SB1.2/SM2 Computational Statistics
Lecture notes: The Bootstrap

François Caron

University of Oxford, Hilary Term 2019

Version of February 5, 2019

This document builds on earlier notes from Nicolai Meinshausen, as well as the following references:

- L. Wasserman. All of Statistics. Springer, 2010.
- G. James, D. Witten, T. Hastie, R. Tibshirani. An Introduction to Statistical Learning with Applications in R. Chapter 5. Springer, 2013.
- B. Efron and R.J. Tibshirani. An introduction to the Bootstrap. Chapman & Hall, 2010.
- A.W. van der Vaart. Asymptotic Statistics. Chapter 23. Cambridge University Press, 1998.
- A. DasGupta. Asymptotic Theory of Statistics and Probability. Chapter 29. Springer, 2008.

The course requires the following notions:

- Estimator; bias, variance and consistency of an estimator; confidence interval [Part A Statistics; SB1.1/SM1]
- Convergence of random variables [Part A Probability]

This document includes some R examples with datasets from the packages MASS, ISLR, BOOT. The algorithms described are easy to implement and only require a few lines of code. The R package boot implements more advanced bootstrap methods. Please report typos to caron@stats.ox.ac.uk.

Contents

The Bootstrap belongs to the wider class of resampling methods. When introduced in 1979 by Bradley Efron, it had a huge impact on statistics, as it offered a principled, simulation-based approach, for assigning measures of accuracy to statistical estimates, and estimating biases, variances or obtaining approximate confidence intervals. It applies to both parametric (when the distribution of the data is known and parametrized by a finite-dimensional parameter) and nonparametric settings. Bootstrap ideas of "resampling the data" are at the heart of subsequent popular statistical methods such as random forests, and still motivate nowadays the development of novel statistical tools for handling massive datasets.

1 Motivating examples

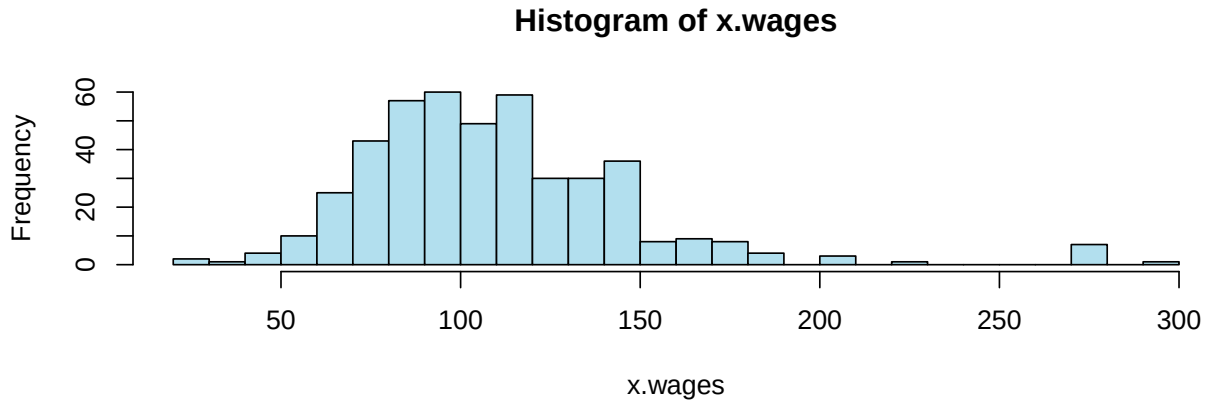
Confidence interval for the median of a population. Let $X_1, \dots, X_n$ be independent identically distributed (iid) random variables from a cumulative distribution function (cdf) F and consider that we are interested in the median $m = F^{-1}(0.5)$ of this distribution. We consider the following estimator

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

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

Example 1 (Wages)**.** *For example, assume that we are interested in the median wage in the mid-Atlantic states of the USA in 2005. We consider the following dataset from the R package ISLR, which consists of the wages of $n = 447$ workers in that region in 2005.*

```
require('ISLR')

## Loading required package:  ISLR

data(Wage)
x.wages = Wage$wage[Wage$year==2005]
head(x.wages, n=10L) # Show the first data

##  [1]  75.04315  89.49248  50.40666 277.60142 101.40205 111.72085  73.77574
##  [8] 200.54326 152.83880  77.73760

hist(x.wages, breaks=20,col='lightblue2')
```



The median estimate is

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

## [1] 104.9215
```

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 probability density function (pdf) f, then the asymptotic

distribution of $\widehat{M}_n$ is[1]

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

as $n \to \infty$. However this does not help much as we cannot evaluate the term $1/(4f^2(m))$, the density f being typically unknown. If we could evaluate $\sigma^2 = \mathbb{V}(\widehat{M}_n)$, 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^2 of the estimator.

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

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

However, the quantiles are in general impossible to obtain analytically. The bootstrap provides estimates of the quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ and thus approximate confidence intervals.

Bias and Variance of an estimator. Let $X_1, \ldots, X_n$ be iid random variables from a pdf $f(x;\theta)$ parametrized by θ. Let

$$\widehat{\theta}_n = t(X_1, \ldots, X_n) \tag{3}$$

be some estimator of θ. In order to evaluate the accuracy of the estimator, we may be interested in evaluating its bias $\mathbb{E}[\hat{\theta}_n] - \theta$ or its variance $\mathbb{V}[\hat{\theta}_n]$. Those quantities may not be analytically available. The bootstrap is a general strategy to obtain estimates for these quantities. Similarly, we may also be interested in building confidence intervals. For maximum likelihood estimators, we can use the asymptotic normality of the estimator and the estimate of the variance to obtain this confidence interval (see Part A statistics). Alternatively, as described above, if one could obtain estimates of the quantiles of the distribution of $\hat{\theta}_n - \theta$, then they could be used to construct confidence intervals for the parameter of interest θ. The bootstrap allows to obtain these estimates.

2 Background material

2.1 Some preliminary comments on notations

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

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

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

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

We will use the unifying notation $\int_{\mathbb{R}^p} x dF(x)$ to define an expectation for both continuous and discrete random variables. This notation has a special meaning, which some of you have seen in the Part A option on Integration and measure theory. However, this course does not assume any background in measure theory, and we will just consider $\int_{\mathbb{R}^p} x dF(x)$ as a convenient notation.

2.2 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 $m = F^{-1}(0.5)$.

[1] The proof is out of the scope of this course.

2.3 Empirical distribution function

Definition 2. *Let $X_1, \ldots, X_n$ be independent and identically distributed real-valued random variables with cdf F. The **empirical cumulative distribution function** is defined as*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) = \frac{|\{\ i \mid X_i \leq x\}|}{n} \tag{4}$$

where

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

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

Proposition 3. *The empirical cdf estimator is*

(i) Unbiased

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

(ii) (Strongly) consistent

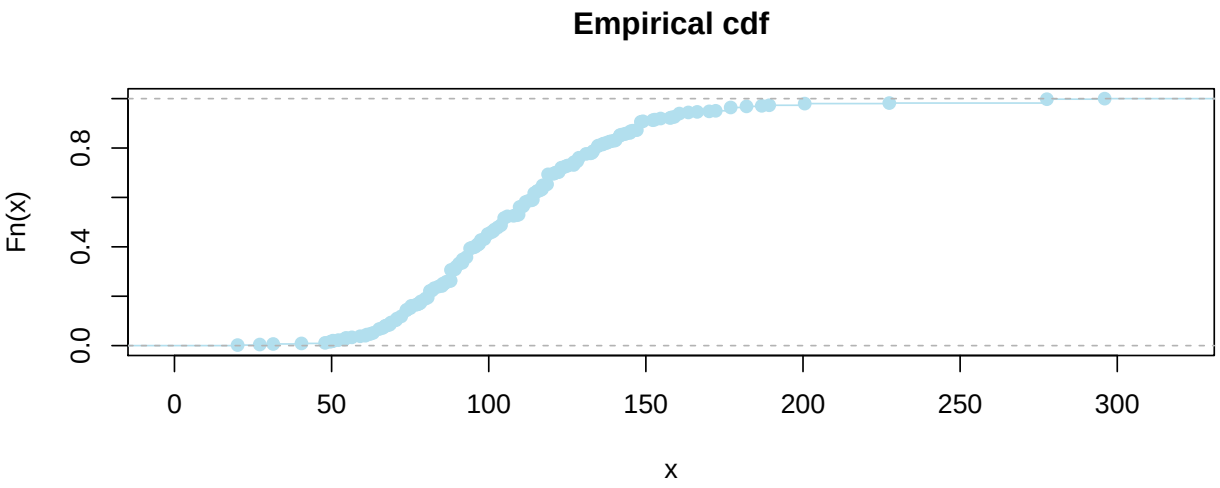
$$F_n(x) \overset{as}{\to} F(x) \text{ for all } x \text{ as } n \to \infty$$

(iii) Asymptotically normal

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

Proof. For any $x \in \mathbb{R}$, the binary random variables $I(X_1 \leq x), \ldots, I(X_n \leq x)$ are independently and identically Bernoulli distributed with success probability $F(x)$. Hence the discrete random variable $nF_n(x)$ is binomially distributed $nF_n(x) \sim \text{Binomial}(n, F(x))$ and (i) follows. (ii) follows from the strong law of large numbers, and (iii) from the central limit theorem. □

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



2.4 Recap: Monte Carlo integration

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

Definition 4 (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.

The Monte Carlo algorithm, which only requires to be able to simulate from G, is as follows.

Algorithm 1 Monte Carlo Algorithm

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

Here are the properties of the Monte Carlo estimator, which follow from the law of large numbers and the central limit theorem (see Part A Simulation for detailed proofs).

Proposition 5. *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

$$\sqrt{B}(\widehat{\eta}_B - \eta) \overset{d}{\to} \mathcal{N}(0, \sigma^2) \text{ as } B \to \infty.$$

For example, the Monte Carlo estimators of the mean and variance of Y are

$$\begin{aligned}
\widehat{\mu}_B &= \frac{1}{B}\sum_{j=1}^{B} Y^{(j)} \overset{as}{\to} \mathbb{E}(Y), \\
\widehat{\sigma}^2_B &= \frac{1}{B}\sum_{j=1}^{B} (Y^{(j)} - \widehat{\mu}_B)^2 \\
&= \frac{1}{B}\sum_{j=1}^{B} (Y^{(j)})^2 - \left(\frac{1}{B}\sum_{j=1}^{B} Y^{(j)}\right)^2 \overset{as}{\to} \mathbb{V}(Y).
\end{aligned}$$

3 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.)"*

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

$$\hat{\theta}_n = t(X_1, \ldots, X_n) \tag{5}$$

for estimating $\theta = \mathcal{F}(F)$ a statistical functional of F. The relation is represented below.

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

Remark 6. *The same notation will be used for random variables and their realizations (similarly for estimators/estimates).*

3.1 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.

An idealized variance estimator. Suppose first that we *could* simulate from the cdf F, that is that we *could* reproduce the Real World and generate new datasets. Then we could use the Monte Carlo method described in section 2.4 to obtain a Monte Carlo estimator for σ^2 as follows.

- For $j = 1, \dots, B$,
  - Let $X_1^{(j)}, \dots, X_n^{(j)}$ be iid from F and $\hat{\theta}_n^{(j)} = t(X_1^{(j)}, \dots, X_n^{(j)})$
- The Monte Carlo variance estimator is $\frac{1}{B}\sum_{j=1}^B \left(\hat{\theta}_n^{(j)} - \frac{1}{B}\sum_{j=1}^B \hat{\theta}_n^{(j)}\right)^2$.

However, the cdf F is unknown and we cannot simulate from it, so we cannot implement this estimator.

The bootstrap for variance estimation. The idea of the bootstrap is to

1. Replace the unknown cdf F by its (known) empirical cdf F_n defined in Equation (4),
2. Use the Monte Carlo method.

More formally, **conditionally on** $X_1, \dots, X_n$, let $X_1^*, \dots, X_n^*$ be iid random variables with cdf F_n, and $\hat{\theta}_n^* = t(X_1^*, \dots, X_n^*)$. The bootstrap mimics the Real World by using the empirical cdf F_n

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

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 the bootstrap estimate $\widehat{\sigma}_{n,B}^2$.

In a nutshell:

$$\mathbb{V}_F(\hat{\theta}_n) \overset{\text{Empirical cdf}}{\simeq} \mathbb{V}_{F_n}(\hat{\theta}_n^*) \overset{\text{Monte Carlo}}{\simeq} \widehat{\sigma}_{n,B}^2 \tag{6}$$

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

Algorithm 2 Bootstrap algorithm for estimating the variance

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

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

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

R implementation and illustration. Consider the median estimator $\widehat{M}_n$ introduced in Equation (1). We will use the bootstrap method and Algorithm 2 to estimate its variance $\mathbb{V}_F(\widehat{M}_n)$.

Here is the R function which takes as input a dataset $X = (X_1, \dots, X_n)$ and the number B of Monte Carlo samples, and returns the bootstrap samples $\widehat{M}_n^{*(1)}, \dots, \widehat{M}_n^{*(B)}$ and the bootstrap variance estimate $\widehat{\sigma}_{n,B}^2$.

```
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 samples
  }
  var.boot <- var(mhat.boot) # Evaluate the bootstrap variance estimate
  return(list(var.boot = var.boot, mhat.boot = mhat.boot))
}
```

We start with a toy example, with a synthetic normal dataset. Let $n = 500$. Consider that the data $X_1, \ldots, X_n$ are normally distributed with mean 0.4 and variance 1. Of course for real applications, we do not know the distribution of the data, but this example is used for illustration purposes.

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

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

```
# Compute the bootstrap variance estimate
B <- 10000
results.gauss = bootstrap_variance_median(X,B)
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')
```

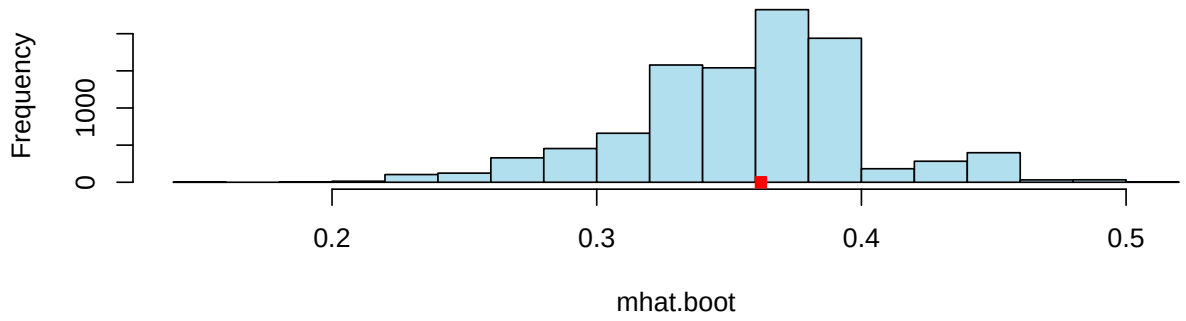


Figure 1: Histogram of the $B = 10000$ bootstrap samples $\widehat{M}_n^{*(1)}, \ldots, \widehat{M}_n^{*(B)}$ of the median estimator and (red square) estimate $\widehat{M}_n$ for the Gaussian synthetic dataset.

```
var.boot.gauss = results.gauss$var.boot
var.boot.gauss

## [1] 0.002028543
```

The $B = 10000$ bootstrap samples $\widehat{M}_n^{*(1)}, \ldots, \widehat{M}_n^{*(B)}$ used to obtain the bootstrap variance estimate are shown in Figure 1. Note we never had to know the distribution of $X_1, \ldots, X_n$ to obtain the bootstrap estimator: all we needed is to sample by replacement from the original dataset.

Example 7 (Wages, continued)**.** *We consider the wages dataset considered in Section 1, 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)
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')
```

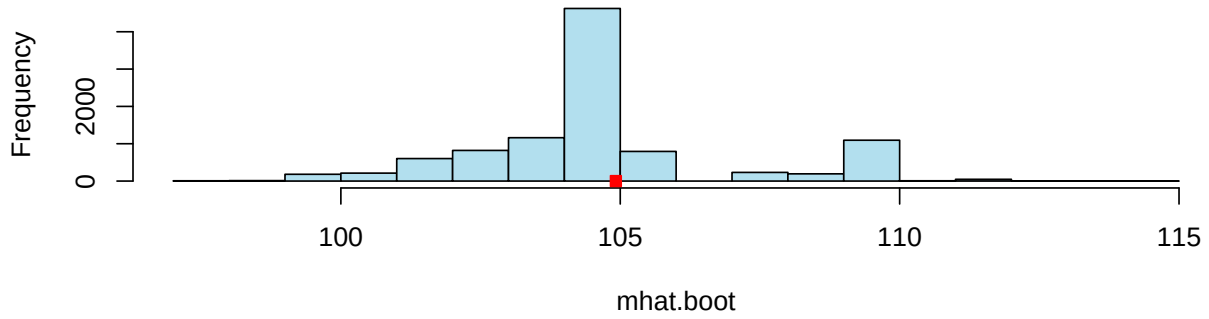


Figure 2: Histogram of the $B = 10000$ bootstrap samples $\widehat{M}_n^{*(1)}, \ldots, \widehat{M}_n^{*(B)}$ of the median estimator and (red square) estimate $\widehat{M}_n$ for the wages dataset.

```
var.boot.wages = results.wages$var.boot
var.boot.wages

## [1] 5.496774
```

3.2 Confidence intervals

3.2.1 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 $\hat{\theta}_n$

$$\frac{\hat{\theta}_n - \theta}{\sqrt{\mathbb{V}_F(\hat{\theta}_n)}} \xrightarrow{d} \mathcal{N}(0,1) \tag{7}$$

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 $1-\alpha$ confidence intervals as

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

Example 8 (Wages (continued))**.** *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] 100.3263 109.5167
```

3.2.2 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

$$H_n^{-1}(q) = \inf\{x : H(x) > q\} \tag{8}$$

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. 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 estimate $\widehat{H}_{n,B}^*$ of the cdf. In a nutshell:

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

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

$$\widehat{H}_{n,B}^*(r) = \frac{1}{B}\sum_{j=1}^{B} I(R_n^{*(j)} \leq r) \tag{9}$$

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

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

and build the bootstrap pivotal $1-\alpha$ confidence interval

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

Note that 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 $(\hat{\theta}_n^{*(1)}, \ldots, \hat{\theta}_n^{*(B)})$. As a conclusion,

Definition 9. *The $1-\alpha$ **bootstrap pivotal confidence interval** is given by*

$$C_n^* = [2\,\hat{\theta}_n - \hat{q}_{1-\alpha/2}^{\theta*}, 2\,\hat{\theta}_n - \hat{q}_{\alpha/2}^{\theta*}]. \tag{10}$$

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 $\hat{\theta}_n^{*(1)}, \ldots, \hat{\theta}_n^{*(B)}$.

Example 10 (Wages (continued))**.** *We will use the bootstrap pivotal approximation to obtain a 95% confidence interval for the median m.*

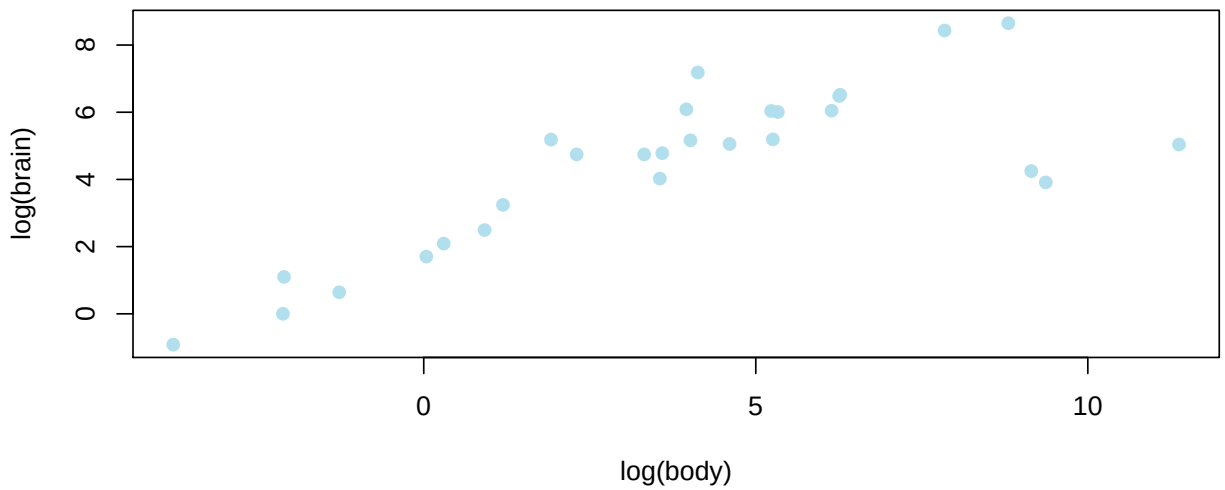
```
ci.bootpivot.wages = c(2*mhat.wages-quantile(mhat.boot.wages, 1-alpha/2,names=FALSE) ,
                       2*mhat.wages-quantile(mhat.boot.wages,alpha/2,names=FALSE))
print(ci.bootpivot.wages)

## [1] 100.0090 109.0787
```

Example 11 (Confidence intervals for correlation)**.** *Consider the following dataset that gives the average brain and body weights for 28 species of land animals.*

```
library('MASS')
data(Animals)
head(Animals)
```

```
##                       body brain
## Mountain beaver       1.35   8.1
## Cow                 465.00 423.0
## Grey wolf            36.33 119.5
## Goat                 27.66 115.0
## Guinea pig            1.04   5.5
## Dipliodocus       11700.00  50.0
```

```
body <- log(Animals[,"body"])
brain <- log(Animals[,"brain"])
plot(body, brain, xlab='log(body)', ylab='log(brain)',pch=21,col='lightblue2', bg='lightblue2')
```

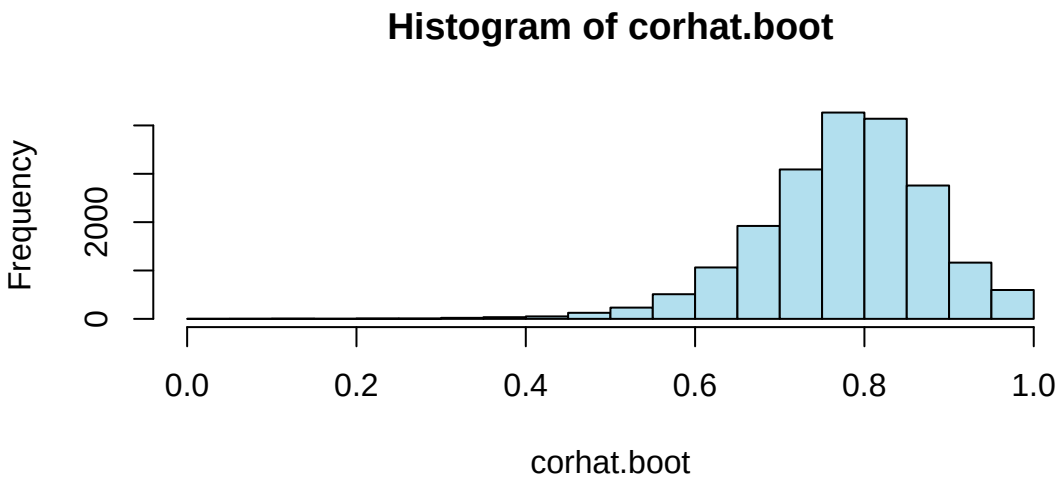


We are interested in the correlation between the log-average brain weight and the log-average body weight, and want to obtain a 95% confidence interval for the correlation using the pivotal bootstrap.

```
corhat = cor(body, brain)
corhat
```

```
## [1] 0.7794935
```

```
B <- 20000
n <- length(body)
corhat.boot <- numeric(B)
for (j in 1:B){
    ind <- sample(1:n,n,replace=TRUE)
    corhat.boot[j] <- cor(body[ind],brain[ind])
}
hist(corhat.boot, col='lightblue2')
```




```
ci.bootpivot.cor <- c(2*corhat - quantile(corhat.boot, probs=1-alpha/2, names=FALSE),
        2*corhat - quantile(corhat.boot, probs=alpha/2, names=FALSE))
ci.bootpivot.cor

## [1] 0.6019583 1.0074239
```

3.3 Bias estimation

One is often also interested in bias estimation, where the 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

$$\hat{b}_{n,B} = \left(\frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_n^{*(j)} \right) - \hat{\theta}_n \,. \tag{11}$$

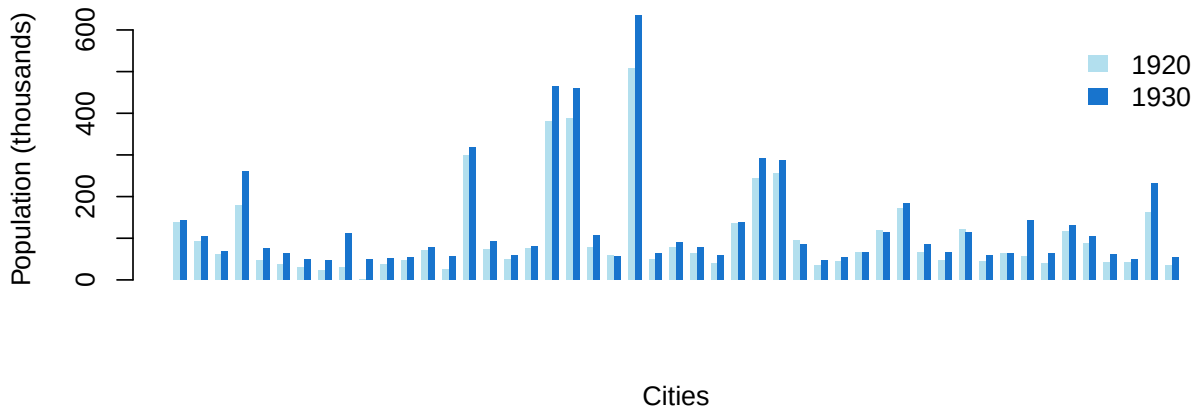
Example 12 (Census data)**.** *The population for 49 major cities in the US, sampled randomly from the 196 largest cities in 1920, is known for the years 1920 and 1930. The ratio of the average population should serve as a scaling information to adjust census information from 1920 for 1930 in the country. Let* $((X_{11}, X_{12}), \ldots, (X_{n1}, X_{n2}))$ *where* X_{i1} *is the population of city* i *in 1920 and* X_{i2} *is the population of the city* i *in 1930.* $(X_{11}, \ldots, X_{n1})$ *are assumed iid from some cdf* F *with unknown mean* μ_1*. Similarly,* $(X_{12}, \ldots, X_{n2})$ *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*

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

```
require(boot) # for the data only

## Loading required package:  boot

head(bigcity)
```

```
##     u   x
## 1 138 143
## 2  93 104
## 3  61  69
## 4 179 260
## 5  48  75
## 6  37  63
```

```
# Plot the data
barplot(t(bigcity),beside=TRUE,ylab='Population (thousands)',xlab='Cities',
   legend.text=c('1920','1930'),axisnames=FALSE,col=c('lightblue2','dodgerblue3'),
   border=NA,args.legend=list(bty='n',border=NA))
```




```
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] 0.00175139
```

4 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.

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; in classical situations $a_n = \sqrt{n}$. $\widetilde{H}_n$ is the fixed cdf of interest. The bootstrap cdf $\widetilde{H}^*_{F_n}$ is a random cdf, as it depends on the original sample $(X_1, \ldots, X_n)$ through F_n.

Definition 13. *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.

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

Theorem 14 (Boostrap consistency for sample mean)**.** *Let $X_1, \ldots, X_n$ 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

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

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, despite its attractiveness.

5 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 $(\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)$

The first approach is the standard nonparametric bootstrap approach. The second approach makes use of the parametric part $g(X_i, \theta)$ of the data generating mechanism and is called semiparametric bootstrap.

Nonparametric paired bootstrap.

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

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

Semiparametric residual bootstrap

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

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

6 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.