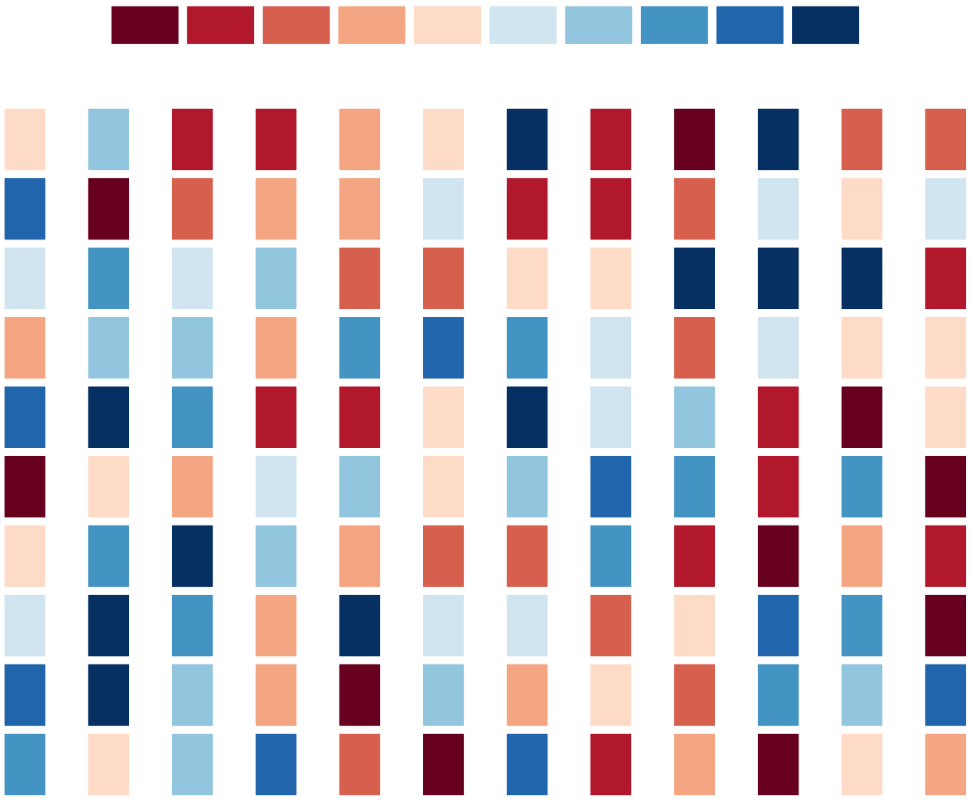
The Bootstrap

Stephan Smeekes*

Maastricht University
Department of Quantitative Economics
s.smeekes@maastrichtuniversity.nl

Current Version: April 17, 2020



 I would like to thank Eric Beutner, Marina Friedrich, Yicong Lin and Hanno Reuvers for their careful and critical reading of the manuscript and for spotting many typos. The remaining errors are, of course, my own. This note is a continuing work in progress; comments and suggestions for improvements are therefore most welcome.

Contents

1 Introduction

These notes provide a general introduction to the bootstrap that can be read as a follow-up to Casella and Berger's (2002) *Statistical Inference* (hereafter C&B). The bootstrap was first developed by the U.S. statistician Bradley Efron in a seminal paper from 1979.[1] In the years since, it has become one of the most important concepts in statistics. In 2005, Efron was awarded the National Medal of Science, the highest scientific honor in the United States, for his work on the bootstrap.

As a further illustration of the importance of the method, that is now a standard tool in all fields of applied statistics, consider Google Scholar (scholar.google.com), the search engine that searches only in academic papers. Searching for 'bootstrap confidence interval' yields 351,000 results, while searching for 'bootstrap hypothesis test' even yields 439,000 results.[2] Concluding, the very brief discussion in Section 10.1.4 of C&B does not do such a popular and prominent technique in modern statistics fully justice. These notes aim to fill this gap and present the bootstrap method in greater detail.

There exist many surveys and introductions about the bootstrap in the literature of varying levels and depths. While we cannot mention them all here, the introductions by Efron and Tibshirani (1994), Davison and Hinkley (1997), Davidson and MacKinnon (2004), Horowitz (2001), Efron and Hastie (2016, Chapters 10 and 11) and Hansen (2019, Chapter 10) are worth mentioning. The material in these notes is loosely based on these sources.

The remainder of Section 1 explains why the bootstrap can be useful. As a preliminary step to the development of the bootstrap, Section 2 discusses a nonparametric estimation technique called the plug-in principle. Section 3 then defines the bootstrap and lays down the foundations. In Section 4 we take a more practical view and discuss how the bootstrap can be applied. The theoretical foundations of the bootstrap are discussed in Section 5. Section 6 contains some exercises.

As not all readers may be equally familiar with the notation used in C&B, we briefly discuss the most important conventions in Appendix A, such that these notes can also be understood as a stand-alone document. Readers familiar with C&B can safely skip this appendix, but readers unfamiliar with the notation in the book are advised to read Appendix A first.

For the course Mathematical Statistics, all proofs of lemmas and theorems, as well as the starred sections, are optional.

[1] The name "bootstrap" derives from the expression "to pull one self up by his own bootstraps" and ultimately from the famous tales of Baron von Münchhausen, who claimed that he pulled himself up out of a swamp by his own bootstraps.

[2] This search was done on March 20, 2020. A regular Google search for these terms even yields 9,310,000 and 19,100,000 hits respectively!

1.1 The shortcomings of asymptotic statistics

In Chapter 10 of C&B we have seen that asymptotic analysis allows us both to simplify analysis where small sample results are complicated to obtain, and to allow for analysis when otherwise small sample analysis would not be possible. As such it is a very powerful tool, however as remarked on in Chapter 10, it is not perfect. The assumption that the sample size increases to infinity, is (almost) never met in practice. Asymptotic results are therefore only approximations to reality. How good are these approximations? For sure the sample size will play an important role. After all, one could say that a large sample size is closer to infinity than a small one! Yet, how large should the sample size be to be called large? Unfortunately there is no straightforward answer to that question.

Example 1. Let $X_1, \ldots, X_n$ be a random sample from an unspecified pdf $f(x)$, for which we only know that $\mathbb{V}\mathrm{ar}\, X < \infty$. We would like to construct a confidence interval for $\mathbb{E}\, X = \mu$. However, under these assumptions we cannot obtain small sample results, and as such, constructing a confidence interval with exact confidence level $(1-\alpha)$ is impossible.

Of course, we know from the central limit theorem that $\sqrt{n}\frac{\overline{X}_n - \mu}{S_n} \xrightarrow{d} N(0,1)$, from which we can construct the *asymptotic* $(1-\alpha)$ confidence interval

$$C_\mu(\mathbf{X}) = \left[\overline{X}_n - z_{\alpha/2}\frac{S_n}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{S_n}{\sqrt{n}}\right].$$

What is the actual confidence level of this interval? We only know that as n increases, it will become closer to $(1-\alpha)$. However, that still means that for a reasonable practical sample size, coverage can be quite far away from the desired level. We can imagine two things that affect the difference between the true confidence level and the asymptotic confidence level: (i) the sample size and (ii) the shape of the distribution $f(x)$.

Let us investigate this in more detail. To find the true confidence level, we need to obtain the distribution of $Q(\mathbf{X}, \mu) = \sqrt{n}\frac{\overline{X}_n - \mu}{S_n}$, in order to calculate the coverage probability

$$\begin{aligned}\gamma(\mu) &= \mathbb{P}_\mu(C_\mu(\mathbf{X}) \ni \mu) = \mathbb{P}_\mu\left(\overline{X}_n - z_{\alpha/2}\frac{S_n}{\sqrt{n}} \leq \mu \leq \overline{X}_n + z_{\alpha/2}\frac{S_n}{\sqrt{n}}\right) \\ &= \mathbb{P}_\mu\left(-z_{\alpha/2} \leq \sqrt{n}\frac{\overline{X}_n - \mu}{S_n} \leq z_{\alpha/2}\right) = \mathbb{P}_\mu\left(-z_{\alpha/2} \leq Q(\mathbf{X}, \mu) \leq z_{\alpha/2}\right).\end{aligned} \tag{1}$$

We can calculate this probability if we choose a specific parametric distribution for $f(x)$; for example, if $f(x)$ is equal to $N(\mu, \sigma^2)$, we know that $Q(\mathbf{X}, \mu) \sim t_{n-1}$ and we need to calculate $\mathbb{P}\left(-z_{\alpha/2} \leq t_{n-1} \leq z_{\alpha/2}\right)$. As in this case $\gamma(\mu)$ does not depend on μ and we can directly conclude that the confidence level is equal to $\inf_\mu \gamma(\mu) = \mathbb{P}\left(-z_{\alpha/2} \leq t_{n-1} \leq z_{\alpha/2}\right)$. Table 1 shows the exact confidence level of $C_\mu(\mathbf{X})$ for different values of n and α if we assume that the sample is normally distributed.

We see that as n increases, the difference between the true confidence level and the asymp-

n / α	0.005	0.010	0.025	0.050	0.100	0.150	0.200
10	0.981	0.972	0.951	0.922	0.869	0.819	0.771
20	0.989	0.982	0.963	0.936	0.884	0.835	0.785
30	0.991	0.985	0.967	0.941	0.890	0.840	0.790
40	0.992	0.986	0.969	0.943	0.892	0.842	0.793
50	0.993	0.987	0.971	0.944	0.894	0.844	0.794
70	0.994	0.988	0.972	0.946	0.896	0.846	0.796
100	0.994	0.989	0.973	0.947	0.897	0.847	0.797
∞	0.995	0.990	0.975	0.950	0.900	0.850	0.800

Table 1: Exact confidence levels of the asymptotic $(1-\alpha)$ confidence interval $C_\mu(\mathbf{X})$ for normally distributed samples of size n.

totic confidence level becomes smaller, as expected. Of course, if we knew that $f(x)$ was the normal distribution, we did not have to use the asymptotic interval $C_\mu(\mathbf{X})$, and could directly use cut-off points from the t_{n-1} distribution.

Therefore it is more interesting to consider another choice for $f(x)$ than the normal. In that case we could perform a similar calculation provided we can derive the distribution of $Q(\mathbf{X}, \mu)$. In general this is very complicated or even impossible though. Instead of attempting an analytical derivation, we can let a computer do the work instead by simulating samples from the chosen distribution. The procedure works as follows:

1. For every simulation $j = 1, \ldots, N$, draw a random sample $\mathbf{X}^{(j)} = X_1^{(j)}, \ldots, X_n^{(j)}$ from the chosen distribution $f(x)$. Importantly, for every simulation j, the sample must be drawn independently from the other simulations.[3]

2. For every simulation $j = 1, \ldots, N$, use the sample $\mathbf{X}^{(j)}$ generated in step 1 to construct the confidence interval $C_\mu(\mathbf{X}^{(j)})$.

3. For every simulation $j = 1, \ldots, N$, check if $\mu \in C_\mu(\mathbf{X}^{(j)})$ and record a 1 if true, and a 0 otherwise. The average of this number is the estimated coverage probability $\hat{\gamma}(\mu)_N$. Formally, $\hat{\gamma}(\mu)_N = \frac{1}{N}\sum_{j=1}^{N} I_{C_\mu(\mathbf{X}^{(j)})}(\mu)$.

This computer-assisted method of investigating properties of statistical methods is called *Monte Carlo simulation*. While $\hat{\gamma}(\mu)_N$ is not equal to the true coverage probability $\gamma(\mu)$, it is close to it for reasonably large values of N, and one can show that if N increases to infinity, it converges to the true coverage probability by the law of large numbers.[4]

[3]Statistical software packages such as R or Gauss have built-in algorithms to generate samples from most popular distributions. Of course these numbers are not truly random (think how difficult it is to generate real random numbers!), but these *pseudo-random* numbers are close enough to use as random numbers for the purposes of statistical analysis. These algorithms also ensure that samples drawn in consecutive simulations can be treated as being mutually independent.

[4]We can see this as follows. Let $Y_j = \sum_{j=1}^{N} I_{C_\mu(\mathbf{X}^{(j)})}$ for $j = 1, \ldots, N$. As Y_j is a function of the random sample $\mathbf{X}^{(j)}$, Y_j is a random variable; more specifically, Y_j is Bernoulli distributed with the probability of

The switch in terminology from "confidence level" to "coverage probability" is intentional. Remember that the confidence level is defined as the infimum over all μ of the coverage probability; yet in a Monte Carlo simulation study, we have to choose a specific value of μ to implement in the simulations, and so we cannot generalize the result to an arbitrary μ, and therefore certainly not to the infimum over all μ. However, by varying the values of μ used in the simulations, we can get an idea of whether the coverage probabilities depend on μ or not.

Figure 1 shows the coverage probabilities of $C_\mu(\mathbf{X})$ with $1 - \alpha = 0.95$ for the means of several distributions as a function of the sample size n. These were obtained by Monte Carlo simulation with $N = 1,000,000$ simulations in R. The coverage probability clearly improves for increasing n for all these distributions and becomes closer to the desired 0.95 level. However, for any given sample size, there are clear differences between the distributions. This illustrates that the shape of the distribution (e.g. skewness, thickness of the tails) matters for the accuracy of the asymptotic confidence interval. As such the value of n for which the coverage becomes satisfactory,[5] differs for each distribution. Figure 1 therefore also demonstrates that the often quoted rule-of-thumb that a sample size of 30 or larger is sufficient to apply the central limit theorem, is nonsense: the coverage probability of $C_\mu(\mathbf{X})$ for the mean of an Exponential(2) distribution at $n = 30$ is just 0.918, compared with 0.941 for the normal distribution. The former seems hardly close enough to 0.95, in particular relative to the latter, to justify the CLT approximation as very accurate.

Example 1 shows that asymptotic inference, though extremely useful, is not perfect. In fact, there are many settings with more complicated models where it still does much worse than in the example. Ideally, we would therefore like to have a way to improve on asymptotic inference. Clearly, if we know the parametric form of the distribution of our sample, we can derive exact results. Typically though we cannot be certain of that distribution. In particular, the common assumption of normality is most often not justified. Although obtaining exact results is therefore typically not feasible, this does not necessarily mean our only choice is to use the standard asymptotic approximation. We may consider alternative methods that, although they are validated using asymptotic theory, may still be more accurate in small samples. One of the most prominent of those methods, the *bootstrap*, is the subject of these notes.[6]

success p equal to $p = \mathbb{E}\, Y_j$. The expected value of Y_j is simply equal to the coverage probability $\gamma(\mu)$, that is $\mathbb{E}\, Y_j = \mathbb{P}_\mu(C_\mu(\mathbf{X}) \ni \mu) = \gamma(\mu)$ for all $j = 1, \ldots, N$, as all the samples $\mathbf{X}^{(j)}$ are identically distributed. Moreover, as $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(N)}$ are independent, the same is true for $Y_1, \ldots, Y_N$. Hence, $Y_1, \ldots, Y_n$ are i.i.d. Bernoulli($\gamma(\mu)$) random variables, and therefore by the weak law of large numbers, $\hat{\gamma}(\mu)_N = \overline{Y}_N \xrightarrow{p} \mathbb{E}\, Y = \gamma(\mu)$.

[5] Defining what is satisfactory is another matter in itself.

[6] People also often consider *Bayesian* statistics as an alternative. However, this is based on very different concepts and foundations, and as such comparing it with standard asymptotic inference is comparing apples and oranges. Moreover, a thorough discussion of Bayesian statistics requires a whole course in itself. We therefore do not discuss it here.

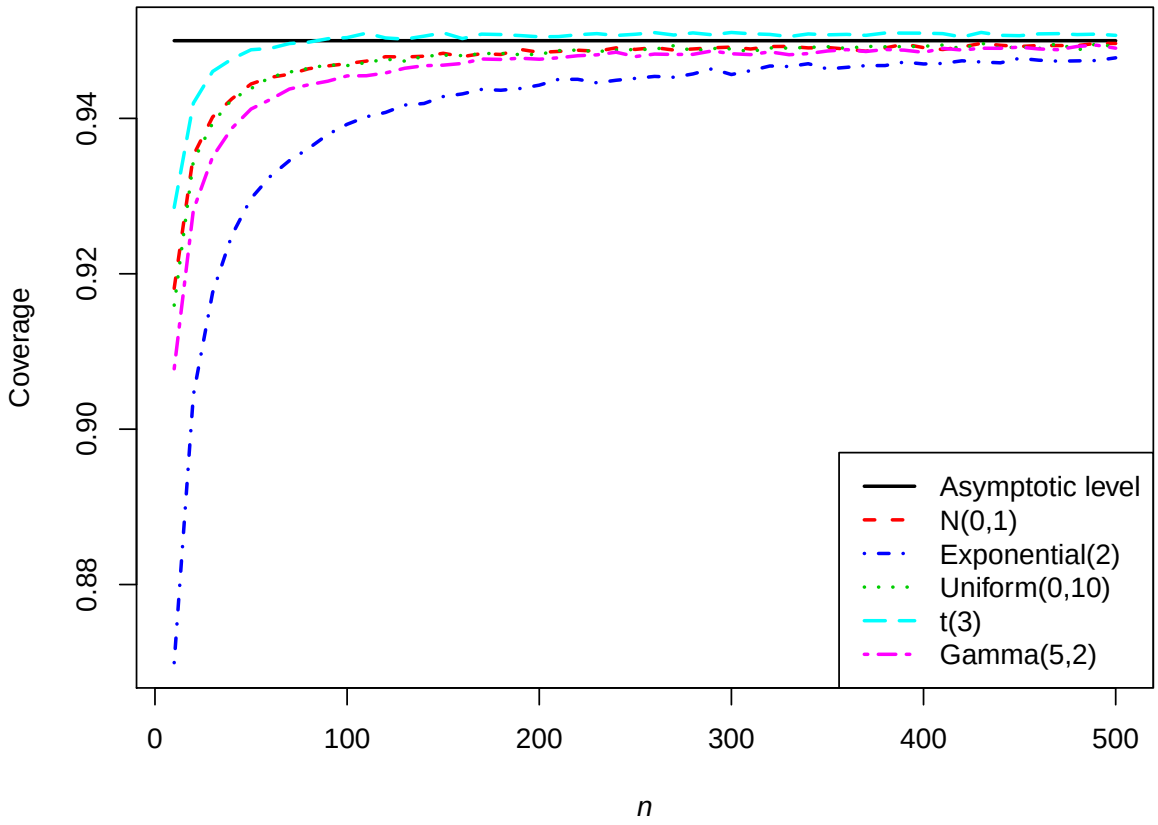


Figure 1: Coverage probabilities for the asymptotic confidence interval $C_\mu(\mathbf{X})$ for the mean of a number of distributions.

1.2 An alternative way to do inference: the bootstrap

The general principle of the bootstrap, originally developed by Efron (1979), is to treat the sample that one observes as the population. That is, one draws new samples from the original sample and approximates the relation between the (unknown) population and the sample with the relation between the sample, considered as a (known) artificial population, and the newly drawn samples. Because it requires the drawing of (many) new samples, it is computationally heavy and requires a computer to implement. With the rise of more powerful computers, its popularity also increased massively. Nowadays a "standard" bootstrap application is easily done on any home computer, and partly because of this, it has become one of the most important and popular methods in modern statistics. Let us illustrate the bootstrap with an example.

Example 2. We return to the setting of Example 1 and investigate the distribution of the quantity $Q_n(\mathbf{X}, \mu) = \sqrt{n}\frac{\overline{X}_n - \mu}{S_n}$, for example needed to construct a confidence interval for the mean. A simple bootstrap algorithm for approximating this distribution looks as follows.

1. Draw n numbers with replacement from the observed sample $x_1, \ldots, x_n$. We call the

collection of these draws the *bootstrap sample* and denote it with stars, i.e. $x_1^*, \ldots, x_n^*$.

2. Using your bootstrap sample $x_1^*, \ldots, x_n^*$, calculate the bootstrap statistics $\overline{x}_n^* = \frac{1}{n}\sum_{i=1}^n x_i^*$ and $s_n^* = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_i^* - \overline{x}_n^*)^2}$. Using these, obtain the bootstrap version of $Q_n(\mathbf{x}, \mu)$, which is

$$Q_n^* = Q_n(\mathbf{x}^*, \overline{x}_n) = \sqrt{n}\frac{\overline{x}_n^* - \overline{x}_n}{s_n^*}.$$

3. Repeat steps 1-2 B times, and store all numbers obtained for Q_n^*. These form the bootstrap approximation to the distribution of $Q(\mathbf{X}, \mu)$.

Note that in step 2, in the definition of the statistic we replaced μ with $\overline{x}_n$. The reason is that the mean of the bootstrap sample is no longer μ, but $\overline{x}_n$. We will come back to this point later.

The bootstrap in the form as described above draws a sample *with replacement* from the original sample. This action, and therefore sometimes the bootstrap in general, is also called *resampling*. Initially it might be more tempting to draw a new sample without replacement, but this does not work. After all, drawing n values from a sample of size n without replacement, simply means we end up with the original sample. Moreover, it violates the assumption that the bootstrap sample (like the original sample) is a random sample.

Once we have obtained the bootstrap distribution of a statistic, as done in Example 2, we can use it to construct bootstrap confidence intervals or to do hypothesis testing. However, we first need to gain an understanding of the rationale behind the bootstrap. Why do we draw samples with replacement from the original sample? And can we somehow justify this? Before we can go into this, we must first build the necessary foundations. For this purpose we initially digress to nonparametric estimation, and we introduce the plug-in principle. This principle will later turn out to be very useful to explain the bootstrap.

2 Nonparametric Estimation: The Plug-in Principle

2.1 Defining parameters without specifying a parametric distribution

In C&B we considered parameters θ as unknown quantities that determine the shape of a parametric distribution $f(x|\theta)$. However, these parameters themselves may not always be the quantities we are interested in. There are exceptions: for the normal distribution $N(\mu, \sigma^2)$, the parameters μ and σ^2 have a clear interpretation as the mean and variance of the distribution. This is however not always the case.

Example 3. Let $X_1, \ldots, X_n$ be a random sample from a Gamma(α, β) distribution. We would typically not be interested in α and β themselves; after all, what do α and β mean?

Rather, we would be more interested in a function of α and β. It is for instance far more likely that we are interested in the mean $\mu = \mathbb{E}\, X = \alpha\beta$.

We can solve this issue by explicitly writing the parameters of interest as a function of the original parameters. In the example above one would then write $\mu = \zeta(\alpha, \beta)$, where the function $\zeta(x, y)$ is defined as $\zeta(x, y) = xy$. We could then base our inference for μ on α and β, linking the two using the function $\zeta(x, y)$. For example, to do maximum likelihood estimation on μ we could use the invariance property of maximum likelihood from which it follows that $\hat{\mu}_{ML} = \zeta(\hat{\alpha}_{ML}, \hat{\beta}_{ML}) = \hat{\alpha}_{ML}\hat{\beta}_{ML}$.

While this provides a way to deal with the interpretation of parameters, a second problem remains, in that parameters are not unique.

Example 4. Let $Y_1, \ldots, Y_n$ be a random sample from the exponential(β) distribution where $f(x|\beta) = \frac{1}{\beta}e^{-x/\beta}$. Now define $\gamma = 1/\beta$ and let $g(x|\gamma) = \gamma e^{-\gamma x}$. Then we can equivalently say that $Y_1, \ldots, Y_n$ is a random sample from the distribution $g(x|\gamma)$. Which parameter is the "true" parameter, β or γ? There is no way to tell. Moreover, the interpretation of the parameter of interest does not solve the issue. Suppose that we need the variance of X, say $\theta = \mathbb{V}\mathrm{ar}\, X = \beta^2$. Then we have that $\theta = \zeta(\beta)$ with $\zeta(x) = x^2$, but equivalently $\theta = \zeta'(\gamma)$ with $\zeta'(x) = 1/x^2$. Hence we have two ways of defining θ that are not discernible from each other.

To deal with these issues in a more coherent way, we consider a different way to define the parameter of interest. Where in Example 3 we wrote the parameter of interest as a function of the original parameters, we can take this one level higher to avoid the notion of original parameters: we write the parameter of interest as a function of the pdf or pmf $f(x)$. Let θ be the parameter of interest. Then we write $\theta = \xi(f)$. Here we need to be careful of what kind of operator $\xi(\cdot)$ is. It is similar to a function, but instead of a real number x it takes as its argument a function f. Hence, it maps the function f to a real number, whereas a function maps one real number to another real number. We call such an operator a *functional*.

Example 5. Let X be a continuous random variable with the pdf $f(x)$ defined on $x \in \mathbb{R}$. Suppose that we are interested in the probability that X is larger than some constant c. We can write $\theta = \mathbb{P}(X > c) = \int_c^\infty f(x)dx = \xi(f)$, where $\xi(g) = \int_c^\infty g(x)dx$. If Y is a discrete random variable with pmf $f(y)$ taking values in $\mathbb{N}$, we can again do the same by letting $\xi(g) = \sum_{y=\lceil c \rceil}^\infty g(y)$. Here we let $\lceil c \rceil$ denote the smallest integer that is larger than c, to make sure that the definition works for any value of c, including those that Y cannot take.

Example 6. Let X be a continuous random variable with the pdf $f(x)$ defined on $x \in \mathbb{R}$. Suppose that we are interested in the mean of x, let us call this parameter θ. Now note that

$$\theta = \mathbb{E}\, X = \int_{-\infty}^{\infty} x f(x)dx = \tau(f), \quad \text{where} \quad \xi_1(g) = \int_{-\infty}^{\infty} x g(x)dx,$$

for any continuous function $g(x)$. Similarly, we can define other moments as $\theta_k = \mathbb{E}\, X^k = \gamma_k(f)$, with $\xi_k(g) = \int_{-\infty}^{\infty} x^k g(x)dx$ for $k = 1, 2, \ldots$.

If Y is a discrete random variable taking values $1, 2, ...$ with pmf $f(y)$, we can write $\theta_k = \mathbb{E}\, X^k = \xi_k(f)$, with $\xi_k(g) = \sum_{x=1}^{\infty} x^k g(x)$.

Note that in none of the examples above we had to make any assumptions on the form of the distribution $f(x)$, other than assuming that the relevant integrals or sums exist. In particular, we do not need to assume a particular parametric family. As we will see in the next subsection, this make it possible to do *nonparametric* inference, that is inference without making a parametric assumption on the distribution of the sample. Note though, that if we do assume a parametric family for $f(x)$, we are back to our familiar results.

Example 7. Consider the function $\xi_1(g)$ defined in Example 6. Now assume that X has a $N(\mu, \sigma^2)$ distribution; that is, let $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$. Then it follows directly that $\theta = \mathbb{E}\, X = \xi_1(f) = \mu$. Similarly, for $\xi(g)$ defined in Example 5 let $f(x) = \frac{1}{\beta} e^{-x/\beta}$. Then $\theta = \mathbb{P}(X > c) = \xi(f) = e^{-c/\beta}$.

We want to make one further step. In the examples we still needed to make a difference between discrete and continuous random variables. However, if we work with the cumulative distribution function $F(x)$ instead of the pdf/pmf $f(x)$, we do not have to make the distinction anymore. To distinguish with the previous functional $\xi(\cdot)$ operating on $f(x)$, we write the functional that operates on $F(x)$ as $\tau(\cdot)$.

For the setting discussed in Example 5 where $\theta = \mathbb{P}(X > c)$, it follows directly for both continuous and discrete random variables that $P(x > c) = 1 - \mathbb{P}(X \leq c) = 1 - F(c)$. Therefore, we can simply define $\tau(g) = 1 - g(c)$ such that we have $\theta = \tau(F)$. For the moments of Example 6 the situation is slightly more complicated. For this purpose we need to introduce *Riemann-Stieltjes integrals.*

Definition 1. Let $a = x_1, \ldots, x_r = b$ be a partition of the interval $[a, b]$, and define $\delta_r = \max_{2 \leq j \leq r} |x_j - x_{j-1}|$. Then, for any two functions $g(x)$ and $h(x)$ that satisfy certain regularity conditions[7] the **Riemann-Stieltjes** integral of $g(x)$ with respect to $h(x)$ is defined as

$$\int_a^b g(x)dh(x) = \lim_{\delta_r \to 0} \sum_{j=2}^{r} g(c_j)[h(x_j) - h(x_{j-1})], \tag{2}$$

where c_j is an arbitrary point in the interval $[x_{j-1}, x_j]$.

Note the difference with the "standard" Riemann integral, which is defined as $\int_a^b g(x)dx = \lim_{\delta_r \to 0} \sum_{j=2}^{r} g(c_j)(x_j - x_{j-1})$. Rather than integrating over x, we essentially weigh x through

[7] These conditions ensure that the limit of the right-hand side of (2) exists and does neither depend on the sequence of partitions nor on the choice of $c_1, \ldots, c_r$. For our purposes these are satisfied.

the function $h(x)$. If $h'(x)$ exists and is continuous, then we find that

$$\int_a^b g(x)dh(x) = \int_a^b g(x)h'(x)dx, \tag{3}$$

which follows directly from writing $h'(x) = \frac{dh(x)}{dx}$, from which we can substitute $dh(x)$ with $h'(x)dx$. The advantage of the Riemann-Stieltjes integral is that we can now deal with situations where $h'(x)$ is not continuous. The main use of the Riemann-Stieltjes integral for us is described in Lemma 1, where we choose $h(x) = F(x)$, and hence for a discrete random variable $h'(x) = f(x)$ is not continuous.

Lemma 1. *Let X be a discrete or continuous random variable with cdf $F(x)$. For any function $g(x)$, we can then write*

$$\mathbb{E}\, g(X) = \int_{-\infty}^{\infty} g(x)dF(x). \tag{4}$$

Proof. Take $h(x) = F(x)$ in (3) such that $h'(x) = f(x)$. It then follows directly from (3) that for a continuous random variable X we get that $\mathbb{E}\, g(X) = \int_{-\infty}^{\infty} g(x)dF(x)$.

We next show that for discrete random variables we can write $\mathbb{E}\, g(X) = \int_{-\infty}^{\infty} g(x)dF(x)$ as well. Consider a discrete random variable X which can take values $a_1, \ldots, a_k$ where $P(X = a_i) = p_i$ with $0 \leq p_i \leq 1$ for all $i = 1, \ldots, k$ as well as $\sum_{i=1}^k p_i = 1$. Let $F(x) = P(X \leq x)$ be the corresponding cdf of X.

Now consider $a_1 = x_1, \ldots, x_r = a_k$ as a partition of the interval $[a_1, a_k]$. As we only need to consider $\lim_{\delta_r} \to 0$ in order to define the Riemann-Stieltjes integral, we can restrict ourselves to those partitions for which $\delta_r \leq \min_{2 \leq i \leq k} |a_i - a_{i-1}|$. In words, we only consider those partitions in which in every interval $x_j - x_{j-1}$, there is at most one a_i. This automatically implies for those partitions $r \geq k$. For such partitions we can consider two scenarios:

1. No value of a_i lies in the interval $[x_{j-1}, x_j]$. In that case $F(x_j) - F(x_{j-1}) = 0$.
2. There is one value of a_i that lies in the interval $[x_{j-1}, x_j]$. In that case $F(x_j) - F(x_{j-1} = P(x_{j-1} < X < x_j) = P(X = a_i) = p_i$.

Note that for any such partition, the second scenario occurs exactly k times, one for each $a_1, \ldots, a_k$. This means that $\sum_{j=2}^r g(c_j)[F(x_j) - F(x_{j-1})]$ can be restricted to only those intervals for which scenario 2 above applies.

For $i = 1, \ldots, k$, define $y_i^- = x_{j-1}$ and $y_i^+ = x_j$ if $a_i \in [x_{j-1}, x_j]$. It then follows that

$$\begin{aligned}\int_a^b g(x)dF(x) &= \lim_{\delta_r \to 0} \sum_{j=2}^{r} g(c_j)[F(x_j) - F(x_{j-1})] \\ &= \lim_{\delta_r \to 0} \sum_{j=2}^{r} g(c_j) \sum_{i=2}^{k} P(X = a_i) I_{[x_{j-1}, x_j]}(p_i) = \lim_{\delta_r \to 0} \sum_{i=2}^{k} g(c_{j_i}) p_i,\end{aligned}$$

where j_i is defined such that $c_{j_i} \in [y_i^-, y_i^+]$. Now note that $\lim_{\delta_r \to 0} c_{j_i} = a_i$. Therefore

$$\int_a^b g(x) dF(x) = \sum_{i=2}^{k} g(a_i) p_i = \mathbb{E}\, g(X).$$

This is the result we set out to show. We can therefore conclude that for any random variable X, whether discrete or continuous, we have that $\mathbb{E}\, X^k = \int_{-\infty}^{\infty} x^k dF(x)$. □

2.2 The plug-in principle

We derived before that we may write the parameter of interest as $\theta = \tau(F)$. We now use this representation to construct estimators of θ. To do this we first estimate the cdf $F(x)$ by an estimator $\hat{F}(x)$ for all x, and then plug in this estimator into the functional $\tau(\cdot)$. This way of constructing an estimator is called the *plug-in principle*.

Definition 2. Let $X_1, \ldots, X_n$ be a random sample from a cdf $F(x)$. Let the function $\hat{F}_n(x)$ be an estimator of $F(x)$, where $\hat{F}_n$ is a cdf itself as well. Then the **plug-in estimator** of $\theta = \tau(F)$ is defined as $\hat{\theta} = \tau(\hat{F}_n)$.

Of course, the key in applying the plug-in principle for estimation is how we choose $\hat{F}_n(x)$. If we assume a parametric family $F(x|\gamma)$, then the plug-in estimator will simply require us to estimate γ.

Example 8. Assume that $X_1, \ldots, X_n$ is a random sample from an exponential(β) distribution and let $\theta = \mathbb{P}(X > c) = \tau(F) = e^{-c/\beta}$. We can then consider the estimator $\hat{F}_n(x)$ of $F(x|\beta) = 1 - e^{-c/\beta}$ defined by

$$\hat{F}_n(x) = F(x|\hat{\beta}_n) = 1 - e^{-c/\hat{\beta}_n}.$$

From this it directly follows that the plug-in estimator of θ is equal to $\hat{\theta} = \tau(\hat{F}_n) = e^{-c/\hat{\beta}_n}$. It then depends on how we estimate β what this estimator looks like. If we estimate β using maximum likelihood, $\hat{\beta}_n = \overline{X}_n$ and $\hat{\theta}_n = e^{-c/\overline{X}_n}$. Of course, in this case this plug-in estimator coincides with the "direct" maximum likelihood estimator of θ because of the invariance property of maximum likelihood estimators.

2.3 The empirical distribution function

Example 8 demonstrates that the plug-in estimator in the setting where we have a parametric family does not really add new insights. It still requires us to come up with an estimator of the parameters of the distribution and does not specify how to do that. Hence in the parametric setting the plug-in principle does not provide guidance to construct estimators of θ. Instead, it is more useful in a nonparametric setting, where we cannot or do not want to

assume a parametric family $f(x|\gamma)$. For this purpose we now define a simple nonparametric estimator of $F(x)$.

Definition 3. Let $X_1, \ldots, X_n$ be a random sample from a cdf $F(x)$. The **empirical distribution function** $\hat{F}_n^E(x)$, also denoted as EDF, is the estimator of F that for every x counts which proportion of $X_1, \ldots, X_n$ is smaller or equal to x. Formally,

$$\hat{F}_n^E(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i) = \frac{1}{n}\sum_{i=1}^{n} I_{[X_i,\infty)}(x).$$

Before we continue, two remarks on notation. First, traditional notation for the EDF is F_n, and therefore in most literature about the bootstrap you will see $F_n(x)$ instead of $\hat{F}_n^E$. We use $\hat{F}_n^E$ instead as it might be confusing that F_n without a hat is still an estimator of F. Second, as can be seen from the definition there are two equivalent ways of expressing the indicator functions in the EDF. Both the conditions $I_{(-\infty,x]}(X_i)$ and $I_{[X_i,\infty)}(x)$ are true whenever $X_i \leq x$. This confirms that no matter how one writes it, the EDF indeed counts how many of $X_1, \ldots, X_n$ are smaller than or equal to x. The two ways of writing are equivalent, and it depends on personal preference which to prefer. Throughout this chapter we will use the first one.

Example 9. We observe a random sample $X_1, \ldots, X_4$ from an unspecified distribution F. Suppose the following values are observed: $x_1 = 1.26$, $x_2 = -2.45$, $x_3 = 0.75$ and $x_4 = -0.27$. The realization of the corresponding EDF $\hat{F}_n^E$ can then be specified as follows:

$$\hat{F}_n^E(x) = \begin{cases} 0 & \text{if } x < -2.45 \\ \frac{1}{4} & \text{if } -2.45 \leq x < -0.27 \\ \frac{1}{2} & \text{if } -0.27 \leq x < 0.75 \\ \frac{3}{4} & \text{if } 0.75 \leq x < 1.26 \\ 1 & \text{if } 1.26 \leq x \end{cases}$$

This EDF is displayed in Figure 2.

It is important to note that the realization of the EDF itself is a valid cdf as it satisfies all conditions of Theorem 1.5.3. Given the importance of this result, we formally state this and prove it in the following theorem.

Theorem 1. *Consider a sequence of real numbers $x_1, \ldots, x_n$ and let $\hat{F}_n^E(x)$ be defined as (the realization of) the empirical distribution function corresponding to these points, that is*

$$\hat{F}_n^E(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(x_i).$$

Then $\hat{F}_n^E(x)$ is a cdf.

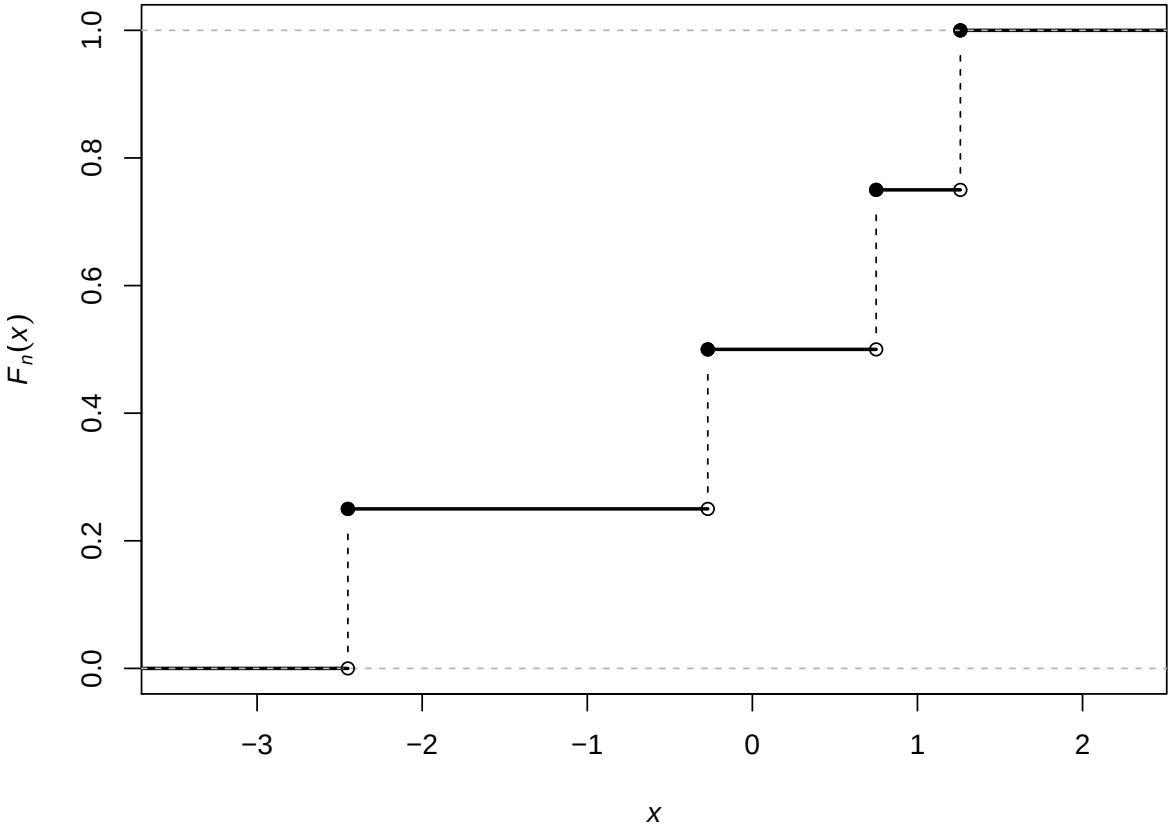


Figure 2: The EDF for the sample $x_1 = 1.26$, $x_2 = -2.45$, $x_3 = 0.75$ and $x_4 = -0.27$

Proof. We verify that $\hat{F}_n^E(x)$ satisfies all conditions of Theorem 1.5.3. For condition (a), note that for any $x < \min_i x_i$, we have that $I_{(-\infty,x]}(x_i) = 0$ for all $i = 1, \ldots, n$ and therefore $\hat{F}_n^E(x) = 0$. As $-\infty < \min_i x_i$, this means $\lim_{x \to -\infty} \hat{F}_n^E(x) = 0$. Similarly, for any $x \geq \max_i x_i$ we have that $I_{(-\infty,x]}(x_i) = 1$ for all $i = 1, \ldots, n$ and therefore $\hat{F}_n^E(x) = 1$, implying that $\lim_{x \to \infty} \hat{F}_n^E(x) = 1$.

For condition (b), take any two points $a, b \in \mathbb{R}$ where $a < b$. Now note that $\hat{F}_n^E(b) - \hat{F}_n^E(a) = \frac{1}{n} \sum_{i=1}^n \left[I_{(-\infty,b]}(x_i) - I_{(-\infty,a]}(x_i) \right]$. We prove that $\hat{F}_n^E(b) - \hat{F}_n^E(a) \geq 0$ for all a and b satisfying $a < b$ by contradiction. Assume that there is some $a < b$ such that $\hat{F}_n^E(b) - \hat{F}_n^E(a) < 0$. For this to be true there must be some i for which $I_{(-\infty,b]}(x_i) - I_{(-\infty,a]}(x_i) < 0$. As the indicator function only takes values 0 and 1, it must be that $I_{(-\infty,b]}(x_i) = 0$ and $I_{(-\infty,a]}(x_i) = 1$. However, these two statements can only be true if $b < x_i \leq a$, which contradicts that $a < b$. This proves condition (b).

To prove that condition (c) holds, let us first order $x_1, \ldots, x_n$ from low to high. Remember from Definition 5.4.1 that we denote these *order statistics* as $x_{(1)} < \ldots < x_{(n)}$, where $x_{(k)}$ corresponds to the k-th smallest among $x_1, \ldots, x_n$. Now note that $\hat{F}_n^E(x) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty,x]}(x_i) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty,x]}(x_{(i)})$. Take an x_0 such that $x_{(k-1)} \leq x_0 < x_{(k)}$ for some $k = 2, \ldots, n$. As exactly $k - 1$ values of $x_1, \ldots, x_n$ are smaller than or equal to x_0, we know

that $\hat{F}_n^E(x_0) = \frac{k-1}{n}$.

Next we consider the limit for $x \downarrow x_0$. Note that $\lim_{x\downarrow x_0} \hat{F}_n^E(x) = \frac{1}{n}\sum_{i=1}^n \lim_{x\downarrow x_0} I_{(-\infty,x]}(x_{(i)})$. As $x > x_0 \geq x_{(k-1)}$, we have that $\lim_{x\downarrow x_0} I_{(-\infty,x]}(x_{(i)}) = 1$. It then remains to show that for $i \geq k$, $\lim_{x\downarrow x_0} I_{(-\infty,x]}(x_{(i)}) = 0$. If we can show that for every $\epsilon > 0$ there exists a $\delta > 0$ such that for all $x_0 < x < x_0 + \delta$, $I_{(-\infty,x]}(x_{(k)}) < \epsilon$, the result then immediately follows for $i > k$. Take δ such that $0 < \delta < x_{(k)} - x_0$. For all $x < x_0 + \delta < x_{(k)}$ we then have that $I_{(-\infty,x]}(x_{(k)}) = 0 < \epsilon$ for all $\epsilon > 0$. This proves that for $i \geq k$, $\lim_{x\downarrow x_0} I_{(-\infty,x]}(x_{(i)}) = 0$ and consequently that $\lim_{x\downarrow x_0} \hat{F}_n^E(x) = \hat{F}_n^E(x_0)$ for $x_{(k-1)} \leq x_0 < x_{(k)}$.

As we assumed an arbitrary $k = 2, \ldots, n$, this result holds for each of these intervals, and therefore for any x_0 such that $x_{(1)} \leq x_0 < x_{(n)}$. In exactly the same way we can prove the result for x_0 in the intervals $(-\infty, x_{(1)})$ and $[x_{(n)}, \infty)$ to conclude the proof.[8] □

Now that we established that $\hat{F}_n^E(x)$ is a cdf (upon observing a realization of a sample), we can also discuss random variables with the distribution $\hat{F}_n^E(x)$. What does such a random variable look like? First note that $\hat{F}_n^E(x)$ is a discrete distribution function, which jumps up by an amount of $\frac{1}{n}$ whenever $x = x_i$ for $i = 1, \ldots, n$. From this it directly follows that the corresponding pmf has the form $\hat{f}_n^E(x) = 1/n$ for $x = x_1, \ldots, x_n$, and $\hat{f}_n^E(x) = 0$ otherwise. The following lemma formalizes this notion and derives some further properties.

Lemma 2. *Let $\hat{F}_n^E(y) = \frac{1}{n}\sum_{i=1}^n I_{(-\infty,y]}(x_i)$ be the EDF corresponding to the sample $x_1, \ldots, x_n$ and define a random variable Y that has cdf $\hat{F}_n^E(y)$. Then*

(a) $\mathbb{P}_{\hat{F}_n^E}(Y = x_i) = \frac{1}{n}$ for $i = 1, \ldots, n$;

(b) $\mathbb{E}_{\hat{F}_n^E} g(Y) = \frac{1}{n}\sum_{i=1}^n g(x_i)$ for any function $g(\cdot)$.

Proof. Part (a) follows directly from the definition of the cdf and corresponding pmf. Part (b) then follows from part (a), as for a discrete random variable as Y we can write

$$\mathbb{E}_{\hat{F}_n^E} g(Y) = \sum_{i=1}^n g(x_i)\, \mathbb{P}_{\hat{F}_n^E}(Y = x_i) = \frac{1}{n}\sum_{i=1}^n g(x_i).$$

We can also use Riemann-Stieltjes integrals to conclude that $\mathbb{E}_{\hat{F}_n^E} g(Y) = \int_{-\infty}^{\infty} g(y) d\hat{F}_n^E(y) = \frac{1}{n}\sum_{i=1}^n g(x_i)$. □

Note that if we draw a variable Y from the distribution $\hat{F}_n^E(y)$, then by Lemma 2(a), this variable takes on one of the values $x_1, \ldots, x_n$ with equal probability. If we draw a random sample of size n, say $Y_1, \ldots, Y_n$, then each of these variables take on one of the values $x_1, \ldots, x_n$ with equal probability. Note that this is equivalent to what we did in Example 2, that is,

[8] Note that it is crucial for condition (c) that we consider intervals of the form $(-\infty, x]$ rather than $(-\infty, x)$ in the indicator functions used for the EDF. If we use open intervals the function is not right-continuous as $1 = \lim_{x\downarrow x_i} I_{(-\infty,x)}(x_i) \neq I_{(-\infty,x_i)}(x_i) = 0$.

drawing a sample of size n from the original sample with replacement. Hence, we can link the bootstrap directly to the EDF. Before we do so formally, we consider the properties of the EDF as a estimator of the true distribution.

2.4 Consistency of the EDF

As a next step we now consider the properties of $\hat{F}_n^E(x)$ as a *random variable*, or more specifically, as an estimator of $F(x)$. In light of our use for the bootstrap, here we restrict ourselves to consistency. Before we considered only properties of $\hat{F}_n^E(x)$ upon observing a particular realization $x_1, \ldots, x_n$; in that case $\hat{F}_n^E(x)$ was an estimate rather than an estimator. Now, in order to study its properties as an estimator, we let $X_1, \ldots, X_n$ be a random sample from a population with pmf or pdf $f(x)$, and corresponding cdf $F(x)$. For every $x \in \mathbb{R}$, it then follows directly from the weak law of large numbers that

$$\hat{F}_n^E(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(X_i) \xrightarrow{p} F(x),$$

and it even holds almost surely by applying the strong law of large numbers.[9]

This result would be sufficient if we were only interested in a single point x, however typically we are not. Consistency of the EDF is simply a stepping stone to show consistency of the plug-in estimator that uses it, or later, to show consistency of the bootstrap. For both purposes we need the function $\hat{F}_n^E(x)$ over all points, rather than at a single point. Therefore a stronger result is needed: that of *uniform* convergence.

Definition 4. Let $f(x)$ be a function and $f_n(x)$ be a sequence of functions defined on $x \in \mathbb{R}$. We say that $f_n(x)$ **converges uniformly** to $f(x)$ if

$$\sup_{x \in \mathbb{R}} |f_n(x) - f(x)| \to 0.$$

as $n \to \infty$. If $f_n(x)$ is a sequence of random function, almost sure uniform convergence (uniform convergence in probability) holds if $\sup_{x \in \mathbb{R}} |f_n(x) - f(x)| \xrightarrow{a.s.} 0$ ($\sup_{x \in \mathbb{R}} |f_n(x) - f(x)| \xrightarrow{p} 0$).

The difference between uniform convergence, and the regular, so-called pointwise convergece, is that we take the supremum over all x. As such uniform convergence is stronger than pointwise convergence. The Glivenko-Cantelli Theorem (here stated without proof) shows that this indeed holds for the EDF.

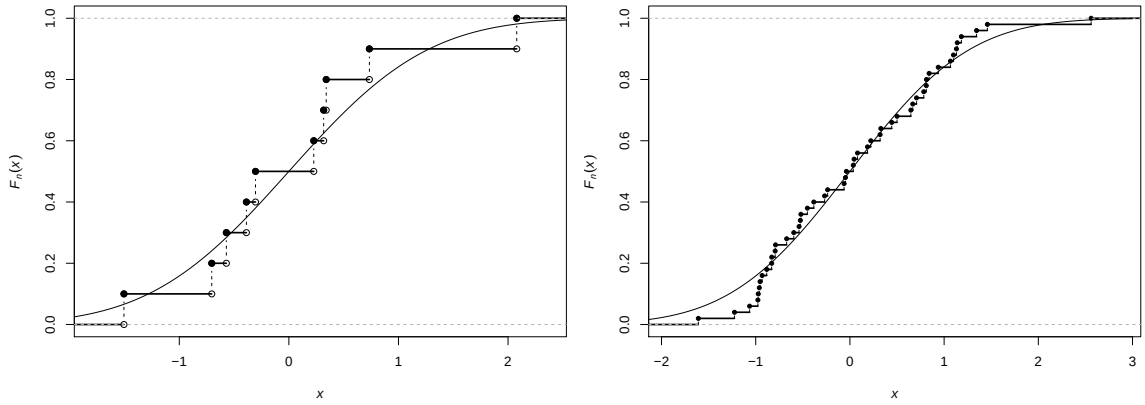
Theorem 2 (Glivenko-Cantelli)**.** *Let $X_1, \ldots, X_n$ be a random sample from a population with pmf or pdf $f(x)$, and corresponding cdf $F(x)$. Let $\hat{F}_n^E(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(X_i)$ be the EDF*

[9] Realize that $Y_i = I_{(-\infty,x]}(X_i)$ is a Bernoulli distributed random variable with success probability $F(x)$, and that $Y_1, \ldots, Y_n$ are independent. It then follows that $\hat{F}_n^E(x) = \overline{Y}_n \xrightarrow{p} F(x)$.

based on this sample. Then

$$\sup_{x\in\mathbb{R}}\left|\hat{F}_n^E(x) - F(x)\right| \xrightarrow{a.s.} 0.$$

Example 10. To illustrate the convergence of $\hat{F}_n^E(x)$ to $F(x)$, consider a random sample $X_1, \ldots, X_n$ from a $N(0,1)$ distribution. Figure 3 shows the EDF for four such samples, based on different sample sizes. The specific samples are drawn using a computer to simulate the $N(0,1)$ distribution. While the top left panel (a) still shows a big discrepancy between the EDF and the true cdf (the smooth line), this rapidly decreases when n increases; for $n = 500$ in the bottom right panel (d), hardly any difference is visible anymore.



(a) $n = 10$ (b) $n = 50$

(c) $n = 100$ (d) $n = 500$

Figure 3: The EDF for samples from a $N(0,1)$ distribution; the smooth line is the true cdf.

This result can now be used to prove consistency of various plug-in estimators. Obviously, for many of them, such as the plug-in moment estimators, there is no need to follow this line of proving consistency, as a direct proof will be easier. However, when we consider more complex applications of the plug-in principle, such as the bootstrap, the Glivenko-Cantelli

Theorem provides a justification why the bootstrap works.

3 The Bootstrap

3.1 Definition through the plug-in principle

We are now ready to formally describe and motivate the bootstrap. The tool we will use to do so is the plug-in principle developed in the previous section. Assume we observe a random sample $X_1, \ldots, X_n$ from a cdf F and let us be interested in the distribution of a certain function of the sample, and possibly of an (unknown) parameter θ. Let us call this quantity $Q'_n(\mathbf{X}, \theta)$. One option for $Q'_n(\mathbf{X}, \theta)$ could simply be a sample statistic, for example $\overline{X}_n$. In that case $Q'_n(\mathbf{X}, \theta) = Q'_n(\mathbf{X})$ is not a function of θ. Alternatively, we can consider an asymptotically pivotal quantity, as in Examples 1 and 2 where $Q'_n(\mathbf{X}, \mu) = \sqrt{n}\frac{\overline{X}_n - \mu}{S_n}$. We make one further change of notation. As explained in the previous section, any population parameter θ is a function of the population distribution F, i.e. $\theta = \tau(F)$. Therefore we write the quantity in terms of F rather than θ as $Q_n(\mathbf{X}, F)$. The two notations can easily be linked as

$$Q'_n(\mathbf{X}, \theta) = Q'_n(\mathbf{X}, \tau(F)) = Q_n(\mathbf{X}, F). \tag{5}$$

We want to obtain (an approximation to) the unknown distribution of $Q_n(\mathbf{X}, F)$, let us call its cdf $G_n(x, F)$, i.e. $G_n(x, F) = \mathbb{P}_F(Q_n(\mathbf{X}, F) \leq x)$. We add the subscript '$n$' to emphasize this is the exact cdf (i.e. not the asymptotic one). If $Q_n(\mathbf{X}, F)$ is an asymptotic pivot, then $\lim_{n\to\infty} G_n(x, F_1) = \lim_{n\to\infty} G_n(x, F_2)$ for different distributions F_1 and F_2; that is, the asymptotic distribution does not depend on F. This is however not necessary to assume at this point.

Example 11. In Examples 1 and 2 we looked at $Q_n(\mathbf{X}, F) = \sqrt{n}\frac{\overline{X}_n - \mu}{S_n}$. In that case $\lim_{n\to\infty} G_n(x, F) = \Phi(x)$, where $\Phi(x)$ is the cdf of a standard normal random variable. So this is indeed an asymptotic pivot. Alternatively, had we considered $Q_n^{\#}(\mathbf{X}, F) = \sqrt{n}(\overline{X}_n - \mu)$, then $\lim_{n\to\infty} G_n^{\#}(x, F) = \Phi(x/\sigma)$, and as such this is not an asymptotic pivot.

Unless strong assumptions are made on F, such as normality, $G_n(x, F)$ is typically unknown and cannot be used for inference. The standard way to proceed is to use $\lim_{n\to\infty} G_n(x, F) = G_\infty(x, F)$ for inference, provided $G_\infty(x, F)$ does not depend on F. Example 1 showed this is not without dangers, as the asymptotic approximation may not be very accurate.

The bootstrap offers an alternative through application of the plug-in principle. Just as we did for estimating parameters, we can plug in an estimate of F, let us call it $\hat{F}_n$, into the formula for G_n. This is made formal in the following definition of the bootstrap.

Definition 5. Consider a random sample $X_1, \ldots, X_n$ from a population with cdf F. Let our quantity of interest be $Q_n(\mathbf{X}, F)$ and denote its cdf by $G_n(x, F)$. Let $\hat{F}_n(x)$ be an estimator

of F that satisfies all conditions for being a cdf. The **bootstrap estimator** of $G_n(x, F)$ is defined as $G_n(x, \hat{F}_n)$.

It might be odd to think of $G_n(x, \hat{F}_n)$ as an estimator. It is however, a function of the sample $X_1, \ldots, X_n$ only, as $\hat{F}_n(x)$ is an estimator, and therefore a function of $X_1, \ldots, X_n$ only. We also call $G_n(x, \hat{F}_n)$ the *bootstrap distribution*, which hides the fact that it is an estimator. It is important for the further development of the bootstrap to keep in mind though that it is, statistically speaking, an estimator.

Unlike the case where we estimate parameters through the plug-in principle, we typically cannot find an analytical expression for $G_n(x, \hat{F}_n)$. However, we can approximate it with arbitrary accuracy using simulation. This is what is typically done using the bootstrap, as described in the algorithm below.

Algorithm 1 (Simulating the bootstrap distribution)**.** Follow the following steps to obtain the approximation to the bootstrap distribution $G_n(x, \hat{F}_n)$. In steps 1 and 2 below, the superscript b, which takes values $1, 2, \ldots, B$, describes one specific bootstrap simulation. These steps 1 and 2 should therefore be repeated B times.

1. Draw a random sample from the cdf $\hat{F}_n$. The realized sample $x_1^{*b}, \ldots, x_n^{*b}$ is your *bootstrap sample*.

2. Calculate the bootstrap version of $Q_n(\mathbf{x}, F)$, which is $Q_n(\mathbf{x}^{*b}, \hat{F}_n)$. We also denote this as short-hand by Q_n^{*b}.

3. After repeating steps 1 and 2 B times, collect all the calculated bootstrap quantities $Q_n^{*1}, Q_n^{*2}, \ldots, Q_n^{*B}$. These form the approximation to the bootstrap distribution.

We can now use $Q_n^{*1}, Q_n^{*2}, \ldots, Q_n^{*B}$ to calculate whatever aspect of $G_n(x, \hat{F}_n)$ we need for inference.

Remark 1. It is common to denote bootstrap quantities with a superscript '$*$', as we did above. Typically people leave out the 'b' which indicates the specific bootstrap simulation iteration, unless it is really needed when describing a computational algorithm. We follow the same convention here,only adding a superscript 'b' when we need to stress the specific simulation it comes from. In addition, we again distinguish between random variables by using capital letters, and specific realizations using small letters. Hence, $X_1^*, \ldots, X_n^*$ denotes a general random sample from the distribution $\hat{F}_n$. In the bootstrap context, when we discuss specific realizations, it is often in the context of a computational algorithm. Therefore, we will only need the superscript 'b' when considering specific realizations. Hence, we will generally write $X_1^*, \ldots, X_n^*$ without superscript 'b' but a specific realization $x_1^{*b}, \ldots, x_n^{*b}$ with superscript 'b', although there are occasions where we need to deviate from this convention.

How the sample in step 1 of Algorithm 1 is drawn, depends on the form of the estimator $\hat{F}_n$. As for parameter estimation, we can make the distinction between parametric estimation, in which case a parametric family $F(x|\theta)$ is assumed and only the parameter θ needs to be estimated, or nonparametric estimation, in which no assumption on the distribution is made and the estimator is the EDF. These lead to the following two bootstrap methods:

1. **Parametric bootstrap** [F *is estimated by* $F(x|\hat{\theta}_n)$]: The bootstrap sample $X_1^*, \ldots, X_n^*$ is drawn from the cdf $F(x|\hat{\theta}_n)$, where $\hat{\theta}_n$ is an estimate of θ based on the original sample $x_1, \ldots, x_n$.

2. **Nonparametric (iid) bootstrap** [F *is estimated by the EDF* $\hat{F}_n^E$]: The bootstrap sample $X_1^*, \ldots, X_n^*$ is drawn from the cdf $\hat{F}_n^E$, which by Lemma 2 implies that $X_1^*, \ldots, X_n^*$ is drawn with replacement from the original sample $x_1, \ldots, x_n$.

When people talk about "the bootstrap", they typically mean the nonparametric bootstrap as described above. The nonparametric bootstrap has become by far the more popular bootstrap version because it does not require one to make an (unrealistic) assumption about the distribution F, and moreover, it typically is not even much less accurate than the parametric bootstrap when the parametric assumption is correct. We will see this illustrated later.

There are many other bootstrap methods that are applicable in settings where the random sample assumption is not appropriate. We will not discuss these here, but to avoid confusion later on, we will call the nonparametric bootstrap *iid bootstrap* from here on. This way we can distinguish it from other nonparametric forms of the bootstrap, highlighting the only real assumption needed to apply it: that the sample is iid.

3.2 Probabilistic calculations in the bootstrap

One of the more complicated and confusing things about the bootstrap is to properly understand and perform probabilistic calculations with bootstrap quantities. In terms of notation, just as done for the bootstrap sample, people typically append a '$*$' to probability or expectation operators to signify that those probabilities should be taken in the bootstrap world. While this looks nice, the simple notation does obscure the complicated matter quite a bit. In particular, it hides the fact that, as discussed below Definition 5, the bootstrap distribution is an estimator and thus a random variable. Similarly, this means that any probability calculation in the bootstrap need to be performed conditionally on observing the sample $\mathbf{X} = \mathbf{x}$. Here we look at this in more detail.

Let us first introduce a bit of notation. For a random variable X with cdf F, we explicity append a subscript 'F' to probabilities or expectations, as in $\mathbb{P}_F(X \leq x)$ and $\mathbb{E}_F(X)$. This reminds us what the relevant distribution is with respect to which we calculate the probabilities. This addition makes it easier to make the transition to the bootstrap world.

Now assume a random sample $X_1, \ldots, X_n$ with cdf F. Suppose we have an estimator $\hat{F}_n$ for F – for example the EDF – and assume that we draw a bootstrap sample $X_1^*, \ldots, X_n^*$ from this distribution $\hat{F}_n$. The distribution of the bootstrap sample is *conditional* on $X_1, \ldots, X_n$, or if we want to condition on an observed outcome, conditional on observing that $X_1 = x_1, \ldots, X_n = x_n$. If we now calculate a probability or expectation for the bootstrap sample – still conditional on the original sample – we need to do so with respect to the distribution $\hat{F}_n$ rather than F. For instance, the expectation of X_i^* – conditional on the sample $\mathbf{X}$ – can be written as

$$\mathbb{E}^* X_i^* = \mathbb{E}_{\hat{F}_n}(X_i^* | \mathbf{X}) = \int x d\hat{F}_n(x).$$

Let us consider the two equality signs in turn. The first equality sign simply tells us what we mean with the notation $\mathbb{E}^*$: (i) the expectation is with respect to the distribution $\hat{F}_n$ and (ii) the expectation is conditional on the original sample. The second equality sign simply uses the definition of an expectation using the Riemann-Stieltjes integral. How we now calculate the expectation, depends on the choice of $\hat{F}_n$. For the iid bootstrap where $\hat{F}_n$ is the EDF $\hat{F}_n^E$, we have from Lemma 2(b) that

$$\mathbb{E}^* X_i^* = \mathbb{E}_{\hat{F}_n^E}(X_i^* | \mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

For a parametric choice of $\hat{F}_n$ one simply considers the expectation of that distribution, but then with an estimated parameter rather than the true one.

For probabilities things work exactly the same. For instance, again with $\hat{F}_n = \hat{F}_n^E$, we have that

$$\mathbb{P}^*(X_i^* \leq c) = \mathbb{P}_{\hat{F}_n^E}(X_i^* \leq c | \mathbf{X}) = \hat{F}_n^E(c) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, c]}(X_i).$$

An important consequence of working conditionally on the sample, is that when we do not condition on a specific outcome $\mathbf{X} = \mathbf{x}$, but on the random variables $\mathbf{X}$, that bootstrap probabilities and expectations *are random variables themselves.* This is true for any conditional probability or expectation.

4 Applications of the Bootstrap

Once one has obtained the bootstrap distribution $G_n(x, \hat{F}_n)$, it can be used for inference. How this is done depends on the goal of the researcher. Here we discuss four possible uses, of which the last two are by far the most important.

4.1 Improving point estimation (bias reduction)

One of the most basic uses of the bootstrap is to improve the quality of point estimators. For a random sample $X_1, \ldots, X_n$, assume we have a certain parameter $\theta = \tau(F)$, and an estimator $\hat{\theta}_n = W_n(\mathbf{X})$ of θ that has a bias $A_n = \mathbb{E}_F(\hat{\theta}_n - \theta)$. The bootstrap can be used to estimate the bias and consequently to reduce the bias of $\hat{\theta}_n$. Define $Q_n(\mathbf{X}, \theta) = \hat{\theta}_n - \theta$, and let $G_n(x, F)$ be the corresponding cdf. Then we can write

$$\mathbb{B}\text{ias}_n = \mathbb{E}_F(\hat{\theta}_n - \theta) = \mathbb{E}_F\, Q_n(\mathbf{X}, F) = \int x dG_n(x, F). \tag{6}$$

Now assume that we have a bootstrap sample $X_1^*, \ldots, X_n^*$ with distribution $\hat{F}_n$ – conditional on the sample $\mathbf{X}$. The bootstrap version of the quantity of interest is $Q_n(\mathbf{X}^*, \hat{F}_n)$. How do we calculate this? Note that $Q_n(\mathbf{X}, \theta) = \hat{\theta}_n - \theta$. The first part is easy: to get the bootstrap version, replace $\hat{\theta}_n = W_n(\mathbf{X})$ which was calculated using the original sample $X_1, \ldots, X_n$, with $\hat{\theta}_n^* = W_n(\mathbf{X}^*)$, the same function but calculated with the bootstrap sample. For the second part we use that $\theta = \tau(F)$, and therefore the bootstrap equivalent is $\theta^* = \tau(\hat{F}_n)$. Putting the two steps above together, we get $Q_n(\mathbf{X}^*, \hat{F}_n) = \hat{\theta}_n^* - \theta^* = W_n(\mathbf{X}^*) - \tau(\hat{F}_n)$. We can then write down the bootstrap approximation to the bias:

$$\mathbb{B}\text{ias}_n^* = \mathbb{E}^*\left[Q_n(\mathbf{X}^*, \hat{F}_n)\right] = \mathbb{E}_{\hat{F}_n}\left[W(\mathbf{X}^*) - \tau(\hat{F}_n) |\, \mathbf{X}\right]. \tag{7}$$

The quantity $\mathbb{E}_{\hat{F}_n}[Q_n(\mathbf{X}^*, \hat{F}) |\, \mathbf{X}]$ typically cannot be calculated analytically. However, we also do not need to as we can use our bootstrap simulations for that! Assume that we can obtain $Q_n^{*1}, \ldots, Q_n^{*B}$ as in Algorithm 1. We can then approximate $\mathbb{B}\text{ias}_n^*$ by the simulation average

$$\mathbb{B}\text{ias}_{n,B}^* = \frac{1}{B}\sum_{b=1}^{B} Q_n^{*b} = \frac{1}{B}\sum_{b=1}^{B} Q_n(\mathbf{X}^{*b}, \hat{F}) = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^{*b} - \theta^* = \frac{1}{B}\sum_{b=1}^{B} W(\mathbf{X}^{*b}) - \tau(\hat{F}_n).$$

A simple application of the weak law of large numbers tells us that $\mathbb{B}\text{ias}_{n,B}^* \xrightarrow{p} \mathbb{B}\text{ias}_n^*$ as $B \to \infty$.

Now that we obtained an estimate for the bias, we can use it to construct a *bias-corrected* estimator. Ideally, if we knew the true bias $\mathbb{B}\text{ias}_n$, we could construct an estimator $\hat{\theta}_n' = \hat{\theta}_n - \mathbb{B}\text{ias}_n$ which by construction has zero bias. We can do the same with estimated bias, in the hope that at least the bias will be less. Hence, we construct the estimator

$$\hat{\theta}_n^{bc} = \hat{\theta}_n - \mathbb{B}\text{ias}_{n,B}^*\,.$$

The method described above requires us to calculate $\theta^* = \tau(\hat{F}_n)$. How to do this depends on the function $\tau(\cdot)$ and the choice of $\hat{F}_n$. For example, if $\theta = \mathbb{E}_F(g(X)) = \int g(x)dF(x)$, then $\theta^* = \mathbb{E}_{\hat{F}_n}(g(X^*) |\, \mathbf{X}) = \int g(x)d\hat{F}_n(x)$. If we then consider the nonparametric bootstrap with

$\hat{F}_n$ equal to the EDF, Lemma 2 tells us that $\theta^* = \frac{1}{n}\sum_{i=1}^n g(X_i)$.

Often the estimator $\hat{\theta}_n$ will coincide with θ^*. This is typically trivially true for the parametric bootstrap, but also for the nonparametric bootstrap. If we do not want to make an assumption on the parametric distributional family, we will most likely estimate θ using the plug-in principle, such that $\hat{\theta}_n = \tau(\hat{F}_n)$ and typically $\hat{F}_n$ is the EDF. In that case $\theta^* = \hat{\theta}_n$ and the estimator reduces to

$$\hat{\theta}_n^{bc} = \hat{\theta}_n - \mathbb{B}\text{ias}^*_{n,B} = \hat{\theta}_n - \left(\frac{1}{B}\sum_{b=1}^B \hat{\theta}^{*b} - \theta^*\right) = \hat{\theta}_n - \frac{1}{B}\sum_{b=1}^B \hat{\theta}^{*b} + \hat{\theta}_n = 2\hat{\theta}_n - \overline{\hat{\theta}^*_n},$$

where $\overline{\hat{\theta}^*_n} = \frac{1}{B}\sum_{b=1}^B \hat{\theta}^{*b}_n$. This is the form of the bootstrap bias-corrected estimator typically encountered in the literature.

4.2 Variance estimation

A second purpose of the bootstrap is variance estimation. Suppose that we know that

$$\sqrt{n}\frac{\hat{\theta}_n - \theta}{\sqrt{\mathbb{V}\text{ar}(\hat{\theta}_n)}} \xrightarrow{d} N(0,1),$$

but that this asymptotic pivot cannot be used, because $\mathbb{V}\text{ar}(\hat{\theta}_n)$ is unknown, and it is difficult to obtain an estimate of the variance. We can then use the bootstrap to estimate the variance.

Letting $Q_n(\mathbf{X}, F) = \hat{\theta}_n$, and consequently $Q_n(\mathbf{X}^*, \hat{F}_n) = \hat{\theta}^*_n$, we can estimate the variance of $\hat{\theta}_n$ by the bootstrap variance

$$\mathbb{V}\text{ar}^*(\hat{\theta}_n) = \mathbb{E}_{\hat{F}_n}\left[\left(\hat{\theta}^*_n - \mathbb{E}_{\hat{F}_n}\hat{\theta}^*_n\right)^2 \middle| \mathbf{X}\right].$$

In practice, having again followed Algorithm 1 to obtain $Q_n^{*1}, \ldots, Q_n^{*B}$, we let

$$\mathbb{V}\text{ar}^*_B(\hat{\theta}_n) = \frac{1}{B}\sum_{b=1}^B \left(\hat{\theta}^{*b}_n - \overline{\hat{\theta}^*_n}\right)^2.$$

While this application of the bootstrap was very popular in the early stages of the development, it is not used often anymore. Typically the estimate of the variance is not the final goal, but only an intermediate step towards for example a hypothesis test or confidence interval. In that case its use is to make the pivot $\sqrt{n}\frac{\hat{\theta}_n - \theta}{\sqrt{\mathbb{V}\text{ar}^*_B(\hat{\theta}_n)}}$ feasible. However, doing so still requires one to use the asymptotic normality of the pivot to do inference. This is a considerable drawback, especially since the bootstrap can be used to avoid this entirely, as we shall see for the next two applications.

4.3 Hypothesis testing

Assume we have a random sample $X_1, \ldots, X_n$ from a cdf F. We wish to perform a hypothesis test on a parameter $\theta = \tau(F)$. In particular, let $H_0 : \theta = \theta_0$, and H_1 either one-sided or two-sided. Let F_0 denote the "null distribution", that is a distribution that satisfies H_0, such that $\theta_0 = \tau(F_0)$. Typically, F_0 is not fully specified, as there may be many different distributions for which $\theta_0 = \tau(F_0)$. For instance, in Example 12 below, if $\theta = \mathbb{E}_F X$, any distribution which has mean θ_0 satisfies H_0.

Assume we have an asymptotic pivot $Q_n(\mathbf{X}, F)$ on which we would like to base the test. In particular, we define the test statistic $T_n(\mathbf{X}, \theta_0) = Q_n(\mathbf{X}, F_0)$ such that its asymptotic distribution $G_n(x, F)$ is fully known if $F = F_0$.[10] However, rather than using the asymptotic distribution, we want to use the bootstrap to obtain critical values, or equivalently, p-values.[11]

Again we can follow Algorithm 1 to obtain the bootstrap quantities, however there is one complication with hypothesis testing. Assuming that $H_1 : \theta > \theta_0$ – the other cases follow similarly – remember from Chapter 8 that we can define the critical value c_α for a test with size α as

$$\alpha = \mathbb{P}_{\theta_0}(T_n(\mathbf{X}, \theta_0) \geq c_\alpha) = \mathbb{P}_{F_0}(Q_n(\mathbf{X}, F_0) \geq c_\alpha) = 1 - G_n(c_\alpha, F_0),$$

and the p-value $p(\mathbf{x})$ for an observed sample $\mathbf{x}$ as

$$p(\mathbf{x}) = \mathbb{P}_{\theta_0}(T_n(\mathbf{X}, \theta_0) \geq T_n(\mathbf{x}, \theta_0)) = \mathbb{P}_{F_0}(Q_n(\mathbf{X}, F_0) \geq Q_n(\mathbf{x}, F_0)) = 1 - G_n(Q_n(\mathbf{x}, F_0), F_0).$$

In either case, what is required is the probability if the null hypothesis is true. Hence, to estimate these using the bootstrap we need to approximate F_0 rather than F. What this practically means is that in the bootstrap we must make sure the null hypothesis is indeed true. We have two options to achieve this:

1. Adapt the null hypothesis in the bootstrap such that it is satisfied. If $\hat{F}_n$ is the estimate of F used, construct the test statistic to test $H_0 : \theta^* = \theta_0^*$, where $\theta_0^* = \tau(\hat{F}_n)$, rather than testing $H_0 : \theta^* = \theta_0$. In this case the "standard" bootstrap quantity $Q_n(\mathbf{X}^*, \hat{F}_n)$ can be used.

2. Adapt your estimator $\hat{F}_n$ to satisfy the null hypothesis. Let $\hat{F}_{0,n}$ denote an estimator of F that satisfies H_0, in the sense that $Q_n(\mathbf{X}^*, \hat{F}_{0,n}) = T_n(\mathbf{X}^*, \theta_0)$. Then $X_1^*, \ldots, X_n^*$ are drawn from $\hat{F}_{0,n}$ and therefore satisfy H_0.

[10] Even if F_0 is not uniquely defined, i.e. there are multiple distributions that satisfy H_0, due to the pivot that we use the asymptotic distribution is still fully known.

[11] We can treat a composite null hypothesis such as $H_0 : \theta \leq \theta_0$ as if it were $H_0 : \theta = \theta_0$ for the purposes of the bootstrap; as explained in Chapter 8, only the parameter value on the border θ_0 matters for obtaining critical values or p-values.

For a parametric bootstrap this is not really an issue. For option 1, if $\hat{\theta}_n$ is the estimated parameter of the assumed parametric family, one can simply take $H_0 : \theta^* = \hat{\theta}_n$ in the bootstrap. For option 2, one can generate the bootstrap sample from $F(x|\theta_0)$.

For the nonparametric bootstrap things are a bit more involved. Typically, the EDF $\hat{F}_n^E$ does not satisfy the null hypothesis and so, if the nonparametric bootstrap is used, one of the two corrections has to be employed. What this entails is best illustrated by an example.

Example 12. Let $X_1, \ldots, X_n$ be a random sample with cdf F. As in Examples 1 and 2, let $\mu = \mathbb{E}_F X$ and assume we want to test $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$. Consider the asymptotic pivot

$$Q_n(\mathbf{X}, F) = \sqrt{n}\frac{\overline{X}_n - \mu}{S_n} \xrightarrow{d} N(0,1).$$

Now let F_0 be such that $\mu_0 = \mathbb{E}_{F_0} X$,[12] and consider the test statistic

$$T_n(\mathbf{X}, \mu_0) = Q_n(\mathbf{X}, F_0) = \sqrt{n}\frac{\overline{X}_n - \mu_0}{S_n},$$

where we reject for large values of $T_n(\mathbf{X}, \mu_0)$.

If we draw the bootstrap sample $X_1^*, \ldots, X_n^*$ from the EDF $\hat{F}_n^E$, then we have seen before that $\mu^* = \mathbb{E}^* X^* = \mathbb{E}_{\hat{F}_n^E}(X^*|\,\mathbf{X}) = \overline{X}_n$. So, by construction, $\mu^* \neq \mu_0$ with probability 1, and a correction has to be applied. For option 1 we adapt the null hypothesis in the bootstrap. Hence we need to test $H_0 : \mu^* = \mu_0^*$ versus $H_1 : \mu^* > \mu_0^*$, where $\mu_0^* = \overline{X}_n$. Note that this statement only makes sense by conditioning on a realization $\mathbf{X}$, otherwise we test if a parameter is equal to a random variable.[13] The bootstrap statistic then looks like

$$T_n(\mathbf{X}^*, \mu_0^*) = Q_n(\mathbf{X}^*, \hat{F}_n^E) = \sqrt{n}\frac{\overline{X}_n^* - \overline{x}_n}{S_n^*}.$$

For option 2, we need to draw the bootstrap sample from a different distribution. One option is the following. Let $Z_i = X_i - \overline{X}_n$ for $i = 1, \ldots, n$, then $\bar{Z}_n = 0$, and define $Y_i = Z_i + \mu_0 = X_i - \overline{X}_n + \mu_0$ for $i = 1, \ldots, n$, which implies that $\overline{Y}_n = \mu_0$.

Now draw the bootstrap sample $X_1^*, \ldots, X_n^*$ from the EDF of $Y_1, \ldots, Y_n$, let us denote it as $\hat{F}_n^{E,y}$. In that case $\mathbb{E}^* X^* = \mathbb{E}_{\hat{F}_n^{E,y}}(X^*|\,\mathbf{X}) = \overline{Y}_n = \mu_0$ and the null hypothesis is satisfied. Hence, we know that $\hat{F}_n^{E,y}$ satisfies H_0, such that we can write $\hat{F}_n^{E,y} = \hat{F}_{0,n}$. We can then use the bootstrap statistic

$$T_n(\mathbf{X}^*, \mu_0) = Q_n(\mathbf{X}^*, \hat{F}_{0,n}) = \sqrt{n}\frac{\overline{X}_n^* - \mu_0}{S_n^*}$$

[12] As explained above, F_0 is not fully specified, as there are many different distributions that have the same mean. Even if we assume a parametric distributional family F_0 may not be fully specified; for instance, any normal distribution with mean μ_0 satisfies H_0, regardless of the value of σ^2.

[13] Formally we should write that conditionally on observing the sample $\mathbf{X} = \mathbf{x}$, we test $H_0 : \mu^* = \overline{x}$.

to perform the test.

A careful inspection of Example 12 reveals that in that specific case the two options are identical – the only difference being the point at which the sample mean is subtracted – but in general this is not the case. There is no general consensus on which approach is better. Typically the difference between the two approaches is small and the preferred option depends on the specific application, and very often simply on which one is easier to implement.

Whichever way we set up $Q_n^* = T_n(\mathbf{X}^*, \theta_0^*) = Q_n(\mathbf{X}^*, \hat{F}_n)$, the bootstrap critical values c_α^* can then be obtained using the relation

$$\mathbb{P}^*(Q_n^* \geq c_\alpha^*) = \mathbb{P}_{\hat{F}_n}(Q_n(\mathbf{X}^*, \hat{F}_n) \geq c_\alpha^* | \, \mathbf{X}) = \alpha.$$

Note that c_α^*, as all quantities with a '$*$', is a random variable as it depends on the sample $\mathbf{X}$. So, formally, we should write $c_{\hat{F}_n,\alpha}(\mathbf{X})$, but to lighten the notation we just write c_α^* instead.

It then depends on the type of test how we proceed:

- In case of a right-tailed test we obtain c_α^* as above and reject H_0 if $T_n(\mathbf{x}, \theta_0) > c_\alpha^*$.
- For a left-tailed test we get the left-tail cutoff point $c_{1-\alpha}^*$ in the same way, and reject H_0 if $T_n(\mathbf{x}, \theta_0) < c_{1-\alpha}^*$
- For a two-tailed test, there are two options. First we consider the *equal-tailed* test. We do not make an assumption about the symmetry of the distribution of the test statistic, and use separate critical values for the lower and upper tail, $c_{1-\alpha/2}^*$ and $c_{\alpha/2}^*$ respectively. Both are obtained in the same way as above. We reject H_0 if either $T_n(\mathbf{x}, \theta_0) > c_{\alpha/2}^*$ or if $T_n(\mathbf{x}, \theta_0) < c_{1-\alpha/2}^*$.
- Alternatively, if one believes the distribution of the test statistic to be symmetric under H_0, one can obtain the critical value as the $1-\alpha$ percentile of the absolute values of Q_n^*. In this case we define c_α^* by the relation

  $$\mathbb{P}^*(|Q_n^*| \geq c_\alpha^*) = \mathbb{P}_{\hat{F}_n}\left(\left|Q_n(\mathbf{X}^*, \hat{F}_n)\right| \geq c_\alpha^* \middle| \, \mathbf{X}\right) = \alpha,$$

  and we reject H_0 if $|T_n(\mathbf{x}, \theta_0)| > c_\alpha^*$.

In practice, we can again approximate c_α^* from the collection $Q_n^{*1}, \ldots, Q_n^{*B} = T_n^{*1}, \ldots, T_n^{*B}$ that we obtained by following Algorithm 1. The approximate bootstrap critical value $c_{\alpha,B}^*$ for a level α test is simply obtained as the appropriate percentile of the $Q_n^{*1}, \ldots, Q_n^{*B}$. In particular,

$$c_{\alpha,B}^* = Q_n^{*([(1-\alpha)B])},$$

where $Q_n^*([(1-\alpha)B])$ is the $(1-\alpha)B$-th order statistic of $Q_n^{*1}, \ldots, Q_n^{*B}$. Exactly the same reasoning can be applied to find the left-tail critical value. For the symmetric two-tailed, one can obtain the critical value as the $1-\alpha$ percentile of the absolute values of $Q_n^{*1}, \ldots, Q_n^{*B}$.

The bootstrap can also be used to obtain p-values. These bootstrap p-value for a right-tailed test is given by

$$p^*(\mathbf{X}) = \mathbb{P}^*(Q_n^* \geq Q_n(\mathbf{x}, F_0)) = \mathbb{P}_{\hat{F}_n}(Q_n(\mathbf{X}^*, \hat{F}_n) \geq Q_n(\mathbf{X}, F_0) | \, \mathbf{X}),$$

with similar definitions for the other tests. In practice, they are simply calculated by counting how many bootstrap statistics are larger (or smaller, depending on the direction of the test) than the original statistic $t_n = T_n(\mathbf{x}, \theta_0)$ for the observed sample $\mathbf{x}$. Specifically, for a

- right-tail test: $p_{r,B}^*(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^B I_{(t_n,\infty)}(Q_n^{*b})$.
- left-tail test: $p_{l,B}^*(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^B I_{(-\infty,t_n)}(Q_n^{*b})$.
- two-tailed test (equal-tailed): $p_{e,B}^*(\mathbf{x}) = 2\min\{p(\mathbf{x})_l^*, p(\mathbf{x})_r^*\}$.
- two-tailed test (symmetric): $p_{s,B}^*(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^B I_{(|t_n|,\infty)}(\left|Q_n^{*b}\right|)$.

Before we end the section on hypothesis testing, let us make one final remark. So far we assumed that the test statistic $T_n(\mathbf{X}, \theta_0)$ was based on an asymptotically pivotal quantity, i.e. its asymptotic distribution does not depend on nuisance parameters. While this is necessary if one wishes to apply asymptotic results, this is in fact not necessary for the bootstrap!

In Example 12 for instance, we could have taken $T_n'(\mathbf{X}, \mu_0) = \sqrt{n}(\overline{X}_n - \mu_0)$. As $T_n'(\mathbf{X}, \mu_0) \xrightarrow{d} N(0, \sigma^2)$ we cannot use this quantity for our standard asymptotic analysis unless we know $\sigma^2 = \mathbb{V}\mathrm{ar}\, X$. However, the bootstrap can deal with this automatically – just put this quantity into Algorithm 1 and a bootstrap distribution comes out regardless of knowing σ^2. This is because implicitly the bootstrap automatically provides us with a (plug-in) estimate of σ^2. While in general it is much preferable to indeed use pivotal quantities if possible, the bootstrap can be used without. In situations where it is difficult to find a pivotal quantity, this is a major advantage of the bootstrap.

Example 13. Let us again consider the setting of Example 12 and test $H_0 : \mu \leq \mu_0$ vs. $H_1 : \mu > \mu_0$. We showed that for the nonparametric bootstrap using the EDF $\hat{F}_n^E$, we have

$$T_n(\mathbf{X}^*, \mu_0^*) = Q_n(\mathbf{X}^*, \hat{F}_n^E) = \sqrt{n}\frac{\overline{X}_n^* - \overline{X}_n}{S_n^*}.$$

The full bootstrap algorithm then looks as follows.

1. For $b = 1, 2, \ldots, B$, draw a sample $x_1^{*b}, \ldots, x_n^{*b}$ with replacement from $x_1, \ldots, x_n$.

2. Calculate the bootstrap quantity

$$Q_n^{*b} = Q_n(\mathbf{x}^{*b}, \hat{F}_n^E) = T_n(\mathbf{x}^{*b}, \mu_0^*) = \sqrt{n}\frac{\overline{x}_n^{*b} - \overline{x}_n}{s_n^{*b}}.$$

3. Repeat steps 1 and 2 B times and collect all bootstrap quantities $Q_n^{*1}, Q_n^{*2}, \ldots, Q_n^{*B}$. Let

$$c_{\alpha,B}^* = Q_n^{*([(1-\alpha)B])}.$$

4. Reject H_0 if $Q_n(\mathbf{x}, F_0) > c_{\alpha,B}^*$.

4.4 Confidence intervals

Next we consider the construction of confidence intervals using the bootstrap. There are a few different intervals one can consider. Here we treat them in turn. For clarity, assume we have a random sample $X_1, \ldots, X_n$ with cdf F and our parameter of interest is $\theta = \tau(F)$. Let $\hat{\theta}_n = W_n(\mathbf{X})$ be an estimator of θ.

4.4.1 Equal-tailed percentile intervals

The first type of interval we consider, the so-called *percentile interval*, is applied directly to $\hat{\theta}_n = W_n(\mathbf{X})$ without attempting to find a pivotal quantity. While not optimal, it is a very easy interval to construct and motivate. Hence, we let $Q_n(\mathbf{X}, F) = \hat{\theta}_n - \theta$.

We first derive the infeasible interval based on knowledge of $G_n(x, F)$, the cdf of $Q_n(\mathbf{X}, F)$. Afterwards we then only have to plug in $\hat{F}_n$ for F.[14] Define the cut-off point c_α such that $\mathbb{P}_F(Q_n(\mathbf{X}, F) \geq c_\alpha) = \alpha$. Now we can derive the interval using the probabilities

$$\begin{aligned}
&\mathbb{P}_F(c_{1-\alpha/2} \leq \hat{\theta}_n - \theta \leq c_{\alpha/2}) = 1 - \alpha \\
&\Leftrightarrow \mathbb{P}_F(-\hat{\theta}_n + c_{1-\alpha/2} \leq -\theta \leq -\hat{\theta}_n + c_{\alpha/2}) = 1 - \alpha \\
&\Leftrightarrow \mathbb{P}_F(\hat{\theta}_n - c_{\alpha/2} \leq \theta \leq \hat{\theta}_n - c_{1-\alpha/2}) = 1 - \alpha.
\end{aligned}$$

Hence, the infeasible interval based on $Q_n(x, F)$ would be $C_\theta(\mathbf{X}) = \left[\hat{\theta}_n - c_{\alpha/2}, \hat{\theta}_n - c_{1-\alpha/2}\right]$.

For the bootstrap, letting $\hat{\theta}_n^* = W_n(\mathbf{X}^*)$ and $\theta^* = \tau(\hat{F}_n)$, we can define the bootstrap quantity $Q_n^* = Q_n(\mathbf{X}^*, \hat{F}_n) = \hat{\theta}_n^* - \theta^*$.[15] We can then define the bootstrap cut-off point c_α^*

[14] Equivalently one can derive the bootstrap percentile interval from the inversion of the acceptance region of the bootstrap test of $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ and the test statistic is $T_n(\mathbf{X}, \theta_0) = \hat{\theta}_n - \theta$, see Exercise B.1

[15] As discussed for the bias reduction case, typically $\theta^* = \hat{\theta}_n$ and therefore $Q_n^*(\mathbf{X}^*, \hat{F}_n) = \hat{\theta}_n^* - \hat{\theta}_n$.

such that

$$\mathbb{P}^*(Q_n^* \geq c_\alpha^*) = \mathbb{P}_{\hat{F}_n}(Q_n(\mathbf{X}^*, \hat{F}_n) \geq c_\alpha^* | \, \mathbf{X}) = \alpha.$$

The bootstrap percentile interval can then be written as

$$C_\theta^*(\mathbf{X}) = \left[\hat{\theta}_n - c_{\alpha/2}^*, \hat{\theta}_n - c_{1-\alpha/2}^*\right]. \tag{8}$$

In practice we again approximate c_α^* by $c_{\alpha,B}^* = Q_n^{*([(1-\alpha)*B])}$ after following Algorithm 1.

4.4.2 An incorrect percentile interval

Efron (1979) originally proposed a different interval. Efron took $\tilde{Q}_n(\mathbf{X}^*, \theta^*) = \hat{\theta}_n^*$, and let $\tilde{c}_\alpha^*$ be defined such that $\mathbb{P}^*(\tilde{Q}_n^* \geq \tilde{c}_\alpha^*) = \alpha$. He then proposed to use the interval $\tilde{C}_\theta^*(\mathbf{X}) = [\tilde{c}_{1-\alpha/2}^*, \tilde{c}_{\alpha/2}^*]$. While this seems to be a logical choice, this is in fact an incorrect interval. It turns out the interval has its tails reversed.

To show that its coverage is not the intended $(1-\alpha)$, first note if we take $Q_n^* = Q_n(\mathbf{X}^*, \hat{F}_n) = \hat{\theta}_n^* - \theta^*$ as above, we can write that

$$\alpha = \mathbb{P}^*(\tilde{Q}_n^* \geq \tilde{c}_\alpha^*) = \mathbb{P}^*(\tilde{Q}_n^* - \theta^* \geq \tilde{c}_\alpha^* - \theta^*) = \mathbb{P}^*(Q_n^* \geq c_\alpha^*).$$

From this we can conclude that $c_\alpha^* = \tilde{c}_\alpha^* - \theta^*$ and represent Efron's interval as $\tilde{C}_\theta^*(\mathbf{X}) = [\theta^* + c_{1-\alpha/2}^*, \theta^* + c_{\alpha/2}^*]$.

First note that this interval only makes sense if $\theta^* = \hat{\theta}_n$, as it is otherwise not centered around the estimator $\hat{\theta}_n$. If we assume that this is indeed true, which is frequently the case, then we have that

$$\tilde{C}_\theta^*(\mathbf{X}) = \left[\hat{\theta}_n + c_{1-\alpha/2}^*, \hat{\theta}_n + c_{\alpha/2}^*\right].$$

It now becomes apparent that this is the interval $C_\theta^*(\mathbf{X})$ with its tails "flipped around". Note that Efron's interval approximates the infeasible interval $[\hat{\theta}_n + c_{1-\alpha/2}, \hat{\theta}_n + c_{\alpha/2}]$, where $\mathbb{P}_F(Q_n(\mathbf{X}, F) \geq c_\alpha) = \alpha$. Then we can calculate that interval's coverage as

$$\begin{aligned}
&\mathbb{P}_F(\hat{\theta}_n + c_{1-\alpha/2} \leq \theta \leq \hat{\theta}_n + c_{\alpha/2}) = \mathbb{P}_F(c_{\alpha/2} \leq \theta - \hat{\theta}_n \leq c_{1-\alpha/2}) \\
&= \mathbb{P}_F(-c_{\alpha/2} \leq Q_n(\mathbf{X}, F) \leq -c_{1-\alpha/2}).
\end{aligned}$$

This probability is typically not equal to $1-\alpha$. The tails are clearly reversed, and only if the distribution of $Q_n(\mathbf{X}, F)$ is symmetric around 0, in which case $c_{1-\alpha/2} = -c_{\alpha/2}$, is this interval appropriate.

4.4.3 Equal-tailed percentile-t interval

The percentile interval is not optimal as it is not based on a pivotal quantity. Due to the central limit theorem, in many instances there will be a pivotal t-ratio available for which

$$\frac{\hat{\theta}_n - \theta}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\hat{\theta}_n)}} = \frac{W_n(\mathbf{X}) - \theta}{\sqrt{V_n(\mathbf{X})}} \xrightarrow{d} N(0,1),$$

where $\hat{\theta}_n = W_n(\mathbf{X})$ is an appropriate estimator of θ, and $\widehat{\mathbb{V}\mathrm{ar}}(\hat{\theta}) = V_n(\mathbf{X})$ is an estimator of the variance of $\hat{\theta}_n$. In this case, an infeasible interval for θ can be derived from the manipulation of

$$\begin{aligned}
&\mathbb{P}_F\left(c_{1-\alpha/2} \leq \frac{\hat{\theta}_n - \theta}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\hat{\theta}_n)}} \leq c_{\alpha/2}\right) = 1-\alpha \\
&\Leftrightarrow \mathbb{P}_F\left(-c_{\alpha/2}\sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\hat{\theta}_n)} \leq \theta - \hat{\theta}_n \leq -c_{1-\alpha/2}\sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\hat{\theta}_n)}\right) = 1-\alpha \\
&\Leftrightarrow \mathbb{P}_F\left(\hat{\theta}_n - c_{\alpha/2}\sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\hat{\theta}_n)} \leq \theta \leq \hat{\theta}_n - c_{1-\alpha/2}\sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\hat{\theta}_n)}\right) = 1-\alpha,
\end{aligned}$$

where c_α is such that $\mathbb{P}_F(Q_n(\mathbf{X}, F) \geq c_\alpha) = \alpha$.

The bootstrap estimator of this interval is based on the bootstrap quantity

$$Q_n^* = Q_n(\mathbf{X}^*, \hat{F}_n) = \frac{\hat{\theta}_n^* - \theta^*}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\hat{\theta}_n^*)}} = \frac{W_n(\mathbf{X}^*) - \theta^*}{\sqrt{V_n(\mathbf{X}^*)}},$$

where $\theta^* = \tau(\hat{F}_n)$, and $\hat{\theta}_n^* = W_n(\mathbf{X}^*)$ and $\widehat{\mathbb{V}\mathrm{ar}}(\hat{\theta}_n^*) = V_n(\mathbf{X}^*)$ are the same functions as for the original sample but calculated using $\mathbf{X}^*$. The bootstrap interval is then

$$C_\theta^*(\mathbf{X}) = \left[\hat{\theta}_n - c_{\alpha/2}^*\sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\hat{\theta}_n)}, \hat{\theta}_n - c_{1-\alpha/2}^*\sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\hat{\theta}_n)}\right], \tag{9}$$

where c_α^* is again such that $\mathbb{P}^*(Q_n^* \geq c_\alpha^*) = \mathbb{P}_{\hat{F}_n}(Q_n(\mathbf{X}^*, \hat{F}) \geq c_\alpha^* | \mathbf{X}) = \alpha$.

Example 14. In Example 1 we saw that the asymptotic confidence interval for the mean did not perform well for all distributions and sample sizes. In particular, for the exponential(2) distribution and sample sizes smaller than $n = 100$, the asymptotic confidence interval did not come close to the desired 95% coverage. We now check if the bootstrap does better. In particular, we investigate if the equal-tailed bootstrap percentile-t interval defined in (9) does better. We consider both the iid bootstrap, where $\hat{F}_n$ is the EDF $\hat{F}_n^E$, and the parametric bootstrap where we (correctly) assume an exponential distribution with parameter $\hat{\mu}_n$, that is where $\hat{F}_n = \int_{-\infty}^x f(y|\hat{\mu}_n)dx = \int_{-\infty}^x \frac{1}{\hat{\mu}_n} e^{-x/\hat{\mu}_n} dx$.

For the estimator of μ we take $\hat{\mu}_n = \overline{X}_n$, and we take $\widehat{\mathbb{V}\mathrm{ar}}(\hat{\mu}_n) = \frac{S_n^2}{n}$, such that $Q_n(\mathbf{X}, F) =$

$\sqrt{n}\frac{\overline{X}_n-\mu}{S_n}$ and $Q_n^* = Q_n(\mathbf{X}^*, \hat{F}_n) = \sqrt{n}\frac{\overline{X}_n^*-\mu^*}{S_n^*}$. It then follows from (9) that the confidence interval is

$$C_\mu^*(\mathbf{X}) = \left[\overline{X}_n - c_{\alpha/2}^*\frac{S_n}{\sqrt{n}}, \overline{X}_n - c_{1-\alpha/2}^*\frac{S_n}{\sqrt{n}}\right].$$

To calculate this interval in practice, we take the following steps:

1. For $b = 1, 2, \ldots, B$, draw a random sample $x_1^{*b}, \ldots, x_n^{*b}$ from $\hat{F}_n$.

2. Calculate the bootstrap quantity

$$Q_n^{*b} = Q_n(\mathbf{x}^{*b}, \hat{F}_n^E) = \sqrt{n}\frac{\overline{x}_n^{*b} - \mu^*}{s_n^{*b}}.$$

   For the nonparametric bootstrap with $\hat{F}_n$ equal to the EDF $\hat{F}_n^E$, we have $\mu^* = \overline{x}_n$, while for the parametric bootstrap we have $\mu^* = \hat{\mu}_n$.

3. Repeat steps 1 and 2 B times and collect all bootstrap quantities $Q_n^{*1}, Q_n^{*2}, \ldots, Q_n^{*B}$. Let

$$c_{\alpha/2,B}^* = Q_n^{*([(1-\alpha/2)B])} \qquad \text{and} \qquad c_{1-\alpha/2,B}^* = Q_n^{*([(\alpha/2)B])}.$$

4. Obtain the interval

$$C_\mu^*(\mathbf{x}) = \left[\overline{x}_n - c_{\alpha/2,B}^*\frac{s_n}{\sqrt{n}}, \overline{xx}_n - c_{1-\alpha/2,B}^*\frac{s_n}{\sqrt{n}}\right].$$

We now use Monte Carlo simulations to investigate the coverage of the interval. The whole procedure now looks as follows:

1. Generate a sample $X_1, \ldots, X_n$ drawn from the exponential distribution with $\mu = 2$.

2. Use the sample generated in step 1 to construct the asymptotic confidence interval as well as the two bootstrap confidence intervals described above.

3. For each confidence interval, check if μ is contained in the interval and record a 1 if true, and a 0 otherwise.

4. Repeat step 1-3 N times; the average of the numbers recorded in step 3 is the estimated coverage.

Figure 4 shows the resulting coverages. The bootstrap intervals massively improve the coverage compared with the asymptotic interval. Perhaps surprisingly, the nonparametric interval is almost as good as the parametric bootstrap interval where we had the distribution right. This is typical for the nonparametric bootstrap in general; it turns out that the loss

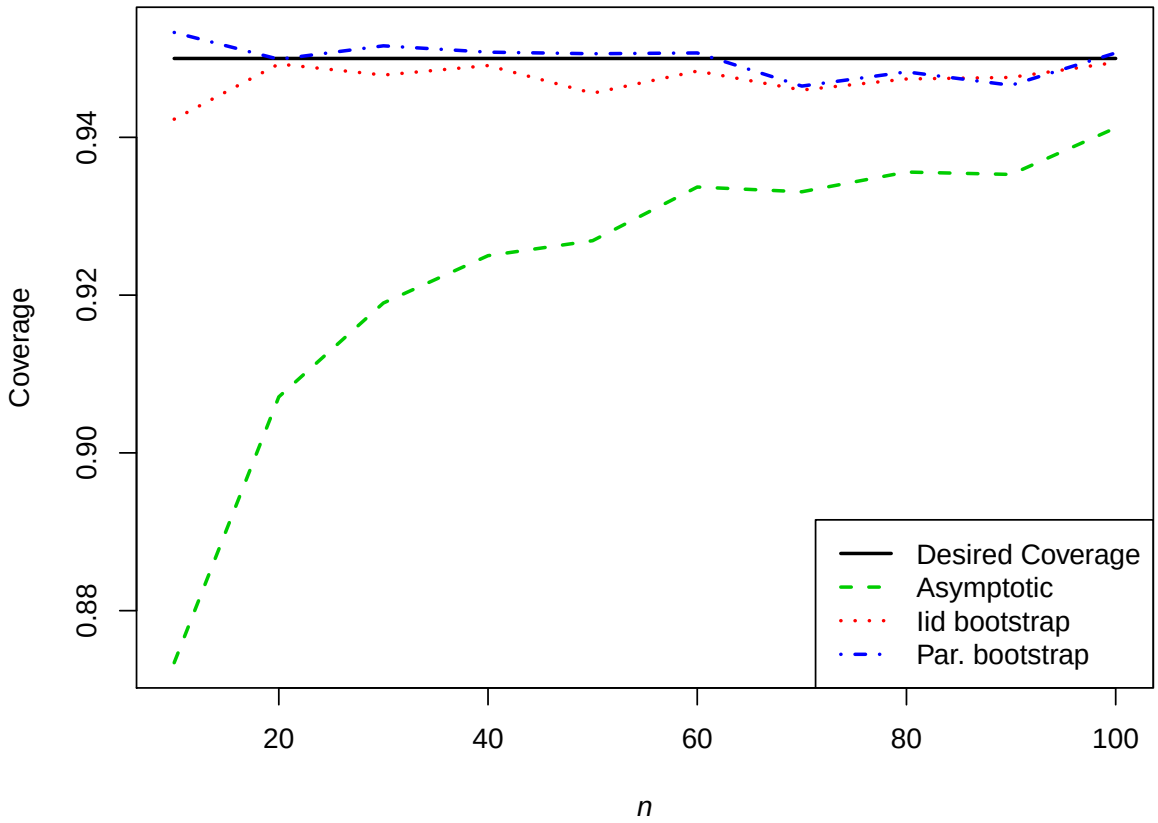


Figure 4: Estimated coverage probabilities as a function of the sample size n for three confidence intervals for the mean of an exponential(2) distribution.

for not making a parametric distributional assumption is rather small in general. Therefore it is usually preferred in most applications.

4.4.4 Symmetric intervals

If we know that the distribution of $Q_n(\mathbf{X}, F)$ is symmetric, we might as well use that explicitly in the construction of the confidence interval. For the percentile interval in that case we can define c_α^* such that $\mathbb{P}^*(|Q_n^*| \geq c_\alpha^*) = \alpha$, and construct the interval as

$$C_\theta^*(\mathbf{X}) = \left[\hat{\theta}_n - c_\alpha^*, \hat{\theta}_n + c_\alpha^*\right].$$

Similarly, the symmetric percentile t-interval uses a critical value c_α^* defined in the same way as above, but now for the t-ratio $Q_n^* = \frac{\hat{\theta}_n^* - \theta^*}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\hat{\theta}_n^*)}}$, which gives the interval

$$C_\theta^*(\mathbf{X}) = \left[\hat{\theta}_n - c_\alpha^* \sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\hat{\theta}_n)}, \hat{\theta}_n + c_\alpha^* \sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\hat{\theta}_n)}\right].$$

4.5 Bootstrap in regression models

As a final application we look at how the bootstrap can be used in regression models. Assume that we have the following regression model for $i = 1, \ldots, n$:

$$Y_i = \alpha + \beta X_i + \varepsilon_i,$$

with $\varepsilon_1, \ldots, \varepsilon_n$ an iid error term. We consider two different bootstrap methods for this setting. Although both can be seen as an extension of the bootstrap as discussed so far, they are based on a different outlook.

4.5.1 Pairs bootstrap

The first bootstrap method we consider, the *pairs bootstrap*, is a direct extension of the iid bootstrap, and is based on the bivariate EDF $\hat{F}_n^E(x, y)$ for the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. To this end the only assumption one needs to make is that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are a random sample with a bivariate cdf F. For example, the bivariate normal model described in Section 11.3.3 in C&B fits this description. No assumption on the regression model actually being the true model has to be made; remember that in this setting we can define $\alpha + \beta X$ as the *best linear predictor* of Y where

$$\beta = \tau(F) = \frac{\mathbb{C}\text{ov}_F(X, Y)}{\mathbb{V}\text{ar}_F(X)} = \frac{\mathbb{E}_F(X - \mathbb{E}_F X)(Y - \mathbb{E}_F Y)}{\mathbb{E}_F(X - \mathbb{E}_F X)^2}.$$

The pairs bootstrap builds the bootstrap sample $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$ by drawing pairs with replacement from the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. It is crucial that (X_i, Y_i) are kept together as a pair, otherwise if $\mathbf{Y}^*$ and $\mathbf{X}^*$ were drawn separately the bootstrap would assume there is no relation between the two.

As an illustration, let us consider constructing a bootstrap equal-tailed percentile-t confidence interval for β. Consider the least squares estimator

$$\hat{\beta}_{n,LS} = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^n (X_i - \overline{X}_n)^2} = \frac{S_{XY}}{S_{XX}}.$$

Consider the asymptotically pivotal quantity

$$Q_n(\mathbf{Y}, \mathbf{X}, F) = Q_n'(\mathbf{Y}, \mathbf{X}, \beta) = \frac{\hat{\beta}_{n,LS} - \beta}{\sqrt{S_n^2 / S_{XX}}},$$

where $S_n^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\alpha}_{n,LS} - \hat{\beta}_{n,LS} X_i)^2$. Its bootstrap version is

$$Q_n^* = Q_n(\mathbf{Y}^*, \mathbf{X}^*, \hat{F}_n) = Q_n'(\mathbf{Y}^*, \mathbf{X}^*, \beta^*) = \frac{\hat{\beta}_{n,LS}^* - \beta^*}{\sqrt{S_n^{*2} / S_{XX}^*}},$$

where

$$\hat{\beta}^*_{n,LS} = \frac{\sum_{i=1}^n (X_i^* - \overline{X}_n^*)(Y_i^* - \overline{Y}_n^*)}{\sum_{i=1}^n (X_i^* - \overline{X}_n^*)^2} = \frac{S^*_{XY}}{S_{XX}},$$

$\beta^* = \tau(\hat{F}_n)$, and $S_n^{*2} = \frac{1}{n-2}\sum_{i=1}^n (Y_i^* - \hat{\alpha}^*_{n,LS} - \hat{\beta}^*_{n,LS} X_i^*)^2$.

For the pairs bootstrap where $\hat{F}_n$ is the EDF $\hat{F}_n^E$, one can show that $\beta^* = \tau(\hat{F}_n^E) = \hat{\beta}_{n,LS}$. Defining c^*_α such that $\mathbb{P}^*(Q_n^* \geq c^*_\alpha) = \alpha$, we can then construct the equal-tailed percentile-t interval for β

$$\left[\hat{\beta}_{n,LS} - c^*_{\alpha/2}\sqrt{\frac{S_n^2}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}}, \hat{\beta}_{n,LS} - c^*_{1-\alpha/2}\sqrt{\frac{S_n^2}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}}\right]. \tag{10}$$

While the pairs bootstrap is simple to use, it may be "too much nonparametric" for our needs. That is, it does not use our (assumed) knowledge that a linear regression model is appropriate for Y and X.

4.5.2 Residual bootstrap

The second bootstrap method we consider, the *residual bootstrap*, does utilize that knowledge more efficiently. Its first step is to calculate the residuals of the regression model:

$$\hat{\varepsilon}_i = Y_i - \hat{\alpha}_n - \hat{\beta}_n X_i \qquad \text{for } i = 1, \ldots, n.$$

We then apply either a parametric bootstrap or the iid bootstrap to the residuals $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$ to obtain the *bootstrap errors* $\varepsilon_1^*, \ldots, \varepsilon_n^*$. If the i.i.d. bootstrap is used, it is important to make sure that the residuals have mean zero, i.e. that $\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i = 0$. In that case $\mathbb{E}^* \varepsilon_i^* = \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i = 0$, which should be the case as error terms should always have mean zero. If a constant is included in the regression, as is the case above, this is automatically satisfied.

We next need to construct the bootstrap regressors $X_1^*, \ldots, X_n^*$. Here we have two options: (i) fix the bootstrap regressors by taking $X_1^* = X_1, \ldots, X_n^* = X_n$; or (ii) draw the bootstrap regressors randomly with replacement from $X_1, \ldots, X_n$. In practice the options yield very similar results. If the actual regressor is actually a fixed deterministic quantity, as we assumed throughout much of Chapter 11 of C&B, then option (i) is clearly preferred.

The final step consists of building the bootstrap sample $Y_1^*, \ldots, Y_n^*$, using the elements we have so far:

$$Y_i^* = \alpha^* + \beta^* X_i^* + \varepsilon_i^*,$$

where for α^* and β^* one could take any value, but usually one takes $\alpha^* = \hat{\alpha}_n$ and $\beta^* = \hat{\beta}_n$.[16]

[16] If we use the bootstrap for an hypothesis test, then one could also take $\beta^* = \beta_0$ to automatically make

Once the bootstrap sample is generated, the remainder of this bootstrap procedure is identical to the pairs bootstrap.

It is not always clear whether the residual or pairs bootstrap should be preferred. Often they give very similar results. As explained above, the residual bootstrap has the advantage of imposing the knowledge of the linear regression model that the pairs bootstrap does not do. On the other hand, the pairs bootstrap is more robust to misspecification, for instance if $\varepsilon_1, \ldots, \varepsilon_n$ are not iid, but the variances differ. In such a setting the pairs bootstrap will be superior. It is therefore not clear in general which method to prefer. In practice the residual bootstrap seems the more popular method.

4.6 Practical implementation

How many bootstrap replications To implement the bootstrap in practice, we first need a value for B, the number of bootstrap replications. If one uses it for hypothesis testing or confidence intervals, it is recommended to take B such that $\alpha(B+1)$ is an integer, where α is the significance level of the test or $1-\alpha$ is the confidence level of the interval, see e.g. Davidson and MacKinnon (2004).

Ideally we want to take B as large as possible, although taking B large of course increases computation time. For applications, a reasonable value is $B = 9,999$. Lower values such as $B = 1,999$ are used if the computation time is too high. For Monte Carlo simulations, which take a lot more time, lower values such as $B = 499$ or $B = 999$ are more common.

Implementation in R The crucial aspect of implementing the bootstrap in a language like R, is how to draw the bootstrap sample. Here we focus on the nonparametric bootstrap where we draw from the EDF $\hat{F}_n^E$. This implies that X_1^* should be drawn randomly from $X_1, \ldots, X_n$. The whole bootstrap sample is then drawn with replacement from $X_1, \ldots, X_n$.

While commands exist in R to draw with replacement, we do not even need those. In fact, all we need is to draw random numbers from a discrete uniform distribution. To see this, note that we can write

$$X_1^* = X_{J_1},$$

where J_1 determines which of $X_1, \ldots, X_n$ we draw. As we are equally likely to pick $X_1, \ldots, X_n$, we can equivalently say that J_1 is equally likely to be any of the numbers $1, \ldots, n$. This simply means that J_1 is uniformly distributed on $\{1, \ldots, n\}$. Similarly, $X_2^* = X_{J_2}$, where J_2 has the same uniform distribution and is independent of J_1, and so on.

To draw a random sample of continuous uniformly distriubed random numbers in R, we can use the function `runif(n, min, max)`. The input `n` indicates how many (independent)

the bootstrap sample satisfy H_0.

random variables one draws, while `min` and `max` are the minimum and maximum of the range of the uniform respectively. Finally, to draw discrete rather than continuous random variables, we just need to round up. Putting everything together, the line

```
J <- ceiling(runif(n, min = 0, max = n))
```

will give us the vector $J_1, \ldots, J_n$ we need. Alternatively, as a "short-cut" we can use the function `sample.int`, which directly draws a sample of integers:

```
J <- sample.int(n,size = n, replace = TRUE)
```

Here the first argument `n` means that we draw from the integers up to the number n, the second argument `size = n` indicates we want to draw a sample of size n, and the third argument `replace = TRUE` means we draw a sample with replacement.

To build the bootstrap sample, we then simply need to access the right elements of the vector $\mathbf{X} = X_1, \ldots, X_n$. Assuming we store the sample as a vector in the variable `X`, we get

```
X.star <- X[J]
```

We can then use `X.star` to obtain whatever quantity we need. If we let `theta.star` contain our bootstrap parameter θ^*, we can get $Q_n^* = Q_n(\mathbf{X}^*, \hat{F}_n^E) = Q_n'(\mathbf{X}^*, \theta^*)$ as

```
Q.star <- Q.func(X.star, theta.star)
```

where `Q.func` is a function that we have to change depending on our specific application.

As we need to store Q_n^* for all $b = 1, \ldots, B$ bootstrap replications, it is best to first construct a vector `Q.star` of dimension B, for instance as a vector of zeroes, and then loop over all bootstrap replications, storing the b-th version of Q_n^* in the b-th element of the vector.

Putting everything together, a code in R to do the bootstrap could look like

```
Q.star <- rep(0, times = B)
for (b in 1:B) {
    J <- sample.int(n, size = n, replace = TRUE)
    X.star <- X[J]
    Q.star[b] <- Q.func(X.star, theta.star)
}
```

Finally, for hypothesis testing or confidence intervals, we need to take out the right elements from `Q.star`. The R function `quantile` can be used for this.

5 Theoretical Properties of the Bootstrap

So far we looked at how to implement the bootstrap, but we did not consider if it is actually appropriate to use it. Also, we did not provide any reason why it would actually improve on

the standard asymptotic approximation. Example 14 is very promising regarding performance of the bootstrap, but we cannot go on one specific example to draw any conclusions on whether the bootstrap is appropriate or not. Instead we need theory to answer that question in a more systematic way. First we consider when it is allowed to use the bootstrap, that is, when we can guarantee it provides at least reasonable results. Next we consider when the bootstrap actually performs better than the asymptotic approximation.

5.1 Consistency

5.1.1 Definition of consistency

In this section we want to establish a necessary condition to be allowed to use the bootstrap. That is, if the bootstrap does not satisfy this condition, its use should always be avoided. The way we do this is link it to the standard asymptotic approximation. We know that this approximation is at least correct for a sample size increasing to infinity. Therefore we may reasonably expect the bootstrap to also be correct when the sample size tends to infinity. We call this *consistency* of the bootstrap. It is formalized in the following definition.

Definition 6. Let $G_n(x, F)$ be the cdf of the quantity $Q_n(x, F)$. Furthermore, let $G_n(x, \hat{F}_n)$ be the cdf of the corresponding bootstrap quantity $Q_n(\mathbf{X}^*, \hat{F}_n)$. Let $\mathcal{F}$ be a set of permissible distributions F. The bootstrap estimator $G_n(x, \hat{F}_n)$ is *consistent* for $G_n(x, F)$ if for every $F \in \mathcal{F}$

$$\sup_{x \in \mathbb{R}} \left| G_n(x, \hat{F}_n) - G_\infty(x, F) \right| \xrightarrow{p} 0. \tag{11}$$

If (11) is satisfied, we also say that the bootstrap is *asymptotically valid*.

The definition states that the bootstrap is asymptotically valid, or consistent, if the bootstrap distribution $G_n(x, \hat{F})$ converges uniformly (with respect to x) in probability to $G_\infty(x, F)$, the true asymptotic distribution of $Q_n(x, F)$. The set $\mathcal{F}$, to which F is restricted, is simply there to be able to rule out some "crazy" distributions.

Generally it turns out that the bootstrap is valid for most "well-behaved" statistics. It is not automatically true that the bootstrap is consistent though. There are several practical applications where the bootstrap is not consistent. Two of these are considered in Example 15 as well as Exercises B.2 and B.3. Horowitz (2001) considers several other cases. As illustrated in Example 15. it always remains important to check if the bootstrap is consistent or not.

Often we do not have to check uniform convergence (using the difficult looking supremum over x), as the following theorem tells us that pointwise convergence is enough when $G_\infty(x, F)$ is a continuous function (which is often the case).

Theorem 3 (Polya's Theorem)**.** *If $G_\infty(x, F)$ is a continuous function over $x \in \mathbb{R}$, and*

$$G_n(x, \hat{F}_n) \xrightarrow{p} G_\infty(x, F) \qquad \text{for all } x \in \mathbb{R}, \tag{12}$$

then $\sup_{x\in\mathbb{R}} \left| G_n(x, \hat{F}_n) - G_\infty(x, F) \right| \xrightarrow{p} 0$.

The pointwise convergence in (12) is easier to verify, but although this check can be done on a case-by-case basis, it is still typically not very efficient nor easy to verify for each single application that the condition (11) is satisfied. In the following (optional) section we therefore consider a general theorem that can be used.

Example 15 (**Bootstrap Invalidity).** In Exercise B.2 it is shown that the nonparametric bootstrap is invalid for inference on the maximum of a uniform distribution. Here we elaborate on the implications of the invalidity in practice.

Let $X_1, \ldots, X_n$ be a random sample from a uniform$(0, \theta)$ distribution. We construct bootstrap percentile confidence intervals based on the iid bootstrap where $\hat{F}_n = \hat{F}_n^E$ the EDF, and a parametric bootstrap where $\hat{F}_n$ is uniform$(0, \hat{\theta}_n)$, where we take the maximum likelihood estimator $\hat{\theta}_n = \max_i X_i$. To calculate the interval, let $Q_n(\mathbf{X}, F) = n(\hat{\theta}_n - \theta)$, and take the following steps:

1. For $b = 1, 2, \ldots, B$, draw a random sample $X_1^{*b}, \ldots, X_n^{*b}$ from $\hat{F}_n$.

2. Calculate the bootstrap quantity

   $$Q_n^{*b} = Q_n(\mathbf{X}^{*b}, \hat{F}_n^E) = n(\hat{\theta}_n^* - \hat{\theta}_n),$$

   where $\hat{\theta}_n^* = \max_i X_i^*$.

3. Repeat steps 1 and 2 B times and collect all bootstrap quantities $Q_n^{*1}, Q_n^{*2}, \ldots, Q_n^{*B}$. Let

   $$c_{1-\alpha,B}^* = Q_n^{*([\alpha B])}.$$

4. Obtain the interval[17]

   $$C_\mu^*(\mathbf{X}) = \left[\overline{X}_n, \overline{X}_n - c_{1-\alpha,B}^*\right].$$

As in Example 14 we perform a Monte Carlo study to investigate the accuracy of the two intervals in finite samples. Figure 5 presents the coverage probabilities when we take $\theta = 1$. We see that, unlike the parametric bootstrap interval, the nonparametric iid bootstrap interval

[17] As Q_n^{*b} can only take negative values (like $Q_n(\mathbf{X}, F)$), we do not use the standard $c_{1-\alpha,B}^*$ and $c_{1-\alpha,B}^*$ cut-off points, but "shift the interval to the left" to ensure that $\hat{\theta}_n$ is included (as the left endpoint).

does not have the correct coverage. Moreover, it does not get better when n increases. This is in line with the theoretical results from Exercise B.2. As the actual coverage is typically below 80%, the interval is very inaccurate and should not be used in practice. This illustrates the need for checking that the bootstrap is indeed consistent.

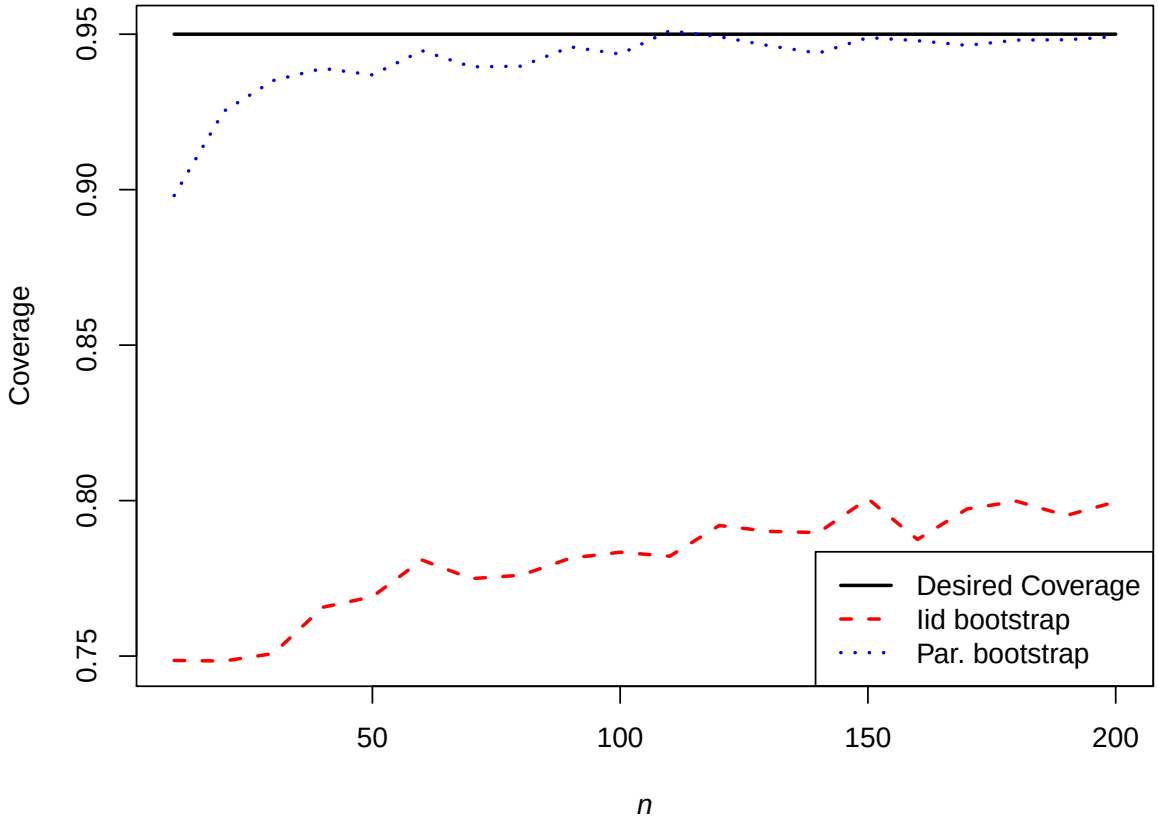


Figure 5: Estimated coverage probabilities as a function of the sample size n for two confidence intervals for the maximum of an uniform$(0, 1)$ distribution.

5.1.2 A General theorem for proving consistency*

Horowitz (2001) discusses a theorem by Beran and Ducharme (1991) which derives sufficient conditions to guarantee the validity of (11). They show that if the following three conditions are satisfied, the bootstrap is indeed consistent.

1. For every $F \in \mathcal{F}$, $\sup_{x \in \mathbb{R}} \left| \hat{F}_n^E(x) - F(x) \right| \xrightarrow{p} 0$. This condition is satisfied for the EDF $\hat{F}_n^E$ by the Glivenko-Cantelli Theorem. For the parametric bootstrap this can be shown to be true if the parametric family chosen $F(x|\theta)$ is the correct one and if $\hat{\theta}_n \xrightarrow{p} \theta$.

2. For every $F \in \mathcal{F}$, $G_\infty(x, F)$ is a continuous function of x. This is true in most applications; very often $G_\infty(x, F)$ is the normal or χ^2 distribution, which is continuous.

3. For every x and every sequence of functions $\tilde{F}_n \in \mathcal{F}$ for which $\tilde{F}_n(x) \to F(x)$, we have that $\lim_{n\to\infty} G_n(x, \tilde{F}_n) = G_\infty(x, F)$. This condition essentially states that for any consistent estimator of F, such as the EDF $\hat{F}_n^E$, the consistency should "carry over" to $G_n(x, \tilde{F}_n)$. If the bootstrap is inconsistent, it is often the case that this condition is violated.

The third condition – which is often the critical one – can still be quite difficult to verify. Many people have come up with sufficient conditions for which the condition is satisfied, and thankfully for most "well-behaved" or regular statistics, the bootstrap is indeed consistent. Horowitz (2001) discusses this in detail. Among the results discussed there, the following theorem, inspired by Mammen (1992), is worth mentioning explicitly.

Theorem 4 (Mammen, 1992)**.** *Let $X_1, \ldots, X_n$ be a random sample from a distribution with cdf F. Let $W_n(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n g_n(X_i)$ for a sequence of functions $g_n(\cdot)$. Let k_n and l_n be sequences of numbers, and define*

$$Q_n(\mathbf{X}, F) = \frac{W_n(\mathbf{X}) - k_n}{l_n}.$$

Let the bootstrap sample $X_1^*, \ldots, X_n^*$ be generated from the EDF $\hat{F}_n^E$, and define

$$Q_n(\mathbf{X}^*, \hat{F}_n^E) = \frac{W_n(\mathbf{X}^*) - W_n(\mathbf{X})}{l_n}.$$

Then the bootstrap is consistent for $Q_n(\mathbf{X}, F)$ if and only if $Q_n(\mathbf{X}, F) \xrightarrow{d} N(0, 1)$.

Many statistics can either be written in the form above, or approximated as such. The theorem proves consistency of the bootstrap for such quantities if they have a standard normal distribution asymptotically. The following example shows how this theorem can be used to prove bootstrap consistency for the sample mean.

Example 16. Let us consider the consistency of the bootstrap for the distribution of

$$Q_n(\mathbf{X}, F) = \sqrt{n}\frac{\overline{X}_n - \mu}{S_n}, \tag{13}$$

where $\mu = \mathbb{E}_F X$. Also let $\sigma^2 = \mathbb{V}\mathrm{ar}\, X < \infty$. However, in order to apply the theorem, we first consider

$$Q_n'(\mathbf{X}, F) = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}.$$

Now take $g_n(x) = x$, such that $W_n(\mathbf{X}) = \overline{X}_n$. Let $k_n = \mu$, and let $l_n = \mathbb{V}\mathrm{ar}(\overline{X}_n) = \sigma^2/n$.

Then

$$Q_n(\mathbf{X}, F) = \frac{W_n(\mathbf{X}) - k_n}{l_n} = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow{d} N(0,1).$$

As $\mu^* = \overline{X}_n = W_n(\mathbf{X})$, we can take $Q'_n(\mathbf{X}^*, \hat{F}_n^E) = \frac{W_n(\mathbf{X}^*) - W_n(\mathbf{X})}{l_n}$. Now all conditions of the theorem are satisfied, and thereby the consistency of the bootstrap for the mean when based on $Q'_n(\mathbf{X}, F) = \sqrt{n}(\overline{X}_n - \mu)/\sigma$ is shown.

Of course, typically we don't know σ. However, we also don't need to in order to apply this result. The bootstrap considered above is the same as the bootstrap of $Q''_n(\mathbf{X}, F) = \sqrt{n}(\overline{X}_n - \mu)$, given that the denominator σ does not change inside the bootstrap. Clearly we may multiply the quantity of interest with any real number, as long as we do it in the bootstrap too, it will not change anything.

Finally, bootstrap consistency for $Q_n(\mathbf{X}, F)$ as defined in (13) now follows from the fact that $S_n \xrightarrow{p} \sigma$, and thus $Q_n(\mathbf{X}, F) \xrightarrow{p} Q'_n(\mathbf{X}, F)$. By proving the same result for S_n^*, bootstrap consistency follows.

The theorem can be used to prove consistency for a much larger class of quantities, that include for example least squares estimators as well as many method of moments and maximum likelihood estimators for exponential families.

5.2 Higher order properties of the bootstrap*

The previous section only discussed conditions under which the bootstrap may be applied. However, all that was required was that asymptotically the bootstrap is the same as the standard approach. This says nothing about whether the bootstrap actually works better than the standard asymptotic approach.

In particular, the section does not explain at all why the bootstrap is often observed to provide a better approximation to the unknown true quantity than the asymptotic approximation, like we saw in Figure 4.

In order to explain this difference, we need to look at the approximation errors made by the bootstrap, compared to the standard approximation. For this we use so-called *higher order* asymptotic analysis. The name refers to the fact that we do not only consider the resulting limiting approximation (the first order term), but we also look at terms in the approximation that vanish at a certain rate (the higher order terms). This works similarly as in a Taylor expansion, and will be explained in more detail below. Before we can do so we first need a way to describe the order of the approximation error.

5.2.1 Stochastic order symbols*

Here we introduce *stochastic order symbols*, which are an extension of the standard order of magnitude symbols. These will be used in the next section to describe approximation errors

for the bootstrap.

Remember, if $x_1, x_2, \ldots$ is any sequence of real numbers, and $a_1, a_2, \ldots$ is any sequence of positive real numbers, then we say that x_n is of the order of magnitude of a_n, written as $x_n = O(a_n)$, if there is a $K < \infty$ such that $|x_n| / a_n < K$ for all n. Similarly, if $|x_n| / a_n \to 0$ as $n \to \infty$, we say that x_n is of a smaller order of magnitude than a_n, and we write $x_n = o(a_n)$.

Example 17. Consider the sequence $x_n = c/n$ for some real number $c > 0$. Take the sequence $a_n = 1$ for all n. Then clearly $|x_n| / a_n = c/n \to 0$ as $n \to 0$. Therefore $x_n = o(1)$. We can be more precise however. Take $a_n = n^{-b}$, for $0 < b < 1$. Then still $|x_n| / a_n = \frac{c/n}{n^{-b}} = cn^{b-1} \to 0$ as $n \to 0$, as $b - 1 < 0$. Therefore, $x_n = o(n^{-b})$ for any $0 < b < 1$. However, if we take $a_n = 1/n$, then $|x_n| / a_n = \frac{c/n}{1/n} = c$, which does not converge to 0 as $n \to 0$. However, we can clearly find a $K < \infty$ such that $|x_n| / a_n < B$ for all n; any number larger than c will do. Therefore, $x_n = O(1/n)$.

Similarly, consider the sequence $y_n = \sqrt{n}\sin(n)$. As $-1 \leq \sin(x) \leq 1$, it follows directly that $|\sin(n)| \leq 1$ for all n. Taking the sequence $b_n = \sqrt{n}$, it follows directly that $|y_n| / b_n = \frac{|\sqrt{n}\sin(n)|}{\sqrt{n}} = |\sin(n)| \leq 1$ for all n. Therefore, $y_n = O(\sqrt{n})$.

Stochastic order symbols extend this definition to random variables.

Definition 7. Let $X_1, X_2, \ldots$ denote a sequence of random variables, and let $a_1, a_2, \ldots$ denote a sequence of positive real numbers. Then, if for any $\epsilon > 0$ there exists a $K_\epsilon > 0$ that does not depend on n but may depend on ϵ, such that

$$\mathbb{P}(|X_n| / a_n > K_\epsilon) < \epsilon \qquad \text{for all } n,$$

we say that X_n is of (at most) the same stochastic order as a_n, written as $O_p(a_n)$.

Similarly, if for every $\epsilon > 0$ we have that $\mathbb{P}(|X_n| / a_n > \epsilon) \to 0$ we say that X_n is of smaller stochastic order than a_n, written as $o_p(a_n)$.

We can use consistency results, for example by the weak law of large numbers, to determine $o_p(\cdot)$ orders.

Example 18. Let $X_1, \ldots, X_n$ be a random sample from an unspecified distribution, with $\mathbb{E}\, X = \mu$ and $\mathbb{V}\mathrm{ar}\, X = \sigma^2 < \infty$. Then the weak law of large numbers tells us that $\overline{X}_n \xrightarrow{p} \mu$. Let $Y_n = \overline{X}_n - \mu$. Then, it follows from the weak law of large numbers that, for any $\epsilon > 0$,

$$\mathbb{P}(|Y_n| > \epsilon) = \mathbb{P}(\left|\overline{X}_n - \mu\right| > \epsilon) \to 0.$$

Therefore, $Y_n = o_p(1)$. Or to write things in a different way, $\overline{X}_n = \mu + Y_n = \mu + o_p(1)$.

To determine $O_p(\cdot)$ orders, we have to use results on convergence in distribution.

Example 19. For the same setting as in Example 18, the central limit theorem tells us that

$$\sqrt{n}Y_n = \sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

This is enough to determine the $O_p(\cdot)$ order of Y_n. First note that for any fixed n, we can always find a constant $K_{\epsilon,n}$ such that $\mathbb{P}(|Y_i| / a_i > K_{\epsilon,n})$ for any sequence a_i for all $i = 1, \ldots, n$ (see the proof of Lemma 3 for details). Hence, we do not have to look at fixed n, but instead to the tail of the sequence, in other words let $n \to \infty$. Yet for that setting, we have by the CLT, for any K,

$$\mathbb{P}\left(\frac{|Y_n|}{1/\sqrt{n}} > K\right) = \mathbb{P}(\sqrt{n}\,|Y_n| > K) \to \mathbb{P}(Z > K).$$

Therefore, if we take $a_n = 1/\sqrt{n}$, we find that there exists a $K_\epsilon > 0$ such that $\mathbb{P}(|Y_n| / a_n > K_\epsilon) < \epsilon$ for all n, and therefore $Y_n = O_p(1/\sqrt{n})$, or alternatively $\overline{X}_n = \mu + O_p(1/\sqrt{n})$.

As can be seen from Example 19, the stochastic order when written in an equation as $\overline{X}_n = \mu + O_p(1/\sqrt{n})$, can be interpreted as a statement on how fast $\overline{X}_n$ converges to μ. This speed of convergence is what we need to discuss asymptotic refinements. Before we do so, we give one more result that extends the approach in Example 19 to find stochastic orders. Although the result is fairly intuitive, a formal proof is not so straightforward to set up.

Lemma 3. *Let $X_1, \ldots, X_n$ be a sequence of random variables, let $m_1, \ldots, m_n$ be a sequence of positive real numbers, and let $T_n(\mathbf{X}) = T_n(X_1, \ldots, X_n)$ be a function of the random variables such that*

$$m_n(T_n(\mathbf{X}) - \theta) \xrightarrow{d} Y, \tag{14}$$

where Y is a random variable with pdf $f(x)$. Then $T_n(\mathbf{X}) = \theta + O_p(1/m_n)$.

Proof. The proof of this lemma is essentially the same proof as used to prove the result that the (non-random) sequence $a_1, \ldots, a_n$ is bounded if $a_n \to a$ as $n \to \infty$ for some $|a| < \infty$. The difference is that we need to adapt the proof to a probabilistic setting.

The convergence result (14) tells us that for every x, we have that $\mathbb{P}(m_n(T_n(\mathbf{X}) - \theta) \leq x) \to \mathbb{P}(Y \leq x)$ as $n \to \infty$. Explicitly writing out the limit, we have that for any $\delta > 0$, there exists an $N_{\delta,x}$ such that for all $n > N_{\delta,x}$ we have that

$$|\mathbb{P}(m_n(T_n(\mathbf{X}) - \theta) \leq x) - \mathbb{P}(Y \leq x)| < \delta/2.$$

As $\mathbb{P}(|Y| > x) = 1 - \mathbb{P}(Y \leq x) + \mathbb{P}(Y < -x)$, we may equivalently write that for any $\delta > 0$, there exists an $N_\delta = \max\{N_{\delta,x}, N_{\delta,-x}\}$ such that for all $n > N_\delta$ we have that

$$|\mathbb{P}(m_n\,|T_n(\mathbf{X}) - \theta| > x) - \mathbb{P}(|Y| > x)| < \delta.$$

Now let $K_{\epsilon,Y}$ be defined such that $\mathbb{P}(|Y| > K_{\epsilon,Y}) < \epsilon/2$. Then, taking $\delta = \epsilon/2$, for any $n > N_{\epsilon/2}$,

$$\begin{aligned}\mathbb{P}(m_n\,|T_n(\mathbf{X}) - \theta| > K_{\epsilon,Y}) \leq\; & |\mathbb{P}(m_n\,|T_n(\mathbf{X}) - \theta| > K_{\epsilon,Y}) - \mathbb{P}(|Y| > K_{\epsilon,Y})| \\ & + \mathbb{P}(|Y| > K_{\epsilon,Y}) < \epsilon/2 + \epsilon/2 = \epsilon.\end{aligned}$$

Next we consider $n \leq N_{\epsilon/2}$. For all $n = 1, \ldots, N_{\epsilon/2}$, let $k_{\epsilon,n}$ be such that $\mathbb{P}(m_n\,|T_n(\mathbf{X}) - \theta| > k_{\epsilon,n}) < \epsilon$. For any specific j such a $k_{\epsilon,j} < \infty$ always exists. Then take $K_{\epsilon,N_{\epsilon/2}} = \max_{1\leq j\leq N_{\epsilon/2}} k_{\epsilon,j}$, such that for any $n \leq N_{\epsilon/2}$,

$$\begin{aligned}&\mathbb{P}(m_n\,|T_n(\mathbf{X}) - \theta| > K_{\epsilon,N_{\epsilon/2}}) = \mathbb{P}(m_n\,|T_n(\mathbf{X}) - \theta| > \max_{1\leq j\leq N_{\epsilon/2}} k_{\epsilon,n}) \\ &\quad \leq \mathbb{P}(m_n\,|T_n(\mathbf{X}) - \theta| > k_{\epsilon,n}) < \epsilon.\end{aligned}$$

Finally, we put the two together. Let $K_\epsilon = \max\{K_{\epsilon,N_{\epsilon/2}}, K_{\epsilon,Y}\}$, then

$$\begin{aligned}&\sup_n \mathbb{P}(m_n\,|T_n(\mathbf{X}) - \theta| > K_\epsilon) \\ &\quad = \max\left\{\sup_{n\leq N_{\epsilon/2}} \mathbb{P}(m_n\,|T_n(\mathbf{X}) - \theta| > K_\epsilon), \sup_{n>N_{\epsilon/2}} \mathbb{P}(m_n\,|T_n(\mathbf{X}) - \theta| > K_\epsilon)\right\} \\ &\leq \max\left\{\sup_{n\leq N_{\epsilon/2}} \mathbb{P}(m_n\,|T_n(\mathbf{X}) - \theta| > K_{\epsilon,N_{\epsilon/2}}), \sup_{n>N_{\epsilon/2}} \mathbb{P}(m_n\,|T_n(\mathbf{X}) - \theta| > K_{\epsilon,Y})\right\} \\ &\leq \max\{\epsilon, \epsilon\} = \epsilon.\end{aligned}$$

This completes the proof. □

5.2.2 Asymptotic refinements*

Now that we defined stochastic order symbols, we can provide a formal definition of the asymptotic condition under which the bootstrap performs better than the standard asymptotic approximation.

Definition 8. Let $X_1, \ldots, X_n$ be a random sample with cdf F. Let $Q_n(\mathbf{X}, F)$ be our quantity of interest, and denote its cdf by $G_n(x, F)$. Let the bootstrap pivot be $Q_n(\mathbf{X}^*, \hat{F}_n)$ with associated cdf $G_n^* = G_n(x, \hat{F}_n)$. Consider doing inference based on a functional of G_n, let us call this $\gamma(G_n)$. Denote the approximation error of the asymptotic distribution to the true finite sample distribution as $\mathcal{E}_{a,n} = \gamma(G_\infty) - \gamma(G_n)$ and the approximation error made by the bootstrap to the true finite sample distribution as $\mathcal{E}_{b,n} = \gamma(G_n^*) - \gamma(G_n)$. Then the bootstrap provides *asymptotic refinements* for $\gamma(G)$ if $\mathcal{E}_{b,n}/\mathcal{E}_{a,n} \to 0$ as $n \to \infty$.

The definition states that the bootstrap offers asymptotic refinements if the approximation error it makes, decreases faster than the asymptotic approximation error. While it does not

directly say anything about the performance in small samples, it can be used as a justification for the bootstrap's superior small sample performance.

We can rephrase the condition for refinements in terms of order symbols. Assume that the asymptotic approximation error is of order $O(a_n)$, while the bootstrap approximation error is of order $O_p(b_n)$. Formally,

$$\gamma(G_\infty) = \gamma(G_n) + O(a_n), \qquad \gamma(G_n^*) = \gamma(G_\infty) + O_p(b_n).$$

Then the bootstrap provides asymptotic refinements for $\gamma(G)$ if $b_n/a_n \to 0$ as $n \to \infty$.

Bias Reduction The theory is best illustrated for the purposes of bias reduction. Let $\theta = \tau(F)$ be our parameter of interest, and $\hat{\theta}_n = \tau(\hat{F}_n)$ the associated estimator. Let $Q_n(x, F) = \hat{\theta}_n - \theta$. Recall from (6) that, if we take $\gamma(G) = \int x dG$ we have

$$\gamma(G_n) = \int x dG_n(x, F) = \mathbb{B}\text{ias}_n = \mathbb{E}_F\, Q_n(\mathbf{X}, F) = \mathbb{E}_F(\hat{\theta}_n - \theta).$$

As a consistent estimator does not have an asymptotic bias, the asymptotic approximation in this case is simply $\tau(G_\infty) = 0$. The asymptotic approximation error is then just (minus) the bias itself. Let the error be of order a_n, that is, $\mathbb{B}\text{ias}_n = O(a_n)$, where $a_n \to 0$.

As in Section 4.1, let $Q_n(\mathbf{X}^*, \hat{F}_n) = \hat{\theta}_n^* - \hat{\theta}_n$. From (7), the bootstrap error then is

$$\mathbb{B}\text{ias}_n^* - \mathbb{B}\text{ias}_n = \mathbb{E}_{\hat{F}}\left[\hat{\theta}_n^* - \theta^* | \, \mathbf{X}\right] - \mathbb{E}_F(\hat{\theta}_n - \theta) = O_p(b_n).$$

The bootstrap then offers asymptotic refinements if $b_n/a_n \to 0$. To find the actual orders a_n and b_n, we need to consider higher order Taylor expansions, which is why this kind of analysis is called higher order asymptotic analysis. We illustrate this with an example.

Example 20. Let $X_1, \ldots, X_n$ be a random sample with cdf F. Let $\mu = \mathbb{E}_F X$, and assume that we are interested in $\theta = e^\mu$. Let $\hat{\theta} = e^{\overline{X}_n}$. The bias is then

$$\mathbb{B}\text{ias}_n = \mathbb{E}_F(e^{\overline{X}_n} - e^\mu).$$

To evaluate this bias we perform a Taylor expansion. Letting $g(x) = e^x$, we have that $g'(x) = g''(x) = e^x$. Then

$$e^{\overline{X}_n} - e^\mu = e^\mu(\overline{X}_n - \mu) + \frac{e^\mu}{2}(\overline{X}_n - \mu)^2 + R_n,$$

where one can show for the remainder term R_n that $\mathbb{E}_F(R_n) = O(n^{-2})$. Now take the expectation:

$$\mathbb{B}\text{ias}_n = \mathbb{E}_F(e^{\overline{X}_n} - e^\mu) = e^\mu\, \mathbb{E}_F(\overline{X}_n - \mu) + \frac{e^\mu}{2}\, \mathbb{E}_F(\overline{X}_n - \mu)^2 + \mathbb{E}_F(R_n) = \frac{e^\mu}{2}\, \mathbb{V}\text{ar}\, \overline{X}_n + O(n^{-2}),$$

as $\mathbb{E}_F(\overline{X}_n - \mu) = 0$ and $\mathbb{E}_F(\overline{X}_n - \mu)^2 = \mathbb{V}\text{ar}\,\overline{X}_n$. As $\mathbb{V}\text{ar}\,\overline{X}_n = \mathbb{V}\text{ar}\,X/n = O(n^{-1})$, we find that $\mathbb{B}\text{ias}_n = O(n^{-1})$.

For the bootstrap, we get similarly

$$e^{\overline{X}_n^*} - e^{\overline{X}_n} = e^{\overline{X}_n}(\overline{X}_n^* - \overline{X}_n) + \frac{e^{\overline{X}_n}}{2}(\overline{X}_n^* - \overline{X}_n)^2 + R_n^*,$$

where $\mathbb{E}_{\hat{F}_n}(R_n^* | \mathbf{X}) = O_p(n^{-2})$. Then, as $\mathbb{E}_{\hat{F}_n}(\overline{X}_n^* - \overline{X}_n | \mathbf{X}) = 0$, we have that

$$\mathbb{B}\text{ias}_n^* = \mathbb{E}_{\hat{F}_n}(e^{\overline{X}_n^*} - e^{\overline{X}_n} | \mathbf{X}) = \frac{e^{\overline{X}_n}}{2}\,\mathbb{V}\text{ar}_{\hat{F}_n}(\overline{X}_n^*) + O_p(n^{-2}).$$

So far, nothing is different for the bootstrap compared with the standard asymptotic approximation. However, now we put things together. Consider the bias-corrected estimator $\hat{\theta}_n^{bc} = e^{\overline{X}_n} - \mathbb{B}\text{ias}_n^*$ and let us calculate the bias of this estimator.

$$\begin{aligned}\mathbb{B}\text{ias}(\hat{\theta}_n^{bc}) &= \mathbb{E}_F(e^{\overline{X}_n} - \mathbb{B}\text{ias}_n^* - \theta) = \mathbb{E}_F(e^{\overline{X}_n} - \theta) - \mathbb{E}_F[\mathbb{E}_{\hat{F}_n}(e^{\overline{X}_n^*} - e^{\overline{X}_n} | \mathbf{X})] \\ &= \frac{e^{\mu}}{2}\,\mathbb{V}\text{ar}\,\overline{X}_n + O(n^{-2}) - \mathbb{E}_F\left[\frac{e^{\overline{X}_n}}{2}\,\mathbb{V}\text{ar}_{\hat{F}_n}(\overline{X}_n^*)\right] + O(n^{-2}).\end{aligned}$$

Using a similar Taylor expansion as above we can show that

$$\mathbb{E}_F\left[\frac{e^{\overline{X}_n}}{2}\,\mathbb{V}\text{ar}_{\hat{F}_n}(\overline{X}_n^*)\right] = \frac{e^{\mu}}{2}\,\mathbb{V}\text{ar}\,\overline{X}_n + O(n^{-2}).$$

It then follows that

$$\mathbb{B}\text{ias}(\hat{\theta}_n^{bc}) = \frac{e^{\mu}}{2}\,\mathbb{V}\text{ar}\,\overline{X}_n - \frac{e^{\mu}}{2}\,\mathbb{V}\text{ar}\,\overline{X}_n + O(n^{-2}) = O(n^{-2}).$$

Hence, the bias of the bootstrap bias-corrected estimator is of smaller order – $O(n^{-2})$ – than the bias of the original estimator, which is $O(n^{-1})$.

Approximation for distributions In a similar way we can make approximations to other functions of $G_n(x, F)$, or to $G_n(x, F)$ itself. This is needed if we want to look at hypothesis testing or confidence intervals. Unfortunately Taylor expansions do not work anymore in that setting. Instead we have to use *Edgeworth expansions*, which work in the same way as Taylor expansions, but then for distribution functions. The theory behind is rather more complicated though, and therefore we do not treat them in detail.

6 Exercises

B.1 The confidence intervals derived in Section 4.4 can alternatively be derived by inverting the acceptance region of an appropriate hypothesis test. Do this for the

(a) equal-tailed percentile interval;

(b) symmetric percentile interval;

(c) equal-tailed percentile-t interval.

B.2 Let $X_1, \ldots, X_n$ be a random sample from a uniform$(0, \theta)$ distribution. Remember (see e.g. X 7.2) that the MLE of θ is equal to $\hat{\theta}_n = \max_i X_i$, and let $Q_n(\mathbf{X}, F) = n(\theta - \hat{\theta}_n)$. Then from X 10.6 we know that

$$G_\infty(x, F) = \lim_{n\to\infty} \mathbb{P}_F(Q_n(\mathbf{X}, F) \leq x) = \mathbb{P}\left(n(\theta - \hat{\theta}_n) \leq x\right) = 1 - e^{-x/\theta}, \qquad x \geq 0.$$

Now consider the nonparametric bootstrap where $\hat{F}_n = \hat{F}_n^E$, the EDF, and the corresponding bootstrap quantity $Q_n(\mathbf{X}, \hat{F}_n^E) = n(\hat{\theta}_n - \hat{\theta}_n^*)$, where $\hat{\theta}_n^* = \max_i(X_1^*, \ldots, X_n^*)$. Let $G_n(x, \hat{F}_n^E)$ denote the cdf of $Q_n(\mathbf{X}, \hat{F}_n^E)$.

(a) Show that $\mathbb{P}^*(X_1^* \leq x) = k/n$, where $k = \sum_{i=1}^n I_{(-\infty,x]}(X_i)$.

(b) Explain why $\mathbb{P}^*(\max_i X_i^* > \max_i X_i) = 0$.

(c) Show that $\mathbb{P}^*(\max_i X_i^* = \max_i X_i) = 1-(1-1/n)^n$. Hint: show that $\mathbb{P}^*(\max_i X_i^* < x) = [\mathbb{P}^*(X_1^* < x)]^n$.

(d) Using part (c), show that $\left|\mathbb{P}^*(\hat{\theta}_n^* = \hat{\theta}_n) - \mathbb{P}(\hat{\theta}_n = \theta)\right| \xrightarrow{p} 1 - e^{-1} \neq 0$.

(e) Explain why (d) implies that the bootstrap is inconsistent.

B.3 Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution with σ^2 known. Consider estimating $\theta = \mu^2$ using its MLE $\hat{\theta}_n = \overline{X}_n^2$.

(a) Use the Delta method to show that, if $\mu \neq 0$,

$$\sqrt{n}\frac{\overline{X}_n^2 - \mu^2}{2\,|\mu|\,\sigma} \xrightarrow{d} N(0, 1).$$

(b) Explain why the standard Delta method does not apply if $\mu = 0$. Instead, show how the *second-order* Delta method (Th. 5.5.26) implies that

$$n\frac{\overline{X}_n^2}{\sigma^2} \xrightarrow{d} \chi_1^2$$

if $\mu = 0$.

(c) For the nonparametric bootstrap with $\hat{F}_n$ equal to the EDF $\hat{F}_n^E$, one can show that, conditionally on the sample $\mathbf{X}$,

$$\sqrt{n}\frac{\overline{X}_n^* - \overline{X}_n}{\sigma^*} \xrightarrow{d} N(0,1), \tag{15}$$

where $\sigma^{*2} = \mathbb{V}\text{ar}_{\hat{F}_n^E} X^* = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. Use (15) and the Delta method to show that, conditionally on the sample $\mathbf{X}$,

$$\sqrt{n}\frac{\overline{X}_n^{*2} - \overline{X}_n^2}{2\left|\overline{X}_n\right|\sigma^*} \xrightarrow{d} N(0,1),$$

if $\overline{X}_n \neq 0$.

(d) We can write the result in part (c)

$$\mathbb{P}_{\hat{F}_n^E}\left(\sqrt{n}\frac{\overline{X}_n^{*2} - \overline{X}_n^2}{2\left|\overline{X}_n\right|\sigma^*} \leq x \middle| \overline{X}_n \neq 0\right) \xrightarrow{p} \mathbb{P}(Z \leq x).$$

Explain why $\mathbb{P}_\mu(\overline{X}_n = 0)$ for any μ, and use it to show that

$$\mathbb{P}_{\hat{F}_n^E}\left(\sqrt{n}\frac{\overline{X}_n^{*2} - \overline{X}_n^2}{2\left|\overline{X}_n\right|\sigma^*} \leq x\right) \xrightarrow{p} \mathbb{P}(Z \leq x).$$

that is, the result in (c) holds even without assuming that $\overline{X}_n \neq 0$.

(e) Explain why it follows from (d) that the bootstrap is valid if $\mu \neq 0$, but invalid if $\mu = 0$.

References

Beran, R. and G. R. Ducharme (1991). *Asymptotic Theory for Bootstrap Methods in Statistics*. Les Publications CRM, Centre de recherches mathématiques, Université de Montréal.

Casella, G. and R. L. Berger (2002). *Statistical Inference* (2nd ed.). Cengage Learning.

Davidson, R. and J. G. MacKinnon (2004). *Econometric Theory and Methods*. Oxford University Press.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics 7*, 1–26.

Efron, B. and T. Hastie (2016). *Computer Age Statistical Inference: Algorithms, Evidence and Data Science*. Cambridge University Press.

Efron, B. and R. Tibshirani (1994). *An Introduction to the Bootstrap*. CRC Press.

Hansen, B. E. (2019). *Econometrics*. http://www.ssc.wisc.edu/ bhansen/econometrics/.

Horowitz, J. L. (2001). The bootstrap. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 5, Chapter 52, pp. 3159–3228. Amsterdam: North Holland Publishing.

Mammen, E. (1992). *When Does Bootstrap Work? Asymptotic Results and Simulations*. New York: Springer.

A Notation

Random variables and realizations Random variables are denoted by upper case letters such as X and Y, while realizations of those random variables are denoted by lower case letters, such as x and y. A sample of n random variables is denoted as $X_1, \ldots, X_n$ or in vector notation as $\mathbf{X}$, while its realization is denoted as $x_1, \ldots, x_n$ or $\mathbf{x}$.

Density and distribution functions A probability density function (pdf) or probability mass function (pmf) is generally denoted by lower case $f(x)$, while its corresponding cumulative distribution function (cdf) is denoted by upper case $F(x)$. When it is not necessary, we don't distinguish between continuous and discrete distributions, and use pdf as abbreviation that could indicate both. When $f(x)$ belongs to some specific parametric family indexed by parameters θ, we often write $f(x|\theta)$ and $F(x|\theta)$ to emphasize this.

Probabilities and expectations Probabilities and expectations are generally denoted by $\mathbb{P}(\cdot)$ and $\mathbb{E}(\cdot)$ respectively. In order to make clear with respect to which distribution the probability or expectation is taken, we often add a subscript 'F' or – in the case of a parametric family $F(x|\theta)$ – 'θ' to the symbol. For example, $\mathbb{P}_F(X \leq c)$ is understood to mean the probability that the random variable X, with cdf F, is smaller than or equal to c; in other words, $\mathbb{P}_F(X \leq c) = F(c)$. Similarly, $\mathbb{E}_\theta Y^2$ is the expectation of Y^2 for a random variable Y with cdf $F(x|\theta)$, where $F(x|\theta)$ belongs to a known parametric family.

Cut-off points and quantiles Generic cut-off points of distributions are denoted as c_α, where c_α is the *right-tail* cut-off points corresponding to probability α. That is, for some distribution F, c_α corresponds to the $(1-\alpha)$-quantile of F, that is, $c_\alpha = \{\inf c : \mathbb{P}(X \leq c) \geq 1-\alpha\}$. When no confusion can arise, we ignore discrete distributions and define c_α as that

(assumed unique) value for which $\mathbb{P}(X \leq c_\alpha) = 1 - \alpha$. Similarly, $c_{1-\alpha}$ is the left-tail cut-off point, defined as the α-quantile of the distribution. For specific, often-used, distributions, the notation is adapted accordingly. For example, z_α is the right-tail cut-off point of the standard normal distribution, with values such as $z_{0.025} = 1.96$ and $z_{0.95} = -z_{0.05} = -1.645$. Similarly, $t_{n,\alpha}$ denotes the right-tail cut-off point of the t-distribution with n degrees of freedom. As explained in the notes, we use a similar convention for bootstrap cut-off points.