# The Bootstrap

This lecture covers the basics of bootstrap procedures for bias-correction and the formation of confidence intervals. Objects of interest such as estimators and test statistics have distributions that depend on the unknown population distribution for the data. The idea of the bootstrap is to replace the population distribution with a consistent estimator in order to approximate the distribution of the statistics of interest. In some settings bootstrap estimates can improve on asymptotic approximations by reducing finite-sample bias or providing higher-order approximations to quantiles. Horowitz (2001) is a comprehensive reference for more details on bootstrap theory.

## 1 Introduction

We observe a random sample of data $X_1, \ldots, X_n$ drawn from a distribution with CDF $F_0(x) = P(X \leq x)$. We are interested in some statistic $T_n = T_n(X_1, \ldots, X_n)$, i.e. a function of the sample. The statistic $T_n$ is a random object since it depends on the random sample of data. Let $G_n(\tau, F_0) = P(T_n \leq \tau)$ denote the distribution of the statistic, given that the data are drawn from $F_0$. In general the distribution $G_n$ will depend on some feature of the distribution of the underlying data, but in some special cases $G_n$ will not depend on $F_0$. We call such statistics *pivotal*. One simple example of such a statistic is the t-statistic for a sample of normally distributed random variables. In this case, the t-statistic has a Student's t distribution regardless of the mean and variance of the normal distribution that the data came from. Such instances a rare in practical examples however.

A more useful class of statistics is those that have *asymptotic* distributions that do not depend on $F_0$. Let $G_\infty(\tau, F_0)$ be the asymptotic distribution of $T_n$ (as $n \to \infty$); then we say that $T_n$ is *asymptotically pivotal* if $G_\infty$ does not depend on the distribution of the underlying data $F_0$. In many instances $G_\infty(\tau, F_0)$ does depend on $F_0$, but only through a few estimable functions of the data. For example, many statistics are asymptotically normal, and so $G_\infty$ depends only on the mean and variance which are typically straightforward to estimate. In these cases, the asymptotic distribution can be used as an approximation to the true finite sample distribution $G_n$. This is the approach to hypothesis testing that we have discussed so far.

Another option for approximating $G_n(\tau, F_0)$ is to substitute the unknown $F_0$ with a consistent estimator $\widehat{F}$. Here there are two possibilities. One is to use a nonparametric estimator, the empirical distribution function

$$F_n(x) = \frac{1}{n}\sum_i 1\{X_i \leq x\}$$

The empirical distribution function is known to be consistent for the true distribution, a consequence of the Glivenko-Cantelli theorem. A second option is a parametric estimator. If it is known that $F_0 = F(\theta_0)$ for some known family of distributions $F$, then we can replace $\theta_0$ with a consistent estimator. If $F$ is continuous in a neighborhood of $\theta_0$ then $F(\widehat{\theta}) \xrightarrow{p} F(\theta_0)$.

Given some estimator of $F_0$, it is straightforward to approximate $G_n(\tau, F_0)$ via simulation as follows:

1. Draw a bootstrap sample of data $(X_1^*, \ldots, X_n^*)$ from $\widehat{F}$. If the nonparametric estimator is used, a sample can be drawn by sampling from $(X_1, \ldots, X_n)$ *with replacement*.
2. Compute $T_b^* = T_n(X_1^*, \ldots, X_n^*)$
3. Repeat steps 1 and 2 some large number of times, say $b = 1, \ldots, B$. Then an estimator for the distribution of $T_n$ is
$$G_n(\tau, \widehat{F}) = \frac{1}{B}\sum_{b=1}^{B} 1\{T_b^* \leq \tau\}$$

## 2 Bootstrap consistency

In order for the bootstrap distribution to provide a valid approximation to the true finite-sample distribution of the statistic, we need to ensure consistency. Here we briefly describe conditions under which the bootstrap will provide a consistent estimator of the sampling distribution. The idea behind the bootstrap is that if $\widehat{F}$ is close to the truth $F_0$ then $G_n(\tau, \widehat{F})$ should be close to $G_n(\tau, F_0)$. This argument requires continuity of $G_n(\tau, F)$ in $F$, which will be a key condition for bootstrap consistency. At the same time, in large samples we should expect $G_n$ to be close to $G_\infty$ so that the bootstrap estimator should become close to the asymptotic distribution. The following theorem makes some of these ideas precise.

**Theorem 1: Consistency of the bootstrap**

Let $\mathcal{F}$ denote the space of permitted distribution functions, and let $\rho$ be some metric on $\mathcal{F}$. If (i) $\lim_{n\to\infty} P(\rho(\widehat{F}, F) > \varepsilon) = 0$ for any $\varepsilon > 0$; (ii) $G_\infty(\tau, F)$ is continuous in $\tau$ for every $F \in \mathcal{F}$; and, (iii) for any $\tau$ and sequence of distributions $\{H_n\} \in \mathcal{F}$ such that $\lim_{n\to\infty} \rho(H_n, F_0) = 0$, we have $G_n(\tau, H_n) \to G_\infty(\tau, F_0)$, then $G_n(\tau, F_n)$ is consistent, i.e.

$$\sup_\tau |G_n(\tau, F_n) - G_\infty(\tau, F_0)| \xrightarrow{p} 0$$

The limiting distribution need not be normal, although in many applications statistics do have normal limits. In the special case that the statistic of interest is a linear functional of $F_0$, i.e. a sample average of some functions of the data, $T_n = (\bar{g}_n - t_n)/s_n$ for $\bar{g}_n = \sum_i g_n(X_i)/n$ and $t_n$ and $s_n$ some sequences of numbers (typically the mean and standard error of $\bar{g}$), it is the case that the bootstrap is consistent if and only if the limiting distribution of $T_n$ is standard normal.

The bootstrap is consistent in most standard applications, but it is worth pointing out some situations in which it is known to be inconsistent. These typically mirror the settings in which standard asymptotic approximations also fail.

- *Heavy-tailed data.* The bootstrap can fail when the data is too heavy-tailed. For example, the sample average of data from a Cauchy distribution is not asymptotically normally distributed and the bootstrap is not consistent in this case.

- *Points of discontinuity or superefficiency.* Consider trying to estimate the square of the mean of some data with mean $\mu$ and variance $\sigma^2$. For $\mu \neq 0$ we can shown that that $\sqrt{n}(\bar{X}^2 - \mu^2)$ is asymptotically normal; however, when $\mu = 0$ the estimator $\bar{X}^2$ actually converges at a faster rate, specifically $n\bar{X}^2/\sigma^2$ is asymptotically chi-squared distributed. This is known as *superefficiency* and occurs in a number of settings with edge cases like this (e.g. a unit root process). Condition (iii) of the theorem actually fails here and so the bootstrap is inconsistent which is generally the case with superefficiency.

- *Distribution of the maximum.* The bootstrap estimator of the maximum of a sample from a continuous distribution is inconsistent. Intuitively, the maximum is drawn with probability zero from the population distribution, but drawn with probability $1 - (1 - 1/n)^n \to 1 - 1/e$ from the empirical CDF.

- *Parameters at the edge of a parameter space.* The bootstrap is also generally inconsistent when the true parameter is on the boundary of some constraint. In this case, limiting distributions are generally truncated due to the boundary, and the bootstrap will not be consistent.

In some cases the validity of the bootstrap can be restored by using alternative sampling procedures, such as drawing subsamples of size $m < n$ or by sampling *without replacement*. See Horowitz (2001) for more details on these methods.

# 3 Using the bootstrap

## 3.1 Bias reduction

Many estimators have some finite sample bias. As long as the bias is $o_p(n^{-1/2})$, it will typically be dominated by the standard error of the estimator and will not affect the asymptotic distribution of the estimator. Nonetheless, we can improve the finite-sample properties of our estimator by attempting to remove this bias. As an example, consider estimating a function of the sample mean, say $\theta_0 = g(\mu)$ with the estimator $\widehat{\theta} = g(\bar{X})$. Although the sample mean is unbiased for the population mean, if $g$ is nonlinear then the $E[\widehat{\theta}] \neq \theta_0$. A Taylor expansion of $\widehat{\theta}$ around $\theta_0$ can be used to show that the bias is of order $O(n^{-1})$.

Now consider the bootstrap version of the statistic $\theta^* = g(\bar{X}^*)$, where $\bar{X}^*$ is the sample mean from a bootstrap sample. This estimator is also biased as an estimator for $\widehat{\theta}$, but since we know the 'true' parameter value corresponding to the empirical CDF $F_n$, we can compute this bias term, $E^*[\theta^*] - \widehat{\theta}$, where $E^*$ is expectation over draws from $F_n$. This can be done to arbitrary accuracy by drawing bootstrap samples and approximating $E^*[\theta^*]$ using $\sum_b \theta_b^*/B$. If $g$ is a smooth function of $\mu$, it can be shown that the bootstrap bias-corrected estimator

$$\widehat{\theta}_{bc} = \widehat{\theta} - \left(E^*[\theta^*] - \widehat{\theta}\right)$$

has bias that is of order $O(n^{-2})$, an improvement from the original estimator.

A general procedure is as follows:

1. Estimate $\widehat{\theta}$ using the full sample
2. Compute $B$ bootstrap versions of the statistic $\theta_b^*$ by drawing samples of size $n$ from the data
3. Estimate the bias as $\frac{1}{B}\sum_b \theta_b^* - \widehat{\theta}$, and construct a bias-corrected estimator as

$$\widehat{\theta}_{bc} = 2\widehat{\theta} - \frac{1}{B}\sum_b \theta_b^*$$

## 3.2 Hypothesis testing and confidence intervals

Consider constructing a test for some hypothesis that rejects whenever $|T_n| > c_{n,\alpha/2}$ where $c_{n,\alpha/2}$ is some critical value that is chosen to ensure that the size of the test is $\alpha$. Such test statistics are

common, for example when $T_n$ is a statistic that is asymptotically normal. The choice of critical value that gives exact finite-sample size is the $c$ that solves

$$G_n(c, F_0) - G_n(-c, F_0) = 1 - \alpha$$

This gives a *symmetric* two-tailed test, by which we mean that the rejection region is placed symmetrically around the null value. If $T_n$ is some (rescaled) parameter estimate, this leads to a symmetric confidence interval around the estimate. We could also form a *equal-tailed* test, which places equal weight in the left and right tails of the distribution and solves

$$\begin{aligned} G_n(c_l, F_0) &= \alpha/2 \\ G_n(c_r, F_0) &= 1 - \alpha/2 \end{aligned}$$

For now we will stick with the symmetric interval, since our arguments will apply to both, and asymptotically the distinction is unimportant for statistics with symmetric limiting distributions like the normal.

Of course we cannot solve the above equation without knowledge of $F_0$. Replacing $G_n$ with its asymptotic distribution $G_\infty$ gives one way of finding approximately correct critical values. If $T_n \Rightarrow N(0, \sigma^2)$ then the choice $c = z_{\alpha/2}\sigma$ gives a test of asymptotically correct size, where $z_{\alpha/2}$ is the critical value from the standard normal distribution.

A bootstrap critical value is found by replacing $F_0$ with the empirical CDF $F_n$.

$$G_n(c^*, F_n) - G_n(-c^*, F_n) = 1 - \alpha$$

This equation can be solved with arbitrary accuracy via simulation. For example, if $T_n = \sqrt{n}(\widehat{\theta} - \theta_0)/s_n$ is a t-statistic for $\theta_0$, we can use the following procedure

1. Estimate $\widehat{\theta}$ and its asymptotic standard error $\widehat{s}$.
2. Generate a bootstrap sample and some estimators for the parameter and asymptotic standard error, $\theta_b^*$ and $s_b^*$. Construct the bootstrap t-statistic $T_b^* = \sqrt{n}(\theta_b^* - \widehat{\theta})/s_b^*$
3. Compute $B$ bootstrap versions of $|T_b^*|$ and set the critical value $c_{1-\alpha}$ equal to the $1 - \alpha$ quantile of this distribution
4. A $(1 - \alpha)$ confidence interval is given by $\widehat{\theta} \pm c_{1-\alpha}\widehat{s}/\sqrt{n}$

**An equal-tailed interval**

The above algorithm gives a symmetric confidence interval. To construct an equal-tailed interval, we follow the same bootstrapping procedure and set

$$c_1 = (\alpha/2) - \text{quantile of } T_b^*$$
$$c_2 = (1-\alpha/2) - \text{quantile of } T_b^*$$

Then an equal-tailed interval is given by

$$[\widehat{\theta} - c_2\widehat{s}/\sqrt{n}, \quad \widehat{\theta} - c_1\widehat{s}/\sqrt{n}]$$

This interval includes an automatic bias correction since the equal-tailed construction accounts for the possible of $E^*[T_b^*]$ being non-zero.

**Accuracy and asymptotic pivotality**

An important question is how accurate is the critical value computed using the bootstrap. It turns out that the answer depends on whether $T_n$ is asymptotically pivotal. Under some regularity conditions, a Taylor style expansion to the distribution $G_n(\tau, F_0)$ is available, centered on the asymptotic distribution. This is known as an *Edgeworth expansion*, and taking such expansions of both the true finite sample distribution and the bootstrap approximation gives

$$G_n(\tau, F_0) = G_\infty(\tau, F_0) + \frac{1}{n^{1/2}} g_1(\tau, F_0) + \frac{1}{n} g_2(\tau, F_0) + \frac{1}{n^3/2} g_3(\tau, F_0) + O(n^{-2})$$
$$G_n(\tau, F_n) = G_\infty(\tau, F_n) + \frac{1}{n^{1/2}} g_1(\tau, F_n) + \frac{1}{n} g_2(\tau, F_n) + \frac{1}{n^3/2} g_3(\tau, F_n) + O(n^{-2})$$

Differencing the two expansions, we find that the largest term in the difference between $G_n(\tau, F_n)$ and $G_n(\tau, F_0)$ is the difference in the asymptotic distributions, $G_\infty(\tau, F_n) - G_\infty(\tau, F_0)$. This term is $O(n^{-1/2})$ (since $F_n$ converges to $F_0$ at this rate) so that the bootstrap makes an error in estimating the CDF of this size. This is of the same order as the error made by using an asymptotic approximation, $G_n(\tau, F_0) - G_\infty(\tau, F_0)$.

This situation changes when we bootstrap a statistic that is asymptotically pivotal. Recall that a statistic is asymptotically pivotal if its asymptotic distribution does not depend on the distribution of the underlying data, this implies that $G_\infty(\tau, F_0) = G_\infty(\tau, F_n)$. Therefore, the approximation error made by bootstrapping an asymptotically pivotal statistic is dominated by the term $\frac{1}{\sqrt{n}}(g_1(\tau, F_n) - g_1(\tau, F_0))$, which is of order $O(n^{-1})$. So the bootstrap is more accurate than the asymptotic approximation in this case.[1] The same arguments follow through to estimation of

[1] Hall (1998) shows that the symmetric confidence interval has even greater coverage accuracy of $O(n^{-2})$, and

critical values and rejection probabilities. The bottom line here is that we should *always bootstrap asymptotically pivotal statistics* when they are available.

### 3.3 Bootstrap for GMM

When using the bootstrap for GMM it is important tor *recenter* the moment conditions at each bootstrap step. The GMM estimator is based on the moment condition $E[g(z,\theta_0)] = 0$, but in overidentified settings, the moment condition will not hold exactly in sample, i.e.

$$E^*[g(z,\widehat{\theta})] = \frac{1}{n}\sum_i g(z_i,\widehat{\theta}) \neq 0$$

where $E^*$ is expectation with respect to the empirical CDF $F_n$. This means that in any individual bootstrap sample, the model is actually misspecified under the empirical distribution $F_n$. This does not affect the validity of bootstrap confidence sets and critical values for $\theta$ (although it does take away the higher accuracy we gain by using asymptotically pivotal statistics), but it does mean that bootstrap tests for the overidentification statistic will fail. We can remedy both of these problems by simply recentering the moment conditions to enforce the validity of moments in sample, i.e

$$\tilde{g}(z,\theta) = g(z,\theta) - \frac{1}{n}\sum_i g(z_i,\widehat{\theta})$$

In some settings, we may also find bootstrapping a slow or computationally intensive process. This can sometimes be the case with extremum estimators with objectives that are slow to optimize over. One solution to this issue is to replace bootstrap estimates with single (or a small number) of Newton or quasi-Newton steps from $\widehat{\theta}$. For GMM, we can use the first-order approximation to $\sqrt{n}(\theta^* - \widehat{\theta})$ to give

$$\theta_b^* = \widehat{\theta} - (\widehat{G}'W\widehat{G})^{-1}\widehat{G}'W\frac{1}{\sqrt{n}}\sum_i \tilde{g}(z_i^*,\widehat{\theta})$$

This produces a valid bootstrap since, in large samples, the approximation above becomes close to the actual estimator we would have obtained by optimizing the GMM objective in the bootstrap sample. We have saved ourselves having to actually optimize over the objective though and so this method may be much faster in practice. We do however lose the higher-order accuracy of the bootstrap when our statistic is asymptotically pivotal.

so may be preferred to the equal-tailed interval in this setting. Note that in settings with meaningful bias this result no longer holds (since if the estimator is asymptotically biased, it is not asymptotically pivotal) and using a bias-corrected interval is obviously important.

# 4 Constructing simultaneous confidence sets

When performing tests on multiple parameters we should construct confidence sets and critical values that ensure correct size *simultaneously* for all parameters. The confidence intervals we have constructed so far are individually valid; each interval covers the corresponding true parameter with probability $1-\alpha$. However, when we construct multiple intervals, all intervals will cover all true parameters simultaneously with probability less than $1-\alpha$ (random samples for which the confidence interval for $\theta_1$ do not cover the true value are not necessarily the same samples for which confidence intervals for $\theta_2$ do not cover).

Consider a vector of $k$ parameters that are jointly asymptotically normal,

$$\sqrt{n}(\widehat{\theta}-\theta_0) \Rightarrow N(0,\Sigma)$$

where $\Sigma$ is a $k \times k$ asymptotic covariance matrix. We aim to form a confidence set $\widehat{S} = [\ell_1, u_1] \times \cdots \times [\ell_k, u_k]$ with the property that

$$P(\theta_0 \in \widehat{S}) = P(\bigcap_{j=1}^{k} \{\theta_j \in [\ell_j, u_j]\}) \to 1-\alpha$$

where $\cap_j\{\theta_j \in [\ell_j, u_j]\}$ is the intersection of the individual coverage events. This is a rectangular confidence set, i.e. the intersection of intervals. There are many other options (e.g. ellipses) but rectangular sets are simple to communicate since the set for each parameters does not depend on the others.

One possibility for constructing simultaneous intervals is to use a Bonferroni correction, which uses the union bound

$$P\big(\bigcup_{j=1}^{k} \{\theta_j \notin [\ell_j, u_j]\}\big) \leq \sum_{j=1}^{k} P(\theta_j \notin [\ell_j, u_j])$$

This suggests simply constructing $k$ individually valid $(1-\alpha/k)$ confidence intervals, which by the above inequality will have coverage no less than $1-\alpha$. This is a valid method, but in general will be conservative since it takes a worst case bound and does not take into account the covariance structure between the different parameter estimates.

Another option is to use the asymptotic joint normality of the parameters. Let $D = diag(\Sigma)$ be the diagonal matrix of variances, then $C = D^{-1/2}\Sigma D^{-1/2}$ is the correlation matrix (a matrix with diagonal entries of 1 and off-diagonal entries containing correlations between estimates). Take the critical value $c_{1-\alpha}$ as

$$c_{1-\alpha} = (1-\alpha) - \text{quantile of } \|N(0,C)\|_\infty$$

that is, the quantile of the distribution of the *maximum* of the absolute value of a vector of jointly

normally distributed variables with covariance $C$. Then, we can construct jointly valid intervals as

$$\widehat{\theta}_j \pm c_{1-\alpha}\sqrt{\hat{\Sigma}_{jj}/n}$$

for $j = 1\ldots, k$. Note that these confidence intervals will be wider than the individually valid ones, since the critical value comes from the distribution of the maximum of $k$ normal random variables, as opposed to just one. The critical value needs to be simulated in practice by repeatedly drawing multivariate random normals with variance $C$ and storing the largest in absolute value.

Of course we can also compute jointly valid confidence sets using the bootstrap. A method for this is as follows

1. Estimate $\widehat{\theta}$ and the asymptotic covariance $\widehat{\Sigma}$. Denote diagonal entries (variances) as $\widehat{s}_j^2 = \widehat{\Sigma}_{jj}$.
2. Draw $B$ bootstrap samples, and for each sample compute the parameter estimate $\theta_b^*$ and the corresponding standard errors $s_{b,j}^* = \sqrt{\Sigma^*}_{jj}$.
3. For each bootstrap, let
$$t_{b,max} = \max_{j=1,\ldots,k} |\sqrt{n}(\theta_{b,j}^* - \widehat{\theta}_j)/s_{b,j}^*|$$
4. Let the critical value $c_{1-\alpha}$ be equal to the $(1-\alpha)$-quantile of $t_{b,max}$.
5. Compute jointly valid intervals for each $j$ as
$$\widehat{\theta}_j \pm c_{1-\alpha} \times \widehat{s}_j/\sqrt{n}$$

# 5 Handling dependent data

With dependent data we cannot simply draw new observations at random from the observed sample since this would not retain any of the dependence structure that exists in the original sample. With dependent (time-series) data, two options are a parametric bootstrap and a block bootstrap. The parametric approach is straightforward to carry out whenever a parametric model for the data exists. For example, assume that the data follow the AR(1) process

$$Y_t = \mu + \rho Y_{t-1} + \varepsilon_t$$

Then, one method for simulating data sets is to: (1) estimate the parameters of the model, in this case $\widehat{\mu}$ and $\widehat{\rho}$, (2) compute estimated residuals $\widehat{\varepsilon}_t = Y_t - \widehat{\mu} - \widehat{\rho}X_t$, (3) draw bootstrap data sets by sampling $\varepsilon_t^*$ with replacement from $\{\widehat{\varepsilon}_t\}_{t=1}^{\infty}$ and constructing data recursively using

$$Y_t^* = \widehat{\mu} + \widehat{\rho}Y_{t-1}^* + \varepsilon_t^*$$

This method can be easily extended to any ARMA process with i.i.d. errors. If the DGP can be represented as an infinite-order AR process then the above method can be used by estimating an AR(p) model, where $p \to \infty$ as the sample size grows. This is known as a *sieve bootstrap* and can result in estimates that are close in accuracy to that available under random sampling.

The second option is to use a nonparametric bootstrap that samples from the data in a way that retains the dependence structure. This is done by using the *block bootstrap*, which involves dividing the data into blocks of consecutive observations and sampling from the blocks at random with replacement. There are several versions of the block bootstrap, including using non overlapping or overlapping blocks, and using fixed or random block size. For consistency, it is important that the block size increases with the sample size.

Finally, with panel data or other data structures that involve clustering, we can perform a nonparametric bootstrap by drawing entire clusters with replacement. For example, with a panel data set we would draw individuals (i.e. all $T$ observations for individual $i$). If we have data on children in different classrooms we could sample entire classrooms with replacement. This type of bootstrap assumes independence over $i$ (or over classroom), but allows for dependence within individuals over time (or between students in the same classroom). See Horowitz (2001) and Cameron and Trivedi (2005), Chapter 11 for more references and details on these alternative data structures.