Bootstrap, Chapter 8

Elizabeth Purdom

This document has last been compiled on Oct 02, 2024.

Contents

We have been focusing on non-parametric estimators based on the plug-in principle. We discussed CI for $\hat{F}_n$, but it is more common to have a specific parameter of F that we are interested in, $\theta = T(F)$. How can we get a CI for a plug-in estimator $T(\hat{F}_n)$?

Let's take the simplest example when we have asymptotic normality of $T(\hat{F}_n)$. In principle this allows us to form a normal-based approximate $1-\alpha$ confidence interval for $T(F)$ of

$$C_n = T(\hat{F}_n) \pm z_{\alpha/2}\widehat{se}(T(\hat{F}_n))$$

where

$$se(T(\hat{F}_n)) = \sqrt{var_F(T(\hat{F}_n)},$$

i.e. the unconditional variance of $T(\hat{F}_n)$.

However, generically we won't necessarily know $se(T(\hat{F}_n))$ except in simple cases, much less be able to estimate it.

Example Let's make this concrete with an example. Suppose that we have data $X_i \sim F$ and we want to estimate

$$\theta = T(F) = E_F(sin^2(|X|))$$

Then

$$\hat{\theta} = T(\hat{F}_n) = E_{\hat{F}_n}(sin^2(|X|)) = \frac{1}{n}\sum_i sin^2(|X_i|)$$

so that

$$var_F(T(\hat{F}_n)) = var_F(\frac{1}{n}\sum_i sin^2(|X_i|))$$

Calculating this quantity, even if we knew F could be complicated.

The bootstrap gives us the ability to non-parametrically[1] estimate this quantity, specifically estimating se with the bootstrap estimator $\widehat{se}_{boot}$ without ever writing down an analytical formula for se.

More generally, the bootstrap is a computer-intensive method for estimating measures of uncertainty in problems for which no analytical solution is available.

In developing the bootstrap we are going to discuss two hurdles the bootstrap overcomes:

[1] There are technically two classes of bootstrap methods: parametric and nonparametric. We will focus on the nonparametric bootstrap.

Hurdle 1 Let's assume we want to estimate

$$se^2 = var_F(T(\hat{F}_n))$$

Recall that $T(\hat{F}_n)$ is a function of our data $X_1, \ldots, X_n$. To make it less confusing, let's write

$$T(\hat{F}_n) = \theta(X_1, \ldots, X_n) = \hat{\theta}_n$$

so that we need to estimate

$$var_F(\hat{\theta}_n)$$

What would be a plug-in estimator of this quantity?

$$var_{\hat{F}_n}(\hat{\theta}_n)$$

This is the variance of $\hat{\theta}_n$ under the *conditional* distribution – i.e. the distribution of $\hat{\theta}_n$ if our n data observations were drawn from $\hat{F}_n$, and

$$\hat{\theta}_n = \theta(X_1^*, \ldots, X_n^*).$$

In the previous section, I emphasized that conditional probability statements like this involve no unknown parameters – they are just functions of the data (and thus valid estimators!).

But there's a sneaky problem here: while functionals of $\hat{F}_n$ are technically just functions of our data with no unknown parameters, it can be hard to actually calculate these functionals of $\hat{F}_n$ from our data. $\hat{F}_n$ is a specific distribution, and

$$var_{\hat{F}_n}(\hat{\theta}_n) = \sum_{u \in \Omega}(u - E_{\hat{F}_n}(\hat{\theta}_n))P(\hat{\theta}_n = u)$$

(where Ω are all the possible values that $\hat{\theta}_n$ can take on). Conceptually this is a concrete definition; practically, how can one possibly get this value?

So the first hurdle is how to calculate quantities like $var_{\hat{F}_n}(\hat{\theta})$?

The first part of this module, we will see how we can approximate these types of quantities with **Monte Carlo** integration – which is a general technique to approximate these quantities for any distribution. But we will apply it to the specific distribution $\hat{F}_n$.

Hurdle 2 We would also want to be able to go further, and not need to know whether our estimator $T(\hat{F}_n)$ is asymptotically normal. We want to find bootstrap CI that do not have the form of a normal approximate confidence interval.

The second part of this module will look at how we can construct CI that make less assumptions about the distribution of $\hat{\theta}$.

1 Monte-Carlo Integration

We are going to remove ourselves from the world of estimation, and go back to classical probability questions where we assume we know our distribution F and we want to calculate things about it.

Suppose we want to calculate $E_F(\sin^2(|X|))$. Let's assume F is a normal, specifically $X \sim N(0, 1/2)$. We want to calculate

$$\int_{-\infty}^{\infty} \sin^2(|X|) \frac{1}{\sqrt{\pi}} e^{-x^2} dx$$

When you start taking complicated integrals, there are various numerical approximations available that rely on computer algorithms. Monte Carlo integration is a method of approximation specific to the setting of integrals that are derived from probability calculations.

Monte Carlo integration relies on the fact that if we have $X_j \overset{iid}{\sim} F$

$$\frac{1}{B} \sum_{b=1}^{B} h(X_b) \xrightarrow{P} E_F[h(X)] \text{ as } B \to \infty,$$

assuming $E[h(X)] < \infty$.[2]

It's pretty easy to simulate B random variables from a $N(0, 1/2)$ using the computer. So we could get an approximation of $E(\sin^2(|X|))$ as

- Simulate 1,000 $X_b \sim N(0, 1/2)$
- Calculate $Z_b = sin^2(|X_b|)$
- Take the average of 1,000 Z_b values

$$E(\sin^2(X)) \approx \frac{1}{B} \sum_{b=1}^{B} Z_b$$

Notice that we can choose B as big as our computer can handle to get a better and better approximation.

Importantly, this same rationale works for other functions, such as the variance:

$$\frac{1}{B} \sum_{b=1}^{B} (X_b - \bar{X})^2 = \frac{1}{B} \sum_{b=1}^{B} X_b^2 - (\frac{1}{B} \sum_{b} X_b)^2 \xrightarrow{P} E(X^2) - (EX)^2 = var(X)$$

[2] More precisely, convergence is almost sure convergence, a.s.

A more complicated example: Suppose we have data $X_1, \dots, X_n \overset{iid}{\sim} Exp(\lambda)$. Use Monte Carlo integration to approximate $var_\lambda[median(X_1, \dots, X_n)]$.

This is more complicated because here our random variable is a statistic of n random variables. I.e., let

$$T_n = g(X_1, \dots, X_n) = median(X_1, \dots, X_n)$$

For a particular n and λ, we need to sample B times from the *distribution of* T_n. This means that for each $b = 1, \dots, B$,

- we sample $X_1^b, \dots, X_n^b \overset{iid}{\sim} Exp(\lambda)$
- Calculate $T_{n,b} = median(X_1^b, \dots, X_n^b)$

Then we calculate the variance as

$$\frac{1}{B} \sum_{b=1}^{B} (T_{n,b} - \bar{T})^2.$$

Don't confuse n and B: n is the sample size of the data, while B is the number of MC samples.

Here's an example of how we might code this for one combination: $n = 10$ and $\lambda = 5$.

```
lambda <- 5
n <- 100
randMedian <- function(lambda = 5, n = 100) {
    x <- rexp(n, rate = 1/lambda)
    z <- median(x)
}
B <- 1e+05
set.seed(1489)
medExp.lambda5.n100 <- replicate(B, randMedian())
var(medExp.lambda5.n100)
```

```
## [1] 0.2494042
## Compare to Asymptotic Variance:
## [1] 0.2525253
```

Often we might want to explore how this changes for different parameter choices. We have two quantities we might consider: λ *and* n. For the analytical calculation above, they are just fixed variables. To see how $var_\lambda[T_n]$ changes with n and λ, we need to use Monte Carlo integration many times for different combinations.

2 Bootstrap

Suppose we have data $X_1, \ldots, X_n \overset{iid}{\sim} F$ and we compute statistic $T_n = g(X_1, \ldots, X_n)$[3].

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

If we knew F, we could use MC integration to approximate $var_F[T_n]$, like we did above for the exponential distribution.

However, we don't in practice know F. But we do have some data drawn from F, so we can approximate F. Specifically, we make an approximation of F with the empirical CDF $\hat{F}_n$. Using $\hat{F}_n$ instead of F, we *can* do the Monte-Carlo integration like before.

We now have two different types of approximations happening here:

$$var_F[T_n] \quad \overset{\substack{\text{ECDF;} \\ \text{depends on } n}}{\approx} \quad var_{\hat{F}_n}(T_n) \quad \overset{\substack{\text{MC integration;} \\ \text{depends on } B}}{\approx} \quad \widehat{var}_{\hat{F}_n}(T_n)$$

The MC integration step we can control by increasing the size of B. However the first approximation depends on how close the approximation of $\hat{F}_n$ is to F.

Sampling from $\hat{F}_n$ Think about the definition of $\hat{F}_n$ – it gives a point mass for every value of x seen in the data. So a single draw from $\hat{F}_n$ is just giving equal probability of drawing any of the $X_1, \ldots, X_n$ observations.

[3] I'm using a generic statistic T_n, but usually $T_n = \hat{\theta} = T(\hat{F})$, i.e. our plug-in estimator for θ

Notice that our $X_1, \dots, X_n \overset{iid}{\sim} F$. So to approximate $var_F[T_n]$ we would draw i.i.d. samples from F. This is also true if we are going to replace F with $\hat{F}_n$, we still need to draw i.i.d samples from $\hat{F}_n$. This means we need to repeatedly draw n samples from $X_1, \dots, X_n$, known as "with replacement."

The algorithm:

1. Repeat the following B times

   (a) Draw $X^*_{1,b}, \dots, X^*_{n,b} \overset{iid}{\sim} \hat{F}_n$. (i.e. sample with replacement from original data)

   (b) Compute $T^*_{n,b} = g(X^*_{1,b}, \dots, X^*_{n,b})$.

   This will result in the **bootstrap samples** of T_n :

   $$T^*_{n,1}, \dots, T^*_{n,B},$$

   an *iid* sample from the sampling distribution for T_n implied by $\hat{F}_n$.

2. Use this sample to approximate $var_{\hat{F}_n}(T_n)$ by MC integration. That is,

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

v_{boot} is our bootstrap estimate of $var(T_n)$

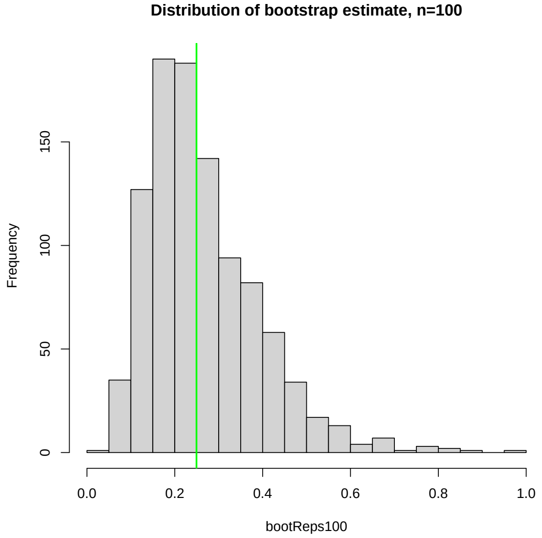
We can return to our simple MC Integration example above and this time assume that we had 100 observations from $exp(5)$ distribution, and try to estimate

$$var(median(X))$$

```
## Bootstrap Estimate of Variance:
## [1] 0.2197
## MC Approximation of Truth:
## [1] 0.2494042
```

Notice that we don't expect to get the same answer! Our MC Integration above was an approximation to the truth. The bootstrap gives an *estimate* of the truth based on a single sample of data. It's very important to understand this distinction between a numerical approximation and an estimate based on a finite data sample.

Indeed, since this is a simulation, I can replicate the process many times (i.e. over and over redo the above simulation), to see how the distribution of the bootstrap estimate of the variance compares the truth.



What quantities do we estimate with the bootstrap? What about if we wanted to estimate $\theta = E(X)$? What would the bootstrap give us? Well, we know that

$$E_{\hat{F}_n} X^* = \bar{X}_n,$$

(and we don't need the computer to get that estimate). Similarly, if we wanted to estimate $median(X)$ or $var(X)$, these are not quantities for which we need the full power of the bootstrap (i.e. computationally resampling, etc).

Estimating $var(median(X_1, \ldots, X_n)) = var(g(X_1, \ldots, X_n))$ is a different thing! In general, the bootstrap is useful for calculating features of the distribution of an *estimate*, not for calculating parameters of the distribution of our *data*.

3 Bootstrap Confidence Intervals

Confidence intervals can also be constructed from the bootstrap samples. There are three standard ways, though there are many variations on this:

3.1 Method 1: Normal-based intervals

We said before that often we have $T(\hat{F}_n) = \hat{\theta}_n \approx N(T(F), se^2)$, which in principle allows us to form a normal-based approximate $1-\alpha$ confidence interval for $\theta = T(F)$ of

$$T(\hat{F}_n) \pm z_{\alpha/2}\widehat{se}$$

However, we do not have a generic formula for a non-parametric estimate of se.

The bootstrap gives us the ability to non-parametrically estimate this quantity, specifically estimating se with $\widehat{se}_{boot}$.

This gives us the confidence intervals in the form of normal confidence intervals, with only the $se(T_n)$ estimated with the bootstrap:

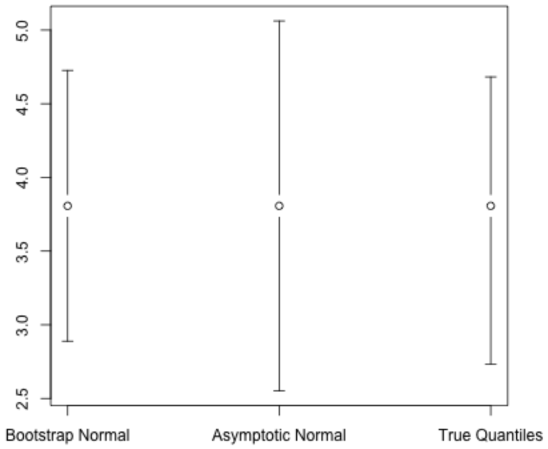
$$C_n = T(\hat{F}_n) \pm z_{\alpha/2}\widehat{se}_{boot}$$

where $\widehat{se}_{boot} = \sqrt{v_{boot}}$.

Generically, this works for any estimate T_n which is asymptotically normal.

Note that asymptotic normality of T_n is a property involving n, not B.

```
## Normal Bootstrap:
## [1] 2.888114 4.725504
## Asymptotic Normal:
## [1] 2.552333 5.061285
## CI Using quantiles of true distribution:
## [1] 2.733640 4.681553
```



3.2 Method 2: Pivotal Intervals

What if you do not have (or know if you have) asymptotically normal estimators?

We never discussed more generally how one can construct confidence intervals other than in a normal case, but you've seen that just knowing the finite distribution of $\hat{\theta}_n$ doesn't always result in a CI, since the interval you construct might depend on your unknown θ.

We are going to consider a slight generality of the normal construction, called **pivotal** CIs; they are not limited to the bootstrap, but are particularly useful in this context.

Suppose we have a parameter θ and a statistic $\hat{\theta}$. Consider the random variable $R_n = \hat{\theta}_n - \theta$, with CDF H. Then if we can find the $\alpha/2$-quantile and $1-\alpha/2$ quantiles of R_n,

$$\begin{aligned} r_{\alpha/2} &= H^{-1}(\alpha/2) \\ r_{1-\alpha/2} &= H^{-1}(1-\alpha/2) \end{aligned}$$

we can use this to state that

$$\begin{aligned} 1-\alpha &= P(r_{\alpha/2} \leq \hat{\theta}_n - \theta \leq r_{1-\alpha/2}) \\ &= P(r_{\alpha/2} - \hat{\theta}_n \leq -\theta \leq r_{1-\alpha/2} - \hat{\theta}_n) \\ &= P(\hat{\theta}_n - r_{1-\alpha/2} \leq \theta \leq \hat{\theta}_n - r_{\alpha/2}) \end{aligned}$$

This results in a confidence interval, $C_n = (\hat{\theta}_n - r_{1-\alpha/2}, \hat{\theta}_n - r_{\alpha/2})$ (notice how the lower and upper quantiles are switched in terms of which is in the lower and upper bounds of the CI)

R_n is called a pivotal statistic.

In parametric models, we might know H, the distribution of R_n, and create finite sample confidence intervals. However, in our setting $\theta = T(F)$ and $\hat{\theta}_n = T(\hat{F})$. So we don't know H.

The idea, then, is to non-parametrically estimate these quantiles using the bootstrap. Specifically, we need to find quantiles, where the p-quantile is

$$\begin{aligned} r_p &= \inf\{u : H(u) \geq p\} \\ &= H^{-1}(p), \text{for continuous } H \end{aligned}$$

where

$$H(u) = P_F(R_n \leq u)$$

To be clear, the CDF H is a very different quantity than F. F is the distribution of our individual X_i, while H is the distribution of $\hat{\theta}_n - \theta$. H clearly depends on F, but is likely a very complicated distribution to evaluate even if you knew F. But if we knew F and could sample X_i from F, we could use Monte-Carlo to get a good approximation of $H(u)$ and/or r_p.

Exercise 1. Describe precisely how you would approximate $H(u)$ using Monte-Carlo.

Exercise 2. Describe precisely how you would approximate r_p using Monte-Carlo.

Of course we do not know F so instead we are going to estimate r_p with our standard non-parametric plug-in estimator:

$$\hat{H}(r) = P_{\hat{F}}(R_n \leq r).$$

and

$$\hat{r}_p \approx \inf\{u : \hat{H}(u) \geq p\}$$

We are going to do this using Monte-Carlo simulations of $X^* = (X_1^*, \ldots, X_n^*)$ each i.i.d. $\hat{F}$ – i.e. the bootstrap.

1. Draw a sample $X^* = (X_1^*, \ldots, X_n^*)$ of n observations each i.i.d. $\hat{F}$
2. Calculate $\hat{\theta}_{n,b}^*$ and
$$R_{n,b}^* = \hat{\theta}_{n,b}^* - \hat{\theta}_n.$$
3. Approximate $\hat{H}(r)$ with
$$\begin{aligned}\hat{H}(r) = P_{\hat{F}}(R_n \leq r) &\approx \frac{1}{B}\sum_{b=1}^{B} I(R_{n,b}^* \leq r) \\ &= \frac{1}{B}\sum_{b=1}^{B} I(\hat{\theta}_{n,b}^* - \hat{\theta}_n) \leq r)\end{aligned}$$
4. $$\hat{r}_p \approx \inf\{u : \hat{H}(u) \geq p\}$$
i.e. if we order our $R_{n,b}$, set $\hat{r}_p$ to be the largest value of $R_{n,b}$ so that $\hat{H}(R_{n,b}) \leq p$ and the next largest value of $R_{n,b}$ has $\hat{H}(R_{n,b}) > p$.

Substituting $\alpha/2$ for p, we get $\hat{r}_{\alpha/2}$ and $\hat{r}_{1-\alpha_2}$ and our confidence interval

$$C_n = (a, b) = (\hat{\theta}_n - \hat{r}_{1-\alpha/2}, \hat{\theta}_n - \hat{r}_{\alpha/2})$$

These are called bootstrap pivotal confidence intervals.

The bootstrap empirical cdf, $\hat{F}^*$ Notice that we are trying to estimate the distribution of

$$R = \hat{\theta} - \theta$$

and we implicitly claim above that the plug-in estimate of that distribution would be to find the distribution of

$$R^* = \hat{\theta}^* - \hat{\theta}$$

(I'm leaving off the n in the subscript for clarity).

Note that

$$R = \hat{\theta} - \theta = T(\hat{F}) - T(F)$$

where our data $X_i \sim F$. So $\hat{F}$ is just a function of the data X, which is the empirical cdf of the data X.

In bootstrap land, we want to work with data $X_i^* \sim \hat{F}$. So if our data is X, we calculate $\hat{F}$, but when our data is X^*, to calculate R^*, we need to calculate the empirical cdf *of X^* our bootstrapped data.* We can call this $\hat{F}^*$. This gives us

$$\begin{gathered} T(F) \to T(\hat{F}) \\ T(\hat{F}) \to T(\hat{F}^*) \\ R = T(\hat{F}) - T(F) = \hat{\theta} - \theta \to R^* = T(\hat{F}^*) - T(\hat{F}) = \hat{\theta}^* - \hat{\theta} \end{gathered}$$

Simplification of pivotal intervals $\hat{r}_p$ is the p-quantile of the values

$$(R^*_{n,1}, \dots, R^*_{n,B}) = (\hat{\theta}^*_{n,b} - \hat{\theta}_n, \dots, \hat{\theta}^*_{n,b} - \hat{\theta}_n)$$

Since $\hat{\theta}_n$ doesn't change across the $R^*_{n,b}$ values, this means that we could equivalently find the p quantiles of

$$(\hat{\theta}^*_{n,1}, \dots, \hat{\theta}^*_{n,B})$$

call it t^*_p and have that

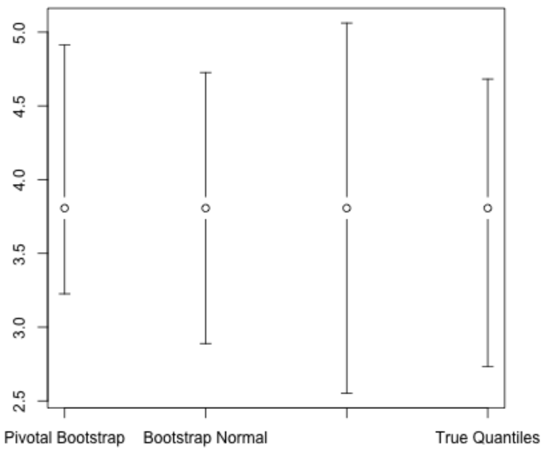
$$\hat{r}_{\alpha/2} = t^*_{\alpha/2} - \hat{\theta}_n.$$

This gives a $1 - \alpha$ bootstrap confidence interval of

$$\begin{aligned} C_n &= (\hat{\theta}_n - \hat{r}_{1-\alpha/2}, \hat{\theta}_n - \hat{r}_{\alpha/2}) \\ &= \left(2\hat{\theta}_n - t^*_{1-\alpha/2}, 2\hat{\theta}_n - t^*_{\alpha/2}\right) \end{aligned}$$

where t^*_p are the quantiles of the bootstrap values $\hat{\theta}^*_{n,b}$.

```
## Pivotal Bootstrap:
## [1] 3.225776 4.912376
## Normal Bootstrap:
## [1] 2.888114 4.725504
## Asymptotic Normal:
## [1] 2.552333 5.061285
## CI Using quantiles of true distribution:
## [1] 2.733640 4.681553
```



Notice that unlike the normal confidence intervals, these confidence intervals are not symmetric around the estimate.

3.2.1 Pivotal Statistics

Why are these called pivotal intervals? Let's look at our basic starting point,

$$1 - \alpha = P(r_{\alpha/2} \leq \hat{\theta}_n - \theta \leq r_{1-\alpha/2})$$

Note that we assumed that the distribution H of the statistic $R_n = \hat{\theta}_n - \theta$ didn't depend on θ (otherwise, $r_{\alpha/2}$ and $r_{1-\alpha/2}$ would be functions of θ, and thus wouldn't give us a confidence interval). We call such a functions R_n a **Pivotal function**,

Definition 3.1 (Pivotal function)**.** A **pivotal function** is a function of only the data and the parameter of interest (θ) whose distribution does not depend on any unknown parameters. An **asymptotically pivotal function** is a function of only the data and the parameter of interest (θ) whose asymptotic distribution does not depend on any unknown parameters.

Pivotal statistics are generally interesting statistics because they allow development of confidence intervals. For example if T_n is asymptotically normal,

$$R_n = \frac{T_n - \theta}{\hat{se}}$$

is asymptotically pivotal since it's limiting distribution is $N(0, 1)$, i.e. doesn't depend on any unknown parameter. That is how we develop CI for asymptotically normal statistics.

What about for the bootstrap where we estimate $h_{\alpha/2}$ from the data – how important is it that our $R_n = \hat{\theta}_n - \theta$ be pivotal? After all, if we want to be non-parametric, we generally won't know; moreover, we don't want to be constrained to only estimating parameters for which the estimate is a pivotal statistic.

The answer is that asymptotically, not much, but finite samples the closer you are to pivotal, the better performance of your confidence interval. Let's break that down.

Bootstrap Pivot Intervals with Pivotal statistics If $R_n = \hat{\theta}_n - \theta$ is pivotal for the true unknown distribution F, then the bootstrap confidence intervals you get will have the correct coverage without having to rely on asymptotic assumptions (i.e. without having to have very large sample sizes).

An example is if θ is a **location parameter**. This means our class of distributions $\mathcal{F}$ are all characterized by the same distribution except shifted by a value θ. The mean in a normal distribution is an example of a location parameter.

But an example of a parameter for which R_n is not pivotal is a **scale parameter**. A scale parameter acts multiplicatively on the density, scaling the density. Pivotal statistics for scale parameters are ratios, not differences, i.e. for a scale parameter τ,

$$\tilde{R}_n = \frac{\hat{\tau}_n}{\tau}$$

is pivotal. So Bootstrap pivot intervals based on $R_n = \hat{\theta}_n - \theta$ will not perform well on such statistics at finite samples sizes.

However, you can also sometimes transform your statistic so that $R_n = \hat{\theta} - \theta$ is pivotal, or more closely so. So if we took $\theta = \log \tau$ and $\hat{\theta}_n = \log \hat{\tau}_n$ then

$$R_n = \hat{\theta} - \theta = \log(\tilde{R}_n)$$

Since the distribution of $\tilde{R}_n$ doesn't depend on τ, then neither does R_n. So doing bootstrap pivotal intervals for $\log \tau$ would work well (and then could be converted to intervals for τ).

Bootstrap Pivot Intervals with Non-Pivotal statistics More generally, it is problematic to rely on $R_n = \hat{\theta}_n - \theta$ being pivotal for your unknown F. However, you still have guarantees of asymptotically good performance of bootstrap pivotal intervals. Specifically, if $\theta = T(F)$ and $\hat{\theta}_n = T(\hat{F}_n)$, then we know that for "well-behaved" functions T the coverage of C_n will approach $1 - \alpha$

$$P_F(T(F) \in C_n) \to 1 - \alpha$$

Well behaved is roughly what we discussed for convergence of $T(\hat{F}_n)$ to $T(F)$, that if two distributions F and G are "close" to each other, then $T(F)$ and $T(G)$ are close to each other.[4]

3.3 Method 3: Percentile intervals

The last common method of making bootstrap confidence intervals is the percentile method.

Again, let $t^*_{\alpha/2}$ and $t^*_{1-\alpha/2}$ be the sample quantiles of our bootstrap sample,

$$\hat{\theta}^*_{n,1}, \ldots, \hat{\theta}^*_{n,B}.$$

Then the percentile method constructs confidence intervals as

$$C_n = \left(t^*_{\alpha/2}, t^*_{1-\alpha/2}\right).$$

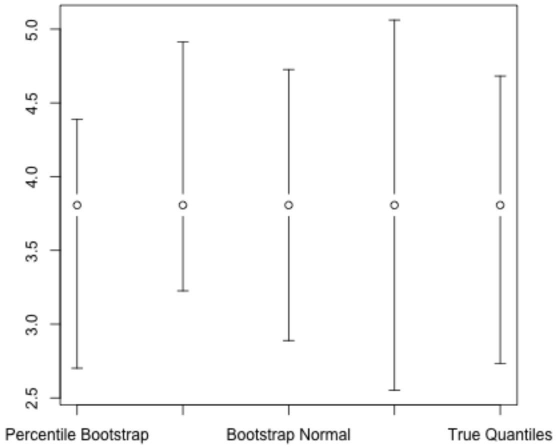
In other words, you just use the distribution of the $\hat{\theta}^*_{n,b}$ to define the limits of the confidence interval.

On the one hand, this seems intuitive – you use the distribution of $\hat{\theta}^*_n$ as a substitution for the distribution of $\hat{\theta}_n$. On the other hand, this is "backward" from the pivotal intervals, with respect to the position of $t^*_{\alpha/2}$, and $t^*_{1-\alpha/2}$.

```
## Percentile Bootstrap:
##     2.5%    97.5%
## 2.701242 4.387842
## Pivotal Bootstrap:
## [1] 3.225776 4.912376
## Normal Bootstrap:
## [1] 2.888114 4.725504
## Asymptotic Normal:
```

[4] It's not mathematically the same issue – you need what is called Hadamard differentiability of T, but the idea is the same.

```
## [1] 2.552333 5.061285
## CI Using quantiles of true distribution:
## [1] 2.733640 4.681553
```



Moreover, when we normally construct CI, we actually don't just use the distribution of $\hat{\theta}_n$ to get the boundaries of the CI – we generally manipulate a probability statement that contains θ in order to get a CI around θ, much like we did in pivotal bootstrap intervals. We seemed to have side-stepped that process. So why does this interval make sense?

Justification For the percentile method to be valid, you must make the assumption that there exists a monotonic tranformation m, so that $m(\hat{\theta}_n) - m(\theta)$ is pivotal (i.e. doesn't depend on θ) and symmetrically distributed around 0. We don't need to actually know m, but just assume that it exists. Without loss of generality, we can assume that m is monotonically *increasing* (otherwise, we just multiply it by a negative.)

Let $U_n = m(\hat{\theta}_n)$ and assume that the distribution of U is symmetric. Then from our work on pivotal statistics earlier, we can create a confidence intervals for $\phi = m(\theta)$

$$C_n = (U_n - r_{1-\alpha/2}, U_n - r_{\alpha/2})$$

where r_α is the α-quantile of $R_n = U_n - \phi$.

But since the distribution of R_n is symmetric, $r_{\alpha/2} = -r_{1-\alpha/2}$ so we could equivalently have a confidence interval

$$(U_n + r_{\alpha/2}, U_n + r_{1-\alpha/2})$$

(This is what happens with our normal confidence intervals, by the way)

If we estimate this with the bootstrap (again, if we knew m), we would get estimates of the quantiles,

$$\hat{r}_p = u^*_p - U_n$$

where u^*_p is the p-quantile of the bootstrapped U^*_b values.

Putting these into the above confidence interval, we get

$$C_n = (U_n + r^*_{\alpha/2}, U_n + r^*_{1-\alpha/2}) = (U_n + u^*_\alpha - U_n, U_n + u^*_{1-\alpha} - U_n) = (u^*_\alpha, u^*_{1-\alpha/2})$$

i.e. the confidence interval for $m(\theta)$ would be given by the α and $1-\alpha$ quantiles of the U^*_b bootstrap samples.

This procedure would require us to identify m. Except for a cute trick. Since m is monotonic, then our bootstrap confidence interval is:

$$\begin{aligned} 1-\alpha &\leq P(u^*_\alpha \leq m(\theta) \leq u^*_{1-\alpha}) \\ &= P(m^{-1}(u^*_\alpha) \leq \theta \leq m^{-1}(u^*_{1-\alpha})) \\ &= P(t^*_\alpha \leq \theta \leq t^*_{1-\alpha}) \end{aligned}$$

This last step relies on the fact that since m is monotone the data point U^*_b which corresponds to the p-quantile of the U^* will be the same data point so that $m^{-1}(U^*_b)$ that gives the p-quantile of the $m^{-1}(U^*)$ values. And $\theta^*_b = m^{-1}(U^*_b)$ so $t^*_\alpha = m^{-1}(u^*_\alpha)$.

This means, we can find valid confidence intervals with out knowing m – we only have to assume such a m exists!

4 When Does the Bootstrap work?

The power of the bootstrap lies not in estimating relatively straight-forward quantities (the mean, the median, ...). These are useful examples to think about because they are simple and so we can carefully understand how the bootstrap works. But there are many theoretical results about how to estimate these kinds of quantities and create confidence intervals, which are often pretty robust to assumptions as well. In these situations, the bootstrap generally replicates the theoretically-derived solutions.

The real appeal is to be able to use the bootstrap for new types of statistics or where it is difficult to work out the theory. But what do we need to keep in mind in using the bootstrap?

We need that the distribution of

$$\hat{\theta}^*_n - \hat{\theta}_n = T(\hat{F}^*_n) - T(\hat{F}_n)$$

be close to the distribution of what we are interested in:

$$\hat{\theta}_n - \theta = T(\hat{F}_n) - T(F),$$

where $\hat{F}_n^*$ is the empirical CDF of the *bootstrap sample.*

Hopefully, after our discussion on creating bootstrap pivotal intervals, you can see why the distribution of these quantities being approximately the same is quite important for using the bootstrap.

To ensure this happens, we need two things:

1. $\hat{F}_n - F$ to have roughly the same distribution as $\hat{F}_n^* - \hat{F}_n$
2. $\theta = T(F)$ be a function T that is "well-behaved", i.e. F and G are "close" to each other, then $T(F)$ and $T(G)$ are close to each other.

   If this holds, we can translate the fact that $\hat{F}_n - F$ and $\hat{F}_n^* - \hat{F}_n$ have roughly the same distribution to mean that $T(\hat{F}_n^*) - T(\hat{F}_n)$ and $T(\hat{F}_n) - T(F)$ have the same distribution.

Let's look at these two requirements.

Sampling Distribution The first requirement, on the distributions of the empirical cdfs, holds asymptotically, i.e. for large n, for univariate data X_i, i.e. a single number measured on each observation.

But to rely on this assumption means we need a sufficient sample size. So the bootstrap is not a method for small sample sizes. We will consider many asymptotic parametric tests as well, which will have the same caveats, so this does not dramatically differentiate the bootstrap from many other methods, and the question becomes how big of n is sufficient, which will depend on the data.

When we have multivariate data X_i, such as in regression, we will have observations X_i that are vectors of p numbers – i.e. many measurements taken on a single observation. In this setting, the size of p relative to the size of n is very important for how well $\hat{F}$ estimates F; if we do not have $n >> p$, then the first property will not always hold.

Well-behaved T We have mentioned many times in our non-parametric results that we need to T to be well-behaved in order thate when two distributions F and G are "close" to each other, then $T(F)$ and $T(G)$ are close to each other.

We saw an example previously that $T_{max}(F)$, which gives the maximum value of the distribution F, is not well-behaved.

We can easily create a reasonably realistic example that shows that the bootstrap fails for the max, regardless of the sample size. We do this by creating a simple example where we can work out what we should get. Let $X_i \overset{iid}{\backsim} Unif[0,\theta]$ and estimate θ with

$$\hat{\theta}_n = T_{max}(\hat{F}) = \max_{1\leq i\leq n} X_1,\ldots,X_n.$$

It's well-known that

$$n(\theta - \hat{\theta}_n) \Rightarrow \text{Exponential}(1/\theta) \ ,$$

i.e asymptotically r.v with density $1/\theta \exp(-t/\theta)$ on positive real numbers.

But $\hat{\theta}_n - \hat{\theta}^*_n$ has non-zero probability of being exactly equal to 0 (a point mass) – any time you have a bootstrap sample that contains the maximum value of the X_i then $\hat{\theta}_n = \hat{\theta}^*_n$. The probability that you draw the max value of the X_i in your bootstrap sample is

$$P(\max\{X_i\} \in X^*_1,\ldots,X^*_n\} = 1 - (1-\frac{1}{n})^n$$

This probability doesn't disappear with large sample sizes,

$$1 - (1-\frac{1}{n})^n \overset{n\to\infty}{\to} (1-1/e) \approx 0.632 \neq 0$$

So this means that the limiting distribution of $n(\hat{\theta}^* - \hat{\theta})$ has a point mass at 0.

This is clearly different from the limiting distribution of $n(\hat{\theta} - \theta)$, since the exponential distribution doesn't have a point mass at zero (its density is not even defined at zero). So no matter how large the sample sizes, the distributions of these two quantities do not converge to the same thing.