these estimates from least to largest. We denote these ordered bootstrap estimates by

$$\hat{\theta}^*_{(1)}, \ldots, \hat{\theta}^*_{(1000)}$$

where the number in parentheses shows the order in terms of size. Thus $\hat{\theta}^*_{(1)}$ is the smallest estimate of the sd found in one of the 1000 bootstrap samples, and $\hat{\theta}^*_{(1000)}$ is the largest.

The spread in these bootstrap estimates tells us how large the effect of chance error is on the estimate $\hat{\theta}$ that we got in our original sample.

How the Bootstrap Works

Suppose we want to set a 95% confidence interval on θ, the true parameter value for the real population F. And suppose we take 1000 bootstrap samples. The bootstrap method suggests that about 95% of the time, the true parameter value for $\hat{F}_n$ falls the 25th largest observation to the 975th largest observation.

Since $\hat{F}_n$ converges to F, the correct confidence interval for the true parameter on $\hat{F}_n$ should converge to the correct confidence interval on the parameter for F. This logic gives the 95% percentile confidence interval, or:

$$L = \hat{\theta}^*_{(.025)} \qquad U = \hat{\theta}^*_{(.975)}.$$

But this does not take full account of the difference between θ for F and $\hat{\theta}$, the true value for $\hat{F}_n$. We can do a bit better.

How the Bootstrap Works

The pivot confidence interval argues that the behavior of $\theta - \hat{\theta}$ is approximately the same as the behavior of $\hat{\theta} - \hat{\theta}^*$. Thus

$$\begin{aligned}
0.95 &\approx \mathrm{P}[\hat{\theta}^*_{(.025)} \leq \hat{\theta}^* \leq \hat{\theta}^*_{(.975)}] \\
&= \mathrm{P}[\hat{\theta}^*_{(.025)} - \hat{\theta} \leq \hat{\theta}^* - \hat{\theta} \leq \hat{\theta}^*_{(.975)} - \hat{\theta}] \\
&= \mathrm{P}[\hat{\theta} - \hat{\theta}^*_{(.025)} \geq \hat{\theta} - \hat{\theta}^* \geq \hat{\theta} - \hat{\theta}^*_{(.975)}] \\
&\approx \mathrm{P}[\hat{\theta} - \hat{\theta}^*_{(.025)} \geq \theta - \hat{\theta} \geq \hat{\theta} - \hat{\theta}^*_{(.975)}] \\
&= \mathrm{P}[2\hat{\theta} - \hat{\theta}^*_{.025} \geq \theta \geq 2\hat{\theta} - \hat{\theta}^*_{(.975)}]
\end{aligned}$$

So

$$L = 2\hat{\theta} - \hat{\theta}^*_{(.975)} \quad U = 2\hat{\theta} - \hat{\theta}^*_{(.025)}.$$

How the Bootstrap Works

$\hat{F}_n$ converges to F. It is not obvious, but one can show that this implies that the chance error in estimating θ for F converges to the chance error in estimating θ^* for $\hat{F}_n$.

In practice, one has to be able to draw many samples from the box model, and calculate an estimate for each. This can be time consuming, and for realistic examples on usually needs the computer.

Before the bootstrap, statisticians had to write all estimates as special kinds of averages and use the Central Limit Theorem to set approximate confidence intervals. But one can show that, as n gets large, the bootstrap is never worse than the Central Limit Theorem approximation and for many parameters it can be much better.

Example

Suppose one wants to estimate the sd in the number of hours that people work in a week. One draws a random sample of size 8, and finds

$$40,\ 35,\ 40,\ 0,\ 0,\ 40,\ 50,\ 10$$

The point estimate for the sd is easy. It is just the sd of the sample, or

$$\sqrt{\frac{1}{8}(40^2 + \ldots + 10^2) - 26.875^2} = 18.864.$$

Example

The bootstrap trick tells us how to put a confidence interval on this estimate.

Suppose we draw 500 bootstrap samples. We might get samples like the following:

Sample Number	Sample	Estimate
1	0, 40, 40, 10, 10, 10, 0, 0	15.762
2	50, 10, 0, 0, 0, 40, 40, 40	20.463
3	0, 10, 40, 35, 0, 0, 10, 0	15.398
4	40, 40, 40, 40, 40, 40, 40, 40	0
5	0, 0, 50, 50, 0, 0, 50, 50	25
etc.		

Note that the largest possible estimate is 25, and the smallest possible estimate is 0.

Example

Suppose we want to use the 500 bootstrap samples to form a 90% confidence interval on the true sd of the number of hours that people work. We shall need to find the 25th largest and the 475th largest values from the previous table, extended to have 500 samples.

Normally we would use a computer. But for tutelary purposes, suppose the 25th largest value was 14.28 and the 475th largest value was 21.62.

Then the percentile confidence interval is (14.28, 21.62). And the pivot confidence interval, which is better, is:

$$L = 2\hat{\theta} - \hat{\theta}^*_{(475)} = 2 * 18.864 - 21.62 = 16.108$$
$$U = 2\hat{\theta} - \hat{\theta}^*_{(25)} = 2 * 18.864 - 14.28 = 23.448.$$