SB1.2/SM2 Computational Statistics
The Bootstrap

François Caron

Hilary Term 2019

SB1.2/SM2 Computational Statistics

- Webpage:
  http://www.stats.ox.ac.uk/~caron/teaching/sb1b/
- Synopsis
  - Smoothing methods and nonparametric inference (GN)
  - Monte Carlo and permutation tests; rank statistics (GN)
  - Bootstrap (FC)
  - Hidden Markov models (FC)

Outline

Outline

Introduction

- The Bootstrap belongs to the wider class of resampling methods
- Introduced in 1979 by Bradley Efron, it had a huge impact on statistics
- Offers a principled, simulation-based approach, for assigning measures of accuracy to statistical estimates
  - Estimate biases, variances of estimator
  - Obtain approximate confidence intervals
- Applies to parametric and nonparametric settings
- Bootstrap ideas of "resampling the data" are at the heart of subsequent popular statistical methods such as random forests
- Still motivates nowadays the development of novel statistical tools for handling massive datasets

Confidence interval for the median of a population

- Let $X_1, \ldots, X_n$ be iid random variables from an (unknown) cdf F
- Nonparametric setting: we do not make any parametric assumption on F (Gaussian, t, etc.)
- We are interested in the median $m = F^{-1}(0.5)$ of this distribution
- Estimator

$$\widehat{M}_n = \begin{cases} X_{((n+1)/2)} & \text{if } n \text{ odd} \\ \frac{1}{2}(X_{(n/2)} + X_{(n/2+1)}) & \text{if } n \text{ even} \end{cases}$$

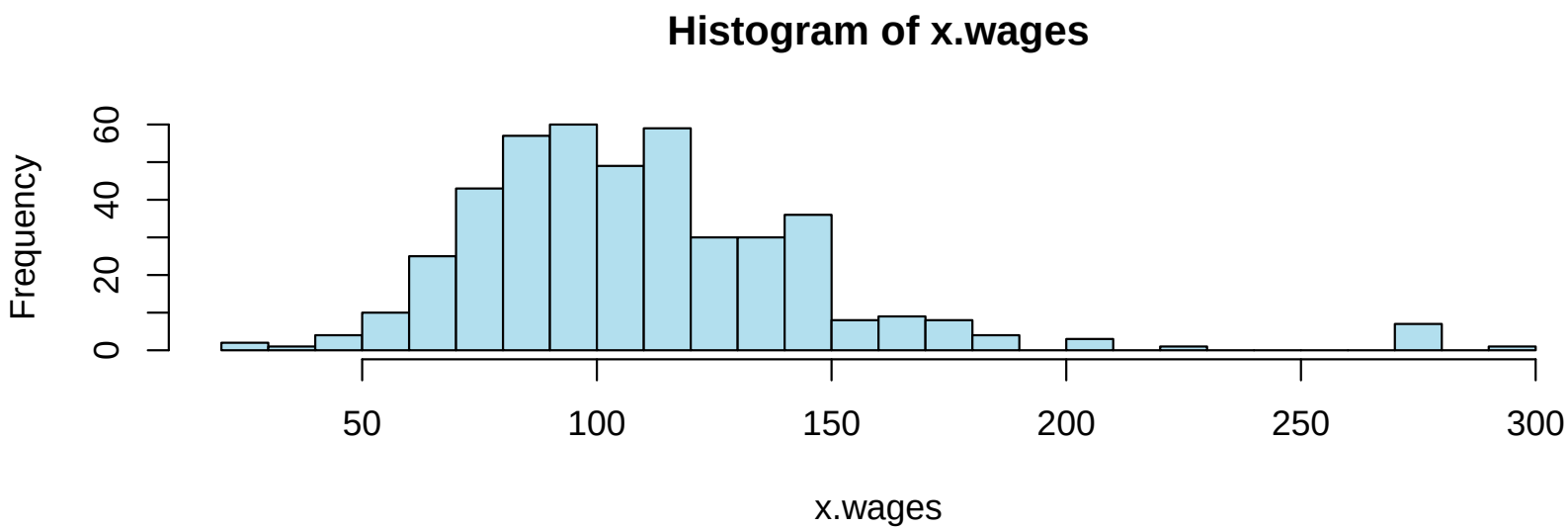
where $X_{(r)}$ is the rth order statistic of the random sample $(X_1, \ldots, X_n)$.

Confidence interval for the median of a population

Example: Median wage in the mid-Atlantic states of the USA in 2005

- Assume that we are interested in the median wage in the mid-Atlantic states of the USA in 2005
- Wages of $n = 447$ workers in that region in 2005.

```
library('ISLR')
data(Wage)
x.wages = Wage$wage[Wage$year==2005]
hist(x.wages, breaks=20,col='lightblue2')
```



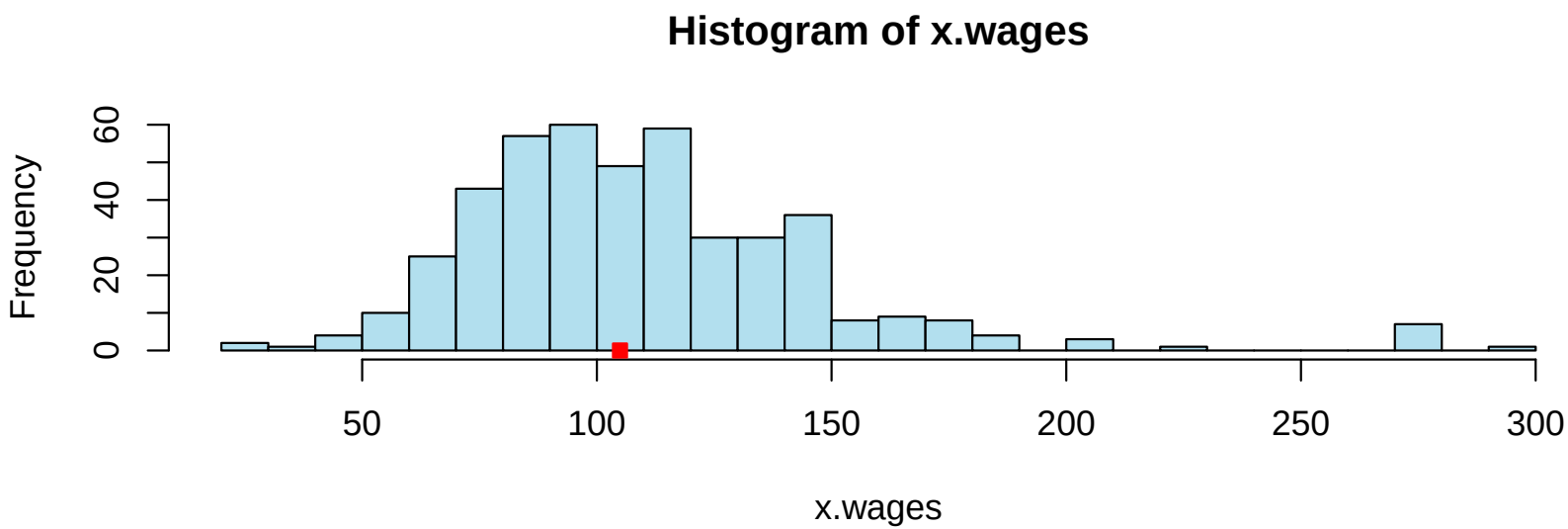
Confidence interval for the median of a population

Example: Median wage in the mid-Atlantic states of the USA in 2005

- The median estimate is

```
mhat.wages = median(x.wages)
mhat.wages
```

```
## [1] 104.9215
```



Confidence interval for the median of a population

Example: Median wage in the mid-Atlantic states of the USA in 2005

- The median estimate is just a point estimate for the population median m.
- How do we get a confidence interval for m?
- If F is differentiable with continuous pdf f

$$\sqrt{n}(\widehat{M}_n - m) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{4f^2(m)}\right) \qquad n \to \infty$$

- Not so helpful: cannot compute $1/(4f^2(m))$
- If we could evaluate $\sigma_n^2 = \mathbb{V}(\widehat{M}_n)$, then we could use the normal approximation to obtain an approximate $1 - \alpha$ confidence interval.
- The bootstrap allows to obtain an estimate of the variance σ_n^2 of the estimator.

Confidence interval for the median of a population

Example: Median wage in the mid-Atlantic states of the USA in 2005

- Alternatively: if we knew the quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ of the distribution of $\widehat{M}_n - m$ where

$$\Pr(\widehat{M}_n - m \leq q_x) = x$$

for $x \in [0, 1]$, then we could use these quantiles to build a $1 - \alpha$ confidence interval $[\widehat{M}_n - q_{1-\alpha/2}, \widehat{M}_n - q_{\alpha/2}]$ for m, as

$$\begin{aligned}
&\Pr(\widehat{M}_n - q_{1-\alpha/2} \leq m \leq \widehat{M}_n - q_{\alpha/2}) \\
&\quad = \Pr(q_{\alpha/2} \leq \widehat{M}_n - m \leq q_{1-\alpha/2}) \\
&\quad = \alpha.
\end{aligned}$$

- But, the quantiles cannot in general be computed
- The bootstrap provides estimates of the quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ and thus approximate confidence intervals.

Parametric setting

Accuracy of an estimator and confidence intervals

- Let $X_1, \ldots, X_n$ be iid random variables from a pdf $f(x; \theta)$ parametrized by θ
- Estimator of θ

$$\widehat{\theta}_n = t(X_1, \ldots, X_n)$$

- In many cases, we cannot evaluate analytically the following quantities
  - Bias $\mathbb{E}[\hat{\theta}_n] - \theta$
  - Variance $\mathbb{V}[\hat{\theta}_n]$
- Confidence intervals for θ:
  - For MLE, we can use the asymptotic normality and the estimate of the variance to construct approximate confidence intervals
  - Bootstrap offers an alternative to construct confidence intervals for θ

Outline

Notations

- Let $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ be a continuous or discrete random variable with cdf

$$F(x) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_p \leq x_p)$$

  where $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$.
- If X is continuous, let f be its probability density function.
- If X is discrete, let $A \subset \mathbb{R}^p$ be the finite (or countably infinite) sample space and f be its probability mass function
- The expectation of X is defined as

$$\mathbb{E}(X) = \int_{\mathbb{R}^p} x dF(x) = \begin{cases} \sum_{x \in A} x f(x) & \text{if } x \text{ is discrete} \\ \int_{\mathbb{R}^p} x f(x) dx & \text{if } x \text{ is continuous} \end{cases}$$

- Unifying notation $\int_{\mathbb{R}^p} x dF(x)$ to define an expectation for both continuous and discrete random variables

Statistical functional

- A statistical functional $\mathcal{F}(F)$ is any function of the cdf F.
- Examples include the mean $\mu = \int_{\mathbb{R}} x dF(x)$, the variance $\sigma^2 = \int_{\mathbb{R}} x^2 dF(x) - \left(\int_{\mathbb{R}} x dF(x)\right)^2$ or the median $m = F^{-1}(0.5)$.

Outline

Empirical distribution function

Definition

Let $X_1, \ldots, X_n$ be iid real-valued random variables with cdf F. The **empirical cumulative distribution function** is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) = \frac{|\{ i \mid X_i \leq x\}|}{n}$$

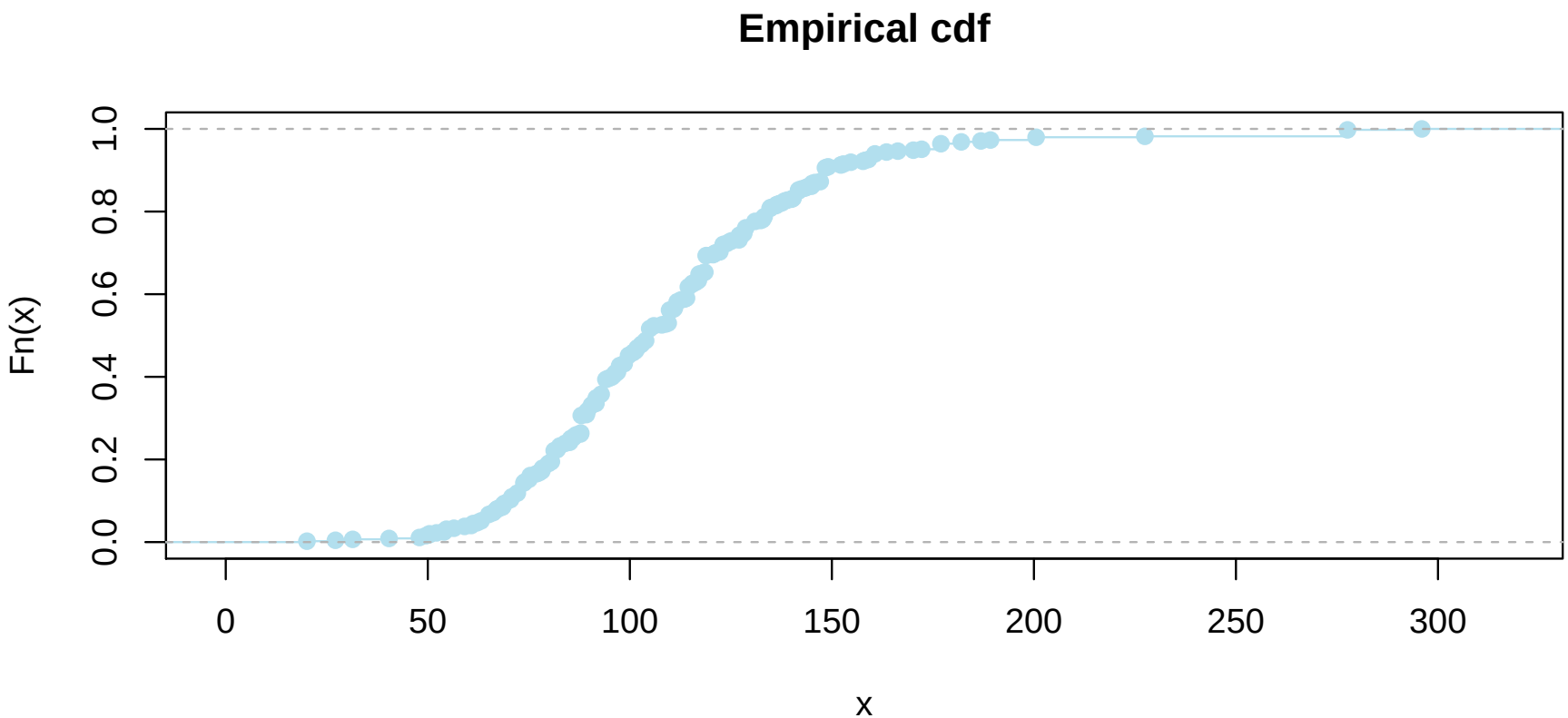
where

$$I(X_i \leq x) = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{otherwise} \end{cases}.$$

Empirical distribution function

Wages dataset

```
ecdf.wages = ecdf(x.wages)
plot(ecdf.wages, xlab = 'x', ylab = 'Fn(x)', main = 'Empirical cdf',
         col='lightblue2')
```



Empirical distribution function

- F_n is a natural estimator for the cdf F when this one is completely unknown.
- It has the following attractive properties.

(i) Unbiased

$$\mathbb{E}[F_n(x)] = F(x) \text{ for all } x \in \mathbb{R} \text{ and } n \geq 1$$

(ii) (Strongly) consistent

$$F_n(x) \overset{\text{as}}{\to} F(x) \text{ for all } x \text{ as } n \to \infty$$

(iii) Asymptotically normal

$$\sqrt{n}(F_n(x) - F(x)) \overset{d}{\to} \mathcal{N}(0, F(x)(1 - F(x))) \text{ for all } x \text{ as } n \to \infty$$

Empirical distribution function

- Proof

Outline

Recap: Monte Carlo integration

- The Monte Carlo method is a way to approximate potentially high-dimensional integrals via simulation.

Definition (Monte Carlo method)

Let $Y \in \mathbb{R}^p$ be a continuous or discrete random variable with cdf G. Consider

$$\eta = \mathbb{E}\left(\phi(Y)\right) = \int_{\mathbb{R}^p} \phi(y) dG(y)$$

where $\phi : \mathbb{R}^p \to \mathbb{R}$. Let $Y^{(1)}, ..., Y^{(B)}$ be iid random variables with cdf G. Then

$$\widehat{\eta}_B = \frac{1}{B} \sum_{j=1}^{B} \phi(Y^{(j)})$$

is called the Monte Carlo estimator of the expectation η.

Recap: Monte Carlo integration

- The Monte Carlo algorithm only requires to simulate from G

Algorithm 1 Monte Carlo Algorithm

- Simulate independent $Y^{(1)}, ..., Y^{(B)}$ with cdf G
- Return $\widehat{\eta}_B = \frac{1}{B} \sum_{j=1}^{B} \phi(Y^{(j)})$.

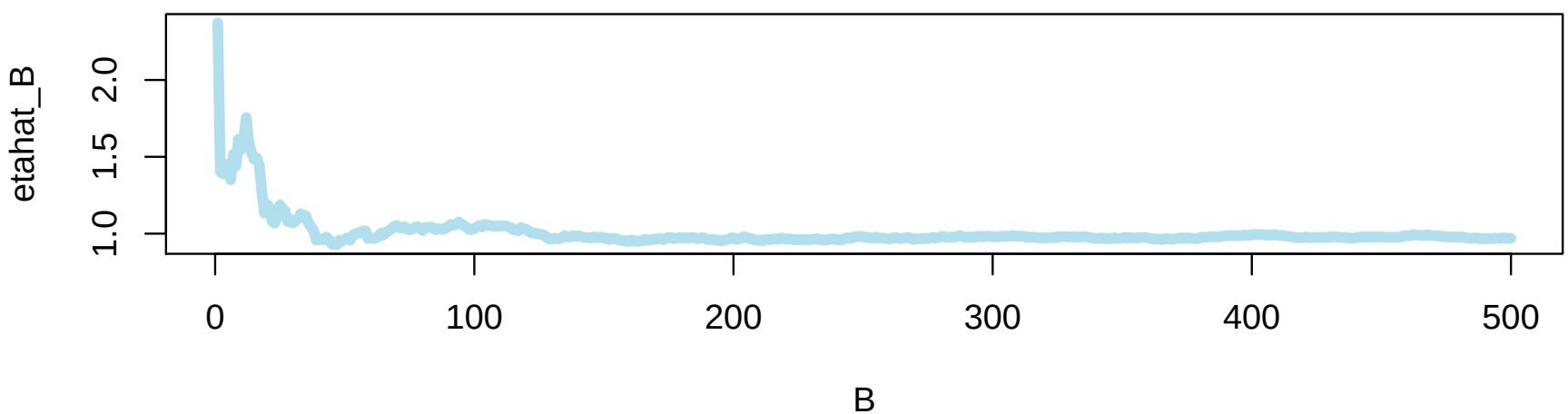
Recap: Monte Carlo integration

- Example: $Y^{(j)} \sim \mathcal{N}(1,1)$, $\hat{\eta}_B = \frac{1}{B}\sum_{j=1}^{B} Y^{(j)}$

```
B=500
Y=rnorm(B,mean=1)
etahat = mean(Y)
etahat

## [1] 0.9699538

etahat.all = cumsum(Y)/c(1:B)
plot(etahat.all, col='lightblue2', type='l', lwd=5, xlab='B', ylab='etahat_B')
```



Recap: Monte Carlo integration

- The Monte Carlo estimator is
  - (i) Unbiased
  $$\mathbb{E}(\widehat{\eta}_B) = \eta$$
  for any $B \geq 1$.
  - (ii) (Strongly) consistent
  $$\widehat{\eta}_B \overset{as}{\to} \eta \text{ as } B \to \infty,$$
  - (iii) Asymptotically normal. If $\sigma^2 = \mathbb{V}(\phi(Y))$ exists
  $$\sqrt{B}(\widehat{\eta}_B - \eta) \overset{d}{\to} \mathcal{N}(0, \sigma^2) \text{ as } B \to \infty.$$

- Proof: direct application of the law of large numbers and the central limit theorem

Recap: Monte Carlo integration

- Monte Carlo estimators of the mean and variance of $Y \sim G$ are

$$\widehat{\mu}_B = \frac{1}{B}\sum_{j=1}^{B} Y^{(j)} \overset{as}{\to} \mathbb{E}(Y),$$

$$\widehat{\sigma}_B^2 = \frac{1}{B}\sum_{j=1}^{B}(Y^{(j)} - \widehat{\mu}_B)^2$$

$$= \frac{1}{B}\sum_{j=1}^{B}(Y^{(j)})^2 - \left(\frac{1}{B}\sum_{j=1}^{B} Y^{(j)}\right)^2 \overset{as}{\to} \mathbb{V}(Y).$$

where $Y^{(1)}, \ldots, Y^{(B)} \sim G$

Outline

Bootstrap

About the name

- The term "bootstrap" was introduced by Bradley Efron in "Bootstrap methods: another look at the jackknife", Annals of Statistics, 7, (1979) 1-26.
- In his book "An Introduction to the Bootstrap" (1993) Efron explained that

> *"the use of the term bootstrap derives from the phrase to pull oneself up by one's own bootstrap, widely thought to be based on one of the eighteenth century '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.)"*

Bootstrap

General setting

- Let $X_1, \ldots, X_n$ be iid random variables with cdf F.
- The distribution F of the data is unknown, and we cannot even simulate from F.
- Consider the estimator

$$\hat{\theta}_n = t(X_1, \ldots, X_n)$$

for estimating $\theta = \mathcal{F}(F)$ a statistical functional of F.

Real World: $F \quad \Rightarrow \quad X_1, \ldots, X_n \quad \Rightarrow \quad \hat{\theta}_n = t(X_1, \ldots, X_n)$

Outline

Bootstrap variance estimation

- We are first interested in estimating the variance $\sigma^2 = \mathbb{V}_F(\hat{\theta}_n)$ of the estimator $\hat{\theta}_n$.
- We write $\mathbb{V}_F$ to emphasize the fact that the variance depends on the unknown cdf F.

Bootstrap variance estimation: An idealized variance estimator

- Suppose first that we *could* simulate from the cdf F
- That is, we *could* reproduce the Real World and generate new datasets
- Monte Carlo method could then be used to obtain an estimator for σ^2
  - For $j = 1, \dots, B$,
    - Simulate $X_1^{(j)}, \dots, X_n^{(j)}$ iid from F and let $\hat{\theta}_n^{(j)} = t(X_1^{(j)}, \dots, X_n^{(j)})$
  - Return $\frac{1}{B} \sum_{j=1}^{B} \left(\hat{\theta}_n^{(j)} - \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_n^{(j)} \right)^2$.
- However, the cdf F is unknown and we cannot simulate from it, so we cannot implement this estimator.

Bootstrap variance estimation

- The idea of the bootstrap is to
  1. Replace the unknown cdf F by the (known) empirical cdf F_n,
  2. Use the Monte Carlo method.
- More formally, conditionally on the data $X_1, \ldots, X_n$, let $X_1^*, \ldots, X_n^*$ be iid random variables with cdf F_n, and $\hat{\theta}_n^* = t(X_1^*, \ldots, X_n^*)$.
- The bootstrap mimics the Real World by using the empirical cdf F_n

$$\begin{array}{rccccc} \text{Real World:} & F & \Rightarrow & X_1, \ldots, X_n & \Rightarrow & \hat{\theta}_n = t(X_1, \ldots, X_n) \\ \text{Bootstrap World:} & F_n & \Rightarrow & X_1^*, \ldots, X_n^* & \Rightarrow & \hat{\theta}_n^* = t(X_1^*, \ldots, X_n^*) \end{array}$$

Bootstrap variance estimation

- The bootstrap relies on two approximations
  1. Approximate $\mathbb{V}_F(\hat{\theta}_n)$ by $\mathbb{V}_{F_n}(\hat{\theta}_n^*)$ by using F_n and
  2. Approximate $\mathbb{V}_{F_n}(\hat{\theta}_n^*)$ by Monte Carlo to obtain $\widehat{\sigma}^2_{n,B}$.
- In a nutshell:

$$\mathbb{V}_F(\hat{\theta}_n) \overset{\text{Empirical cdf}}{\simeq} \mathbb{V}_{F_n}(\hat{\theta}_n^*) \overset{\text{Monte Carlo}}{\simeq} \widehat{\sigma}^2_{n,B}$$

Bootstrap variance estimation

- Simulating $X_1^*, \ldots, X_n^*$ from F_n is easy.
- F_n puts mass $1/n$ at each value $\{X_1, \ldots, X_n\}$.
- $X_1^*, \ldots, X_n^*$ are thus discrete random variables, and we can simulate from F_n by sampling with replacement from the original dataset $\{X_1, \ldots, X_n\}$.

Bootstrap variance estimation

Algorithm 2 Bootstrap algorithm for estimating the variance

Let $X_1, \ldots, X_n$ be some data and $\hat{\theta}_n = t(X_1, \ldots, X_n)$.

- For $j = 1, \ldots, B$
  - Simulate $X_1^{*(j)}, \ldots, X_n^{*(j)} \overset{\text{iid}}{\sim} F_n$ by sampling with replacement from $\{X_1, \ldots, X_n\}$
  - Evaluate $\hat{\theta}_n^{*(j)} = t(X_1^{*(j)}, \ldots, X_n^{*(j)})$
- Return the bootstrap variance estimate

$$\widehat{\sigma}^2_{n,B} = \frac{1}{B} \sum_{j=1}^{B} \left(\hat{\theta}_n^{*(j)} - \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_n^{*(j)} \right)^2 .$$

Bootstrap variance estimation

R implementation

- Consider the median estimator $\widehat{M}_n$
- Bootstrap variance estimate of $\mathbb{V}_F(\widehat{M}_n)$.

```r
bootstrap_variance_median <- function(X, B)
{
  # Bootstrap variance estimate for the median estimator
  # X: Data
  # B: Number of Monte Carlo samples

  n <- length(X)
  mhat.boot <- numeric(B)
  for (j in 1:B){
    X.boot <- sample(X,n,replace=TRUE) # Sample bootstrap data from Fn
    mhat.boot[j] <- median(X.boot) # Median bootstrap sample
  }
  var.boot <- var(mhat.boot) # bootstrap variance estimate
  return(list(var.boot = var.boot, mhat.boot = mhat.boot))
}
```

Bootstrap variance estimation

Illustrative synthetic example

- $X_1, \ldots, X_n \sim \mathcal{N}(0.4, 1)$, $n = 500$

```
n <- 500
X <- rnorm(n,mean=0.4) # Generate the data
mhat.gauss <- median(X) # Median estimate
```

- We now estimate the variance of the median estimator using the bootstrap.

```
B <- 10000
results.gauss = bootstrap_variance_median(X,B) # bootstrap variance estimate
var.boot.gauss = results.gauss$var.boot
var.boot.gauss
```

```
## [1] 0.003382708
```

Bootstrap variance estimation

Illustrative synthetic example

- Bootstrap samples and estimate

```
mhat.boot.gauss = results.gauss$mhat.boot
hist(mhat.boot.gauss, xlab='mhat.boot', main='',col='lightblue2', breaks=20)
points(mhat.gauss,0,pch = 22,col='red',bg='red')
```

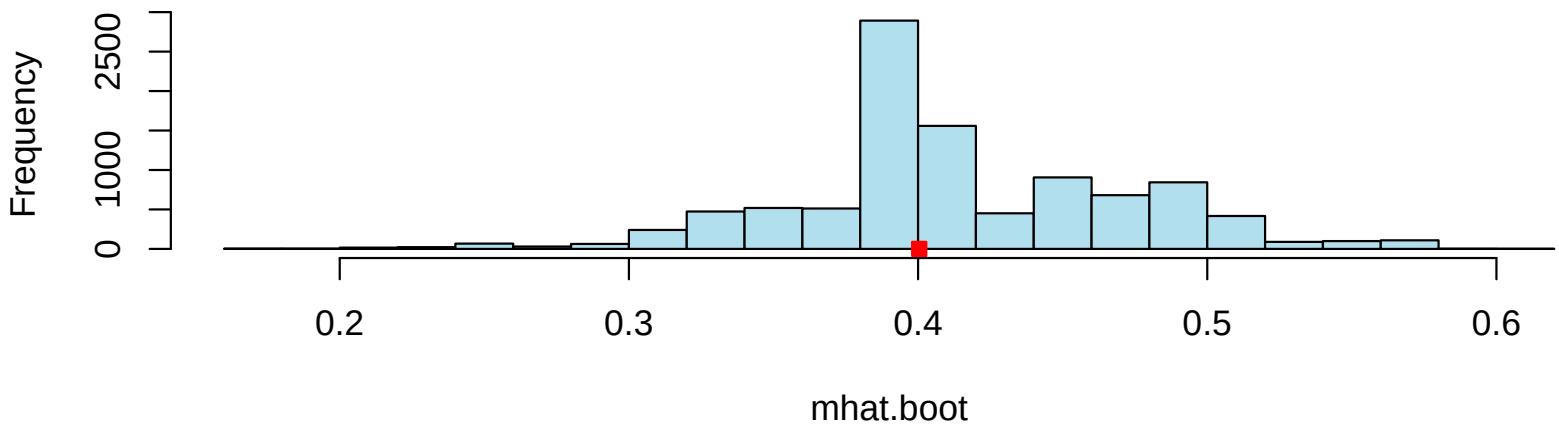


Figure: Histogram of the $B = 10000$ bootstrap samples $\widehat{M}_n^{*(1)}, \ldots, \widehat{M}_n^{*(B)}$ of the median estimator and (red square) estimate $\widehat{M}_n$ for the Gaussian synthetic dataset.

Bootstrap variance estimation

Example: Wages in the US

- Consider again the wages dataset and use the bootstrap estimator to estimate the variance of the sample median.

```
# Compute the bootstrap variance estimate
B <- 10000
results.wages = bootstrap_variance_median(x.wages,B)
var.boot.wages = results.wages$var.boot
var.boot.wages
```

```
## [1] 5.32797
```

Bootstrap variance estimation

Example: Wages in the US

```
mhat.boot.wages = results.wages$mhat.boot
hist(mhat.boot.wages, xlab='mhat.boot', main='',col='lightblue2', breaks=20)
points(mhat.wages,0,pch = 22,col='red',bg='red')
```

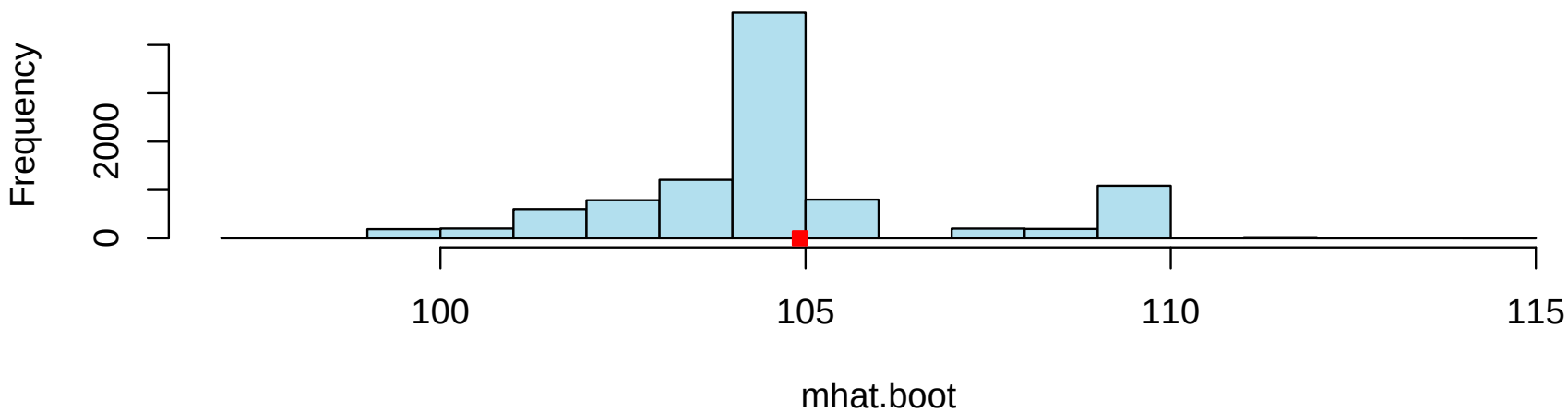


Figure: Histogram of the $B = 10000$ bootstrap samples $\widehat{M}_n^{*(1)}, \ldots, \widehat{M}_n^{*(B)}$ of the median estimator and (red square) estimate $\widehat{M}_n$ for the wages dataset.

Outline

Normal confidence intervals

- For $\alpha \in (0,1)$, we recall that z_α is defined by $\Phi(z_\alpha) = 1-\alpha$ where Φ is the cdf of a $\mathcal{N}(0,1)$ random variable.
- If the central limit theorem holds for the estimator $\hat{\theta}_n$

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

or more generally if $\frac{\hat{\theta}_n - \theta}{\sqrt{\mathbb{V}_F(\hat{\theta}_n)}} \overset{d}{\simeq} \mathcal{N}(0,1)$, then we can use the bootstrap variance estimate $\widehat{\sigma}^2_{n,B} \simeq \mathbb{V}_F(\hat{\theta}_n)$ to form $1-\alpha$ approximate confidence intervals

$$\begin{aligned}
\mathbb{P}(\hat{\theta}_n - z_{\alpha/2}\,\widehat{\sigma}_{n,B} \leq \theta \leq \hat{\theta}_n + z_{\alpha/2}\,\widehat{\sigma}_{n,B}) &= \mathbb{P}\left(\frac{\hat{\theta}_n - \theta}{\widehat{\sigma}_{n,B}} \in [-z_{\alpha/2}, +z_{\alpha/2}]\right) \\
&\simeq \mathbb{P}\left(\frac{\hat{\theta}_n - \theta}{\sqrt{\mathbb{V}_F(\hat{\theta}_n)}} \in [-z_{\alpha/2}, +z_{\alpha/2}]\right) \\
&\simeq 1-\alpha
\end{aligned}$$

Normal confidence intervals

- We use the normal approximation to obtain a 95% confidence interval for the population median m in the mid-Atlantic states of the USA in 2005.

```
alpha = 0.05
ci.bootnormal.wages = c(mhat.wages - sqrt(var.boot.wages)* qnorm(1-alpha/2),
                        mhat.wages + sqrt(var.boot.wages)* qnorm(1-alpha/2))
ci.bootnormal.wages

## [1] 100.3974 109.4456
```

Pivotal confidence intervals

- Consider the random variable $R_n = \hat{\theta}_n - \theta$, called the **pivot**.
- Let $H_n(r) = \mathbb{P}(R_n \leq r)$ be the (unknown) cdf of R_n, H_n^{-1} the associated quantile function

$$H_n^{-1}(q) = \inf\{x : H(x) > q\} \tag{1}$$

  and let $q_\alpha = H_n^{-1}(\alpha)$ be the α quantile of H_n.
- If we knew $q_{\alpha/2}$ and $q_{1-\alpha/2}$, we could trivially use those to construct a $1-\alpha$ confidence intervals as

$$\mathbb{P}(\hat{\theta}_n - q_{1-\alpha/2} \leq \theta \leq \hat{\theta}_n - q_{\alpha/2}) = \mathbb{P}(R_n \in [q_{\alpha/2}, q_{1-\alpha/2}]) = 1-\alpha$$

- However $q_{\alpha/2}$ and $q_{1-\alpha/2}$ are unknown.
- We can use the bootstrap to obtain an approximation of these quantiles.

Pivotal confidence intervals

- The idea is to use the distribution

$$H_n^*(r) = \mathbb{P}_{F_n}(R_n^* \leq r)$$

  of $R_n^* = \hat{\theta}_n^* - \hat{\theta}_n$ conditional on $(X_1, \ldots, X_n)$ as an approximation to the distribution H_n of $R_n = \hat{\theta}_n - \theta$.
- H_n^* can be approximated by Monte Carlo methods to obtain the bootstrap estimator $\widehat{H}_{n,B}^*$ of the cdf. In a nutshell:

$$H_n \overset{\text{Empirical cdf}}{\simeq} H_n^* \overset{\text{Monte Carlo}}{\simeq} \widehat{H}_{n,B}^*$$

Pivotal confidence intervals

- Let $\hat{\theta}_n^{*(1)}, \ldots, \hat{\theta}_n^{*(B)}$ be the bootstrap samples
- The bootstrap estimator of H_n is

$$\widehat{H}^*_{n,B}(r) = \frac{1}{B}\sum_{j=1}^{B} I(R_n^{*(j)} \leq r)$$

  where $R_n^{*(j)} = \hat{\theta}_n^{*(j)} - \hat{\theta}_n$.
- From $\widehat{H}^*_{n,B}$ we can obtain estimators of statistical functionals of H_n, such as the quantiles

$$\hat{q}^*_\alpha = \widehat{H}^{*\,-1}_{n,B}(\alpha)$$

- Bootstrap pivotal $1-\alpha$ confidence interval

$$C_n^* = [\hat{\theta}_n - \hat{q}^*_{1-\alpha/2}, \hat{\theta}_n - \hat{q}^*_{\alpha/2}].$$

Pivotal confidence intervals

- We can rearrange this expression as a function of the quantiles of the bootstrap samples $(\hat{\theta}_n^{*(1)}, \ldots, \hat{\theta}_n^{*(B)})$.
- As $R_n^{*(j)} = \hat{\theta}_n^{*(j)} - \hat{\theta}_n$, we have

$$\hat{q}_\alpha^* = \hat{q}_\alpha^{\theta *} - \hat{\theta}_n,$$

where $\hat{q}_\alpha^{\theta *}$ is the α quantile of the bootstrap samples $(\hat{\theta}_n^{*(1)}, \ldots, \hat{\theta}_n^{*(B)})$.

Definition

The $1-\alpha$ **bootstrap pivotal confidence interval** is given by

$$C_n^* = [2\,\hat{\theta}_n - \hat{q}_{1-\alpha/2}^{\theta *}, 2\,\hat{\theta}_n - \hat{q}_{\alpha/2}^{\theta *}].$$

where $\hat{q}_{\alpha/2}^{\theta *}$ and $\hat{q}_{1-\alpha/2}^{\theta *}$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the bootstrap samples $\hat{\theta}_n^{*(1)}, \ldots, \hat{\theta}_n^{*(B)}$.

Pivotal confidence intervals

- Consider again the wages data
- We will use the bootstrap pivotal approximation to obtain a 95% confidence interval estimate for the median m.

```
ci.bootpivot.wages = c(2*mhat.wages-quantile(mhat.boot.wages, 1-alpha/2,names=FALSE) ,
                 2*mhat.wages-quantile(mhat.boot.wages,alpha/2,names=FALSE))
print(ci.bootpivot.wages)


## [1] 100.0090 109.0787
```

Outline

Bias estimation

- Let $\hat{\theta}_n = t(X_1, \ldots, X_n)$ be an estimator of θ.
- Bias of an estimator $\hat{\theta}_n$ is

$$b := \mathbb{E}_F(\hat{\theta}_n) - \theta$$

- Bias estimation is often difficult but can be achieved with the bootstrap with

$$\hat{b}_n = \mathbb{E}_{F_n}(\hat{\theta}_n^*) - \hat{\theta}_n$$

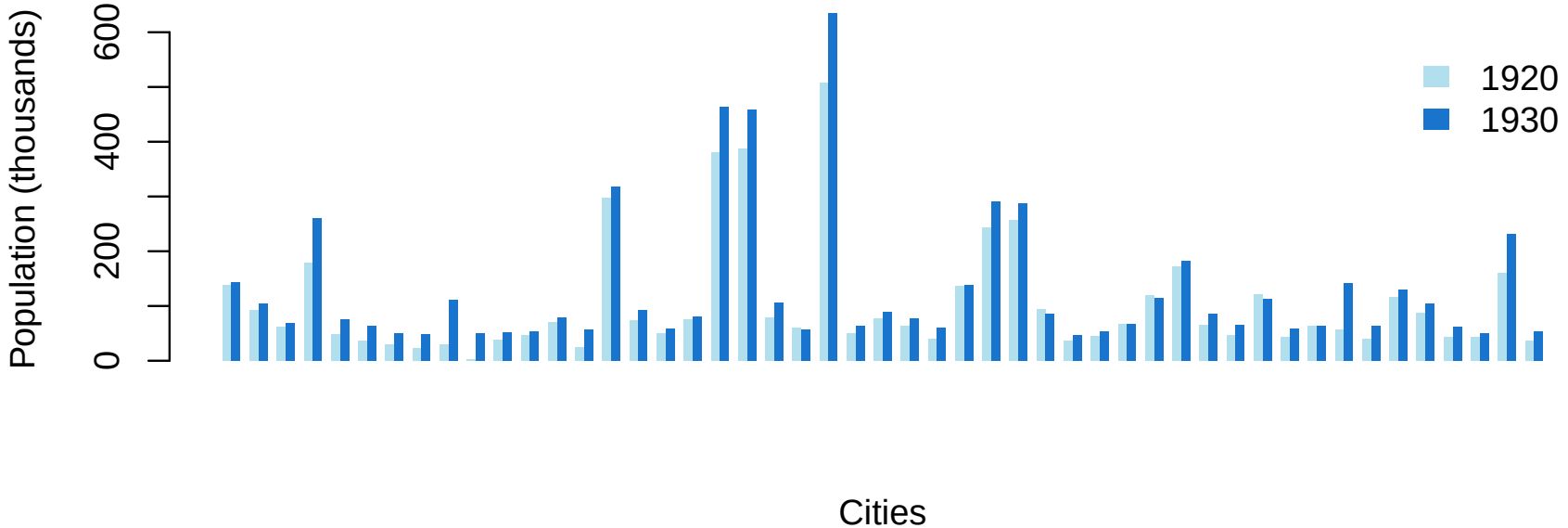
  and using bootstrap samples $\hat{\theta}_n^{*(j)}$ for $j = 1, \ldots, B$ to approximate $\mathbb{E}_{F_n}(\hat{\theta}_n^*)$, just as in bootstrap variance estimation.
- The bootstrap bias estimator is therefore

$$\hat{b}_{n,B} = \left(\frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_n^{*(j)} \right) - \hat{\theta}_n \,.$$

Bias estimation

Census data

- The population for 49 major cities in the US, sampled randomly, is known for the years 1920 and 1930.
- The ratio of the average population should serve as a scaling information to adjust census information from 1920 for 1930



Bias estimation

Census data

- Let $((X_{11}, X_{12}), \dots, (X_{n1}, X_{n2}))$ where X_{i1} is the population of city i in 1920 and X_{i2} is the population of the city i in 1930.
- $(X_{11}, \dots, X_{n1})$ are assumed iid from some cdf F with unknown mean μ_1.
- Similarly, $(X_{12}, \dots, X_{n2})$ are assumed iid from some cdf G with unknown mean μ_2.
- We are interested in estimating the ratio of the average population $\xi = \frac{\mu_2}{\mu_1}$ and use as an estimator the ratio of the empirical means

$$\hat{\xi}_n = \frac{\sum_{i=1}^n X_{i2}}{\sum_{i=1}^n X_{i1}}.$$

- What is the bias of this estimator?

Bias estimation

Census data

```
library(boot)# for the data
pop1920 = bigcity$u
pop1930 = bigcity$x
ratiohat <- mean(pop1930)/mean(pop1920)
ratiohat

## [1] 1.239019

# Bootstrap bias estimate
n <- nrow(bigcity)
B <- 10000
ratiohat.boot <- rep(0, B)
for (j in 1:B) {
  ind <- sample(n, replace=TRUE)
  pop1920mean.boot <- mean(pop1920[ind])
  pop1930mean.boot <- mean(pop1930[ind])
  ratiohat.boot[j] <- pop1930mean.boot / pop1920mean.boot
}
bias.boot <- mean(ratiohat.boot) - ratiohat
bias.boot

## [1] 0.001791172
```

Outline

Properties of the bootstrap

- The bootstrap introduces two approximations: the empirical cdf approximation and the Monte Carlo approximation
- The Monte Carlo estimator is consistent, and the Monte Carlo error can be made arbitrarily small by taking a sufficiently large number of Monte Carlo samples B
- In analysing the properties of the bootstrap, the Monte Carlo error is therefore usually ignored, and we focus on the error introduced by the use of the empirical cdf F_n instead of the true cdf F
- Consistency is studied as the number n of data goes to infinity

Properties of the bootstrap

- The bootstrap uses the distribution of $R_n^* = \hat{\theta}_n^* - \hat{\theta}_n$ as an approximation to the (unknown) distribution of $R_n = \hat{\theta}_n - \theta$.
- To study consistency, we need to appropriately scale these random variables
- Let

$$\widetilde{H}_n(x) := \mathbb{P}_F(a_n R_n \leq x)$$
$$\widetilde{H}^*_{F_n}(x) := \mathbb{P}_{F_n}(a_n R_n^* \leq x)$$

  where a_n is some appropriate scaling (classically $a_n = \sqrt{n}$)
- $\widetilde{H}_n$ is the fixed cdf of interest.
- The bootstrap cdf $\widetilde{H}^*_{F_n}$ is a random cdf, as it depends on the original sample $(X_1, \ldots, X_n)$.

Properties of the bootstrap

Definition (Bootstrap consistency)

Let ρ be some metric on the space of cdfs. The bootstrap is called to be (weakly) consistent for $\hat{\theta}_n$ under the metric ρ if

$$\rho(\widetilde{H}_n, \widetilde{H}^*_{F_n}) \xrightarrow{p} 0 \qquad n \to \infty$$

and strongly consistent if the result holds almost surely.

Properties of the bootstrap

As an example, the following result, proved by Singh in 1981, holds for sample mean estimators.

Theorem (Boostrap consistency for sample mean)

Let $X_1, \dots, X_n$ be iid real-valued random variables from a cdf F with $\mathbb{E}_F(X_i^2) < \infty$ and $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Let K be the Kolmogorov metric

$$K(F, G) = \sup_{x \in \mathbb{R}} |F(x) - G(x)|.$$

The bootstrap is (strongly) consistent under the Kolmogorov metric K

$$K(\widetilde{H}_n, \widetilde{H}^*_{F_n}) \xrightarrow{as} 0 \qquad n \to \infty.$$

Properties of the bootstrap

Illustration on the median estimator

- $X_1, \ldots, X_n \sim \mathcal{N}(0.4, 1)$
- Median estimator
- For different data sample sizes from $n = 1$ to $n = 10000$
  - Plot a realization $\widetilde{H}^*_{F_n}(x)$ (blue)
  - Plot $\widetilde{H}_n(x)$ (red, obtained via simulation)

Properties of the bootstrap

Illustration on the median estimator

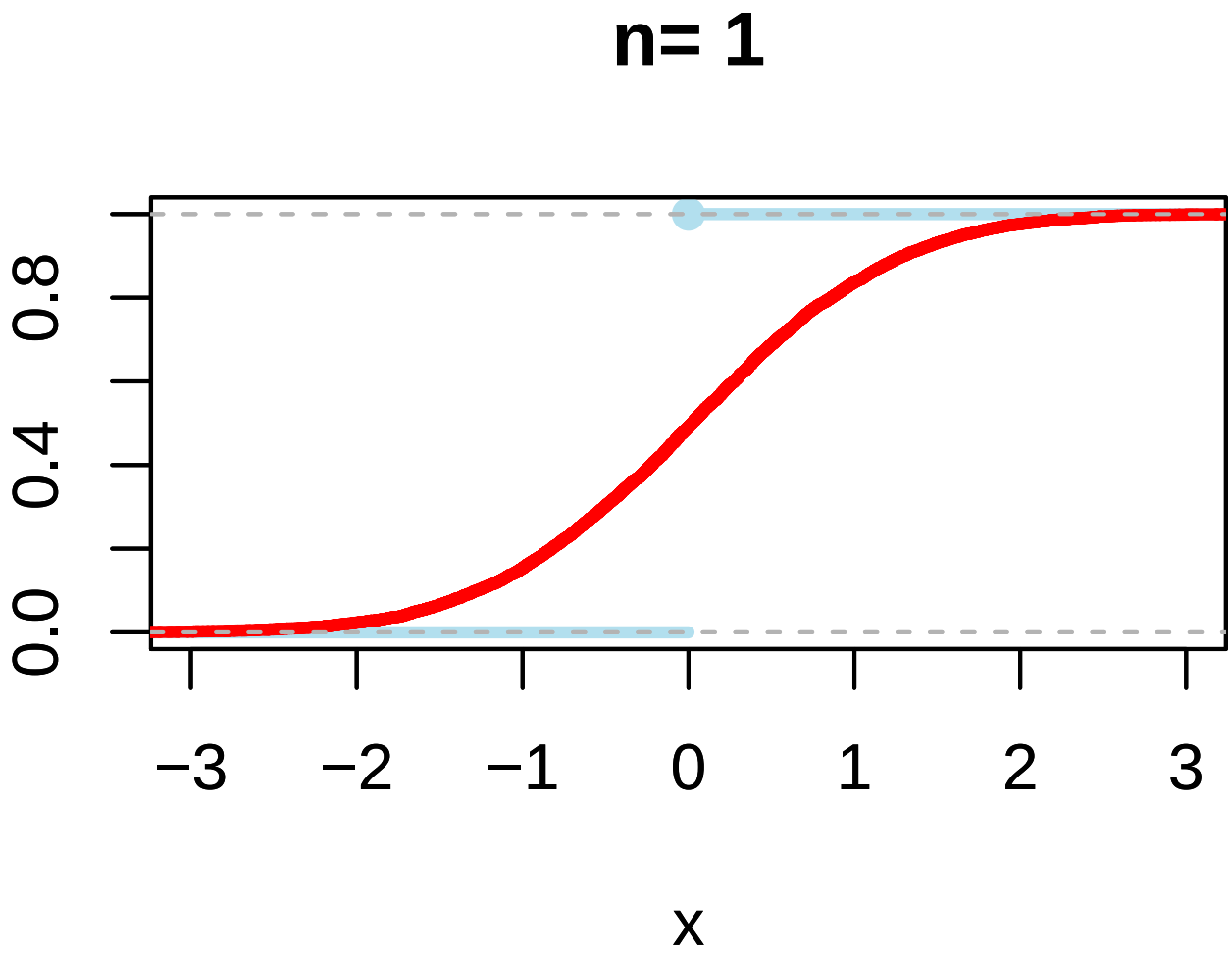
```
n.all <- c(1, 10, 50, 100, 500, 1000, 5000, 10000)
B = 10000 # Number of Monte Carlo samples for the bootstrap
Nsim = 10000 # Number of Monte Carlo samples for the cdf of the pivot
mhat.sim.gauss = rep(0, Nsim)
for (i in 1:length(n.all)) {
  n = n.all[i]

  # Empirical cdf of the bootstrap pivot
  X <- rnorm(n,mean=0.4)
  mhat.gauss <- median(X)
  results.gauss = bootstrap_variance_median(X,B)
  mhat.boot.gauss = results.gauss$mhat.boot
  plot.ecdf(sqrt(n)*(mhat.boot.gauss-mhat.gauss), xlim=c(-3,3),ylim=c(0,1),
            main=paste('n=',n),lwd=3, col='lightblue2',ylab='')


  # Monte Carlo estimate of the cdf of the pivot by direct simulation
  for (j in 1:Nsim)
  {
    X <- rnorm(n,mean=0.4)
    mhat.sim.gauss[j] <- median(X)
  }
  par(new=TRUE)
  plot.ecdf(sqrt(n)*(mhat.sim.gauss-0.4), xlim=c(-3,3),ylim=c(0,1),
            main=paste('n=',n),lwd=3, col='red',ylab='')
  }
```

Properties of the bootstrap

Illustration on the median estimator



Properties of the bootstrap

- Consistency of the bootstrap usually implies that the variance can be consistently estimated,

$$\frac{\mathbb{V}_{F_n}(\hat{\theta}_n^*)}{\mathbb{V}_F(\hat{\theta}_n)} \xrightarrow{p} 1 \qquad n \to \infty.$$

- The same holds for the bias,

$$\frac{\mathbb{E}_{F_n}(\hat{\theta}_n^*) - \hat{\theta}_n}{\mathbb{E}_F(\hat{\theta}_n) - \theta} \xrightarrow{p} 1 \qquad n \to \infty.$$

- Note that the bootstrap is not always consistent!

Outline

Bootstrap for regression

- Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be paired iid random variables assumed to be drawn from
$$Y_i = g(X_i, \theta) + \varepsilon_i$$
where $\varepsilon_1, \ldots, \varepsilon_n$ are iid from some unknown cdf F and $\theta \in \mathbb{R}^p$.
- g is some known function, for example $g(X_i, \theta) = X_i\theta$.
- Let $\hat{\theta}_n = t((X_1, Y_1), \ldots, (X_n, Y_n))$ be some estimator of θ, for example the least square estimator.
- For $i = 1, \ldots, n$, let
$$\hat{\varepsilon}_i = Y_i - g(X_i, \hat{\theta}_n).$$
be the fitted residuals.
- Two bootstrap strategies are possible here:
  1. Resample the data $((X_1, Y_1), \ldots, (X_n, Y_n))$
  2. Resample the residuals $(\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)$

Bootstrap for regression: Nonparametric paired bootstrap

- For $j = 1, \dots, B$
  - Simulate $((X_1^{*(j)}, Y_1^{*(j)}), \dots, (X_n^{*(j)}, Y_n^{*(j)}))$ by sampling with replacement from $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$
  - Evaluate $\hat{\theta}_n^{*(j)} = t((X_1^{*(j)}, Y_1^{*(j)}), \dots, (X_n^{*(j)}, Y_n^{*(j)}))$
- Return the bootstrap estimate. For example, for $\theta \in \mathbb{R}$, the variance estimate is

$$\widehat{\sigma}_{n,B}^2 = \frac{1}{B} \sum_{j=1}^{B} \left(\hat{\theta}_n^{*(j)} - \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_n^{*(j)} \right)^2 .$$

Bootstrap for regression: Semiparametric residual bootstrap

$X_1, \dots, X_n$ are assumed to be fixed

- For $j = 1, \dots, B$
  - Simulate $(\hat{\varepsilon}_1^{*(j)}, \dots, \hat{\varepsilon}_n^{*(j)})$ by sampling with replacement from $(\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)$
  - For $i = 1, \dots, n$, Set $Y_i^{*(j)} = g(X_i, \hat{\theta}_n) + \hat{\varepsilon}_i^{*(j)}$
  - Evaluate $\hat{\theta}_n^{*(j)} = t((X_1, Y_1^{*(j)}), \dots, (X_n, Y_n^{*(j)}))$
- Return the bootstrap estimate. For example, for $\theta \in \mathbb{R}$, the variance estimate is

$$\widehat{\sigma}_{n,B}^2 = \frac{1}{B} \sum_{j=1}^{B} \left(\hat{\theta}_n^{*(j)} - \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_n^{*(j)} \right)^2 .$$

Outline

Parametric bootstrap

- The methods seen so far are nonparametric, in the sense that no assumption has been made on the cdf F.
- In the parametric setting, we assume that $X_1, \ldots, X_n$ are iid random variables with cdf F_θ where θ is an unknown parameter of the model we wish to estimate.
- $\hat{\theta}_n = t(X_1, \ldots, X_n)$ is some estimator of θ, for example a maximum likelihood estimator.
- The parametric bootstrap proceeds similarly to the nonparametric bootstrap described above, except that it uses the fitted cdf $F_{\hat{\theta}_n}$ instead of the empirical cdf to obtain the bootstrap samples.
- If the parametric model is correctly specified, this will lead to superior performances compared to the nonparametric bootstrap.