STAT 200B 2019 Week03 (Bootstrap)

Soyeon Ahn

An Introduction to the Bootstrap (Efron & Tibshirani, 1993).

1 Random samples (Section 3.2)

It is easiest to visualize random samples in terms of finite population or universe U of individual units $U_1, U_2, \ldots, U_N$, any one of which is equally likely to be selected in a single random draw. The population of units might be all the registered voters in an area undergoing a political survey, all the men that might conceivably be selected for a medical experiment, all the high schools in the United States, etc. The individual units have properties we would like to learn, like a political opinion, a medical survival time, or a graduation rate. It is difficult and expensive to examine every unit in U, so we welect for observation a random sample of manageable size.

A *random sample of size n* is defined to be a collection of n units $u_1, u_2, \ldots, u_n$ selected at random from U. In principle the sampling process goes as follows: a random number device independently selects integers $j_1, j_2, \ldots, j_n$, each of which equals any value between 1 and N with probability $1/N$. These integers determine which members of U are selected to be in the random sample, $u_1 = U_{j1}, u_2 = U_{j2}, \ldots, u_n = U_{jn}$. In practice the selection process is seldom this neat, and the population U may be poorly defined, but the conceptual framework of random sample is still useful for understanding statistical inference.

Our definition of random sampling allows a single unit U_i to appear more than once in the sample. We could avoid this by insisting that the integers $j_1, j_2, \ldots, j_n$ be distinct, called *sampling without replacement.* It is a little simpler to allow repetitions, that is to sample with replacement. If the size n of the random sample is much smaller than the population size N, as is usually the case, the probability of sample repetitions will be small anyway. Random sampling always means sampling *with* replacement in what follows, unless otherwise stated.

Having selected a random sample $u_1, u_2, \ldots, u_n$, we obtain one or more measurements of interest for each unit. Let x_i indicate the measurements for unit u_i. The observed data are the collection of measurements $x_1, x_2, \ldots, x_n$. Sometimes we will denote the observed data $(x_1, x_2, \ldots, x_n)$ by the single symbol x.

We can imaging making the measurements of interest on every member $U_1, U_2, \ldots, U_N$ of U, obtaining values $X_1, X_2, \ldots, X_N$. This would be called a census of U.. The symbol $\mathcal{X}$ will denote the census of measurements $X_1, X_2, \ldots, X_N$. We will also refer to $\mathcal{X}$ as the population of measurements, or simply the population, and call x a random sample of size n from $\mathcal{X}$. In fact, we usually can?t afford to conduct a census, which is why we have taken a random sample. The goal of statistical inference is to say what we have learned about the population $\mathcal{X}$ from the observed data x. In particular, we will use the bootstrap to say how accurately a statistic calculated from $x_1, x_2, \ldots, x_n$ (for instance the sample median) estimates the corresponding quantity for the whole population.

2 Random sampling (Section 3.3)

Random sampling with replacement guarantees independence: if $x = (x_1, x_2, \ldots, x_n)$ is a random sample of size n from a population $\mathcal{X}$, then all n observations x_i are identically distributed and mutually independent of each other. In other words, all of the x_i have the same probability distribution F, and

$$E_F[g_1(x_1)g_2(x_2)\ldots g_n(x_n)] = E_F[g_1(x_1)]\ldots E_F[g_n(x_n)]$$

For any functions $g_1, g_2, \ldots, g_n$. (This is almost a definition of what random sampling means.) We will write $F \rightarrow (x_1, \ldots, x_n)$ to indicate that $x = (x_1, x_2, \ldots, x_n)$ is a random sample of size n from a population with probability distribution F. This is sometimes written as

$$x_i \overset{iid}{\sim} F$$

where iid stands for independent and identically distributed.

3 Parameters & statistics (Section 4.2)

Discussions of statistical inference are phrased in terms of parameters of statistics. A parameter is a function of the probability distribution F. A statistic is a function of the sample x.

We will sometimes write parameters directly as functions of F, say

$$\theta = t(F).$$

This notation emphasizes that the value θ of the parameter is obtained by applying some numerical evaluation procedure $t()$ to the distribution function F. For example F is the probability distribution in the real line, the expectation can be thought of as the parameter

$$\theta = t(F) = E_F(x)$$

Here $t(F)$ gives θ by the expectation process, that is the average value of x weighted according to F.

4 Plug-in principle (Section 4.3)

The plug-in principle is a simple method of estimating parameters from samples. The pug-in estimate of a parameter $\theta = t(F)$ is defined to be

$$\hat{\theta} = t(\hat{F})$$

In other words, we estimate the function $\theta = t(F)$ of the probability distribution F by the same function of the empirical distribution $\hat{F}$. Statistics that are used to estimate paramours are sometimes called summary statistics, as wells as estimates, and estimators)

We will use the bootstrap to study the bias an standard error of the plug-in estimate $\theta = t(\hat{F})$. The bootstrap's virtue is that it produces biases and standard errors in the automatic way, no matter how complicated the functional mapping $\theta = t(F)$ may be.

5 The bootstrap estimate of standard error (Section 6.2 and textbook)

Bootstrap methods depend on the notation of a bootstrap sample. Let $\hat{F}$ be the empirical distribution, putting probability $\frac{1}{n}$ on each of the observed values x_i. A bootstrap sample is defined to be a random sample of size n drawn from $\hat{F}$, say $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$,

$$\hat{F} \to (x_1^*, x_2^*, \ldots, x_n^*).$$

The star notation indicates that x^* is not the actual data set x, but rather a randomized, or resampled, version of x. The bootstrap data points $(x_1^*, x_2^*, \ldots, x_n^*)$ are a random sample of size n drown with replacement from the population of n objects $(x_1, x_2, \ldots, x_n)$.

The bootstrap data set $(x_1^*, x_2^*, \ldots, x_n^*)$ consists of members of the original data set $(x_1, x_2, \ldots, x_n)$,

Corresponding to a bootstrap data x^* is a bootstrap replication of $\hat{\theta}$,

$$\hat{\theta} = s(x^*) = g(x^*)$$

(g is the notation from the textbook).

The bootstrap estimate of standard error of a statistic $\hat{\theta}$ is a plug-in estimate that uses the empirical distribution function $\hat{F}$ in place of the unknown distribution F.

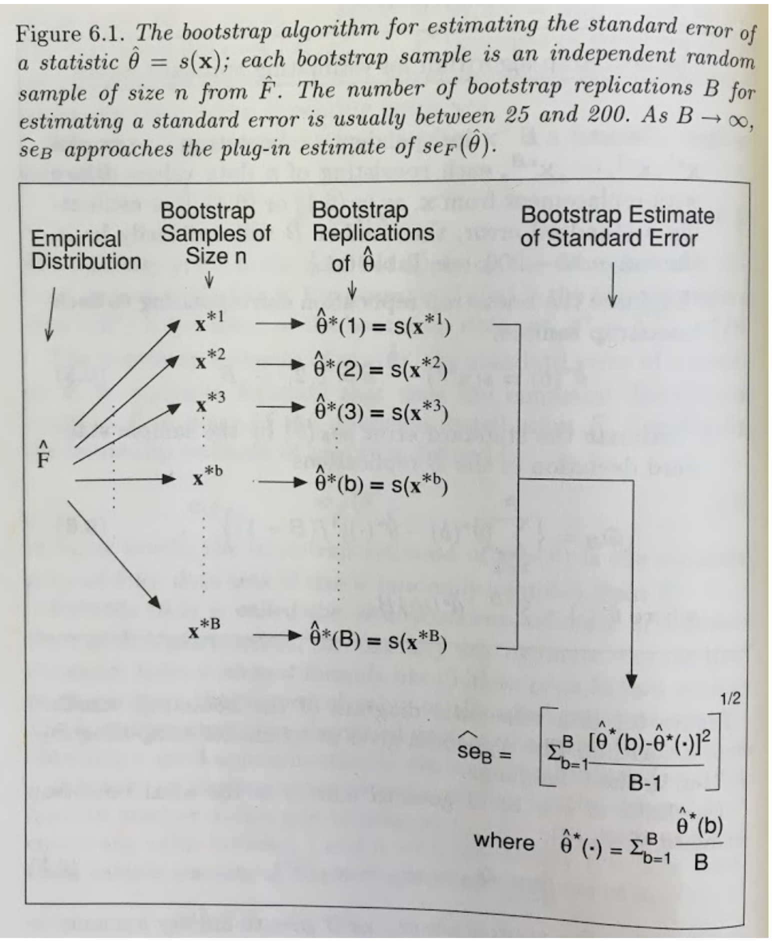

Figure 6.1. *The bootstrap algorithm for estimating the standard error of a statistic $\hat{\theta} = s(\mathbf{x})$; each bootstrap sample is an independent random sample of size n from $\hat{F}$. The number of bootstrap replications B for estimating a standard error is usually between 25 and 200. As $B \to \infty$, $\widehat{se}_B$ approaches the plug-in estimate of $se_F(\hat{\theta})$.*

Empirical Distribution
Bootstrap Samples of Size n
Bootstrap Replications of $\hat{\theta}$
Bootstrap Estimate of Standard Error
$\hat{F}$
$\mathbf{x}^{*1}$
$\mathbf{x}^{*2}$
$\mathbf{x}^{*3}$
$\mathbf{x}^{*b}$
$\mathbf{x}^{*B}$
$\hat{\theta}^*(1) = s(\mathbf{x}^{*1})$
$\hat{\theta}^*(2) = s(\mathbf{x}^{*2})$
$\hat{\theta}^*(3) = s(\mathbf{x}^{*3})$
$\hat{\theta}^*(b) = s(\mathbf{x}^{*b})$
$\hat{\theta}^*(B) = s(\mathbf{x}^{*B})$
$\widehat{se}_B = \left[\frac{\Sigma_{b=1}^{B} [\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2}{B-1} \right]^{1/2}$
where $\hat{\theta}^*(\cdot) = \Sigma_{b=1}^{B} \frac{\hat{\theta}^*(b)}{B}$

The bootstrap algorithm works by drawing many independent bootstrap samples, evaluating the corresponding bootstrap replications, and estimating the standard error of $\hat{\theta}$ by the empirical standard deviation of the replications.

Suppose we have data $X_1, \ldots, X_n$ and we compute statistic $T_n = g(X_1, \ldots, X_n)$.

It's not always possible to calculate $V_F[T_n]$ analytically, which is where the bootstrap comes in.

If we knew F, we could use MC integration to approximate $V_F[T_n]$. However, we don't in practice, so we make an initial approximation of F with the empirical CDF $\hat{F}_n$.

$$V_F[T_n] \overset{\substack{\text{ECDF;} \\ \text{depends on } n}}{\approx} V_{\hat{F}_n}(T_n) \overset{\substack{\text{MC integration;} \\ \text{depends on } B}}{\approx} \widehat{V}_{\hat{F}_n}(T_n)$$

Sampling from $\hat{F}_n$ is easy: just draw one observation at random from $X_1, \ldots, X_n$. Repeated sampling is "with replacement."

The algorithm:

1. Repeat the following B times to obtain $T^*_{n,1}, \ldots, T^*_{n,B}$, an iid sample from the sampling distribution for T_n implied by $\hat{F}_n$.
   (a) Draw $X^*_1, \ldots, X^*_n \sim \hat{F}_n$.
   (b) Compute $T^*_n = g(X^*_1, \ldots, X^*_n)$.
2. Use this sample to approximate $V_{\hat{F}_n}(T_n)$ by MC integration. That is, let

$$v_{boot} = \widehat{V}_{\hat{F}_n}(T_n) = \frac{1}{B} \sum_{j=1}^{B} \left(T^*_{n,j} - \frac{1}{B} \sum_{k=1}^{B} T^*_{n,k} \right)^2$$