Week 10: Bootstrap

MATH-517 Statistical Computation and Visualization

Tomas Masak

November 25th 2022

Intro

- population F
- random sample $\mathcal{X} = \{X_1, \ldots, X_N\}$ from F

Goal of Statistics: Extract information about F using $\mathcal{X}$.

- characteristic of interest $\theta = \theta(F)$

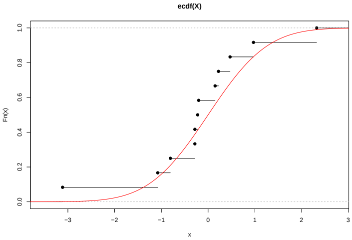
Running Ex.: The mean $\theta = \mathbb{E}X_1 = \int x \, dF(x)$. Δ

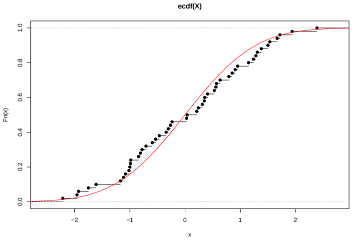
F can be estimated:

- 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 $\widehat{\lambda}$ estimator of the parameter vector λ
- non-parametrically
  - by the ECDF, i.e. $\widehat{F} = \widehat{F}_N$ where $\widehat{F}_N(x) = \frac{1}{N} \sum_{n=1}^N \mathbb{I}_{[X_n \leq x]}$

ECDF

```r
edf_plot <- function(N){
  X <- rnorm(N)
  EDF <- ecdf(X)
  plot(EDF)
  x <- seq(-4,4,by=0.01)
  points(x,pnorm(x),type="l",col="red")
}
set.seed(517)
edf_plot(12)
edf_plot(50)
```





Intro

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

Running Ex.: The mean $\theta = \mathbb{E}X_1 = \int x \, dF(x)$.

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

Intro

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

Running Ex.: 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, $\widehat{\theta} \sim \mathcal{N}(\theta, \frac{\sigma^2}{N})$ is the sampling distribution
  - without Gaussianity, there is still a sampling distribution, we just don't know what it is Δ

Intro

Statistics is about the **sampling distribution**, which is given by the sampling process (part of which is F itself)

- if we controlled the sampling process, we would approximate the sampling distribution by Monte Carlo

The Bootstrap Idea: Resampling process from $\widehat{F}$ can mimic the sampling process from F itself.

- since $\widehat{F}$ is known, the resampling distribution can be studied
  - or approximated by Monte Carlo

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

Running Ex.

- $X_1, \dots, X_N \overset{\perp\!\!\!\perp}{\sim} F$ and $\theta = \theta(F) = \int x \, dF$
- we want $\widehat{\theta}(\alpha)$ such that $P(\theta \leq \widehat{\theta}(\alpha)) = \alpha$.

1. **Exact CI.** Assuming Gaussianity,

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

and so we get a CI with exact coverage by expressing θ from the inequality $T \leq t_{n-1}(\alpha)$:

$$\theta \leq \bar{X}_N + \frac{\widehat{\sigma}}{\sqrt{N}} t_{n-1}(\alpha) =: \widehat{\theta}(\alpha).$$

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

$$P(\theta \leq \widehat{\theta}(\alpha)) \approx \alpha \quad \text{for} \quad \widehat{\theta}(\alpha) = \bar{X}_N + \frac{\widehat{\sigma}}{\sqrt{N}} z(\alpha),$$

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

- set up the bootstrap statistic $T^\star = \sqrt{N}\frac{\bar{X}_N^\star - \bar{X}_N}{\widehat{\sigma}^\star}$
- denote by $q^\star(\alpha)$ the quantile of $T^\star$
- instead of $\widehat{\theta}(\alpha) = \bar{X}_N + \frac{\widehat{\sigma}}{\sqrt{N}} z(\alpha)$, consider $\widehat{\theta}^\star(\alpha) = \bar{X}_N + \frac{\widehat{\sigma}}{\sqrt{N}} q^\star(\alpha)$

From Edgeworth expansions (complicated!):

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

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

Running Ex. - Coverage Comparison

2. **Asymptotic CI.** By the Berry-Essen theorem

$$P(T \leq x) - P(\mathcal{N}(0,1) \leq x) = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) \quad \text{for all } x$$

$$\Rightarrow \quad P\Big(\theta \leq \underbrace{\bar{X}_N + \frac{\widehat{\sigma}}{\sqrt{N}} z(\alpha)}_{=\widehat{\theta}(\alpha)}\Big) = \alpha + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right).$$

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

3. **Bootstrap CI.** From Edgeworth expansions

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

$$\Rightarrow \quad P\Big(\theta \leq \underbrace{\bar{X}_N + \frac{\widehat{\sigma}}{\sqrt{N}} q^\star(\alpha)}_{=\widehat{\theta}^\star(\alpha)}\Big) = \alpha + \mathcal{O}\left(\frac{1}{N}\right),$$

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

How is this possible?

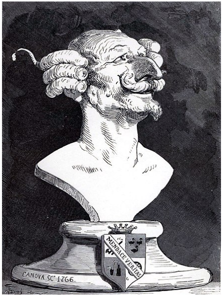
- we got a better interval than that from CLT by resampling our data once
  - resampling once $\equiv$ discarding information

How is this possible?

- we got a better interval than that from CLT by resampling our data once
  - resampling once $\equiv$ discarding information
- however, we did "theoretical" resampling
- in practice, we don't know $q^\star(\alpha)$, we have to approximate it
  - e.g. by Monte Carlo $\equiv$ resampling many times
  - but still, how can we gain information by resampling?

Baron Munchausen (half-fictional character)

- rode a cannonball
- traveled to the Moon (18th century)
- got out from the bottom of the lake by pulling his *bootstraps*



Another Example

- $X_1, \ldots, X_N$ 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
- bootstrap: estimate the bias $b := \mathrm{bias}(\widehat{\theta}) = \mathbb{E}\widehat{\theta} - \theta$ as $\widehat{b}^\star$
- bias-corrected estimator: $\widehat{\theta}_b^\star = \widehat{\theta} - \widehat{b}^\star \ldots$ provably smaller bias?

Another Example

- $X_1, \ldots, X_N$ i.i.d. with $\mathbb{E}|X_1|^3 < \infty$
- characteristic of interest: $\theta = \mu^3$, where $\mu = \mathbb{E}X_1$
- estimator: $\widehat{\theta} = \big(\int x \, d\widehat{F}_N \big)^3 = \big(\bar{X}_N\big)^3$ is biased

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

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

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

- bias-corrected estimator: $\widehat{\theta}_b^\star = \widehat{\theta} - \widehat{b}^\star$ ... provably smaller bias?

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

Bootstrap combines

- the plug-in principle, i.e. estimating the unknowns, and
- Monte Carlo principle, i.e. simulation instead of analytic calculations

What are the unknowns?

- parameters $\Rightarrow$ parametric bootstrap
- the whole F via ECDF $\Rightarrow$ the (standard/non-parametric) bootstrap

The (standard/non-parametric) Bootstrap

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

The (standard/non-parametric) Bootstrap

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

The workflow of the bootstrap is as follows for some $B \in \mathbb{N}$ (e.g. $B = 1000$):

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

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

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

Running Ex. Again

- $X_1, \dots, X_N \overset{\perp\!\!\!\perp}{\sim} F$ and $\theta = \theta(F) = \int x \, dF$
- we want $\widehat{\theta}(\alpha)$ such that $P(\theta \leq \widehat{\theta}(\alpha)) = \alpha$.

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

- set up the bootstrap statistic $T^\star = \sqrt{N}\frac{\bar{X}_N^\star - \bar{X}_N}{\widehat{\sigma}^\star}$
- denote by $q^\star(\alpha)$ the quantile of $T^\star$
- take $\Big(-\infty, \bar{X}_N + \frac{\widehat{\sigma}}{\sqrt{N}} q^\star(\alpha)\Big)$

In practice, $q^\star(\alpha)$ approximated by Monte Carlo:

Data / Resamples

$$\mathcal{X} = \{X_1, \dots, X_N\} \quad \Longrightarrow \quad \left\{ \begin{array}{ccc} \mathcal{X}_1^\star = \{X_{1,1}^\star, \dots, X_{1,N}^\star\} & \Longrightarrow & T_1^\star \\ \vdots & & \vdots \\ \mathcal{X}_B^\star = \{X_{B,1}^\star, \dots, X_{B,N}^\star\} & \Longrightarrow & T_B^\star \end{array} \right.$$

$\Rightarrow$ take $q^\star(\alpha)$ as the sample quantile of $T_1^\star, \dots, T_B^\star$

Running Ex. Specific

```
lambda <- 2
N <- 100
X <- rexp(N,lambda)

( CI_asyptotic <- mean(X) + qnorm(0.95)*sd(X)/sqrt(N) )
```

```
## [1] 0.5645591
```

```
Tstar <- function(Xstar,X){
  return( (mean(Xstar)-mean(X))/sd(Xstar)*sqrt(N))
}
B <- 10^3
boot_ind <- sample(1:N, size=N*B, replace=T)
boot_data <- matrix(X[boot_ind],ncol=B)
Tstars <- rep(0,B)
for(b in 1:B){
  Tstars[b] <- Tstar(boot_data[,b],X)
}
( CI_boot <- mean(X) + quantile(Tstars,0.95)*sd(X)/sqrt(N) )
```

```
##       95%
## 0.5557836
```

The Bootstrap

- now we know what the bootstrap is
    - the scheme is very simple, though a bit mysterious, spawning questions:
- when does it work? ("work" = consistency)
- when does it give us something extra? (e.g. faster rates)

Consistency for Smooth Transformation of the Mean

Bootstrap setup in practice:

- T is the scaled estimator with unknown distribution F_T
- the bootstrap statistic $T^\star$ has $F_T^\star$ also unknown
- the Monte Carlo proxy $F_{T,B}^\star$ is used instead of $F_T^\star$

Glivenko-Cantelli:

$$\sup_x \left| \widehat{F}_{T,B}^\star(x) - F_T^\star(x) \right| \overset{a.s.}{\to} 0 \quad \text{as} \quad B \to \infty.$$

Question: $\sup_x \left| F_T^\star(x) - F_T(x) \right| \to 0$ for $N \to \infty$?

Theorem: Let $\mathbb{E}X_1^2 < \infty$ 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

$$\sup_x \left| F_T^\star(x) - F_T(x) \right| \overset{a.s.}{\to} 0 \quad \text{as} \quad N \to \infty.$$

Remarks

- bootstrap should not be used blindly
  - verification via theory
  - and/or via simulations
- folk knowledge
  - bootstrap "works" when we have non-degenerate asymptotic normality
  - bootstrap "doesn't work" when working with order statistics, extremes, non-smooth transformations, non-i.i.d. regimes (e.g. time series), etc.
- 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 happends e.g. when working with a skewed distribution
- how many Monte Carlo draws needed?
  - $B = 10^2$ is enough for variance estimation (next week)
  - $B = 10^3$ is taken most commonly
  - $B = 10^4$ better for small/large quantiles