Notes on Bootstrap Inference

Jörg Stoye

April 21, 2025

1 Introduction

The bootstrap (introduced by the statistician Brad Efron in a seminal paper in 1979) has become an important alternative to asymptotic approximation. We will first try to understand how it works and then discuss its advantages and limitations.[1]

At the most general level, the bootstrap can be understood as extending the analog or "plug-in" principle from estimation to inference. To understand what I mean by plug-in principle, consider any quantity of interest that can be defined as $\theta = g(F)$, where $g(\cdot)$ is a known function and F is the true distribution of the data. Obvious examples are the mean $\mu \equiv \mathbb{E}X$ of a r.v. or the linear projection $\beta = (\mathbb{E}(XX'))^{-1}\,\mathbb{E}XY$. Imagine one has available an estimator $\hat{F}$ of F. An obvious such estimator, but not the only conceivable one, is the empirical distribution F_n, i.e. the discrete distribution whose probability mass function coincides with sample frequencies. Then it would be natural to estimate θ by $g(\hat{F})$. If $\hat{F}$ is consistent for F in a sufficiently strong sense and g is continuous, the estimator will be consistent. This idea can be used to motivate plug-in estimators of μ and β; in both examples just given, the resulting estimators should look familiar.

Now, in principle $g(\cdot)$ could also be the standard deviation or c.d.f. of a given test statistic at a specific sample size n. Then a plug-in estimator of this c.d.f. could be used to estimate a standard error or critical value. That is the basic idea of bootstrap inference. I'll now explain it with some more notation. We will continue to assume that data are i.i.d., but the bootstrap has been extended beyond that setting.

We will focus on two estimands. One is the (scaled) standard deviation of an estimator. Indeed, recall from the previous section that standard errors in quantile regression are usually bootstrapped

[1] This section owes much to Bruce Hansen's lecture notes, to chapter 1 of Politis, Romano, and Wolf's "Subsampling" textbook, and to Joel Horowitz' *Handbook of Econometrics* chapter. Note that this section is about bootstrap-based inference, not about other uses of bootstrap techniques. We will only cursorily mention bootstrap bias correction and not at all bootstrap averaging of estimators ("bagging"), which is an important part of some Machine Learning estimators but is not motivated for estimators discussed in this lecture.

for practical reasons. The other one is a general distribution

$$J_n(t, F) = \Pr(T_n \leq t|F)$$

of a sample statistic (in practice a test statistic, hence the notation)

$$T_n(W_1, \ldots, W_n, F).$$

Here, F is the population distribution of observables. Note that both T_n and J_n are indexed by sample size n.

The idea is to estimate J_n by a plug-in estimator using $\hat{F}$. Introducing the conventional "asterisk" notation for bootstrap analogs, we write

$$J_n^* = J_n(t, \hat{F}).$$

Again, the simple nonparametric bootstrap estimates F by the empirical distribution F_n, and we will initially stick with that for exposition, but it is not essential and in many cases not the best way to implement the bootstrap. Somewhat more formally,

$$\begin{aligned} J_n^*(t) &= J_n(t, F_n), \\ F_n(w) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{W_i \leq w\}. \end{aligned}$$

By the Glivenko-Cantelli Theorem, $F_n(w) - F(w) \overset{a.s.}{\to} 0$ uniformly over $w \in \mathbb{R}^k$. The obvious hope is that this makes J_n^* a good estimator of J_n. This is far from obvious because T_n is typically scaled by $\sqrt{n}$ and $\sqrt{n}\,|F_n(w) - F(w)| = O_P(1)$. We will later investigate conditions under which it is nonetheless true.

This looks easy enough, but beware that some bootstrap concepts can initially be confusing. The bootstrap distribution of a sample statistic is itself a sample statistic and therefore nonrandom given the data. However, being a distribution, it characterizes a random variable. This random variable will be important and is the bootstrap analog of the sample statistic in question. We will denote bootstrap analogs by asterisks if we refer to abstract bootstrap analogs and by b superscripts if it is necessary to count over $b = 1, \ldots, B$ bootstrap replications. Thus, the bootstrap analog of our observable variable W is the r.v. W^* distributed uniformly on the empirical data $\{W_1, \ldots, W_n\}$. Note that $\mathbb{E}(W^*) = \overline{W}$ and $\text{var}(W^*) = \frac{1}{n}\sum_{i=1}^{n}(W_i - \overline{W})(W_i - \overline{W})'$; that is, the sample moments of W are the population moments of W^*. The bootstrap analog of $\bar{W}$ is the r.v. $\overline{W}^* \equiv \frac{1}{n}\sum_{i=1}^{n} W_i^*$ whose distribution is implied; similarly, any test statistic T_n will have a bootstrap analog T_n^*. However, the randomness in these bootstrap quantities is *not* driven by the randomness of the original data. You may find it conceptually easiest to just think of $J_n(\cdot)$ (and implied critical values etc.) as estimand and bootstrap random variables as a by-product. However, we will see that implementation of the bootstrap will typically involve simulation of bootstrap random variables.

1.1 Implementing the (Simple, Nonparametric) Bootstrap

Our definition precisely characterizes J_n^*, and so in principle it can be computed explicitly. In particular, imagine an $(n^n \times n)$ array $(W_i^b)_{i=1,\ldots,n}^{b=1,\ldots,n^n}$ whose rows correspond to all possible (and equally likely) bootstrap samples $(W_1^*, \ldots, W_n^*)$, or in other words to all possible sequences of n draws from the uniform distribution on $\{W_1, \ldots, W_n\}$. Then for a statistic $T_n(w_1, \ldots, w_n, F)$ we have

$$J_n^*(t) = \Pr(T_n^* \leq t) = \frac{1}{n^n} \sum_{b=1}^{n^n} \mathbf{1}\{T_n(W_1^b, \ldots, W_n^b, F_n) \leq t\}.$$

Evaluating this would lead to a complete, or idealized, bootstrap. Asymptotic theory of the bootstrap is strictly speaking asymptotic theory of this complete bootstrap; that is, we take $J_n(t, F_n)$ to be a known function of F_n and then show that as $n \to \infty$, $J_n(\cdot, F_n)$ and $J_n(\cdot, F)$ become similar in some stochastic sense.

But in practice, $J_n(\cdot, F_n)$ can typically not be computed because n^n is too large.[2] Other than in toy examples and examples that allow closed-form analysis, the actual estimator is, therefore, a simulation-based approximation of $J_n^*(t)$. In particular, one would randomly select $B \ll n^n$ rows from the above array and then compute the above expression only for this smaller array.

For the classic nonparametric bootstrap that we currently discuss, this *Monte Carlo bootstrap* is easy to implement, and thinking through the implementation may aid the intuition. For a size B Monte Carlo bootstrap, simulate B bootstrap samples by drawing $B \times n$ data points from the original, empirical sample *with* replacement. (Why would it be pointless do do this without replacement?) You may figuratively think of putting the n observed data points into an urn and drawing from that urn $B \times n$ times with replacement. Next, compute the bootstrap test statistic T_n^b for each of the B simulated samples. Your simulated distribution of T_n^*, which in turn is your estimated distribution of T_n, is discrete and uniform over the B (not necessarily distinct) support points you just computed, so that $\Pr(T_n \leq t)$ is estimated by $\frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\{T_n^b \leq t\}$ and other features of the distribution of T_n, e.g. its standard deviation, are similarly estimated by bootstrap analogs.

Almost all bootstraps reported in practice are Monte Carlo bootstraps. This is usually suppressed in notation, and we will ignore it henceforth. For the remainder of this subsection only, let $J_n^{MC}(\cdot)$ explicitly refer to a simulation estimate of $J_n^*(\cdot)$. Then we note the following:

- We emphasized that $J_n^*(\cdot)$ is a sample statistic, i.e. it is nonrandom given the data. In contrast, $J_n^{MC}(\cdot)$ is random even given F_n because of the random selection of bootstrap samples. This is usually ignored because in principle, simulation error can be made arbitrarily small by increasing B.

[2] While in most applications, many different data vectors, e.g. permutations of the same vector, will give the same T_n, this effect is not nearly large enough to resolve the problem.

- An obvious question is what B is big enough in practice. You may get away with B on the order of a few hundred in exploratory analysis, if computation is truly costly, or if you are after bootstrap standard errors (which tend to converge much faster than quantiles or p-values). If you bootstrap a quantile, e.g. to get a critical value, and computation is not a big concern, a bootstrap size of $B = 10000$ (or maybe $B = 9999$ because the bootstrap quantile then coincides with a support point of the bootstrap distribution) would be reasonably expected.[3]

- For internal and exploratory analyses, it can be necessary to set B low and useful to make sure that random number seeds change. This will automatically alert you to too low B as you continue exploring. For replicable final versions of your analyses, choose B as large as is computationally reasonable and set specific seeds.

- Make sure you don't get confused by the multiple layers of randomness involved in the bootstrap. To reiterate: (i) The data W_i, and all sample statistics, are random in just the way they always were. (ii) The bootstrap analog W_i^*, and all test statistics based on it, are fictitious random variables whose distributions, e.g. $J_n^*(\cdot)$, are sample statistics, i.e. nonrandom given the data. (But the distributions *are* themselves random from a pre-sample point of view.) (iii) In practice, inference is based on simulated bootstrap objects like $J_n^{MC}(\cdot)$ that *are* random given the data, but we ignore this layer by presuming that B is large. Asymptotic results presented below are strictly speaking about the idealized bootstrap.

Here are some examples.

Example 1.1 *You observed three realizations* $(0, 1, 2)$ *of a random variable* X*, thus the sample average is* 1*. Then* F_n *is the uniform distribution on* $\{0, 1, 2\}$*. This distribution characterizes the bootstrap analog* X^**, which has expectation* $\mathbb{E}(X^*) = \overline{X} = 1$*. The complete bootstrap distribution of a size 3 resample* (X_1^*, X_2^*, X_3^*) *is uniform on* $\{0, 1, 2\}^3$*, i.e. it has* 27 *equally sized mass points. We can derive the bootstrap distribution of the sample average* $\overline{X}^*$*: Evaluating all possible bootstrap samples, it turns out that this average is supported on* $(0, 1/3, 2/3, 1, 4/3, 5/3, 2)$ *with probability masses* $(1/27, 3/27, 6/27, 7/27, 6/27, 3/27, 1/27)$.

Example 1.2 *Same as above, but you observed six realizations* $(0, 1, 2, 3, 4, 5)$*. Now* F_n *is the uniform distribution on* $\{0, 1, 2, 3, 4, 5\}$*. The bootstrap distribution of* $(X_1^*, X_2^*, X_3^*, X_4^*, X_5^*, X_6^*)$ *has* $6^6 = 46656$ *mass points. A complete bootstrap cannot be done manually, but implementing it on the computer would not be a problem.*

[3] For a rough intuition about simulation error, say you want to estimate the 95th quantile of T_n. Letting the true quantile be $c_{95\%}$, a ballpark estimate informed by properties of the binomial distribution is $\sqrt{B}\,(F_n^*(c_{95\%}) - 95\%) \approx N(0, .05)$, i.e. the standard deviation of the simulated probability is about $1/\sqrt{20B}$.

Example 1.3 *Same as above, but you have a realistic sample size. Now you have to do a Monte Carlo bootstrap.*

Example 1.4 ***Estimating a Proportion.***

We explain this example in some detail as we will return to it. The r.v. X is distributed Bernoulli with parameter $\pi = \Pr(X = 1) \in (0, 1)$. You observe a size n i.i.d. sample. The obvious estimator of π is $\hat{\pi} = \bar{X}$. Its exact sampling distribution is characterized by $n\bar{X}$ being Binomial with parameters (n, π). An empirical sample consists of n data points, of which $n\bar{X}$ are successes and $n(1 - \bar{X})$ are failures. The empirical distribution F_n therefore equals $F_n(x) = 1-\bar{X}+\bar{X}\mathbf{1}\{x = 1\}$, or in other words, the bootstrap variable X^* is distributed Bernoulli with parameter $\hat{\pi}$. A bootstrap sample $(X_1^*, \ldots, X_n^*)$ is i.i.d. from that distribution. Define the bootstrap estimator $\hat{\pi}^* = \bar{X}^* = \frac{1}{n}\sum_{i=1}^n X_i^*$, then $n\bar{X}^*$ is distributed Binomial with parameters $(n, \hat{\pi})$. In this particular case, the exact bootstrap distribution can be computed up to sample sizes where the error in Normal approximation is negligible. Also, in this case it is intuitively clear that, due to consistency of $\hat{\pi}$, the procedure "works," i.e. the bootstrap c.d.f. becomes similar to the true sampling c.d.f.

1.2 Basic Concepts in Bootstrap Inference

We next go over the most straightforward, but also widely used, examples of bootstrap based inference. Throughout this discussion, we'll assume that $\hat{\theta}$ is a plug-in estimator or more specifically, that the estimator equals the bootstrap population's true value.

Standard Errors As a warm-up, let's think through computation of bootstrap standard errors. Consider an estimator that is unbiased, at least to first order of approximation. (For any other estimator, forming confidence intervals by adding and subtracting standard errors is not motivated.) In particular, we have

$$SE^* = \left(\mathbb{E}^*(\hat{\theta}^* - \mathbb{E}^*\hat{\theta}^*)^2\right)^{1/2},$$

where the expectation $\mathbb{E}^*(\cdot)$ is with respect to the (bootstrap) distribution of $\hat{\theta}^*$. Note that in cases where we know that $\mathbb{E}^*\hat{\theta}^* = \hat{\theta}$, this simplifies to $\left(\mathbb{E}^*(\hat{\theta}^* - \hat{\theta})^2\right)^{1/2}$. In particular, this is the case when (i) the estimator is unbiased, (ii) the empirical estimate is the bootstrap population true value; both are true for $\theta = \mathbb{E}X$.

As before, this can be approximated to arbitrary degree of precision by choosing a high B in

$$SE^{MC} = \left(\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^b - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^b\right)^2\right)^{1/2},$$

which may simplify to

$$SE^{MC} = \left(\frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}^b - \hat{\theta})^2 \right)^{1/2},$$

where $\hat{\theta}^b$ is the b'th bootstrap realization of $\hat{\theta}^*$. We will ignore the difference between complete and Monte Carlo bootstrap henceforth. To repeat, the hard theoretical question is not whether SE^{MC} approximates SE^*, but whether SE^* estimates SE.

Many estimators are demonstrably $\sqrt{n}$-consistent and asymptotically normal but with hard-to-estimate asymptotic variances. (A simple example is the sample median or more generally quantile regression.) In such cases, it can be convenient to bootstrap their standard errors and then report Wald confidence intervals. On the other hand, simple bootstrap standard errors require that your statistic not only is asymptotically normal, but also has moments (because else, what are you estimating here?). This can fail in relevant cases, e.g. IV or TSLS estimators need not have moments. In such cases, asymptotic variances can typically be estimated by trimmed bootstrap variances; see Hansen for details.

Percentile Interval The bootstrap equal-tailed percentile confidence interval is

$$CI_{\alpha}^{perc-2} = \left[\hat{\theta} - q_n^*(1-\alpha/2), \hat{\theta} - q_n^*(\alpha/2) \right],$$

where q^* denotes bootstrap analog (hence the asterisk) quantiles of $(\hat{\theta} - \theta_0)$, i.e. $q^*(\alpha) = \inf\{t : \Pr(\hat{\theta}^* - \hat{\theta} \leq t) \geq \alpha\}$. To motivate it, let's first write out a so-called "oracle" confidence interval. This is the interval that we'd like to use if all population quantities were known. It should by construction attain a size of exactly 95%. (It is, of course, a pure thought device because if we actually knew all population quantities, we'd not need a confidence interval to begin with.) Letting $T_n = \hat{\theta} - \theta_0$ with exact quantile function q_n, this confidence interval would be

$$CI^{oracle} = \left[\hat{\theta} - q_n(1-\alpha/2), \hat{\theta} - q_n(\alpha/2) \right]$$

because

$$\begin{aligned} \Pr(\theta_0 \in CI^{oracle}) &= \Pr\left(\hat{\theta} - q_n(1-\alpha/2) \leq \theta_0 \leq \hat{\theta} - q_n(\alpha/2) \right) \\ &= \Pr\left(\hat{\theta} - q_n(\alpha/2) \leq \hat{\theta} - \theta_0 \leq q_n(1-\alpha/2) \right) \\ &= 1 - \alpha, \end{aligned}$$

where the last step assumes for simplicity that $\hat{\theta} - \theta_0$ is continuously distributed. The bootstrap percentile interval is a plug-in estimator of this object, i.e. q_n is replaced with its bootstrap analog q_n^*. The interval is exceedingly easy to compute:

Algorithm 1.1 *1. Generate B bootstrap realizations $\hat{\theta}^1, ..., \widehat{\theta}^B$. Let the vector $(\hat{\theta}^{[1]}, ..., \hat{\theta}^{[B]})$ collect them in increasing order.*

2. Up to integer constraints, the bootstrap percentile $(1-\alpha)$ confidence interval is

$$\left[2\hat{\theta} - \hat{\theta}^{[(1-\alpha/2)B]}, 2\hat{\theta} - \hat{\theta}^{[\alpha B/2]}\right].$$

Knowing α, you can of course choose B to avoid integer constraints. A popular trick is to use $B = 9999$ or similar, so the relevant quantiles of the bootstrap distribution are unique.

The construction contrasts with the historically oldest bootstrap percentile interval, which is

$$CI_\alpha^{perc-1} = \left[\hat{\theta} + q_n^*(\alpha/2), \hat{\theta} + q_n^*(1-\alpha/2)\right].$$

This confidence interval is even (marginally) easier to compute: In the preceding algorithm's notation, it is just $[\hat{\theta}^{[\alpha B/2]}, \hat{\theta}^{[(1-\alpha/2)B]}]$, i.e. we draw B bootstrap estimates, sort them, and report the numbers indexed $\alpha B/2$ and $B - \alpha B/2$. It is also translation invariant. However, it is awkwardly motivated. Being the bootstrap analog of

$$\left[\hat{\theta} + q_n(\alpha/2), \hat{\theta} + q_n(1-\alpha/2)\right],$$

at first glance there is no reason to expect that it will cover correctly. The two percentile intervals (asymptotically) coincide if the distribution of T_n is (asymptotically) symmetric around 0, so that $q_n(\alpha/2) \approx q_n(1-\alpha/2)$. This is why CI_α^{perc-1} does the job in many practical applications. Indeed, we will learn that in practice, the simple nonparametric bootstrap can only be used to approximate asymptotically normal test statistics anyway, and for those cases the Efron interval is very simple to compute, may be good enough, and has the advantage of being translation invariant.

Percentile-t Intervals These intervals are based on the idea that most confidence regions can be thought of as lower contour sets of sample test statistics. In other words, think about the problem of testing $H_0 : \theta = \theta^*$ and define your confidence region as nonrejection region of the test.

For simplicity, consider first the one-sided interval, thus we test $H_0 : \theta \leq \theta^*$ vs. $H_1 : \theta > \theta^*$ using test statistic $T_n = (\hat{\theta} - \theta^*)/SE(\hat{\theta})$. (We could also bootstrap the distribution of a non-studentized test statistic, but we will later encounter good reasons not to do so.) We want to reject the null if $T_n > c_{1-\alpha}$, where $c_{1-\alpha}$ is the $(1-\alpha)$-quantile of T_n, i.e. $c_{1-\alpha} = \inf\{c : \Pr(T_n \leq c) \geq 1-\alpha\}$, under the null. Exact computation of $c_{1-\alpha}$ is typically elusive in practice, but we can again think of $c_{1-\alpha}$ as an "oracle" quantity that we estimate by its bootstrap analog. Thus, define $c_{1-\alpha}^* = \inf\{c : \Pr(T_n^* \leq c) = 1-\alpha\}$, where $T_n^* = (\hat{\theta}^* - \hat{\theta})/SE^*(\hat{\theta})$. This gives rise to the one-sided percentile-t interval. The two-sided $(1-\alpha)$-interval is the intersection of the one-sided $(1-\alpha/2)$ intervals from above and below. It is *equal-tailed* because, up to approximation error, the same noncoverage probability (of $\alpha/2$) is incurred

at either end of the interval. The interval is *not* in general symmetric around $\hat{\theta}$. The interval is also difficult to generalize to parameters that are not scalars.

The latter problem is avoided by the *symmetric* percentile-t interval. In the scalar case, this interval is based on $|T_n|$, thus it cannot be used for one-sided testing. On the other hand, it naturally generalizes to higher-dimensional confidence regions, e.g. to the equivalent of confidence ellipsoids in $\mathbb{R}^2$, by thinking in terms of a quadratic form that generalizes T_n^2. The idea is to test $H_0 : \theta = \theta^*$ vs. $H_1 : \theta \neq \theta^*$ and to (ideally) reject H_0 if $T_n > c_{1-\alpha}$, where $T_n = (\hat{\theta} - \theta^*)' \hat{\boldsymbol{V}}^{-1} (\hat{\theta} - \theta^*)$ and $c_{1-\alpha} = \inf\{c : \Pr(T_n \leq c) = 1 - \alpha\}$ under the null. Again, $c_{1-\alpha}$ is replaced with its bootstrap analog $c^*_{1-\alpha}$. In sum, we reject if $T_n > c^*_{1-\alpha}$, where

$$c^*_{1-\alpha} = \inf\{c : \Pr(T^*_n \leq c) = 1 - \alpha\}, \qquad T^*_n = (\hat{\theta}^* - \hat{\theta})' (\hat{\boldsymbol{V}}^*)^{-1} (\hat{\theta}^* - \hat{\theta}).$$

Note in particular that T^*_n evaluates distance from the estimated value $\hat{\theta}$ and not from the hypothesized value θ^*. This is because the former is the true parameter value in the bootstrap population.

1.3 Summary

The above examples illustrate a good method for constructing a bootstrap test or confidence interval: First, imagine that you know the population distribution of observables and therefore the exact distribution, at sample size n, of any estimators and test statistics. Imagine further that you were charged with constructing a confidence region, i.e. a data-dependent set that covers θ_0 with probability $1 - \alpha$. With the full "oracle" knowledge that you imagine to have, you should typically be able to construct an exact confidence region, i.e. a set whose coverage probability is exactly $1 - \alpha$. Next, replace all unknown population quantities in the confidence region with bootstrap analogs, and similarly for tests.

More generally, your bootstrap constructions should aim to mimic as closely as possible the true sampling process. If your data are a clustered sample (as is true for many surveys), your bootstrap procedure should resample clusters. If you have time series dependence, the bootstrap should mimic that. (In practice, this is done by the "block bootstrap" which resamples entire pieces of your time series at a time.) If you test nulls, remember that you want to simulate the behavior of your test statistic on the null, so you better ensure the null is true in the bootstrap population. (With very simple nulls, this is automatic by using $\hat{\theta}$ for the hypothesized value. In more complicated cases, this may require to manipulate the empirical distribution so that it fulfills the null.) And so on and so forth, though details are far beyond the scope of this lecture.

Is the Bootstrap Only Used for Inference? No. Under certain conditions, the bootstrap distribution of an estimator estimates its sampling distribution so precisely that bias in the estimator can be diagnosed. In principle, this can be used to debias the estimator. Also, in some contexts, notably

Machine Learning, it may be beneficial to replace estimators with their own bootstrap averages; this is called bootstrap aggregating or *bagging*. However, bootstrap inference as explained above is the by far most frequent use of the bootstrap in econometrics and also has by far the easiest theoretical justification, and we therefore focus on it.

2 Asymptotic Theory for the Bootstrap

Other than getting the implementation wrong – frequently by confusing population quantities and bootstrap analogs –, the most frequent mistake of practitioners us to trust that the bootstrap "always" works, even where relatively simple (of the sort we did in this lecture) asymptotic approximation theory fails (and is not merely intractable).

In fact, if you know that asymptotic approximation would fail, notably due to discontinuity of limit distributions as functions of parameters, you should assume that the bootstrap also fails. For a simple intuition, go back to the very beginning of the previous chapter. Would you expect plug-in estimation to work if g is not continuous?

An additional, interesting fact is that, if limit distributions do not depend on underlying parameters – as is the case with most test statistics considered so far in this lecture! – the bootstrap may be better than asymptotic approximation. This phenomenon is analyzed in the literature on *bootstrap refinement*. We won't have time for that but see the provided readings.

2.1 Consistency of Bootstrap Inference in a Simple Example

We first reconsider Example 1.4, i.e. estimation of a binomial proportion. In this example, the step-by-step algorithm for computing a (nonparametric, nonstudentized, percentile) MC bootstrap confidence interval is as follows.

Algorithm 2.1 ***Simple Bootstrap Inference for a Proportion***

1. *Generate B bootstrap resamples $(X_1^b, \dots, X_n^b), b = 1, \dots, B$, by resampling the empirical data with replacement.*

2. *Estimate $J_n(t) = \Pr(\bar{X}_n - \pi_0 \leq t)$ by $J_n^*(t) = F_n(\bar{X}_n^b - \bar{X}_n \leq t) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\{\bar{X}^b - \bar{X} \leq t\}$. Here, $\bar{X}^b = \frac{1}{n}\sum_{i=1}^{n} X_i^b$. Note that we really need only $q_n^*(\alpha/2)$ and $q_n^*(1-\alpha/2)$, where q_n^* is the quantile function corresponding to J_n^*. In practice, they can be estimated by ordering the bootstrap realizations of $(\bar{X}^b - \bar{X})$ and reading out the $aB/2$ and $(1-\alpha/2)B$ position (up to integer issues).*

3. The bootstrap CI is $[\bar{X} - q_n^*(1-\alpha/2), \bar{X} - q_n^*(\alpha/2)]$.

I describe an MC bootstrap here because that is what we will have to do in realistic applications. In the specific example, as discussed earlier, a complete bootstrap may actually be feasible.

Also, in writing down the algorithm, I followed a typical bootstrap convention and avoided scaling by $\sqrt{n}$. This does not matter for implementation and is often dropped from code (but be careful to be consistent there!). To show that this bootstrap construction works, however, we want to look at the scaled test statistic $T_n = \sqrt{n}(\bar{X} - \pi_0)$ with distribution $J_n(t) = \Pr(\sqrt{n}(\bar{X} - \pi_0) \leq t)$. If $\tilde{q}_n(\cdot)$ is the quantile function corresponding to $J_n(t)$ and $\tilde{q}_n^*(\cdot)$ its bootstrap analog, then $q_n^*(\cdot) = \tilde{q}_n^*(\cdot)/\sqrt{n}$, so the algorithm could have equivalently be described in terms of estimating J_n by J_n^*.

In the example, that the bootstrap r.v. $\sqrt{n}(\bar{X}^b - \bar{X})$ is likely to be distributed similarly to $\sqrt{n}(\bar{X} - \pi_0)$ is intuitively clear and could also be verified with ad hoc arguments using that the binomial distribution is very well understood. We will now formalize "likely to be distributed similarly" and then provide a proof that is more instructive and generalizable than the ad hoc argument.

Theorem 2.2 *The above estimator of* J_n *is consistent:*

$$\sup_{t \in \mathbb{R}} |J_n^*(t) - J_n(t)| \overset{a.s.}{\to} 0 \quad as \quad n \to \infty.$$

Proof. We prove this by invoking the Berry-Esseen Theorem: For i.i.d. realizations of a r.v. Y, if $\mathbb{E}(Y, Y^2, |Y|^3) = (0, \sigma^2, \gamma)$ and G_n is the c.d.f. of $\sqrt{n}\bar{Y}_n/\sigma$, then

$$\sup_{r \in \mathbb{R}} |G_n(r) - \Phi(r)| \leq \frac{C\gamma}{\sigma^3\sqrt{n}}$$

for a universal constant C. The important thing is that this *uniformly* bounds the rate at which a Central Limit Theorem "kicks in" depending only on the values of a few moments (these values have to be finite).

Next, write

$$J_n(t) = \Pr(\sqrt{n}(\bar{X}_n - \pi_0) \leq t) = \Pr\left(\sqrt{\tfrac{n}{\pi_0(1-\pi_0)}}(\bar{X}_n - \pi_0) \leq t/\sqrt{\pi_0(1-\pi_0)}\right)$$

Define $Y = X - \pi_0$, then $\mathbb{E}(Y, Y^2, |Y|_i^3) = (0, \pi_0(1-\pi_0), \pi_0(1-\pi_0)^3 + (1-\pi_0)\pi_0^3)$. The r.h. probability in the above display therefore is the c.d.f. of $\sqrt{n}\bar{Y}_n/\sigma$ evaluated at $r = t/\sqrt{\pi_0(1-\pi_0)}$. Noting that $\pi_0(1-\pi_0)^3 + (1-\pi_0)\pi_0^3 < 1/2$, we can invoke Berry-Esseen to write

$$\sup_{t \in \mathbb{R}} \left| J_n(t) - \Phi\left(t/\sqrt{\pi_0(1-\pi_0)}\right)\right| \leq \frac{C}{(\pi_0(1-\pi_0))^{3/2}\sqrt{n}}$$

and similarly that

$$\sup_{t \in \mathbb{R}} \left| J_n^*(t) - \Phi\left(t/\sqrt{\bar{X}_n(1-\bar{X}_n)}\right)\right| \leq \frac{C}{(\bar{X}_n(1-\bar{X}_n))^{3/2}\sqrt{n}}.$$

We wrap up by writing

$$\begin{aligned}
& \sup_{t\in\boldsymbol{R}} |J_n(t) - J_n^*(t)| \\
= & \sup_{t\in\boldsymbol{R}} \left| J_n(t) - \Phi\left(\frac{t}{\sqrt{\pi_0(1-\pi_0)}}\right) + \Phi\left(\frac{t}{\sqrt{\pi_0(1-\pi_0)}}\right) - \Phi\left(\frac{t}{\sqrt{\bar{X}_n(1-\bar{X}_n)}}\right) + \Phi\left(\frac{t}{\sqrt{\bar{X}_n(1-\bar{X}_n)}}\right) - J_n^*(t) \right| \\
\leq & \sup_{t\in\boldsymbol{R}} \left| J_n(t) - \Phi\left(\frac{t}{\sqrt{\pi_0(1-\pi_0)}}\right) \right| + \sup_{t\in\mathbb{R}} \left| \Phi\left(\frac{t}{\sqrt{\pi_0(1-\pi_0)}}\right) - \Phi\left(\frac{t}{\sqrt{\bar{X}_n(1-\bar{X}_n)}}\right) \right| + \sup_{t\in\mathbb{R}} \left| \Phi\left(\frac{t}{\sqrt{\bar{X}_n(1-\bar{X}_n)}}\right) - J_n^*(t) \right|.
\end{aligned}$$

Now, first think of $\bar{X}_n$ as a nonstochastic sequence converging to π_0, then the first and last supremum in the last line above vanish by the Berry-Esseen theorem; the middle one vanishes by elementary arguments. Of course, $\bar{X}_n$ is in fact random, but because $\bar{X}_n \overset{a.s.}{\to} \pi_0$, the argument applies with probability 1. (See Hansen's discussion of "almost sure convergence in probability.") ■

Corollary 2.3 *The confidence intervals defined earlier achieve asymptotic size control:*

$$\Pr\big(\pi_0 \in [\bar{X}_n - q_n^*(1-\alpha/2), \bar{X}_n - q_n^*(\alpha/2)]\big) \to 1-\alpha$$
$$\Pr\big(\pi_0 \in [\bar{X}_n + q_n^*(\alpha/2), \bar{X}_n + q_n^*(1-\alpha/2)]\big) \to 1-\alpha.$$

The proof effectively went through an asymptotic approximation: We showed that the (nonstochastic) sequence of distributions of $\sqrt{n}(\bar{X}_n - \pi_0)$ and the (stochastic) sequence of distributions of $\sqrt{n}(\bar{X}_n^* - \bar{X}_n)$ converge to the same limit. This foreshadows a typical feature of bootstrap consistency proofs: Despite the fact that we estimate J_n by J_n^* and therefore do not seem to "send n to ∞," proofs of the method usually do just that. This is also a hint that, if CLT-based approximation fundamentally fails, bootstrap inference might be affected too.

As a final note, this was a relatively involved proof for an intuitively straightforward application of the bootstrap. While the proof is a bit more general than the example (it actually applies to means and sample averages of well-behaved r.v.'s), this is a first pointer that proofs of results regarding the bootstrap quickly escalate in complexity.

2.2 General Results

We next consider general conditions under which the bootstrap works. Consider the following table.

	actual sample size	**asymptotic limit**
bootstrap	$J_n(t, \hat{F})$	$J_\infty(t, \hat{F})$
population	$J_n(t, F)$	$J_\infty(t, F)$

Here, $J_\infty(t, F) = \lim_{n\to\infty} J_n(t, F)$. The true sampling distribution of interest is on the bottom left of this table. Its bootstrap estimator is on the top left. Asymptotic approximations roughly correspond

to the top right entry: We usually show that we know the limiting distribution of an object up to some parameters (e.g., an asymptotic variance) which we can estimate. This is like the top right entry, though $\hat{F}$ is not necessarily the nonparametric estimator.

Our hope is that $J_n(t, \hat{F}) - J_n(t, F) = o_P(1)$. However, whenever this can be shown, the typical chain of proof is $J_n(t, \hat{F}) - J_\infty(t, \hat{F}) = o_P(1)$, $J_\infty(t, \hat{F}) - J_\infty(t, F) = o_P(1)$, $J_\infty(t, F) - J_n(t, F) = o_P(1)$. Therefore, despite the fact that we seem to estimate a finite sample distribution by a finite sample distribution, *the bootstrap is an asymptotic approximation.* That said, there is an intuition that $J_n(t, \hat{F})$ might pick up additional features of the sampling distribution, e.g. skewness, even if the limiting distribution is normal. Could this imply that the bootstrap approximation is better than the asymptotic one? The answer is 'yes' in especially well-behaved cases, namely if the top right and bottom right entries of the table are in fact the same; else, this additional approximation step invalidates the idea.

To formalize this, it will be important to distinguish the following three cases.

1. A statistic T_n is an *asymptotic pivot*, meaning that $J_\infty(t, F)$ is constant in F. Classic examples are all statistics that you can show to be asymptotically distributed as $N(0, 1)$.[4]
2. $J_\infty(t, F)$ is continuous in F.
3. $J_\infty(t, F)$ is not continuous in F.

The take-home message is that the bootstrap improves on asymptotic approximation in case 1, is consistent in 2, and fails (at least without further modification) in 3. Not coincidentally, asymptotic approximation may also fail in 3.

2.2.1 Consistency

For linear functionals of F_0, a classic result due to Mammen (1992) specifies *necessary and sufficient* conditions for the simple nonparametric bootstrap (and also close variations of it, though we do not discuss that here) to consistently estimate the sampling distribution of a statistic. In a nutshell, these state that such statistics can be boostrapped if and only if they are subject to a Central Limit Theorem. Thus, the simple nonparametric bootstrap does *not* have a fundamental advantage of general validity over CLT-based asymptotic approximation. Indeed, this powerful theorem predicts some failures of the bootstrap that we will discuss later.

[4]To motivate the term asymptotic pivot, note the following definition: A statistic T_n is a *pivot* if $J_n(t, F)$ is constant in F. Of course, this is possible only with restrictions on F. Classic examples are the t- and F-test for OLS under an assumption of normal errors and the Kolmogorov-Smirnov statistic for continuous F.

Theorem 2.4 ***Necessary and Sufficient Conditions for Bootstrap Consistency***

Assume i.i.d. sampling. Fix sequences of functions $\{g_n\}$ *and numbers* $\{t_n, \sigma_n\}$. *Write* $\bar{g}_n = \frac{1}{n}\sum_i g_n(W_i)$ *and* $T_n = (\bar{g}_n - t_n)/\sigma_n$ *with bootstrap analogs* $\bar{g}_n^* = \frac{1}{n}\sum_i g_n(W_i^*)$ *and* $T_n^* = (\bar{g}_n^* - \bar{g}_n)/\sigma_n$. *Then*

$$\lim_{n\to\infty} \Pr(\sup_t |J_n(t, F_n) - J_\infty(t, F_0)| > \varepsilon) = 0$$

if, and only if, $T_n \xrightarrow{d} N(0,1)$.

Example 2.1 ***Sample Mean***

Let the r.v. X *have finite expectation* μ *and variance* σ^2. *Let* $T_n = (\bar{X}_n - \mu)/\sigma$ *(i.e.,* (g_n, t_n, σ_n) *are constant sequences) and* $T_n^* = (\bar{X}_n^* - \bar{X}_n)/\sigma$. *Then Theorem 2.4 applies and implies consistency of the bootstrap.*

Note that this example is about the non-studentized sample mean. We standardized T_n and T_n^* only so that Theorem 2.4 applies as written; the conclusion extends to the same quantities multiplied by σ. The result is also true if the test statistic gets studentized, i.e. divided by a standard error.

Example 2.2 ***Sample Median***

Let X *have density* $f(\cdot)$ *that is strictly positive at the median* $m \equiv \text{med}(X_i)$. *The plug-in estimator of* m *is* $\hat{m} = q_n(1/2)$, *or more intuitively, the* $(n/2+1)$*-largest sample realization if* n *is even and the* $(n+1)/2$*-largest realization otherwise. It can be shown that* $\sqrt{n}(\hat{m} - m) \xrightarrow{d} N\big((0, 1/(4f(m)^2)\big)$.

Explicit estimation of this asymptotic variance would require estimation of a density. There are tools for that, but it is not recommended for this application. Instead, we could: (i) use simple nonparametric bootstrap confidence intervals, (ii) use bootstrap standard errors. Note that (i) is not justified by Theorem 2.4, illustrating that that result's strength owes to a very structured setting. However, this bootstrap has been formally justified as long as f is continuous around m and $f(m) > 0$ (Bickel and Freedman, 1981). The more common confidence interval is actually (ii), which usually computes rapidly, though its validity technically needs the additional condition that $\mathbb{E}|X|^\alpha < \infty$ for some $\alpha > 0$ (Ghosh, Parr, Singh, and Babu, 1984).

2.2.2 Limitations of the Bootstrap

We will now consider some limitations of the bootstrap, specifically by examining a few salient examples where it fails. The examples are related to important sources of bootstrap failure: discontinuity of J_∞, nonstandard convergence rate of an estimator, and nonnormal limiting distributions.

One important class of examples are parameter-on-the-boundary problems. Recall that standard asymptotic approximations do not apply in such cases because they assume interiority of θ_0, or more

generally smoothness of the estimation problem. Furthermore, in the specific example below, it is easy to see that normal approximation fails: In the critical case of $\mu = 0$, the test statistic's limiting distribution has a mass point, and convergence to normal is not uniform as $\mu \to 0$. So it would be great if the bootstrap worked because this would make it more general than asymptotic approximation; but for that same reason there are also red flags.

Example 2.3 ***Sample Mean on the Boundary (Andrews 2000)***

Revisit the example of a sample mean from i.i.d. data. For simplicity, let $F_0 = N(\mu_0, 1)$. *Suppose the researcher knows that* $\mu_0 \geq 0$. *The obvious (and, for the record, ML) estimator is* $\hat{\mu}_n = \max\{\bar{X}_n, 0\}$ *with bootstrap analog* $\hat{\mu}_n^* = \max\{\bar{X}_n^*, 0\}$.

For any fixed $\mu > 0$, this problem becomes similar to the simple sample mean as $n \to \infty$ and Theorem 2.4 will apply. However, say $\mu_0 = 0$, thus $\sqrt{n}(\hat{\mu} - \mu_0)$ is exactly distributed as $\max\{Z, 0\}$, where $Z \sim N(0, 1)$. The bootstrap will be inconsistent in this case. Andrews shows this through the following algebra: Fix any $c > 0$, then

$$
\begin{aligned}
& \Pr\left(\sqrt{n}\left(\hat{\mu}^* - \hat{\mu}\right) \leq t | \sqrt{n}\bar{X}_n \geq c\right) \\
= \; & \Pr\left(\sqrt{n}\left(\max\{\bar{X}_n^*, 0\} - \max\{\bar{X}_n, 0\}\right) \leq t | \sqrt{n}\bar{X}_n \geq c\right) \\
= \; & \Pr\left(\sqrt{n}\left(\max\{\bar{X}_n^* - \bar{X}_n, -\bar{X}_n\}\right) \leq t | \sqrt{n}\bar{X}_n \geq c\right) \quad \text{using } \bar{X}_n \geq 0 \Rightarrow \max\{\bar{X}_n, 0\} = \bar{X}_n \\
= \; & \Pr\left(\max\{\sqrt{n}\left(\bar{X}_n^* - \bar{X}_n\right), -\sqrt{n}\bar{X}_n\} \leq t | \sqrt{n}\bar{X}_n \geq c\right) \\
\geq \; & \Pr\left(\max\{\sqrt{n}\left(\bar{X}_n^* - \bar{X}_n\right), -c\} \leq t | \sqrt{n}\bar{X}_n \geq c\right) \quad \text{using } -\sqrt{n}\bar{X}_n \leq -c \\
\to \; & \Pr\left(\max\{Z, -c\} \leq t | \sqrt{n}\bar{X}_n \geq c\right) \quad \text{using } \sqrt{n}(\bar{X}_n^* - \bar{X}_n) \xrightarrow{d} N(0,1) \text{ by Berry-Esseen as above} \\
\geq \; & \Pr\left(\max\{Z, 0\} \leq t | \sqrt{n}\bar{X}_n \geq c\right).
\end{aligned}
$$

The last inequality is strict if $-c < t < 0$, and for any given c, $\Pr(\sqrt{n}\bar{X}_n \geq c) \to \Phi(-c)$ because we assumed $\mu_0 = 0$. Therefore, the distribution of $\sqrt{n}(\hat{\mu} - \mu_0)$ is not correctly estimated, and a percentile t-interval will be invalid.

Alternatively, let's think through what the bootstrap distribution will be. By Berry-Esseen as shown above, the (bootstrap) distribution of $\sqrt{n}(\bar{X}_n^* - \bar{X}_n)$ does consistently estimate the one of $\sqrt{n}(\bar{X}_n - \mathbb{E}X)$ (i.e., $N(0, 1)$). However,

$$
\begin{aligned}
\sqrt{n}(\hat{\mu}^* - \hat{\mu}) \quad &= \quad \sqrt{n}(\max\{\bar{X}_n^*, 0\} - \max\{\bar{X}_n, 0\}) \\
&\underset{\text{if } \bar{X}_n \geq 0}{=} \quad \sqrt{n}\max\{\bar{X}_n^* - \bar{X}_n, -\bar{X}_n\} \\
&= \quad \max\{Z, -\sqrt{n}\bar{X}_n\} + o_P(1)
\end{aligned}
$$

and

$$\begin{aligned} \dots \quad \underset{\text{if } \bar{X}_n \leq 0}{=} \quad & \sqrt{n}\max\{\bar{X}_n^*, 0\} \\ = \quad & \sqrt{n}\max\{\bar{X}_n^* - \bar{X}_n + \bar{X}_n, 0\} \\ = \quad & \max\{Z + \sqrt{n}\bar{X}_n, 0\} + o_P(1). \end{aligned}$$

Either approximation differs from the correct limit distribution $\max\{Z, 0\}$ through a term involving $\sqrt{n}\bar{X}_n$, a quantity that does not vanish and whose marginal distributions is itself (asymptotically) standard normal. Thus the bootstrap distribution does not converge to *any* nonstochastic limit.

In the specific example, the problem is easy to spot. The deeper point, however, is that the bootstrap fails if the true parameter value is on the boundary of parameter space, and this, as well as ad hoc fixes, will not always be so obvious as in this example. For example, a generalization of this problem occurs if the population distribution occurs at a nondifferentiability of the population objective problem's value function. This could be literally a boundary point of parameter space but also a "switching point" of KKT conditions, i.e. a corner of a feasible set. In short, we should bootstrap m-estimators only if we know that the population problem has a smooth maximum. Of course, in any such setting we could in principle (though maybe not in practice) use asymptotic approximation as well.

The example also relates to our previous insights because $J_\infty(\cdot)$ is discontinuous in F. To see this, note that $\sqrt{n}(\hat{\mu} - \mu_0)$ converges to Z for any $\mu_0 > 0$ but to $\max\{Z, 0\}$ if $\mu_0 = 0$. In contrast, $J_n(\cdot)$ is continuous in F. These claims are consistent because convergence of J_n to J_∞ is not uniform over F. Thus, for any n, there are true parameter values – notably at and near 0 – for which J_n does not look similar to J_∞. As a result, $J_n(t, F_n) \to J_\infty(t, F_0)$, though true for constant sequences $\{F_n\}$ (with limit $F_0 = F_n$), need not hold over nonconstant sequences $\{F_n\}$. Theorem 2.4 does not apply because $\hat{\mu}$ cannot be expressed as sample average, but it is suggestive of inconsistency at $\mu = 0$ because T_n does not converge to $N(0, 1)$.

Example 2.4 ***Infinite Variance (Athreya 1987)***

Consider again the same example of a sample mean from i.i.d. data. The assumption that X has a finite variance is essential for the result. Indeed, if the variance of X is infinite, $J_n^*(t)$ converges to a nondegenerate random variable, hence not to $J_\infty(t)$.

The proof of this example is the centerpiece of an *Annals of Statistics* paper, so we'll take it on faith, though we note that Mammen's result reported earlier (but historically subsequent) strongly suggests it.

Example 2.5 ***Binomial Proportion near Zero***

Return to Example 1.4 but say the true parameter value drifts toward zero: $\pi_n = \gamma/n$ for some $\gamma \in (0, 1]$. Then it is easily verified that $n\bar{X}_n$ converges in distribution to a Poisson distribution with parameter γ. Hence, $n(\hat{\pi}_n - \pi_n)$ converges to the same distribution translated by $(-\gamma)$ and therefore centered at 0. The distribution of $n\hat{\pi}_n^*$ is binomial with random parameters $(n, n\hat{\pi}_n)$. Recalling that $n\hat{\pi}_n$ remains stochastically bounded as $n \to \infty$, the distribution of $n\hat{\pi}_n^*$ for large n is well approximated by a Poisson distribution with random parameter $n\hat{\pi}_n$. As $n\hat{\pi}_n$ itself is well approximated by a Poisson distribution with parameter γ, the bootstrap distribution does not converge to any limit and in particular not to the estimand.

This example is important because estimation of small probabilities is a relevant problem. For example, rare events are important in economics, finance, and medical and biostatistics, and propensity scores near 0 or 1 are a frequent issue in estimation of treatment effects.

We note that, from arguments made in the example, $T_n = n(\hat{\pi}_n - \pi_n)/\sqrt{\gamma}$ converges to a r.v. with unit variance, but this r.v. is not normal. Because Theorem 2.4 is otherwise applicable, it establishes the bootstrap failure that we also directly verified. Indeed, that the bootstrap distribution fails to converge anywhere would have been predictable from a close look at Mammen's proof.[5]

In this example, we can furthermore verify that percentile confidence intervals will typically undercover. The empirical distribution is degenerate at 0 with probability $(1 - \gamma/n)^n \to e^{-\gamma}$. Whenever that happens, the bootstrap percentile confidence interval is the degenerate interval $[0, 0]$. Hence, for any nominal size, the true size of this confidence interval converges to at most $1 - e^{-\gamma}$. For example, $1 - e^{-1} \approx .63$, so if $\gamma = 1$, the interval undercovers compared to any conventional level.

Example 2.6 ***Absolute Value of Mean.***

Consider an i.i.d. sample of X with finite expected value $\mu_0 = \mathbb{E}X$ and variance σ^2. We are interested in $\theta_0 = |\mu_0|$, which we estimate by $\hat{\theta} = |\bar{X}|$. The question is if bootstrap inference works; in particular, if we can use $J_n^*(t) = \Pr(\sqrt{n}\hat{\theta}^* - \hat{\theta} \leq t)$ as estimator of $J_n(t) = \Pr(\sqrt{n}\hat{\theta} - \theta_0 \leq t)$.

We first note the true limiting distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$: it is $N(0, \sigma^2)$ if $\theta_0 \neq 0$ but it is the distribution of $|Z|$, where $Z \sim N(0, \sigma^2)$, if $\theta_0 = 0$. This discontinuity, and also the nonnormality of the limiting distribution at θ_0, are clear signs that something might go wrong at that parameter value (which is furthermore on the boundary of the relevant parameter space). We also note and use freely that by the Berry-Esseen theorem, $\sqrt{n}(\hat{\mu}^* - \hat{\mu}) \xrightarrow{d} N(0, \sigma^2)$.

The bootstrap distribution $J_n^*(\cdot)$ is the distribution of

$$\sqrt{n}(\hat{\theta}^* - \hat{\theta}) = \sqrt{n}(\max\{\hat{\mu}^*, -\hat{\mu}^*\} - \max\{\hat{\mu}, -\hat{\mu}\}).$$

[5] An important step in that proof can be roughly verbalized as follows: If the bootstrap converges anywhere, then no single observation is influential in the limit, and then a CLT applies.

Suppose first $\hat{\mu} > 0$. Then we can write

$$\begin{aligned}
& \sqrt{n}(\max\{\hat{\mu}^*, -\hat{\mu}^*\} - \max\{\hat{\mu}, -\hat{\mu}\}) \\
= \; & \sqrt{n}(\max\{\hat{\mu}^*, -\hat{\mu}^*\} - \hat{\mu}) \\
= \; & \sqrt{n}(\max\{\hat{\mu}^* - \hat{\mu}, -\hat{\mu}^* - \hat{\mu}\}) \\
= \; & \sqrt{n}(\max\{\hat{\mu}^* - \hat{\mu}, -2\hat{\mu} - (\hat{\mu}^* - \hat{\mu})\}) \\
= \; & \max\{\sqrt{n}(\hat{\mu}^* - \hat{\mu}), -2\sqrt{n}\hat{\mu} - \sqrt{n}(\hat{\mu}^* - \hat{\mu}))\} \\
\xrightarrow{d} \; & \max\{Z, -2\sqrt{n}\hat{\mu} - z_i\} = \max\{Z, -2\sqrt{n}\hat{\theta} - Z\},
\end{aligned}$$

For fixed positive value of μ, this case is the relevant one with probability approaching 1 and furthermore the r.h. argument of the max will diverge to $-\infty$, so we are left (with high probability) with Z as desired. Similarly, for $\hat{\mu} < 0$, we get

$$\begin{aligned}
& \sqrt{n}(\max\{\hat{\mu}^*, -\hat{\mu}^*\} - \max\{\hat{\mu}, -\hat{\mu}\}) \\
= \; & \sqrt{n}(\max\{\hat{\mu}^*, -\hat{\mu}^*\} + \hat{\mu}) \\
\xrightarrow{d} \; & \max\{-Z, 2\sqrt{n}\hat{\mu} + Z\} = \max\{Z, -2\sqrt{n}\hat{\theta} - Z\},
\end{aligned}$$

where we used that in this case, $\hat{\theta} = -\hat{\mu}$ and that the distribution of Z is symmetric, so it does not change the expression to replace Z with $-Z$.

We now see that if $\mu_0 \neq 0$, we are asymptotically in exactly one of the two cases and the limiting distribution is just the one of Z. However, if $\mu = 0$, then $\sqrt{n}\hat{\theta}$ converges not to any number but is itself asymptotically distributed as $|Z|$. The max operators in the above expression then remain relevant, and the bootstrap distribution does not converge anywhere; also, its support generally includes some negative numbers, which is obviously wrong as $\hat{\theta} - \theta \geq 0$ in this case.

Example 2.7 ***Order Statistic.***

Let us reconsider the maximum of a uniform distribution. Thus, X is distributed $U[0, \alpha_0]$, and we want to estimate α_0. Recall that the obvious estimator is the sample maximum, but also that the usual $\sqrt{n}$-asymptotics for m-estimation do not apply due to discontinuity of the objective function. Indeed, the true rate of convergence is n, that is, the estimator is superconsistent, and in particular we have

$$\begin{aligned}
\Pr(n(\hat{\alpha} - \alpha_0) \leq t) &= \Pr\big(\max_{i=1,\ldots,n}\{X_i\} \leq \alpha_0 + t/n\big) = \Pr\big(X_1 \leq \alpha_0 + t/n, \ldots, X_n \leq \alpha_0 + t/n\big) \\
&= \Pr\big(X_1/\alpha_0 \leq 1 + t/(n\alpha_0), \ldots, X_n/\alpha_0 \leq 1 + t/(n\alpha_0)\big) = (1 + t/(n\alpha_0))^n \to e^{t/\alpha_0}.
\end{aligned}$$

Could be bootstrap this? Informally, both superconsistency and failure of asymptotic normality

are red flags. To see that the bootstrap fails, write

$$\begin{aligned}\Pr(n(\hat{\alpha}^* - \hat{\alpha}) = 0) = \Pr\left(\hat{\alpha}^* = \hat{\alpha}\right) = \Pr\Big(\max_{i=1,\ldots,n}\{X_i^*\} = \max_{i=1,\ldots,n}\{X_i\}\Big) \\ = 1 - \Pr\Big(\max_{i=1,\ldots,n}\{X_i^*\} < \max_{i=1,\ldots,n}\{X_i\}\Big) = 1 - \Pr\Big(X_i^* < \max_{i=1,\ldots,n}\{X_i\}, \ldots, X_n^* < \max_{i=1,\ldots,n}\{X_i\}\Big) \\ = 1 - (1 - 1/n)^n \to 1 - e^{-1} \approx 0.63.\end{aligned}$$

So this distribution has a (rather large) mass point at 0 including in the limit, whereas the true asymptotic distribution is continuous. Indeed, it is a "tedious but trivial" exercise to characterize the entire random c.d.f. of $n(\hat{\alpha}^* - \hat{\alpha})$. This distribution is supported on $n(X_1 - \max_{i=1,\ldots,n}\{X_i\}, \ldots, X_n - \max_{i=1,\ldots,n}\{X_i\})$. The highest of these mass points is at 0 and has mass $1 - (1 - 1/n)^n$ as per our algebra, the next one has mass $(1 - 1/n)^n \times (1 - 1/(n-1))^n$, and the lowest one has mass n^{-n}. While these numbers are not random, the location of the mass points is. Once again, the bootstrap distribution therefore does not converge to any limit.

This example is more removed from the above theorem because that result's assumptions obviously do not apply. It is, however, practically important and also an example of normal approximation failure, so that it would be very nice for the simple bootstrap to work. Unfortunately, we again get bootstrap nonconvergence, illustrating that higher generality compared to normal approximation is *not* an advantage of the bootstrap.

Example 2.8 ***(Smoothed) Maximum Score***

Let's reconsider the maximum score estimator. Analysis of the bootstrap for this example is very involved. It turns out that the simple nonparametric bootstrap is inconsistent here (Abrevaya and Huang, 2005) and fails to converge (contrary to a claim in the paper just cited). Valid inference for this estimator using modified bootstrap techniques is the subject of current research.

The Smoothed maximum Score estimator is much better behaved. Not only can it be bootstrapped; Horowitz (2002) shows that for test statistics based on a studentized estimator, the bootstrap achieves asymptotic refinement.