13.0 Bootstrap Confidence Intervals

- Answer Questions
- Some History
- How the Bootstrap Works
- Example

13.1 Some History

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?

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 the ratio of two expected values, provided that both the numerator and the denominator are from independent normal random variables. But for many applications, this is untrue—income per hour worked is an example.

In 1979, Brad Efron invented a revolutionary new statistical procedure called the **bootstrap**. This is a computer-intensive procedure that substitutes fast computation for theoretical math. Surprisingly, the idea is quite simple.

The main benefit of the bootstrap is that it allows statisticians to set confidence intervals on parameters without having to make unreasonable assumptions.

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

I urge everyone to become familiar with a programming language or a statistical package. Stata is one of many; R and SAS are also popular.

13.2 How the Bootstrap Works

Recall the probability histograms from Lecture 3.1. In the limit, these give the probability of particular outcomes.

Also recall that if one samples from a population with 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.

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 stacking rectangles of area $1/n$, centered 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 value is more likely.

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

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

Let f be the unknown real probability histogram for the population. And let $\hat{f}_n$ be the histogram you have built from a sample of n random draws. The f may be viewed as a density function, and the $\hat{f}_n$ is a discrete histogram that approximates it.

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 $\hat{\theta} = T(\{X_i\})$. (Here $\{X_i\}$ is the set representing the sample values you observed, $X_1, \ldots, X_n$.)

Critically, the distribution of $\hat{\theta}$ will depend upon the unknown true probability histogram (density) f. And you need that distribution to set a confidence interval.

The bootstrap idea is elegantly simple. First, you gather the sample and use it to find $\hat{\theta}$, your estimate the unknown parameter θ.

Then 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 for $\hat{f}_n$. Denote this new estimate by $\hat{\theta}_1^*$.

Since we know the box perfectly, we can draw as many bootstrap samples of size n as we want, obtaining M estimates $\hat{\theta}_1^*, \ldots, \hat{\theta}_M^*$.

Since $\hat{f}_n$ approximates f, the unknown distribution of $\hat{\theta}$, based on f, is approximately the distribution of $\hat{\theta}_1^*, \ldots, \hat{\theta}_M^*$, based on $\hat{f}_n$. That approximating distribution is used to set confidence intervals.

The bootstrap idea is illustrated in the diagram below. The propagation of the chance variation in the initial sample into the estimator $\hat{\theta}$ is approximated by the observed impact of chance variation in the resampled data (with replacement) upon the bootstrap estimates.

$$\begin{tikzcd}
f \arrow[r] & \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \arrow[r, Rightarrow] \arrow[d] & \hat{f} \arrow[r] & \begin{bmatrix} X^*_{1(\mathbf{1})} \\ X^*_{2(\mathbf{1})} \\ \vdots \\ X^*_{n(\mathbf{1})} \end{bmatrix}, \arrow[d] & \cdots , & \begin{bmatrix} X^*_{1(M)} \\ X^*_{2(M)} \\ \vdots \\ X^*_{n(\mathbf{M})} \end{bmatrix} \arrow[d] \\
 & \hat{\theta} & & \hat{\theta}^*_{\mathbf{1}}, & \cdots , & \hat{\theta}^*_M
\end{tikzcd}$$

For example, suppose we use a computer to draw $M = 1000$ bootstrap samples of size n. For each such sample it calculates 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 (approximately) how large is the effect of chance error in the original sample upon the variation in the estimate $\hat{\theta}$. The approximation improves as n increases.

Suppose we want to set a 95% confidence interval on θ, the true parameter value for the real population f. And suppose we take $M = 1000$ bootstrap samples. The bootstrap method suggests that approximately 95% of the time, the true parameter value for $\hat{f}_n$ falls between the 2.5th percentile of the bootstrap samples and the 97.5th percentile. (Recall percentile definitions in Lecture 2.)

Since $\hat{f}_n$ converges to f, the correct confidence interval for the true parameter for $\hat{f}_n$ should converge to the correct confidence interval on the true parameter for f.

This logic gives the 95% **percentile** confidence interval, or:

$$L = \hat{\theta}^*_{(0.025)} \qquad U = \hat{\theta}^*_{(0.975)}.$$

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.

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

$$\begin{aligned}
0.95 &\approx \mathbf{P}[\hat{\theta}^*_{(.025)} \le \hat{\theta}^* \le \hat{\theta}^*_{(0.975)}] \\
&= \mathbf{P}[\hat{\theta}^*_{(0.025)} - \hat{\theta} \le \hat{\theta}^* - \hat{\theta} \le \hat{\theta}^*_{(0.975)} - \hat{\theta}] \\
&= \mathbf{P}[\hat{\theta} - \hat{\theta}^*_{(0.025)} \ge \hat{\theta} - \hat{\theta}^* \ge \hat{\theta} - \hat{\theta}^*_{(0.975)}] \\
&\approx \mathbf{P}[\hat{\theta} - \hat{\theta}^*_{(0.025)} \ge \theta - \hat{\theta} \ge \hat{\theta} - \hat{\theta}^*_{(0.975)}] \\
&= \mathbf{P}[2\hat{\theta} - \hat{\theta}^*_{0.025} \ge \theta \ge 2\hat{\theta} - \hat{\theta}^*_{(0.975)}]
\end{aligned}$$

So

$$L = 2\hat{\theta} - \hat{\theta}^*_{(0.975)} \quad U = 2\hat{\theta} - \hat{\theta}^*_{(0.025)}.$$

We already know (the book says so) that $\hat{f}_n$ 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.

13.3 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. Using the MLE, it is just the sd of the sample (dividing by n), or

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

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.

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 midpoint of the 25th and 26th largest values and the midpoint of the 475th and 476th largest values from the previous table (as extended to have 500 samples).

Normally we would use a computer. But for tutelary purposes, suppose the 5th percentile was 14.28 and the 95th percentile was 21.62.

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

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