Week 9: Bootstrap

MATH-517 Statistical Computation and Visualization

Linda Mhalla

2024-11-15

Introduction

- population F
- random sample $\mathcal{X} = \{X_1, \ldots, X_N\}$ from F
- characteristic of interest $\theta = \theta(F)$

Goal: Extract information about θ using $\mathcal{X}$ and find reliable frequentist assessment of uncertainty

Leading Example: The mean $\theta = \mathbb{E}(X_1) = \int x \, dF(x)$ Δ

F can be estimated:

- parametrically
  - assuming $F \in \{F_\lambda \mid \lambda \in \Lambda \subset \mathbb{R}^p\}$ for some integer p, take $\widehat{F} = F_{\widehat{\lambda}}$ for an estimator $\widehat{\lambda}$ of the parameter vector λ obtained by, e.g., MLE
- non-parametrically
  - by the ECDF, i.e., $\widehat{F} = \widehat{F}_N$ where $\widehat{F}_N(x) = \frac{1}{N} \sum_{n=1}^N \mathbb{1}_{[X_n \leq x]}$

Introduction

- population F
- random sample $\mathcal{X} = \{X_1, \dots, X_N\}$ from F
- characteristic of interest $\theta = \theta(F)$

Leading Example: The mean $\theta = \mathbb{E}X_1 = \int x \, dF(x)$

- parametrically: $\widehat{\theta} = \int x dF_{\widehat{\lambda}}(x)$
- non-parametrically: $\widehat{\theta} = \int x d\widehat{F}_N(x) = \frac{1}{N}\sum_{n=1}^{N} X_n$ Δ

Key questions

- How does $\widehat{\theta}$ behave when samples are repeatedly taken from F?
- How can we use knowledge of this to learn about θ?

Introduction: Thought Experiment

Imagine F is known. Then, we could answer the questions by

- analytical calculation
- Monte Carlo simulation

For $r = 1, \dots, R$:

- generate random sample $x_1^*, \dots, x_N^* \overset{\text{i.i.d.}}{\sim} F$
- compute $\hat{\theta}_r^*$ using $x_1^*, \dots, x_N^*$
- output after R iterations:

$$\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_R^*$$

Use $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_R^*$ to estimate **sampling distribution** of $\hat{\theta}$

$\Rightarrow$ If $R \to \infty$, then get perfect match to theoretical calculation (if available), i.e., Monte Carlo error disappears completely. In practice R is finite, so some error remains

Introduction

- population F
- random sample $\mathcal{X} = \{X_1, \ldots, X_N\}$ from F
- characteristic of interest $\theta = \theta(F)$ (emphasize dep. on F)
- sample characteristic $\widehat{\theta} = \theta(\widehat{F})$
- **sampling distribution** of $\widehat{\theta}$
  - bias or MSE needed to rate the estimator - all characteristics of sampling distribution
  - quantiles of sampling distribution needed for CIs or testing on θ

Leading Example: The mean $\theta = \mathbb{E}(X_1) = \int x dF(x)$

- non-parametrically: $\widehat{\theta} = \int x d\widehat{F}_N(x) = \frac{1}{N}\sum_{n=1}^{N} X_n$
- if F is Gaussian, then $\widehat{\theta} \sim \mathcal{N}(\theta, \frac{\sigma^2}{N})$ is the sampling distribution
  - without Gaussianity, there is still a sampling distribution, we just don't know what it is Δ

Introduction

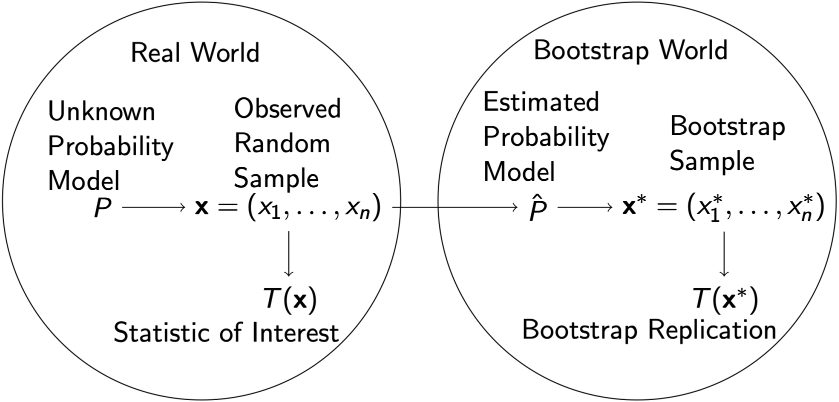
Inference about θ is based on the **sampling distribution**, which is given by the sampling process

- If we control the sampling process, we can approximate the sampling distribution by Monte Carlo
- F unknown but $\widehat{F}$ is known. Then, the (re)sampling distribution can be studied/approximated by Monte Carlo

The Bootstrap Idea: The (re)sampling process from $\widehat{F}$ can mimic the sampling process from F itself

$$\begin{array}{rl} \text{Sampling (real world):} & F \Longrightarrow X_1, \ldots, X_N \Longrightarrow \hat{\theta} = \theta(\widehat{F}) \\ \text{Resampling (bootstrap world):} & \widehat{F} \Longrightarrow X_1^\star, \ldots, X_N^\star \Longrightarrow \hat{\theta}^\star = \theta(\widehat{F}^\star) \end{array}$$

Illustration



$\Rightarrow$ removes need for mathematical skills but still perform well in practice (usually!)

Principle of the Non-Parametric Bootstrap

Bootstrapping an estimator $\hat{\theta} = g(X_1, \dots, X_N)$ can be done as follows

- Generate a **bootstrap sample**

$$X_1^\star, \dots, X_N^\star \overset{\text{i.i.d.}}{\sim} \hat{F}_N$$

(take N uniform random draws with replacement from the original dataset $\{X_1, \dots, X_N\} \Rightarrow$ **resampling the data**)

- Compute the bootstrapped estimator

$$\hat{\theta}^\star = g(X_1^\star, \dots, X_N^\star)$$

- Repeat the first two steps B times to obtain $\hat{\theta}^{\star 1}, \dots, \hat{\theta}^{\star B}$

As $N \to \infty$ and $B \to \infty$, bootstrap sample moments of $\hat{\theta}^{\star 1}, \dots, \hat{\theta}^{\star B}$ converge to the corresp. sample moments of sampling distribution of $\hat{\theta}$

Question: What about the parametric bootstrap?

Principle of the Non-Parametric Bootstrap

Bootstrapping an estimator $\widehat{\theta} = g(X_1, \ldots, X_N)$ can be done as follows

- Generate a **bootstrap sample**

$$X_1^\star, \ldots, X_N^\star \overset{\text{i.i.d.}}{\sim} \widehat{F}_N$$

(take N uniform random draws with replacement from the original dataset $\{X_1, \ldots, X_N\} \Rightarrow$ **resampling the data**)

- Compute the bootstrapped estimator

$$\widehat{\theta}^\star = g(X_1^\star, \ldots, X_N^\star)$$

- Repeat the first two steps B times to obtain $\widehat{\theta}^{\star 1}, \ldots, \widehat{\theta}^{\star B}$

As $N \to \infty$ and $B \to \infty$, bootstrap sample moments of $\widehat{\theta}^{\star 1}, \ldots, \widehat{\theta}^{\star B}$ converge to the corresp. sample moments of sampling distribution of $\widehat{\theta}$

Question: What about the parametric bootstrap? replace $\widehat{F}_N$ by a parametric estimate $\widehat{F}$

Using the $\widehat{\theta}^{\star b}$ to estimate Standard Errors

Bootstrap replicates $\widehat{\theta}^{\star b}$ used to assess quality of $\widehat{\theta}$

- Variance of $\widehat{\theta}$ as estimator of θ is

$$\mathsf{Var}(\widehat{\theta}) = \mathbb{E}_F[\{\widehat{\theta} - \mathbb{E}_F(\widehat{\theta})\}^2]$$

Moving from the real world to the bootstrap world,

$$\mathsf{Var}(\widehat{\theta}) \approx \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta}^{\star b} - \bar{\widehat{\theta}}^{\star}\right)^2,$$

i.e., the sample variance of the bootstrap replicates estimates the variance of the estimator (real world)

Using the $\hat{\theta}^{\star b}$ to estimate the Bias

Bootstrap replicates $\hat{\theta}^{\star b}$ used to estimate properties of $\hat{\theta}$

- Bias of $\hat{\theta}$ as estimator of θ is

$$\text{bias}(\hat{\theta}) = \text{bias}(F) = \mathbb{E}(\hat{\theta} \mid X_1, \dots, X_N \overset{\text{i.i.d.}}{\sim} F) - \theta(F)$$

estimated by replacing unknown F by known estimate $\hat{F}$

$$\begin{aligned} \text{bias}(\hat{F}) &= \mathbb{E}(\hat{\theta} \mid X_1, \dots, X_N \overset{\text{i.i.d.}}{\sim} \hat{F}) - \theta(\hat{F}) \\ &= \mathbb{E}(\hat{\theta}^{\star}) - \hat{\theta} \end{aligned}$$

- Replace theoretical expectation by empirical average

$$\widehat{\text{bias}(\hat{\theta})} = \text{bias}(\hat{F}) \approx \bar{\hat{\theta}}^{\star} - \hat{\theta} = B^{-1} \sum_{b=1}^{B} \hat{\theta}^{\star b} - \hat{\theta}$$

Question: How can we use this to improve inference?

Bias Correction: Another Example

- $X_1, \dots, X_N$ i.i.d. with $\mathbb{E}|X_1|^3 < \infty$
- characteristic of interest: $\theta = \mu^3$, where $\mu = \mathbb{E}(X_1)$
- empirical estimator: $\widehat{\theta} = \big(\int x \, d\widehat{F}_N \big)^3 = \big(\bar{X}_N \big)^3$ is biased
  - bias $b := \mathrm{bias}(\widehat{\theta}) = \mathbb{E}(\widehat{\theta}) - \theta$ of order N^{-1}
- bootstrap: estimate the bias b as $\widehat{b}^\star$
- bias-corrected estimator

$$\widehat{\theta}_b^\star = \widehat{\theta} - \widehat{b}^\star$$

  has smaller order bias (order N^{-2})

Something similar happens more generally for $\theta = g(\mu)$ when g is sufficiently smooth

Bias Correction: Another Example

- $X_1, \dots, X_N$ i.i.d. with $\mathbb{E}|X_1|^3 < \infty$
- Interest in $\theta = \mu^3$, where $\mu = \mathbb{E}(X_1)$, $\sigma^2 = \mathbb{E}(X_1 - \mu)^2$, and $\gamma = \mathbb{E}(X_1 - \mu)^3$
- estimator: $\widehat{\theta} = \big(\int x \, d\widehat{F}_N \big)^3 = \big(\bar{X}_N \big)^3$ is biased

$$\mathbb{E}_F(\widehat{\theta}) = \mathbb{E}_F(\bar{X}_N^3) = \mathbb{E}\Big[\mu + N^{-1} \sum_{n=1}^{N} (X_n - \mu)\Big]^3 = \mu^3 + \underbrace{N^{-1} 3\mu\sigma^2 + N^{-2}\gamma}_{=b=\mathcal{O}(N^{-1})}$$

- bootstrap: estimate the bias $b := \operatorname{bias}(\widehat{\theta}) = \mathbb{E}\widehat{\theta} - \theta$ as $\widehat{b}^\star$

$$\begin{aligned} \mathbb{E}_{\widehat{F}_N} \widehat{\theta}^\star = \mathbb{E}_{\widehat{F}_N} \{(\bar{X}_N^\star)^3\} &= \mathbb{E}_{\widehat{F}_N} \Big\{ \bar{X}_N + N^{-1} \sum_{n=1}^{N} (X_n^\star - \bar{X}_N) \Big\}^3 \\ &= \bar{X}_N^3 + \underbrace{N^{-1} 3 \bar{X}_N \widehat{\sigma}^2 + N^{-2} \widehat{\gamma}}_{=\widehat{b}^\star} \end{aligned}$$

- bias-corrected estimator: $\widehat{\theta}_b^\star = \widehat{\theta} - \widehat{b}^\star$ has smaller order bias

$$\mathbb{E}_F(\widehat{\theta}_b^\star) = \mu^3 + N^{-1} 3 \underbrace{\big\{\mu\sigma^2 - \mathbb{E}_F\big(\bar{X}_N \widehat{\sigma}^2\big)\big\}}_{\mathcal{O}(N^{-1})} + N^{-2} \underbrace{\big\{\gamma - \mathbb{E}_F(\widehat{\gamma})\big\}}_{\mathcal{O}(N^{-1})}$$

Leading Example: Using the $\widehat{\theta}^{*b}$ for CI

- $X_1, \dots, X_N \overset{\text{i.i.d.}}{\sim} F$ and $\theta = \theta(F) = \int x dF$
- $\hat{\theta} = \bar{X}_N$ and $\hat{\sigma}^2 = (N-1)^{-1} \sum_{n=1}^{N} (X_i - \bar{X}_N)^2$
- we want θ_α such that $P\{\theta \geq \theta_\alpha\} = 1 - \alpha$, for $0 < \alpha < 1$

1. **Exact CI. (rare)** Assuming Gaussianity,

$$T = \sqrt{N} \frac{\bar{X}_N - \theta}{\hat{\sigma}} \sim t_{N-1} \quad \Rightarrow \quad P\{T \leq t_{N-1}(1-\alpha)\} = 1 - \alpha$$

   and so we get a CI with exact coverage

$$\theta \geq \bar{X}_N - \frac{\hat{\sigma}}{\sqrt{N}} t_{N-1}(1-\alpha) := \hat{\theta}_\alpha$$

2. **Asymptotic CI.** Assuming only $\mathbb{E} X_1^2 < \infty$, $T \xrightarrow{d} \mathcal{N}(0,1)$ and thus

$$P\{\theta \geq \hat{\theta}_\alpha\} \approx 1 - \alpha \quad \text{for} \quad \hat{\theta}_\alpha = \bar{X}_N - \frac{\hat{\sigma}}{\sqrt{N}} z(1-\alpha)$$

Leading Example: Using the $\widehat{\theta}^{*b}$ for CI

3. **Bootstrap CI.** Let $\mathbb{E}X_1^2 < \infty$ and $X_1^\star, \ldots, X_N^\star$ be a bootstrap sample from the ECDF $\widehat{F}_N$

- get $\overline{X}_N^\star = N^{-1} \sum_{n=1}^N X_n^\star$ and $\hat{\sigma}^{\star 2} = \frac{1}{N-1} \sum_{n=1}^N (X_n^\star - \bar{X}_N^\star)^2$
- set up the bootstrap statistic $T^\star = \sqrt{N} \frac{\overline{X}_N^\star - \overline{X}_N}{\widehat{\sigma}^\star}$
- B bootstrap copies $T_1^\star, \ldots, T_B^\star$ used to estimate the dist. of T

Data | Resamples

$$\mathcal{X} = \{X_1, \ldots, X_N\} \quad \Longrightarrow \quad \begin{cases} \mathcal{X}_1^\star = \{X_{1,1}^\star, \ldots, X_{1,N}^\star\} & \Longrightarrow \quad T_1^\star \\ \quad \vdots & \qquad \vdots \\ \mathcal{X}_B^\star = \{X_{B,1}^\star, \ldots, X_{B,N}^\star\} & \Longrightarrow \quad T_B^\star \end{cases}$$

- take $q^\star(1-\alpha)$ the sample $(1-\alpha)-$quantile of $T_1^\star, \ldots, T_B^\star$
- instead of $\widehat{\theta}_\alpha = \bar{X}_N - \frac{\widehat{\sigma}}{\sqrt{N}} z(1-\alpha)$, consider

$$\widehat{\theta}_\alpha^\star = \bar{X}_N - \frac{\hat{\sigma}}{\sqrt{N}} q^\star(1-\alpha)$$

Leading Example: Coverage Comparison

2. **Asymptotic CI.** $T = \sqrt{N}\frac{\bar{X}_N - \theta}{\widehat{\sigma}} \dot{\sim} \mathcal{N}(0,1)$

By the Berry-Esseen theorem

$$P_F(T \leq x) - \Phi(x) = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) \quad \text{for all } x$$

$$\Rightarrow \quad P\Big(\theta \geq \underbrace{\bar{X}_N - \frac{\hat{\sigma}}{\sqrt{N}} z(1-\alpha)}_{=\hat{\theta}_\alpha}\Big) = P\{T \leq z(1-\alpha)\}$$

$$= 1 - \alpha + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$$

I.e., the coverage of the asymptotic CI is exact up to $\mathcal{O}(N^{-1/2})$

Leading Example: Coverage Comparison

3. **Bootstrap CI.** (assuming "ideal" bootstrap with infinite nbr of replicates)

From Edgeworth expansions (complicated!):

$$P_F(T \leq x) = \Phi(x) + \frac{1}{\sqrt{N}} a(x)\phi(x) + \mathcal{O}\left(\frac{1}{N}\right)$$

$$P_{\widehat{F}_N}(T^\star \leq x) = \Phi(x) + \frac{1}{\sqrt{N}} \hat{a}(x)\phi(x) + \mathcal{O}\left(\frac{1}{N}\right)$$

where $\hat{a}(x) - a(x) = \mathcal{O}(N^{-1/2})$

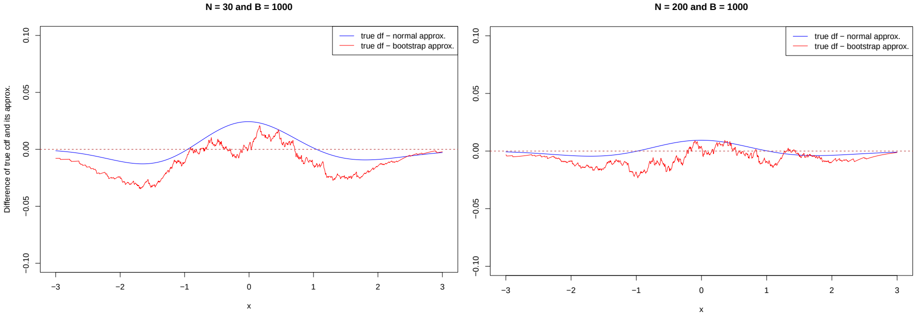
Hence, $P_F(T \leq x) - P_{\widehat{F}_N}(T^\star \leq x) = \mathcal{O}\left(\frac{1}{N}\right)$ and

$$\Rightarrow P\Big(\theta \geq \underbrace{\bar{X}_N - \frac{\hat{\sigma}}{\sqrt{N}} q^\star(1-\alpha)}_{=\hat{\theta}^\star_\alpha}\Big) = P_F\{T^* \leq q^*(1-\alpha)\} + \mathcal{O}\left(\frac{1}{N}\right)$$

$$= 1 - \alpha + \mathcal{O}\left(\frac{1}{N}\right)$$

I.e. the coverage of the bootstrap CI is exact up to $\mathcal{O}(N^{-1})$: faster conv. rate

Leading Example: Sampling Distribution



Problem (1) with the non-parametric bootstrap

Use non-parametric bootstrap to estimate characteristics of the **median**

For a sample of size $N = 2m + 1$, possible distinct values of $\hat{\theta}^\star$ are $X_{(1)} < \cdots < X_{(N)}$, and

$$P\left(\hat{\theta}^\star > X_{(l)}\right) = \sum_{r=0}^{m} \left(\begin{array}{c} N \\ r \end{array} \right) \left(\frac{l}{N}\right)^r \left(1 - \frac{l}{N}\right)^{N-r}$$

- exact calculations of mean, variance (etc.) of bootstrap distribution are possible and converge to correct values (as $N \to \infty$)

$\Rightarrow$ consistency holds

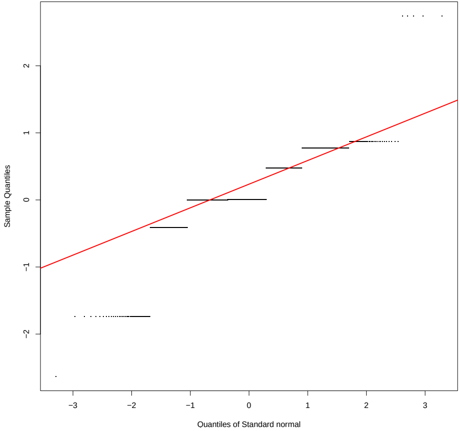
- but $\hat{\theta}^\star$ concentrated on sample values and very vulnerable to unusual values

$\Rightarrow$ discreteness makes convergence very slow

E.g., bootstrap variance of the median can be very poor for heavy-tailed distributions and small sample sizes

Problem (1) with the non-parametric bootstrap

- Simulate from a sample with $N = 11$ from (standard) Cauchy
- Compute medians from $B = 1000$ bootstrap samples and center with true median (=0)



Problem (2) with the non-parametric bootstrap

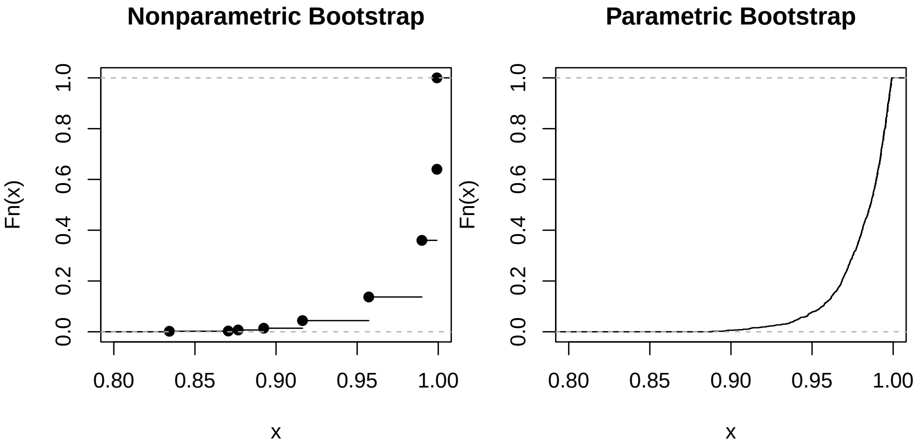
- $X_1, \dots, X_N \sim U(0, \theta)$ i.i.d., $\theta > 0$
- MLE: $\hat{\theta} = \max(X_1, \dots, X_N)$
  - $T = N(\theta - \hat{\theta})/\theta \dot{\sim} Exp(1)$
- Non-parametric bootstrap: $X_1^*, \dots, X_N^*$ sampled indep. from $X_1, \dots, X_N$ with replacement
- Bootstrap estimate $\hat{\theta}^* = \max(X_1^*, \dots, X_N^*)$
  - $T^\star = N(\hat{\theta} - \hat{\theta}^*)/\hat{\theta}$

- Large probability mass at $\hat{\theta}$. In fact $P\left(\hat{\theta}^* = \hat{\theta}\right) = 1 - (1 - 1/N)^N \overset{N \to \infty}{\longrightarrow} 1 - e^{-1} \approx .632$

$\Rightarrow$ the limiting distribution of T^* cannot be $Exp(1)$

Bootstrap fails here and we will see why (consistency fails!)

Problem (2) with the non-parametric bootstrap



(Non-parametric) Bootstrap: Summary

- let $\mathcal{X} = \{X_1, \dots, X_N\}$ be a random sample from F
- quantity of interest: $\theta = \theta(F)$
- (plug-in) estimator: $\hat{\theta} = \theta(\widehat{F}_N)$
  - write $\hat{\theta} = \theta[\mathcal{X}]$, since $\widehat{F}_N$ and thus the estimator depends on the sample
- the distribution $F_{T,N}$ of a scaled estimator $T = g(\hat{\theta}, \theta) = g(\theta[\mathcal{X}], \theta)$ is of interest, e.g., $T = \sqrt{N}(\hat{\theta} - \theta)$

(Non-parametric) Bootstrap: Summary

- let $\mathcal{X} = \{X_1, \dots, X_N\}$ be a random sample from F
- quantity of interest: $\theta = \theta(F)$
- (plug-in) estimator: $\hat{\theta} = \theta(\widehat{F}_N)$
  - write $\hat{\theta} = \theta[\mathcal{X}]$, since $\widehat{F}_N$ and thus the estimator depends on the sample
- the distribution $F_{T,N}$ of a scaled estimator $T = g(\hat{\theta}, \theta) = g(\theta[\mathcal{X}], \theta)$ is of interest, e.g., $T = \sqrt{N}(\hat{\theta} - \theta)$

The workflow of the bootstrap is as follows for some $B \in \mathbb{N}$:

$$
\begin{array}{ccccc}
\text{Data} & & \text{Resamples} & & \\
\mathcal{X} = \{X_1, \dots, X_N\} & \Longrightarrow & \left\{ \begin{array}{l} \mathcal{X}_1^\star = \{X_{1,1}^\star, \dots, X_{1,N}^\star\} \\ \quad \vdots \\ \mathcal{X}_B^\star = \{X_{B,1}^\star, \dots, X_{B,N}^\star\} \end{array} \right. & \begin{array}{c} \Longrightarrow \\ \\ \Longrightarrow \end{array} & \begin{array}{l} T_1^\star = g(\theta[\mathcal{X}_1^\star], \theta[\mathcal{X}]) \\ \quad \vdots \\ T_B^\star = g(\theta[\mathcal{X}_B^\star], \theta[\mathcal{X}]) \end{array}
\end{array}
$$

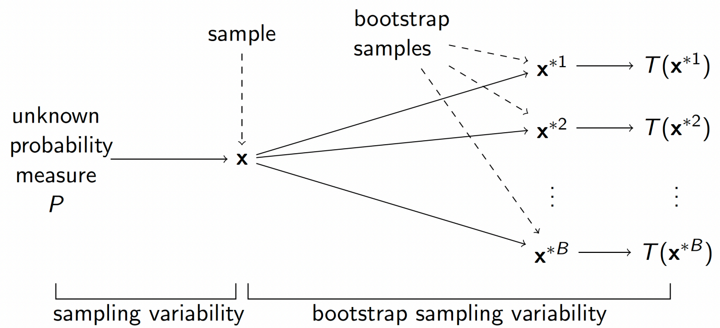
$F_{T,N}$ now estimated by $\widehat{F}^\star_{T,B}(x) = B^{-1} \sum_{b=1}^{B} \mathbb{I}_{[T_b^\star \le x]}$

- any characteristic of $F_{T,N}$ can be estimated by the char. of $\widehat{F}^\star_{T,B}(x)$

Bootstrap: Summary

Bootstrap combines

- the plug-in principle: sample is used to estimate F ($\approx \hat{F}$)
- Monte Carlo principle: simulation replaces theoretical calculation
- two sources of variability
  - sampling variability (we only have a sample of size N)
  - bootstrap resampling variability (only B bootstrap samples)



Bootstrap: Common Questions

- How many bootstraps/Monte Carlo draws?
  - $B \geq 200$ to estimate bias or variance (next week)
  - $B = 10^3$ is taken most commonly
  - $B \geq 10^4$ better for small/large quantiles

Bootstrap: Common Questions

- How many bootstraps/Monte Carlo draws?
  - $B \geq 200$ to estimate bias or variance (next week)
  - $B = 10^3$ is taken most commonly
  - $B \geq 10^4$ better for small/large quantiles

- Why take resamples of size N?
  - to mimic sampling properties of samples like the original one
  - sometimes we take $m < N$ to achieve validity of bootstrap, e.g., for extreme quantiles or median (to avoid discreteness)

Bootstrap: Common Questions

- How many bootstraps/Monte Carlo draws?
  - $B \geq 200$ to estimate bias or variance (next week)
  - $B = 10^3$ is taken most commonly
  - $B \geq 10^4$ better for small/large quantiles
- Why take resamples of size N?
  - to mimic sampling properties of samples like the original one
  - sometimes we take $m < N$ to achieve validity of bootstrap, e.g., for extreme quantiles or median (to avoid discreteness)
- Why resample from the EDF?
  - Non-parametric MLE of F, so it's natural when no restrictions on F
  - Smooth estimate of the EDF (KDE) can be used when discreteness is severe, e.g. the case of the median

Bootstrap: Common Questions

- How many bootstraps/Monte Carlo draws?
  - $B \geq 200$ to estimate bias or variance (next week)
  - $B = 10^3$ is taken most commonly
  - $B \geq 10^4$ better for small/large quantiles
- Why take resamples of size N?
  - to mimic sampling properties of samples like the original one
  - sometimes we take $m < N$ to achieve validity of bootstrap, e.g., for extreme quantiles or median (to avoid discreteness)
- Why resample from the EDF?
  - Non-parametric MLE of F, so it's natural when no restrictions on F
  - Smooth estimate of the EDF (KDE) can be used when discreteness is severe, e.g. the case of the median
- When does the bootstrap work ("work" $=$ consistency)?

Consistency

Bootstrap setup:

- $T = g(X_1, \dots, X_N \mid F)$ is a scaled estimator with unknown (wanted) distribution $F_{T,N}$, with $g(X_1, \dots, X_N \mid \cdot)$ continuous
- bootstrap statistic $T^\star = g(X_1^\star, \dots, X_N^\star \mid \hat{F})$ has $F_{T,N}^\star$ also unknown
- the Monte Carlo proxy $\widehat{F_{T,B}^\star}$ is used instead of $F_{T,N}^\star$

Glivenko-Cantelli:

$$\sup_x \left| \widehat{F_{T,B}^\star}(x) - F_{T,N}^\star(x) \right| \stackrel{a.s.}{\to} 0 \quad \text{as} \quad B \to \infty$$

Question: Under which conditions the bootstrap "works" (gives mathematically correct answers), i.e.,

$$F_{T,N}^\star \to F_{T,N}, \quad \text{as } N \to \infty$$

Consistency

1. $F_{T,N}$ must converge weakly to some continuous limit $F_{T,\infty}$

$$\int h(t) dF_{T,N}(t) \to \int h(t) dF_{T,\infty}(t) \quad \text{as } n \to \infty \text{ and } \forall h \text{ integrable}$$

   $\Rightarrow$ to ensure that the wanted distribution converges to a non-degenerate limit

2. the convergence must be uniform

$\Rightarrow$ to ensure that $F^{\star}_{T,N}$ approaches $F_{T,\infty}$ for all possible sequences of $\hat{F}$ (which changes as N increases)

Then, the bootstrap is consistent, i.e., $\forall t$ and $\epsilon > 0$

$$P\{\mid F^{\star}_{T,N}(t) - F_{T,\infty}(t) \mid > \epsilon\} \stackrel{n \to \infty}{\to} 0$$

Remark: second condition fails in the case of the maximum of a uniform!

Consistency for Smooth Transformation of the Mean

Conditions that ensure consistency of the bootstrap are guaranteed for smooth transformations of the sample mean

Theorem: Let $X_1, \dots, X_N$ be i.i.d. s.t. $\mathbb{E}(X_1^2) < \infty$ and $T = h(\bar{X}_N)$, where h is continuously differentiable at $\mu = \mathbb{E}(X_1)$ and such that $h(\mu) \neq 0$. Then

$$\sup_x \left| F^\star_{T,N}(x) - F_{T,N}(x) \right| \overset{a.s.}{\to} 0 \quad \text{as} \quad N \to \infty$$

Remarks

- bootstrap should not be used blindly
  - verification via theory
  - and/or via simulations
- folk knowledge
  - typically "works" when T asymptotically normal and data i.i.d.
  - "doesn't work" when working with
    - statistics that do not exist (mean of Cauchy distribution)
    - non-smooth transformations of the sample (sample quantiles): non-parametric bootstrap still valid but may not work well for finite samples /Bootstrap not consistent for order statistics
    - non-i.i.d. regimes (e.g. time series): see block bootstrap or bootstrap in regression settings
- bootstrap replaces analytic calculations (in particular the Delta method), but showing that it actually works requires even deeper analytic calculations
- faster rates can be achieved by bootstrap
  - hard to prove, but often happens, e.g., when working with a skewed distribution

References

Davison & Hinkley (2009) Bootstrap Methods and their Application

Wasserman (2005) All of Nonparametric Statistics

Shao & Tu (1995) The Jackknife and Bootstrap

Hall (1992) The Bootstrap and Edgeworth Expansion