2025 BIOSTAT 523 Statistical Inference for Biometry II Slide Set 14: The Bootstrap

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington
jonno@uw.edu

Outline

Motivation

Motivation and Overview

We have seen that the jackknife sampling method can produce estimates of bias and variance, by resampling from the observed data.

But, the jackknife does not work for all statistics, for example, it does not work for the median.

Conformal prediction is a method for a particular type of prediction problem.

We now study the bootstrap, which is more computationally expensive than the jackknife to use, but offers some advantages.

The bootstrap is extremely popular, but is sometimes used, when it shouldn't (because there are rules!).

The Bootstrap

Bootstrap Methods

The bootstrap is a popular resampling technique that is useful in situations in which:

- We wish to relax the assumptions of a parametric modeling approach.
- The asymptotic sampling distribution of the estimator is difficult to derive.

References: Efron and Tibshirani (1993); Davison and Hinkley (1997); Chernick (2011); Efron and Hastie (2016).

Bootstrap Methods

We may be interested in bootstrap methods for:

- Estimating the variance of an estimator.
- Constructing CIs for parameters.

For simplicity, suppose we are in a population setting where we have independent and identically distributed (iid) data $Y_1, \ldots, Y_n$, and denote the (unknown) cumulative distribution function of Y by F.

Parameters of Interest

Let $\theta = T(F)$ be a parameter of interest.

Examples:

1. Suppose $Y_1, \ldots, Y_n \sim F$ where $F \in (F_\theta, \theta \in \Theta)$ and $\widehat{\theta}_n$ be the MLE of θ. We would like to estimate $\text{var}(\widehat{\theta}_n)$ and a $1 - \alpha$ confidence interval (CI) for θ.

2. Suppose $Y_1, \ldots, Y_n \sim F$, and $\theta = T(F)$ is the mean of F, i.e., $\theta = \mathsf{E}[Y] = \int y \, dF(y)$ and $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ is the sample mean. Again, we would like to estimate $\text{var}(\widehat{\theta}_n)$ and a $1 - \alpha$ CI for θ.

3. Suppose $Y_1, \ldots, Y_n \sim F$, and $\theta = T(F)$ is the median of F, i.e., $\Pr(Y_i \leqslant \theta) = \Pr(Y_i > \theta) = 1/2$ and $\widehat{\theta}_n$ is the sample median. Again, we would like to estimate $\text{var}(\widehat{\theta}_n)$ and a $1 - \alpha$ CI for θ.

Parameters of Interest

In 1., θ is a parameter in a parametric model – in this case applying the delta method may be cumbersome, and we might also want to find a CI that doesn't depend on the model being correct (and evaluating the sandwich would be tricky).

In 2. and 3., we are in a nonparametric situation, as we have no specific model.

Inference for the Sample Median

For the sample mean, a $1-\alpha$ confidence interval for the population mean μ is

$$\overline{Y}_n \pm z_{1-\alpha/2} \times \frac{s_n}{\sqrt{n}},$$

where $\overline{Y}_n$ and s_n are the sample mean and sample standard deviation.

Can we do the same thing for the sample median, which we denote by $\widehat{\theta}_n$?

Note that if θ is the population median (and F^{-1} has a continuous derivative in a neighborhood of 1/2):

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightarrow_d \mathrm{N}\left(0, \frac{1}{4f^2(\theta)}\right),$$

where $f(\cdot)$ is the density of Y. This is not straightforward to use in practice, because of the dependence on the unknown $f(\cdot)$.

Monte Carlo Methods

If F is known, it is straightforward to mimic frequentist inference; for simplicity, suppose we have a univariate parameter θ.

For $b = 1, \ldots, B$ samples:

- Generate a random sample $y_1^{\star(b)}, \ldots, y_n^{\star(b)} \sim_{iid} F$.
- Compute $\widehat{\theta}_n^{\star(b)}$ using $y_1^{\star(b)}, \ldots, y_n^{\star(b)}$.

Use $\widehat{\theta}_n^{\star(b)}, b = 1, \ldots, B$, to estimate the sampling distribution of $\widehat{\theta}_n$.

If $B \to \infty$, we approach the theoretical sampling distribution of $\widehat{\theta}_n$.

Of course, in practice, F is unknown and the idea behind the bootstrap is to resample datasets from an estimate of F.

Bootstrap Methods: Unknown F

Two obvious choices for estimating F:

- Use the empirical distribution function of the data, which we denote $\widehat{F}_n$ – this is by far the most common approach used.
- If one has some faith in the assumed model then we may use this model, call this $F_{\widehat{\theta}_n}$, where the notation emphasizes that the distribution function is fully specified by the parameter θ, which we estimate by $\widehat{\theta}_n$.

Then we may:

- Sample data with replacement from $\widehat{F}_n$ to give the **nonparametric** bootstrap.
- Sample data from $F_{\widehat{\theta}_n}$ to give the **parametric** bootstrap.

The Empirical Distribution Function

The empirical distribution function (EDF) is an estimator of the cumulative distribution function (CDF).

The CDF at a fixed value y_0 is,

$$F(y_0) = \Pr(Y_i \leqslant y_0),$$

for $i = 1, \ldots, n$, so that $F(y_0)$ is the probability of the event $\{Y_i \leqslant y_0\}$.

The natural estimator of this probability is the empirical proportion:

$$\widehat{F}_n(y_0) = \frac{\text{Number of } Y_i \leqslant y_0}{\text{Total number of observations}} = \frac{\sum_{i=1}^{n} I(Y_i \leqslant y_0)}{n}.$$

We can do this for all y to give the EDF (this is also the nonparametric MLE of F).

In Figure 1 we see that as n increases we approach the true CDF.

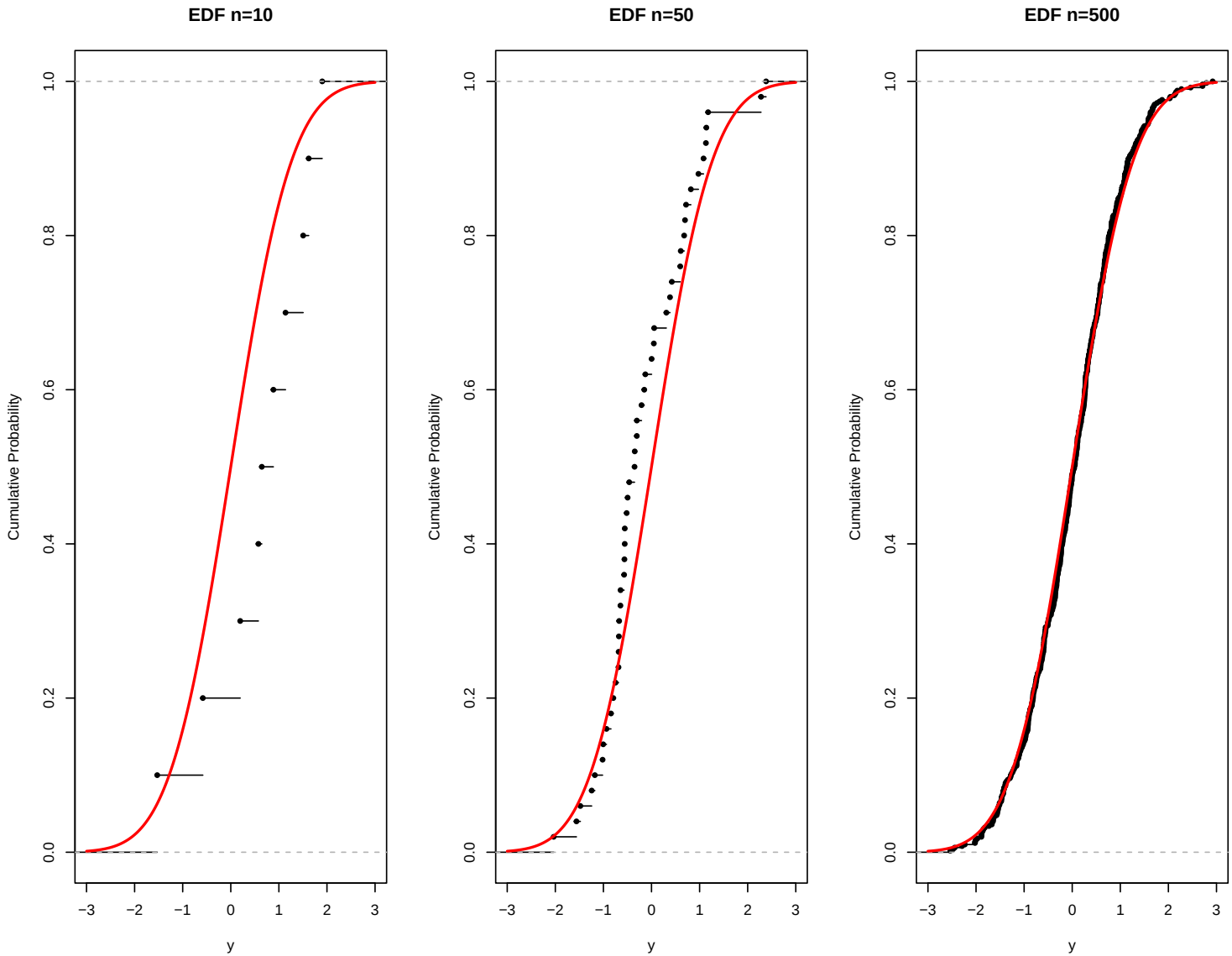

FIGURE 1: Empirical distribution function estimate along with true CDF (in red). There are jumps of $1/n$ at every data point, and between data points, the EDF $\widehat{F}_n(y)$ is flat.

The Empirical DF

The EDF at y is an average of $Z_i = I(Y_i \leqslant y)$ so that

$$Z_i = \begin{cases} 1 & \text{if } Y_i \leqslant y \\ 0 & \text{if } Y_i > y \end{cases}$$

is such that $Z_i \sim \text{Bernoulli}(F(x))$ so that

$$\begin{aligned} \mathrm{E}[I(Y_i \leqslant y)] &= \mathrm{E}[Z_i] = F(y) \\ \mathrm{var}(I(Y_i \leqslant y)) &= \mathrm{var}(Z_i) = F(y)(1 - F(y)) \end{aligned}$$

for a given y.

Recall that $\widehat{F}_n(y) = \sum_{i=1}^n I(Y_i \leqslant y) = \sum_{i=1}^n Z_i$, so

$$\begin{aligned} \mathrm{E}[\widehat{F}_n(y)] &= \mathrm{E}[I(Y_1 \leqslant y)] = F(y) \\ \mathrm{var}(\widehat{F}_n(y)) &= \frac{\sum_{i=1}^n \mathrm{var}(Z_i)}{n^2} = \frac{F(y)(1 - F(y))}{n} \end{aligned}$$

The Empirical DF as an Estimator of $F(y)$

Properties as an estimator,

$$\text{Bias}\left(\widehat{F}_n(y)\right) = \text{E}\left[\widehat{F}_n(y)\right] - F(y) = 0.$$

The variance converges to 0 as $n \to \infty$ so that

$$\widehat{F}_n(y) \to_p F(y),$$

so that $\widehat{F}_n(y)$ is a consistent estimator of $F(y)$.

In addition,

$$\sqrt{n}\left(\widehat{F}_n(y) - F(y)\right) \to_d \text{N}(0, F(y)(1 - F(y))).$$

These are pointwise properties of the estimator, i.e., at y.

It can also be shown that the complete empirical DF converges to F, i.e.,

$$\sqrt{n}(\widehat{F}_n - F) \to_d B,$$

where B is a Brownian bridge.

The Non-Parametric Bootstrap

Again, consider a univariate parameter, $\theta = T(F)$.

For $b = 1, \ldots, B$ samples the non-parametric bootstrap samples:

- Generate a random sample (known as a bootstrap sample) $y_1^{\star(b)}, \ldots, y_n^{\star(b)} \sim_{iid} \widehat{F}_n$ – this is equivalent to drawing n observations, with replacement from the original data $\{Y_n, \ldots, Y_n\}$.
- Compute $\widehat{\theta}_n^{\star(b)}$ using $y_1^{\star(b)}, \ldots, y_n^{\star(b)}$.

Use $\widehat{\theta}_n^{\star(b)}, b = 1, \ldots, B$, to estimate the sampling distribution of $\widehat{\theta}_n$.

Note that we can't enumerate all possible bootstrap samples, as there are $\binom{2n-1}{n}$ of them!

There are two approximations/sources of error in the bootstrap:

1. Statistical: $\widehat{F}_n \neq F$.

2. Simulation: $B \neq \infty$, but we can take B large.

For 1., if n is small, the approximation will be poorer when we use $\widehat{F}_n$.

For 2., for some targets such as the bias and variance, we can get away with smaller B (e.g., $B \geqslant 200$), but for others such as confidence intervals we need larger B (e.g., $B \geqslant 1000$).

Lehmann (1999, p.426), refers to $T(\widehat{F}_n^\star)$ as an approximator of $T(\widehat{F}_n)$ rather than an estimator.

Theoretical: The population is to the sample

Bootstrap: The sample is to the bootstrap sample

As an example, the theoretical bias of as estimator is

$$\mathrm{E}[\widehat{\theta}(\boldsymbol{Y})|F] - \theta.$$

The bootstrap attempts to estimate this by

$$\underbrace{\mathrm{E}[\widehat{\theta}(\boldsymbol{Y}^\star)|\widehat{F}_n]}_{\substack{\text{Average over}\\ \text{Samples}}} - \underbrace{\theta}_{\text{Population}}.$$

And in practice this is estimated by

$$\overbrace{\underbrace{\frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}^{\star(b)}}_{\substack{\text{Average over}\\ \text{Bootstrap Samples}}}}^{\widehat{F}_n^\star} - \underbrace{\widehat{\theta}(\boldsymbol{y})}_{\text{Sample Estimate}}.$$

Heuristic of why the Bootstrap works for the Median

- We have $Y_1, \ldots, Y_n \sim F$.
- The distribution of the median, based on a sample of size n, will be denoted $F_{M_n}(y)$ which we write as,

$$F_{M_n}(y) = \Psi(y; F, n),$$

where Ψ is some complicated function.

- We know that $\widehat{F}_n \approx F$ when n is large so as long as Ψ is smooth with respect to F then $\Psi(y; \widehat{F}_n, n)$ will be similar to $\Psi(y; F, n)$, i.e.,

$$\begin{aligned} \widehat{F}_n \approx F \quad \Longrightarrow \quad F_{M_n^{\star(b)}}(y) &= \Psi(y; \widehat{F}_n, n) \\ &\approx \Psi(y; F, n) = F_{M_n}(y), \end{aligned}$$

where $M_n^{\star(b)}$ is the function evaluated for the b-th bootstrap sample.

Heuristic of why the Bootstrap works for the Median

- Hence, since the bootstrap samples are all from $\widehat{F}_n$ and we have

$$\Psi(y; F, n) = F_{M_n^{\star(1)}}(y) = F_{M_n^{\star(2)}}(y) = \cdots = F_{M_n^{\star(B)}}(y).$$

- What this means is:

The CDF of the median based on the bootstrap samples, $F_{M_n^\star}(y)$ is approximating the CDF of the true CDF of the median in a sample of size n, $F_{M_n}(y)$.

Overview of the Theory

There is a great deal of theory on when the bootstrap does and does not work.

A key observation is to note that, while the bootstrap is applicable in many situations, it is not valid in all situations, and so care should be taken in when it is applied.

A starting rule is that if we are in a situation where the delta method is valid, then the bootstrap will also work – this needs asymptotic normality and smoothness.

From Section 4 of Chapter 18 at:

`https://sites.stat.washington.edu/jaw/COURSES/580s/581/lectnotes.18.html`

Overview of the Theory

The theory needs to show that the asymptotic behavior of the distribution of the nonparametric bootstrap "mimics" the behavior of the original estimator in probability or almost surely (a.s.).

If we are estimating $T(F)$ by $T(\widehat{F}_n)$ and we know (perhaps from a delta method argument) that

$$\sqrt{n}(\underbrace{T(\widehat{F}_n) - T(F)}_{\widehat{\theta}_n - \theta}) \to_d \mathrm{N}(0, \mathrm{var}_F(T)),$$

then we need to show that the bootstrap estimator, $T(\widehat{F}_n^\star)$, satisfies:

$$\sqrt{n}(T(\widehat{F}_n^\star) - T(\widehat{F}_n)) \to_d \mathrm{N}(0, \mathrm{var}_F(Y)), \qquad \text{in probability or a.s.}$$

Overview of the Theory

For the sample mean of a distribution F on $\mathbb{R}$, if $Y \sim F$ and $\mathrm{E}[Y^2] < \infty$ then for

$$T(F) = \int y dF(y) = \mu(F)$$

we know that

$$\sqrt{n}(T(\widehat{F}_n) - T(F)) = \sqrt{n}(\overline{Y}_n - \mu(F)) \to_d \mathrm{N}(0, \mathrm{var}(Y)).$$

For the bootstrap, the corresponding statement is: If $\mathrm{E}[Y^2] < \infty$ then for $Y_1, Y_2, \ldots,$

$$\sqrt{n}(T(\widehat{F}_n^\star) - T(\widehat{F}_n)) = \sqrt{n}(\overline{Y}_n^\star - \overline{Y}_n) \to_d \mathrm{N}(0, \mathrm{var}(Y)).$$

This can be proved using a central limit theorem, see Bickel and Freedman (1981).

Variance Estimation via the Bootstrap

Bootstrap Variance Estimator

The bootstrap variance estimator is

$$\widehat{\text{var}}(\widehat{\theta}_n^\star) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta}_n^{\star(b)} - \overline{\theta}^\star\right)^2,$$

where $\overline{\theta}^\star = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_n^{\star(b)}$.

When B is large the sample variance of the bootstrap estimators

$$\widehat{\text{var}}(\widehat{\theta}_n^\star) \approx \text{var}(\widehat{\theta}_n^\star|\widehat{F}_n), \tag{1}$$

where we have conditioned on $\widehat{F}_n$ being fixed.

Bootstrap Variance Estimator

To argue that the bootstrap variance is a good estimate of the target variance $v(\widehat{\theta}_n)$, we need to have

$$\widehat{\text{var}}(\widehat{\theta}_n^\star) \approx \text{var}(\widehat{\theta}_n^\star | \widehat{F}_n) \approx \text{var}(\widehat{\theta}_n),$$

but because in (1) can be controlled with B large what really matters is

$$\text{var}(\widehat{\theta}_n^\star | \widehat{F}_n) \approx \text{var}(\widehat{\theta}_n),$$

or, more formally,

$$\frac{\text{var}(\widehat{\theta}_n^\star | \widehat{F}_n)}{\text{var}(\widehat{\theta}_n)} \to_p 1 \quad \text{as} \quad n \to \infty.$$

The ratio is often used when both quantities converge to 0 as $n \to \infty$.

Not too tricky to show this for many statistics $\theta = T(F)$.

Sandwich Estimation and the Bootstrap

We heuristically show why we would often expect sandwich and bootstrap variance estimates to be in close correspondence.

For simplicity, we consider a univariate parameter θ, and let $\widehat{\theta}_n$ denote the MLE arising from a sample of size n.

In a change of notation we denote the score by

$$\boldsymbol{S}(\theta) = [S_1(\theta), \ldots, S_n(\theta)]^{\mathsf{T}},$$

where $S_i(\theta) = d\ell_i/d\theta$ is the contribution to the score from observation Y_i, $i = 1, \ldots, n$.

Hence,

$$S(\theta) = \sum_{i=1}^{n} S_i(\theta) = \boldsymbol{S}(\theta)^{\mathsf{T}}\mathbf{1}$$

where $\mathbf{1}$ is an $n \times 1$ vector of 1's.

The sandwich form of the asymptotic variance of $\widehat{\theta}_n$ is

$$\text{var}(\widehat{\theta}_n) = \frac{1}{n}\frac{B_1}{A_1^2}$$

where

$$A_1(\theta) = \text{E}\left[\frac{\partial S_1}{\partial \theta}\right], \qquad B_1(\theta) = \text{E}\left[S_1(\theta)^2\right]$$

which may be empirically estimated via

$$\begin{aligned}
\widehat{A}_n &= \frac{1}{n}\frac{dS}{d\theta}\bigg|_{\widehat{\theta}_n} = \frac{1}{n}\sum_{i=1}^{n}\frac{dS_i}{d\theta}\bigg|_{\widehat{\theta}_n} \\
\widehat{B}_n &= \frac{1}{n}\boldsymbol{S}(\theta)^{\mathsf{T}}\boldsymbol{S}(\theta)\bigg|_{\widehat{\theta}_n} = \frac{1}{n}\sum_{i=1}^{n} S_i(\theta)^2\bigg|_{\widehat{\theta}_n}.
\end{aligned}$$

A convenient representation of a bootstrap sample is $\boldsymbol{Y}^\star = \boldsymbol{Y} \times \boldsymbol{D}$ where $\boldsymbol{D} = \text{diag}(D_1, \ldots, D_n)$ is a diagonal matrix consisting of multinomial random variables

$$\begin{bmatrix} D_1 \\ \vdots \\ D_n \end{bmatrix} \sim \text{Multinomial}\left[n, \left(\frac{1}{n}, ..., \frac{1}{n}\right)\right]$$

with

$$\begin{aligned} \text{E}\left([D_1, \ldots, D_n]^\text{T}\right) &= \mathbf{1} \\ \text{var}([D_1, \ldots, D_n]^\text{T}) &= \text{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\text{T} \to \text{I}_n \end{aligned}$$

as $n \to \infty$.

The MLE of θ in the bootstrap sample is denoted $\widehat{\theta}_n^\star$ and satisfies $S^\star(\widehat{\theta}_n^\star) = 0$, where $S^\star(\theta)$ is the score corresponding to $\boldsymbol{Y}^\star$. Note that,

$$S^\star(\theta) = \sum_{i=1}^{n} S_i^\star(\theta) = \sum_{i=1}^{n} S_i(\theta) D_i.$$

We consider a Taylor series expansion

$$0 = S^\star(\widehat{\theta}_n^\star) \approx S^\star(\widehat{\theta}_n) + (\widehat{\theta}_n^\star - \widehat{\theta}_n) \left. \frac{dS^\star}{d\theta} \right|_{\widehat{\theta}_n}$$

which leads to the one-step approximation

$$\widehat{\theta}_n^\star = \widehat{\theta}_n - \frac{S^\star(\widehat{\theta})}{\frac{d}{d\theta} S^\star(\theta)|_{\widehat{\theta}_n}}.$$

The bootstrap score evaluated at $\widehat{\theta}_n$ is

$$\sum_{i=1}^n S_i^\star(\widehat{\theta}_n) = \sum_{i=1}^n S_i(\widehat{\theta}_n) D_i \neq 0,$$

unless the bootstrap sample coincides with the original sample, i.e., unless $\boldsymbol{D} = \mathsf{I}_n$.

We replace $S^\star(\widehat{\theta})\left[\frac{d}{d\theta}S^\star(\theta)\big|_{\widehat{\theta}_n}\right]^{-1}$ by its limit

$$\mathrm{E}\left[\frac{d}{d\theta}S^\star(\theta)\Big|_{\widehat{\theta}_n}\right] = \mathrm{E}\left[\sum_{i=1}^{n}\frac{d}{d\theta}S_i(\theta)D_i\Big|_{\widehat{\theta}_n}\right] = \frac{d}{d\theta}S(\theta)\Big|_{\widehat{\theta}_n}\mathrm{E}[\boldsymbol{D}] = n\times\widehat{A}_n$$

where $\widehat{A}_n = \frac{d}{d\theta}S(\theta)\big|_{\widehat{\theta}_n}$.

Therefore, the one-step bootstrap estimator is approximated by

$$\widehat{\theta}^\star_n \approx \widehat{\theta}_n - \frac{\boldsymbol{S}(\widehat{\theta}_n)\boldsymbol{D}}{n\widehat{A}_n}$$

and is approximately unbiased as an estimator since

$$\mathrm{E}[\widehat{\theta}^\star_n - \widehat{\theta}_n] \approx -\frac{\boldsymbol{S}(\widehat{\theta}_n)\mathrm{E}[\boldsymbol{D}]}{n\widehat{A}_n} = -\frac{\boldsymbol{S}(\widehat{\theta}_n)\mathbf{1}}{n\widehat{A}_n} = 0$$

and, recall, $\widehat{\theta}_n$ is being held constant.

The variance is

$$\begin{aligned}\text{var}(\widehat{\theta}_n^\star - \widehat{\theta}_n) &\approx \frac{\boldsymbol{S}(\widehat{\theta}_n)\text{var}([D_1, ..., D_n])\boldsymbol{S}(\widehat{\theta}_n)}{(n\widehat{A}_n)^2} = \frac{\boldsymbol{S}(\widehat{\theta}_n)\left(\mathsf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T}\right)\boldsymbol{S}(\widehat{\theta}_n)}{(n\widehat{A}_n)^2} \\ &\approx \frac{\boldsymbol{S}(\widehat{\theta}_n)\mathsf{I}\boldsymbol{S}(\widehat{\theta}_n)}{(n\widehat{A}_n)^2} = \frac{n\widehat{B}_n}{(n\widehat{A}_n)^2} = \frac{\widehat{B}_n}{n\widehat{A}_n^2},\end{aligned}$$

which is the sandwich estimator.

Hence, $\text{var}(\widehat{\theta}_n^\star - \widehat{\theta}_n)$ approximates $\text{var}(\widehat{\theta}_n - \theta)$, which is a fundamental link to the bootstrap.

For a more theoretical treatment, see Arcones and Giné (1992) and Section 10.3 of Kosorok (2008).

Constructing a Confidence Interval via the Bootstrap

Bootstrap Methods: CI Estimation

For CI construction, many improvements on the above normal-based method have been suggested:

- Wald-type CIs.
- The percentile method – pick the appropriate sample quantiles.
- Various improved and bias corrected versions have been proposed. Figure 2 is from Section 3.1 of Chernick (2011). See also Sections 11.3–11.5 of Efron and Hastie (2016).

Table 3.1 Four Methods of Setting Approximate Confidence Intervals for a Real Valued Parameter θ

Method	Abbreviation	α-Level Endpoint	Correct if
1. Standard	$\theta_S[\alpha]$	$\hat{\theta} + \hat{\sigma} z^{(\alpha)}$	$\hat{\theta} \approx N(\theta, \sigma^2)$ σ is constant
2. Percentile	$\theta_P[\alpha]$	$\hat{G}^{-1}(\alpha)$	There exists a monotone transformation such that $\hat{\phi} = g(\hat{\theta})$, where, $\phi = g(\theta)$, $\hat{\phi} \approx N(\phi, \tau^2)$ and τ is constant.
3. Bias-corrected	$\theta_{BC}[\alpha]$	$\hat{G}^{-1}([\phi[2z_\alpha + z^{(\alpha)}])$	There exists a monotone transformation such that $\hat{\phi} \approx N(\phi - z_0\tau, \tau^2)$ and z_0 and τ are constant.
4. BC_a	$\theta_{BC_a}[\alpha]$	$\hat{G}^{-1}\left(\phi\left[z_0 + \frac{[z_0 + z^{(\alpha)}]}{1 - a[z_0 + z^{(\alpha)}]}\right]\right)$	There exists a monotone transformation such that $\hat{\phi} \approx N(\phi - z_0\tau_\phi, \tau_\phi^2)$, where $\tau\phi = 1 + a\phi$ and z_0 and a are constant.

Note: Each method is correct under more general assumptions than its predecessor. Methods 2, 3, and 4 are defined in terms of the percentile of G, the bootstrap distribution.

Source: Efron and Tibshirani (1986, Table 6) with permission from The Institute of Mathematical Statistics.

FIGURE 2: From Chernick (2011).

Bootstrap Methods: Variance and CI Estimation

Recall

$$\widehat{\text{var}}(\widehat{\theta}_n^\star) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta}_n^{\star(b)} - \overline{\theta}^\star\right)^2,$$

where $\overline{\theta}^\star = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_n^{\star(b)}$.

If n is sufficiently large that asymptotic normality of the estimator may be appealed to and a CI estimate may be based upon

$$\widehat{\theta}_n \pm z_{1-\alpha/2} \times \sqrt{\widehat{\text{var}}(\widehat{\theta}_n^\star)}.$$

This is a Wald-type CI, and is not invariant to reparametrization.

Bootstrap Confidence Interval: The Percentile Method

Let

$$\widehat{G}(t) = \frac{1}{B}\sum_{b=1}^{B} I\left(\sqrt{n}(\widehat{\theta}_n^{\star(b)} - \widehat{\theta}_n) \leqslant t\right).$$

Then a $1-\alpha$ bootstrap percentile method confidence interval is

$$C_n = \left[\widehat{\theta}_n - \frac{t_{1-\alpha/2}}{\sqrt{n}}, \widehat{\theta}_n - \frac{t_{\alpha/2}}{\sqrt{n}}\right],$$

where $t_{\alpha/2} = \widehat{G}^{-1}(\alpha/2)$ and $t_{1-\alpha/2} = \widehat{G}^{-1}(1-\alpha/2)$.

Under appropriate regularity conditions,

$$\Pr(\theta \in C_n) = 1 - \alpha - O\left(\frac{1}{\sqrt{n}}\right).$$

Note that this method, unlike the Wald-type interval given earlier, is invariant to the parameterization adopted.

Failure of the Bootstrap

Let $Y_1, \ldots, Y_n \sim_{iid} \text{Uniform}(0, \theta)$, and $\widehat{\theta}_n = \min\{Y_1, \ldots, Y_n\}$, be the minimum of the sample and corresponds to the MLE.

It can be shown that

$$n(\theta - \widehat{\theta}_n) \to_d \text{Exponential}(\theta),$$

i.e., converges to a exponential distribution with mean θ.

However,

$$\begin{aligned}
\Pr(\widehat{\theta}_n^\star = Y_{(n)} | \widehat{F}_n) &= 1 - \Pr(Y_{(n)}^\star < Y_{(n)} | \widehat{F}_n) \\
&= 1 - \Pr(\text{ all } Y_i^\star < Y_{(n)} | \widehat{F}_n) = 1 - \left(\frac{n-1}{n}\right)^n \\
&= 1 - \left(1 - \frac{1}{n}\right)^n \to 1 - e^{-1} \approx 0.632.
\end{aligned}$$

Thus, the bootstrap will select the maximum in the observed data a big chunk of the time, and is not close to the exponential distribution.

Figure 3 compares the non-parametric and parametric bootstraps.

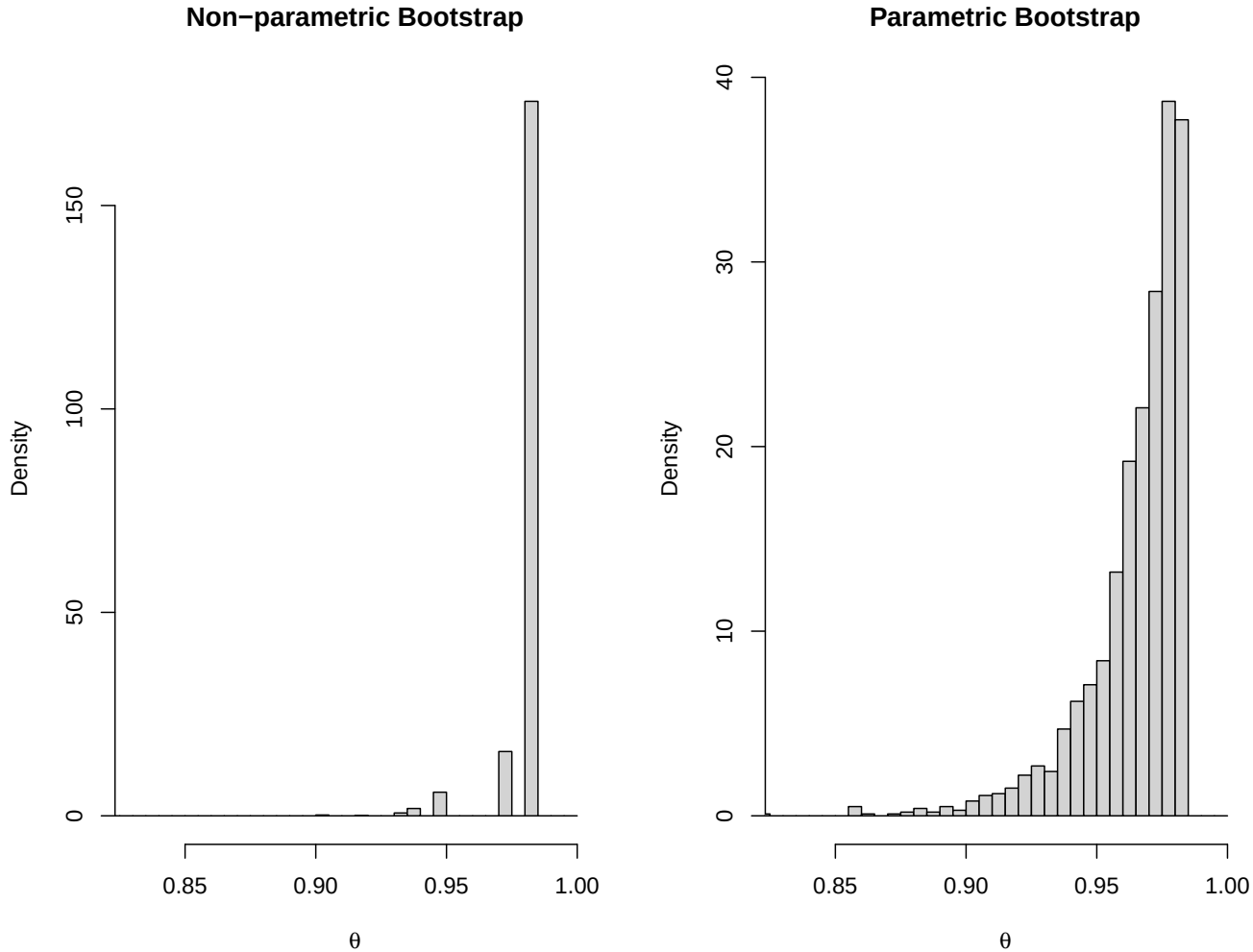

FIGURE 3: Simulated data from Uniform$(0, \theta)$ with $\theta = 1$ and $n = 50$. Bootstrap samples of size $B = 2000$ were obtained for a non-parametric bootstrap (left) and a parametric bootstrap (right), i.e., from Uniform$(0, \widehat{\theta})$. The problem here is that the empirical distribution is not a good approximation to the true distribution in the tail.

Example: Lung Cancer and Radon

For the lung cancer and radon example we implement the nonparametric bootstrap resampling, with replacement, $B = 1000$ sets of n case triples $[Y^{\star}_{bi}, E^{\star}_{bi}, x^{\star}_{bi}]$, $b = 1, \ldots, B$, $i = 1, \ldots, n$.

Figure 4 displays the histogram of estimators arising from the bootstrap samples, along with the asymptotic normal approximations to the sampling distribution of the estimator under the Poisson and quasi-Poisson models.

We see that the distribution under the quasi-likelihood model is much wider than that under the Poisson model.

This is not surprising since we have already seen that the lung cancer data are overdispersed relative to a Poisson distribution.

The bootstrap histogram and quasi-Poisson sampling distribution are very similar, however.

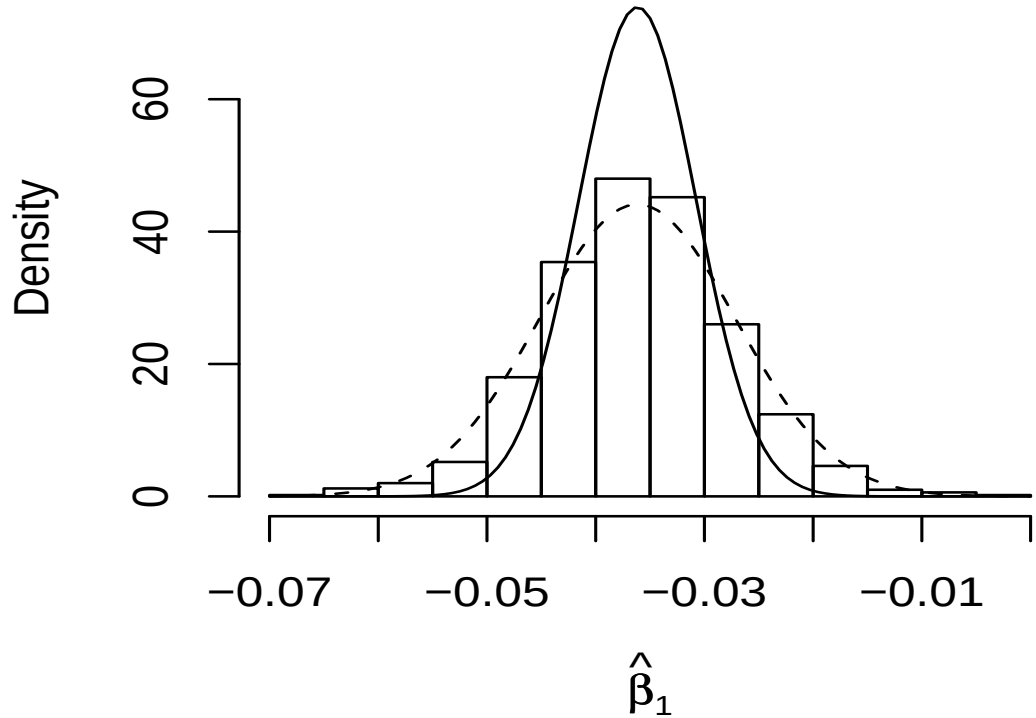

FIGURE 4: Sampling distribution of $\widehat{\beta}_1$ arising from the nonparametric bootstrap samples. The solid curve is the asymptotic distribution of the MLE under the Poisson model, and the dashed line is the asymptotic distribution under the quasi-Poisson model.

Example: Lung Cancer and Radon

Table 1 summarizes inference for β_1 for a number of different methods, and again confirms the similarity of asymptotic inference and the parametric bootstrap under the Poisson model.

The parametric bootstrap cannot be used with a quasi-likelihood model since there is no probability distribution for the data.

Point estimates from the Poisson, quasi-likelihood and sandwich approaches are identical.

Example: Lung Cancer and Radon

Inferential Method	$\widehat{\beta}_1$ $(\times 10^3)$	s.e.$(\widehat{\beta}_1)$ $(\times 10^4)$	95% confidence interval for $e^{10\beta_1}$
Poisson	-0.036	0.0054	0.954, 0.975
Quasi-Likelihood	-0.036	0.0090	0.947, 0.982
Quadratic Variance	-0.030	0.0085	0.955, 0.987
Sandwich Estimation	-0.036	0.0080	0.949, 0.980
Bootstrap Normal	-0.036	0.0087	0.948, 0.981
Bootstrap Percentile	-0.036	0.0087	0.949, 0.981

Table 1: Comparison of inferential approaches for the lung cancer example.

Failure of the Bootstrap

Failure of the Bootstrap

- **Dependence:**
  - If there are "natural" structure to the data that lead to correlated outcomes, such as over time, space, over networks or within families, or "induced" structure due to the experimental design, or complex survey structure. In these cases, the bootstrap distribution will not correspond to the true asymptotic distribution, unless the bootstrap sampling respects the data structure.
- **Lack of Smoothness (cube-root asymptotics):**
  - A number of estimators converge at a rate of $n^{-1/3}$, rather than the more usual $n^{-1/2}$ rate. For example, the least median of squares estimator in a linear regression. For other examples of cube root asymptotics, see Kim and Pollard (1990). The limit distributions are non-normal, and the bootstrap does not work.

Failure of the Bootstrap

- **Machine Learning Methods:**
  - Sparse estimators, such as the lasso (Tibshirani, 1996), are not amenable to being bootstrapped, because zero is a special case for the regression coefficients. Dezeure *et al.* (2015) with respect to the bootstrap and high-dimensional inference say, "...the asymptotic distribution of the Lasso has point mass at zero. This implies, because of noncontinuity of the distribution, that standard bootstrapping and subsampling schemes are delicate to apply and uniform convergence to the limit seems hard to achieve".
  - Chatterjee and Lahiri (2013) discuss the residual bootstrap in which bootstrap samples are generated from,

$$y_i^\star = \boldsymbol{x}^\mathsf{T}\widehat{\boldsymbol{\beta}} + e_i^\star, \qquad i = 1, \ldots, n,$$

    and where $\{e_i^\star\}_{i=1}^n$, are sampled with replacement from the centered residuals obtained from the initial lasso fit.

Failure of the Bootstrap

- **Machine Learning Methods:**
  - Dezeure *et al.* (2015) critique this approach since there is non-uniform convergence to the limiting distribution and problems with the CIs both when $\beta_j = 0$ (zero-length intervals at 0) and when $\beta_j \neq 0$ (poor coverage and wide CIs).
  - To obtain improved asymptotic behavior there has been much research on the debiased (or de-sparsified) lasso estmator that makes an adjustment to the original estimator (Van de Geer *et al.*, 2014).

Failure of the Bootstrap

When does the bootstrap fail?

- **Extrema:** We have already discussed the example in which $Y \sim_{iid} \text{Uniform}(0, \theta)$, and the bootstrap points $\theta^{\star(b)}$ put mass 0.632 on the largest point.

- **Small n:** The justification for the bootstrap follows an asymptotic argument, but will be accurate when we have "small" n.

This was based in part on

`https://notstatschat.rbind.io/2017/02/01/when-the-bootstrap-doesnt-work/`

Other examples can be found on this page.

Discussion

Discussion

- Check the small print when you want to use the bootstrap!

References

Arcones, M. and Giné, E. (1992). On the bootstrap of M-estimators and other statistical functionals. In R. LePage and L. Billard, editors, *Exploring the Limits of Bootstrap*, New York. Wiley.

Bickel, P. and Freedman, D. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics*, **9**, 1196–1217.

Chatterjee, A. and Lahiri, S. N. (2013). Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics*, **41**, 1232–1259.

Chernick, M. R. (2011). *Bootstrap Methods: A Guide for Practitioners and Researchers*. John Wiley & Sons.

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

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: confidence intervals, *p*-values and R-software `hdi`. *Statistical Science*, **30**, 533–558.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall/CRC, Boca Raton.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics*, pages 191–219.

Kosorok, M. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York.

Lehmann, E. L. (1999). *Elements of large-sample theory*. Springer.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, **58**, 267–288.

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, **42**, 1166–1202.