Computational statistics

Chapter 6: Bootstrapping

Thierry Denœux
Université de technologie de Compiègne

August 2026

Overview

The bootstrap

- The bootstrap is a flexible and powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator.
- Main applications:
  - Estimate the bias and standard error of an estimator
  - Compute a confidence interval for a parameter
  - Test a hypothesis about a parameter

Origin of the term "bootstrap"

The use of the term bootstrap derives from the phrase "to pull oneself up by one's bootstraps", widely thought to be based on the 18th century book "The Surprising Adventures of Baron Munchausen" by Rudolph Erich Raspe:

The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps.

Notations

- Let $\theta = T(F)$ be an interesting feature of a distribution function, F, expressed as a functional of F. For example,

$$T(F) = \int x dF(x)$$

  is the mean of the distribution.
- Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be data observed as a realization of the random variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim$ i.i.d. F. In this chapter, we use $\boldsymbol{X} \sim F$ to denote that $\boldsymbol{X}$ is distributed with density function f having corresponding cumulative distribution function F.
- Let $\mathcal{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ denote the entire dataset.
- If $\widehat{F}$ is the empirical distribution function of the observed data, then an estimate of θ is $\widehat{\theta} = T(\widehat{F})$. For example, when θ is a univariate population mean, the estimator is the sample mean,

$$\widehat{\theta} = \int \boldsymbol{x} d\widehat{F}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i.$$

Problem statement

- Statistical inference questions are usually posed in terms of $T(F)$ or some $R(\mathcal{X}, F)$, a function of the data and their unknown distribution function F called a root.
- For example, $R(\mathcal{X}, F)$ might be

$$R(\mathcal{X}, F) = \widehat{\theta} - \theta$$

The expectation of $R(\mathcal{X}, F)$ is then the bias of $\widehat{\theta}$, and its standard deviation is the standard error of $\widehat{\theta}$.

- Other example:

$$R(\mathcal{X}, F) = \frac{\widehat{\theta} - \theta}{\widehat{\text{se}}}$$

where $\widehat{\text{se}}$ is an estimate of the standard error of $\widehat{\theta}$. (If pivotal or approximately pivotal, this statistic can be used to construct an approximate confidence interval on θ).

Bootstrap principle

- The distribution of the random variable $R(\mathcal{X}, F)$ may be intractable or altogether unknown. This distribution also may depend on the unknown distribution F.
- The bootstrap provides an approximation to the distribution of $R(\mathcal{X}, F)$ derived from the empirical distribution function $\widehat{F}$ of the observed data (itself an estimate of F).
- A bootstrap sample (pseudo-dataset) is a sample

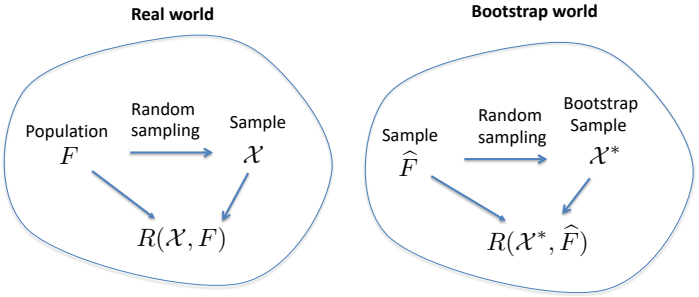
$$\mathcal{X}^* = (\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_n^*)$$

obtained by drawing n values from $\mathcal{X}$ with replacement. It is an i.i.d. sample from $\widehat{F}$.

Bootstrap principle (continued)

Definition (Bootstrap principle)

Approximate the distribution of $R(\mathcal{X}, F)$ by the conditional distribution of $R(\mathcal{X}^*, \widehat{F})$ given the data, where $\mathcal{X}^* = (\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_n^*)$ is an i.i.d. sample from $\widehat{F}$. We write P^* for probability computed under this conditional distribution.



The population is to the sample as the sample is to the bootstrap sample.

Example

- Suppose $n = 3$ univariate data points, namely $\mathcal{X} = \{x_1, x_2, x_3\} = \{1, 2, 6\}$, are observed as an i.i.d. sample from a distribution F that has mean θ.
- At each observed data value, $\widehat{F}$ places mass $1/3$. Suppose the estimator to be bootstrapped is the sample mean $\widehat{\theta}$, which we may write as $T(\widehat{F})$ or $R(\mathcal{X}, F)$, where R does not depend on F in this case.
- Let $\mathcal{X}^* = (X_1^*, X_2^*, X_3^*)$ consist of elements drawn i.i.d. from $\widehat{F}$. There are $3^3 = 27$ possible outcomes for $\mathcal{X}^*$. Let $\widehat{F}^*$ denote the empirical distribution function of such a sample, with corresponding estimate $\widehat{\theta}^* = T(\widehat{F}^*)$.
- Since $\widehat{\theta}^*$ does not depend on the ordering of the data, it has only 10 distinct possible outcomes, listed in the following table.

Example

- The bootstrap principle is to equate the distributions of $R(\mathcal{X}, F)$ and $R(\mathcal{X}^*, \widehat{F})$. Here, we base inference on the distribution of $\widehat{\theta}^*$.
- For example, a simple bootstrap 25/27 (roughly 93%) confidence interval for θ is $[4/3, 14/3]$ using quantiles of the distribution of $\widehat{\theta}^*$.

$\mathcal{X}^*$	$\widehat{\theta}^*$	$P^*\left[\widehat{\theta}^*\right]$
1 1 1	3/3	1/27
1 1 2	4/3	3/27
1 2 2	5/3	3/27
2 2 2	6/3	1/27
1 1 6	8/3	3/27
1 2 6	9/3	6/27
2 2 6	10/3	3/27
1 6 6	13/3	3/27
2 6 6	14/3	3/27
6 6 6	18/3	1/27

Overview

Overview

Nonparametric bootstrap

- For realistic sample sizes the number of potential bootstrap samples is very large, so complete enumeration of the possibilities is not tractable.
- Instead, B independent random bootstrap pseudo-datasets are drawn from the empirical distribution function of the observed data, namely $\widehat{F}$. Denote these

$$\mathcal{X}_b^* = (\boldsymbol{X}_{b1}^*, \ldots, \boldsymbol{X}_{bn}^*), \qquad b = 1, \ldots, B.$$

- The empirical distribution of the $R(\mathcal{X}_b^*, \widehat{F})$ for $b = 1, \ldots, B$ is used to approximate the distribution of $R(\mathcal{X}, F)$, allowing inference.
- The simulation error introduced by avoiding complete enumeration of all possible pseudo-datasets can be made arbitrarily small by increasing B.

Advantages and conditions of use

- Using the nonparametric bootstrap frees the analyst from making parametric assumptions to carry out inference, and provides answers to problems for which analytic solutions are impossible.
- A fundamental requirement of bootstrapping is that the data to be resampled must have originated as an i.i.d. sample. If the sample is not i.i.d., the distributional approximation of $R(\mathcal{X}, F)$ by $R(\mathcal{X}^*, \widehat{F})$ will not hold.
- Methods for bootstrapping with dependent data will be described later.

Parametric bootstrap

- The ordinary nonparametric bootstrap described above generates each pseudo-dataset $\mathcal{X}^*$ by drawing $\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_n^*$ i.i.d. from $\widehat{F}$.
- When the data are modeled to originate from a parametric distribution, so $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim$ i.i.d. $F(\boldsymbol{x}; \theta)$, another estimate of F may be employed.
- Suppose that the observed data are used to estimate θ by $\widehat{\theta}$. Then each parametric bootstrap pseudo-dataset $\mathcal{X}^*$ can be generated by drawing $\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_n^* \sim$ i.i.d. $F(\boldsymbol{x}; \widehat{\theta})$.
- When the model is a good representation of reality, the parametric bootstrap can be a powerful tool. But if the model is not a good fit to the mechanism generating the data, the parametric bootstrap can lead inference to wrong conclusions.

Overview

Bootstrap bias correction

- A particularly interesting choice for bootstrap analysis when $T(F) = \theta$ is the quantity $R(\mathcal{X}, F) = T(\widehat{F}) - T(F) = \widehat{\theta} - \theta$.
- The mean of $R(\mathcal{X}, F)$ is the bias of $\widehat{\theta}$: $\text{bias} = \mathbb{E}_F[\widehat{\theta} - \theta] = \mathbb{E}_F(\widehat{\theta}) - \theta$.
- The bootstrap estimate of the bias is

$$\mathbb{E}_{\widehat{F}}[\widehat{\theta}^* - \widehat{\theta}] = \mathbb{E}_{\widehat{F}}(\widehat{\theta}^*) - \widehat{\theta},$$

which can be estimated by drawing B bootstrap samples:

$$\widehat{\text{bias}} = \frac{1}{B}\sum_{b=1}^{B} \widehat{\theta}_b^* - \widehat{\theta} = \overline{\theta}^* - \widehat{\theta}$$

where $\widehat{\theta}_b^* = T(\widehat{F}_b^*)$ is the estimate of θ obtained from the b-th bootstrap sample, and $\overline{\theta}^*$ is the mean of these estimates.

- Bias-corrected estimate: $\widehat{\theta} - \widehat{\text{bias}}$.
- Empirically, it is sufficient to take B in the range 20–50.

Standard error estimation

- Problem: estimate the standard error (s.e.) of $\widehat{\theta} = T(\widehat{F})$:

$$\text{se}(F) = \sqrt{\mathbb{E}_F\left[\left(\widehat{\theta} - \mathbb{E}_F(\widehat{\theta})\right)^2\right]}$$

- Theoretical bootstrap estimate: $\text{se}(\widehat{F}) = \sqrt{\mathbb{E}_{\widehat{F}}[(\widehat{\theta}^* - \mathbb{E}_{\widehat{F}}(\widehat{\theta}^*))^2]}$.
- Simulated bootstrap estimate:

$$\widehat{\text{se}}_B = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\theta}_b^* - \overline{\theta}^*)^2}$$

- $B = 50$ is often good enough. Very rarely do we need $B > 200$.

Overview

Overview

Confidence intervals from normal approximation

- Under many circumstances, the distribution of a statistic $\widehat{\theta} = T(\widehat{F})$ becomes more and more normal for large n, with mean $\theta = T(F)$ and standard deviation $\mathrm{se}(F)$.
- Let $\widehat{\mathrm{se}} = \mathrm{se}(\widehat{F})$ be an estimate of $\mathrm{se}(F)$. Usually, we have approximately

$$R(\mathcal{X}, F) = \frac{\widehat{\theta} - \theta}{\widehat{\mathrm{se}}} \sim \mathcal{N}(0, 1)$$

We can then write

$$P\left[u_{\alpha/2} \leq \frac{\widehat{\theta} - \theta}{\widehat{\mathrm{se}}} \leq u_{1-\alpha/2}\right] \approx 1 - \alpha$$

where u_α is the α-quantile of the standard normal distribution, from which we get the approximate CI: $\widehat{\theta} \pm u_{1-\alpha/2}\,\widehat{\mathrm{se}}$.

Student t approximation

- For small n, a common alternative is
$$\frac{\widehat{\theta} - \theta}{\widehat{\text{se}}} \sim \mathcal{T}_{n-1}$$
where $\mathcal{T}_{n-1}$ is the Student t distribution with $n-1$ d.f., from which we get the CI: $\widehat{\theta} \pm t_{n-1;1-\alpha/2}\,\widehat{\text{se}}$, where $t_{n-1;\alpha}$ is the α-quantile of $\mathcal{T}_{n-1}$.
- Beware: this is exact only for the mean of a normal sample. For a general $\widehat{\theta}$ it is a rule of thumb, not a theorem.
- In any case, the use of the t-distribution does not adjust for skewness of the underlying population or other errors.

Bootstrap t intervals

- Bootstrap approach: instead of assuming $R(\mathcal{X}, F) \sim \mathcal{N}(0,1)$ or $R(\mathcal{X}, F) \sim \mathcal{T}_{n-1}$, approximate the distribution of $R(\mathcal{X}, F)$ (assumed to be approximately pivotal) by that of $R(\mathcal{X}^*, \widehat{F})$.
- Let

$$Z^* = R(\mathcal{X}^*, \widehat{F}) = \frac{\widehat{\theta}^* - \widehat{\theta}}{\widehat{\text{se}}^*}$$

where $\widehat{\theta}^*$ and $\widehat{\text{se}}^*$ are the estimates of θ and the standard error of $\widehat{\theta}$ computed from bootstrap sample $\mathcal{X}^*$.

- The α-th percentile of Z^* can be estimated by drawing B realizations $z_1^*, \ldots, z_B^*$ and computing $z_B^{*(\alpha)}$ such that

$$\frac{1}{B} \#\{z_b^* \leq z_B^{*(\alpha)}\} = \alpha$$

- We then have the approximate CI:

$$\left[\widehat{\theta} - z_B^{*(1-\alpha/2)} \widehat{\text{se}}, \widehat{\theta} - z_B^{*(\alpha/2)} \widehat{\text{se}}\right]$$

Applicability of the method

- The method is particularly applicable to location parameters (such that a shift in the distribution results in the same shift for the parameter), e.g., the mean, the median, the trimmed mean, or a sample percentile.
- To determine $\widehat{\text{se}}$ and $\widehat{\text{se}}^*$, we need either
  - An analytical formula for $\text{se}(F)$: we take $\widehat{\text{se}} = \text{se}(\widehat{F})$ and $\widehat{\text{se}}^* = \text{se}(\widehat{F}^*)$, or
  - To use the bootstrap: in that case, $\widehat{\text{se}}$ is estimated as explained before (slide 18), and we need two nested bootstrap loops to compute $\widehat{\text{se}}_b^*$: for each $\mathcal{X}_b^*$, generate B' bootstrap samples $\mathcal{X}_{b,b'}^{**}$, $b' = 1, \ldots, B'$ and compute $\widehat{\text{se}}_b^*$ as

$$\widehat{\text{se}}_b^* = \sqrt{\frac{1}{B'-1} \sum_{b'=1}^{B'} (\widehat{\theta}_{b,b'}^{**} - \overline{\theta}_b^{**})^2}$$

Limits of the method

- Since we use the tails of the distribution of $R(\mathcal{X}^*, \widehat{F})$, we need to take B very large, $B \geq 1000$. If we estimate $\widehat{\text{se}}$ by the bootstrap with $B' = 20$, we need $\geq 20{,}000$ bootstrap samples.
- The method can produce poor results when applied to a parameter that is not a location parameter. The next method is more general and more reliable.

Overview

Principle

- The simplest method for drawing inference about a univariate parameter θ using bootstrap simulations is to construct a confidence interval using the percentile method.
- This amounts to reading percentiles off the histogram of $\widehat{\theta}_b^*$ values, $b = 1, \ldots, B$ produced by bootstrapping.
- If $\widehat{\theta}_B^{*(\alpha)}$ denotes the α percentile of the bootstrap estimates, the CI at level $1 - \alpha$ is

$$\left[\widehat{\theta}_B^{*(\alpha/2)}, \widehat{\theta}_B^{*(1-\alpha/2)}\right]$$

- It is intuitive, but why does it work?

Justification

- Assume there exists a monotonically increasing transformation φ that perfectly normalizes the distribution of $\widehat{\theta}$:

$$\varphi(\widehat{\theta}) \sim \mathcal{N}(\varphi(\theta), 1)$$

- Then,

$$P\left[-u_{1-\alpha/2} \leq \varphi(\widehat{\theta}) - \varphi(\theta) \leq u_{1-\alpha/2}\right] = 1-\alpha \tag{1}$$

$$P\left[\varphi(\widehat{\theta}) - u_{1-\alpha/2} \leq \varphi(\theta) \leq \varphi(\widehat{\theta}) + u_{1-\alpha/2}\right] = 1-\alpha$$

$$P\left[\varphi^{-1}\left(\varphi(\widehat{\theta}) - u_{1-\alpha/2}\right) \leq \theta \leq \varphi^{-1}\left(\varphi(\widehat{\theta}) + u_{1-\alpha/2}\right)\right] = 1-\alpha \tag{2}$$

- Now let us apply the bootstrap principle to (1).

Justification (continued)

- We have

$$P^*\left[-u_{1-\alpha/2} \leq \varphi(\widehat{\theta}^*) - \varphi(\widehat{\theta}) \leq u_{1-\alpha/2}\right] \approx 1-\alpha$$

$$P^*\left[\varphi(\widehat{\theta}) - u_{1-\alpha/2} \leq \varphi(\widehat{\theta}^*) \leq \varphi(\widehat{\theta}) + u_{1-\alpha/2}\right] \approx 1-\alpha$$

$$P^*\left[\underbrace{\varphi^{-1}\left(\varphi(\widehat{\theta}) - u_{1-\alpha/2}\right)}_{\approx \widehat{\theta}_B^{*(\alpha/2)}} \leq \widehat{\theta}^* \leq \underbrace{\varphi^{-1}\left(\varphi(\widehat{\theta}) + u_{1-\alpha/2}\right)}_{\approx \widehat{\theta}_B^{*(1-\alpha/2)}}\right] \approx 1-\alpha \tag{3}$$

- The bounds of (3) are identical to those of (2). Hence we may simply read off the quantiles for $\widehat{\theta}^*$ from the bootstrap distribution and use these as the confidence limits for θ.

Remarks

- Note that the percentile method is equivariant to monotone transformations: the percentile method confidence interval for a monotone transformation of θ is the same as the transformation of the interval for θ itself.
- Although the justification is based on a normalizing transformation φ, this transformation is implicit (it need not be specified).

Percentile vs. bootstrap-t

	Percentile	**Bootstrap-t**
Needs $\widehat{\text{se}}$ for each resample	no	yes
Transformation-respecting	yes	no
Range-preserving	yes	no
One-sided coverage error	$O(n^{-1/2})$	$O(n^{-1})$
Stability	good	can be erratic

Overview

Overview

Model

- Consider the ordinary multiple regression model,

$$Y_i = x_i^T \beta + \epsilon_i, \quad i = 1, \ldots, n$$

  where the ϵ_i are assumed to be i.i.d. random variables with zero mean and constant variance σ^2.
- Here, the Y_i are not i.i.d.: we cannot bootstrap the y_i values.
- To determine the correct bootstrap approach, we must determine which data are i.i.d.
- Two main approaches:
  1. Bootstrapping the residuals
  2. Bootstrapping the cases

Bootstrapping the residuals

- If the x_i are considered to be fixed, then a suitable approach is to bootstrap the residuals.
- We know that the errors $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. These variables are not observed, but we can replace them by the residuals $\widehat{\epsilon}_1, \ldots, \widehat{\epsilon}_n$ with
$$\widehat{\epsilon}_i = y_i - x_i^T \widehat{\beta}$$
- Careful: the raw residuals are too small on average, since $\mathrm{Var}(\widehat{\epsilon}_i) = \sigma^2(1 - h_{ii})$, where h_{ii} is the i-th diagonal entry of the hat matrix. It is preferable to resample the rescaled residuals $\widehat{\epsilon}_i / \sqrt{1 - h_{ii}}$, recentred to have zero mean.
- Let $\widehat{\epsilon}_1^*, \ldots, \widehat{\epsilon}_n^*$ be a bootstrap set of residuals. We can construct a bootstrap set of pseudo-responses by
$$y_i^* = x_i^T \widehat{\beta} + \widehat{\epsilon}_i^*, \quad i = 1, \ldots, n$$
- Regressing y_i^* on x_i yields a bootstrap estimate $\widehat{\beta}^*$. By repeating the process B times, we get an empirical distribution $\widehat{\beta}_1^*, \ldots, \widehat{\beta}_B^*$.

Remarks

- This approach is most appropriate for designed experiments or other data where the x_i values are fixed in advance.
- The strategy of bootstrapping residuals is at the core of simple bootstrapping methods for other models such as autoregressive models, nonparametric regression, and generalized linear models.
- Bootstrapping the residuals is reliant on the chosen model providing an appropriate fit to the observed data, and on the assumption that the residuals have constant variance. Without confidence that these conditions hold, the next bootstrapping method is probably more appropriate.

Bootstrapping the cases

- If the pairs (x_i, y_i) are observed for n individuals taken at random from a population, then the n observations $(x_1, y_1), \ldots, (x_n, y_n)$ are a realization from an i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- We may then construct bootstrap samples

$$(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$$

  from which we can estimate $\widehat{\beta}^*$.
- This approach is more robust to violations of the regression assumptions than the "bootstrapping the residuals" approach.

Overview

Stationary data

- Assume that data $x_1, \ldots, x_n$ are a partial realization from a stationary time series of random variables $X_1, \ldots, X_n, \ldots$
- For a time series $(X_1, \ldots, X_n, \ldots)$, stationarity means that, for any $k \geq 0$, the joint distribution of $\{X_t, X_{t+1}, \ldots, X_{t+k}\}$ does not depend on t.
- Let $\mathcal{X} = (X_1, \ldots, X_n)$ denote the time series we wish to bootstrap. Since the elements of $\mathcal{X}$ are dependent, it is inappropriate to apply the ordinary bootstrap for i.i.d. data.
- Several bootstrap methods have been developed for dependent data. Bootstrap theory and methods for dependent data are more complex than for the i.i.d. case. We give only some basic ideas here.

Model-based approach

- As in the case of regression, one method is to postulate a model and bootstrap the residuals.
- For instance, assume an $AR(1)$ model

$$x_t = \beta x_{t-1} + \epsilon_t$$

where $|\beta| < 1$ and the ϵ_t are i.i.d. with zero mean and constant variance.

- After using the standard method to estimate β, we compute the estimated innovations

$$\widehat{e}_t = x_t - \widehat{\beta} x_{t-1}, \quad t = 2, \ldots, n$$

We can then recenter the $\widehat{e}_t$ to obtain the estimated residuals

$$\widehat{\epsilon}_t = \widehat{e}_t - \frac{1}{n-1} \sum_{i=2}^{n} \widehat{e}_i$$

Model-based approach (continued)

- Bootstrap iterations should then resample $n+1$ values in the set $\{\widehat{\epsilon}_2, \ldots, \widehat{\epsilon}_n\}$, to obtain $n+1$ innovations $\{\widehat{\epsilon}_0^*, \ldots, \widehat{\epsilon}_n^*\}$.
- We then get the pseudo time series as

$$x_0^* = \widehat{\epsilon}_0^*$$
$$x_t^* = \widehat{\beta} x_{t-1}^* + \widehat{\epsilon}_t^*, \quad t = 1, \ldots, n$$

- It can be shown that the data generated in that way are not stationary. One remedy is to sample a larger number of pseudo-innovations and start generating the series earlier:

$$\underbrace{x_{-k}^*, x_{-k+1}^*, \ldots, x_0^*}_{\text{burn-in period}}, x_1^*, \ldots, x_n^*$$

- As in the case of regression, this method can only yield sensible results if the model fits the data.

Block-bootstrap approach

- The block-bootstrap method works by splitting the time series $\mathcal{X} = (x_1, \ldots, x_n)$ into N non-overlapping blocks $\mathcal{B}_1, \ldots, \mathcal{B}_N$ of size ℓ such that $N\ell = n$.
- We then sample N blocks $\mathcal{B}_1^*, \ldots, \mathcal{B}_N^*$ from $\{\mathcal{B}_1, \ldots, \mathcal{B}_N\}$ with replacement to get a pseudo time series $\mathcal{X}^* = (\mathcal{B}_1^*, \ldots, \mathcal{B}_N^*)$.
- Each bootstrap estimate $\widehat{\theta}_b^*$ is then computed from $\mathcal{X}_b^*$, $b = 1, \ldots, B$.

Choice of block size

- The idea is to choose a block size ℓ large enough so that observations more than ℓ time steps apart will be nearly independent, while retaining the correlation present in the observations less than ℓ time steps apart.
- The block-bootstrap approach has the advantage of being less model-dependent than the model-based approach. However, the choice of block size ℓ is critical. Automatic selection rules exist (Hall, Horowitz & Jing, 1995; Politis & White, 2004, implemented in the R function `np::b.star`), but the choice remains delicate.

Moving-block bootstrap

- Within this approach, all blocks of size ℓ are considered, regardless of whether the blocks overlap.
- We thus have $n - \ell + 1$ blocks of size ℓ:

$$\mathcal{B}_k = (x_k, \ldots, x_{k+\ell-1}), \quad k = 1, \ldots, n - \ell + 1.$$

- N blocks are resampled with replacement, yielding $\mathcal{B}_1^*, \ldots, \mathcal{B}_N^*$ (we assume again that $N\ell = n$). We then get the pseudo time series $\mathcal{X}^* = (\mathcal{B}_1^*, \ldots, \mathcal{B}_N^*)$.
- As before, each bootstrap estimate $\widehat{\theta}_b^*$ is then computed from $\mathcal{X}_b^*$, $b = 1, \ldots, B$.
- The moving-block bootstrap approach is considered to be superior to the non-moving block bootstrap.

Overview

Consistency

Definition (Consistency)

Let $R_n = R(\mathcal{X}, F)$ be a real-valued root and $R_n^* = R(\mathcal{X}^*, \widehat{F})$ its bootstrap counterpart. The bootstrap is consistent for the root R if

$$\sup_{t \in \mathbb{R}} \left| P^*(R_n^* \leq t) - P(R_n \leq t) \right| \xrightarrow[n \to \infty]{P} 0.$$

(The left-hand term is random: it depends on the data through $\widehat{F}$.)

- For the basic root this reads: the distribution of the bootstrap error $\widehat{\theta}^* - \widehat{\theta}$ mimics that of the real error $\widehat{\theta} - \theta$.
- Consistency is a property of the pair (root, resampling scheme), *not* of "the bootstrap" in the abstract.
- The scaling matters. With the unscaled root $\widehat{\theta} - \theta$, both distributions collapse to a point mass at 0 and the definition holds trivially. We therefore use $\sqrt{n}(\widehat{\theta} - \theta)$, or whatever rate makes the limit non-degenerate.

Two ingredients for consistency

Think of the sampling distribution as a function of the unknown F:

$$L_n : F \longmapsto \mathrm{Law}_F(R_n).$$

The bootstrap simply evaluates this function at $\widehat{F}$ instead of F: it reports $L_n(\widehat{F})$ and hopes it is close to $L_n(F)$.

Approximation of F, plus continuity of L_n

1. $\widehat{F}$ is close to F: Glivenko–Cantelli, $\|\widehat{F} - F\|_\infty \to 0$ a.s. This is free, and holds for *any* F.
2. L_n is smooth (continuous) at F: a small perturbation of F must produce a small perturbation of the law of R_n. This is *not* free, and it is where the bootstrap can break down.

Both are needed: a plug-in estimate is useless if the map you plug it into is wildly discontinuous.

Case of the mean

For $\widehat{\theta} = \overline{X}$ and $\theta = \mu = \mathbb{E}_F(X)$, the root and its bootstrap version are

$$R_n = \sqrt{n}(\overline{X} - \mu), \qquad R_n^* = \sqrt{n}(\overline{X}^* - \overline{X}).$$

Theorem (Bickel & Freedman; Singh, 1981)

If $\mathrm{Var}(X_1) < \infty$, then

$$\sup_{t \in \mathbb{R}} |P^*(R_n^* \leq t) - P(R_n \leq t)| \xrightarrow[n \to \infty]{\text{a.s.}} 0.$$

Thus, the bootstrap is consistent, and the convergence is almost sure: it holds for almost every realisation of the data sequence $X_1, X_2, \ldots$.

This is stronger than convergence in probability. It follows from the proof because both the Glivenko–Cantelli theorem and $\widehat{\sigma}^2 \to \sigma^2$ hold almost surely.

Beyond the mean

All the statements below concern the standardised root $R_n = \sqrt{n}(\widehat{\theta} - \theta)$. In each case, the bootstrap is consistent:

- Smooth functions of means: $\widehat{\theta} = g(\overline{X})$ with g continuously differentiable and $g'(\mu) \neq 0$. The delta method linearises g around μ, so we are back to the case of the mean, up to the factor $g'(\mu)$. The condition $g'(\mu) \neq 0$ is essential: at a critical point the linear term vanishes and consistency fails.
- General functionals $\theta = T(F)$, with $\widehat{\theta} = T(\widehat{F})$: it suffices that T be Hadamard-differentiable at F (a functional version of "differentiable"). This covers quantiles at points of positive density, trimmed means, correlation, M-estimators under regularity, many rank statistics.
- The studentised root $(\widehat{\theta} - \theta)/\widehat{\mathrm{se}}$ inherits these results, provided $\widehat{\mathrm{se}}$ is itself consistent.

When it fails

Smoothness is essential — the bootstrap is inconsistent for

- extreme order statistics (max, min): the resampled maximum cannot exceed the observed one, so the bootstrap only ever looks "to one side";
- non-differentiable functionals, e.g. $\theta = |\mu|$ at $\mu = 0$: the map has a kink exactly where we stand, and $\sqrt{n}(|\overline{X}| - |\mu|)$ has the non-Gaussian limit $|Z|$, with $Z \sim \mathcal{N}(0, \sigma^2)$;
- infinite variance: no CLT in the bootstrap world, and the bootstrap law is random even in the limit.

Example: the sample maximum

$X_i \sim \text{Unif}([0,\theta])$, $\widehat{\theta} = X_{(n)}$. Here the right rate is n, not $\sqrt{n}$:

$$R_n = n\big(\theta - X_{(n)}\big) \xrightarrow{d} \text{Exp}(1/\theta), \qquad R_n^* = n\big(X_{(n)} - X_{(n)}^*\big),$$

where $\text{Exp}(1/\theta)$ has rate $1/\theta$ and mean θ.

Now $R_n^* \geq 0$ always, and $R_n^* = 0$ exactly when $X_{(n)}$ is redrawn at least once:

$$P^*\big(R_n^* = 0\big) = P^*\big(X_{(n)}^* = X_{(n)}\big) = 1 - (1 - 1/n)^n \longrightarrow 1 - e^{-1} \approx 0.632.$$

- So the bootstrap law of R_n^* has an atom at 0 of mass ≈ 0.632, whereas the target law is continuous, with $P(R_n \leq 0) = 0$.
- Taking $t = 0$ in the definition, the supremum stays above 0.632: no consistency, and no amount of bootstrap replications B will fix it.
- Root cause: $X_{(n)}^* \leq X_{(n)}$ *always*. The functional $T(F) = \sup\{x : F(x) < 1\}$ is not differentiable, and $\widehat{F}$ carries no information beyond the largest observation.

Final remarks

- All the bootstrap methods described in this chapter rely on the principle that the bootstrap distribution should approximate the true distribution for a quantity of interest — i.e. that plugging $\widehat{F}$ into $F \mapsto \text{Law}_F(R_n)$ is harmless.
- Remember there are two sources of error: the statistical one (replacing F by $\widehat{F}$), which no computation can remove, and the Monte Carlo one (finitely many replications B), which you control by increasing B.
- The two conditions to keep in mind are i.i.d. data (otherwise: bootstrap the residuals, or use blocks) and smoothness (counterexamples: extremes, boundaries, kinks).
- The bootstrap can also behave poorly for heavy-tailed distributions: the resampled mean is then driven by how many times the few largest points happen to be redrawn.
- We have only scratched the surface of the asymptotic theory. Rates of convergence, Edgeworth expansions, and the remedies for the failures above (m out of n bootstrap, subsampling) are beyond the scope of this course.