UNIVERSITÀ
CATTOLICA
del Sacro Cuore

COMPUTATIONAL STATISTICS II

Professor: Alessia Pini

PhD program in Economics and Statistics (ECOSTAT)

UNIVERSITÀ
CATTOLICA
del Sacro Cuore

POINT ESTIMATE OF THE STANDARD ERROR: THE BOOTSTRAP

BOOTSTRAP

Bootstrap estimate of the variance.

We start from a sample $\mathbf{x}$ of size n, with $x_i \sim F$, $\forall i = 1, \ldots, n$. Assume that we want to estimate the parameter $\theta = t(F)$ with the estimator $\hat{\theta} = s(\mathbf{x})$ (not necessarily the plug-in estimator).
The aim is to understand how variable is $\hat{\theta}$, without knowing the data distribution F.

BOOTSTRAP

Bootstrap estimate of the variance.

We start from a sample $\mathbf{x}$ of size n, with $x_i \sim F$, $\forall i = 1, \dots, n$. Assume that we want to estimate the parameter $\theta = t(F)$ with the estimator $\hat{\theta} = s(\mathbf{x})$ (not necessarily the plug-in estimator).
The aim is to understand how variable is $\hat{\theta}$, without knowing the data distribution F.

→ Nonparametric Bootstrap

BOOTSTRAP

Bootstrap estimate of the variance.

We start from a sample $\mathbf{x}$ of size n, with $x_i \sim F$, $\forall i = 1, \ldots, n$. Assume that we want to estimate the parameter $\theta = t(F)$ with the estimator $\hat{\theta} = s(\mathbf{x})$ (not necessarily the plug-in estimator).
The aim is to understand how variable is $\hat{\theta}$, without knowing the data distribution F.

Boostrap idea: even though we do not know the data distribution F, we can try to estimate it using the empirical distribution $\hat{F}$, that is a consistent estimate.

Then, we can proceed like in Monte Carlo simulation, generating samples of size n from the empirical distribution $\hat{F}$:

$$\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$$

Bootstrap sample.
It is a sample of size n drawn with replacement from the original one, whose elements can appear zero times, once, twice, ...

BOOTSTRAP

Bootstrap estimate of the variance.

We start from a sample $\mathbf{x}$ of size n, with $x_i \sim F$, $\forall i = 1, \ldots, n$. Assume that we want to estimate the parameter $\theta = t(F)$ with the estimator $\hat{\theta} = s(\mathbf{x})$ (not necessarily the plug-in estimator).
The aim is to understand how variable is $\hat{\theta}$, without knowing the data distribution F.

Boostrap idea: even though we do not know the data distribution F, we can try to estimate it using the empirical distribution $\hat{F}$, that is a consistent estimate.

Then, we can proceed like in Monte Carlo simulation, generating samples of size n from the empirical distribution $\hat{F}$:

$$\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$$

$$\boxed{\hat{\theta}^* = s(\mathbf{x}^*)}$$

Bootstrap replication.
For each Bootstrap sample we obtain a replication of the estimated value of theta.

BOOTSTRAP

Bootstrap estimate of the variance.

Finally, we obtain the Bootstrap estimate of the variance (and standard error) of $\hat{\theta}$:

$$\widehat{\mathrm{Var}}_B = \mathrm{Var}_{\hat{F}}(\hat{\theta}^*)$$

$$\widehat{\mathrm{se}}_B = \mathrm{se}_{\hat{F}}(\hat{\theta}^*).$$

Ideal Bootstrap estimates.
Standard error and variance under the empirical distribution. A closed formula is usually not available!

Bootstrap estimate of the variance.

Finally, we obtain the Bootstrap estimate of the variance (and standard error) of $\hat{\theta}$:

$$\widehat{\mathrm{Var}}_B = \mathrm{Var}_{\hat{F}}(\hat{\theta}^*)$$
$$\widehat{\mathrm{se}}_B = \mathrm{se}_{\hat{F}}(\hat{\theta}^*).$$

- In the absence of a closed formula we can directly compute the standard error and variance of all possible Bootstrap replications (finite number).
- In total we have n^n Bootstrap datasets (exact computation can be extremely time consuming).
- There is only a lower number of distinct samples (samples giving a different Bootstrap replication).
- If the data assume n distinct values, the total number of distinct Bootstrap replications is $m = \binom{2n-1}{n}$ (combinations with repetition from n elements in groups of n).

Bootstrap estimate of the variance.

Finally, we obtain the Bootstrap estimate of the variance (and standard error) of $\hat{\theta}$:

$$\widehat{\mathrm{Var}}_B = \mathrm{Var}_{\hat{F}}(\hat{\theta}^*)$$

$$\widehat{\mathrm{se}}_B = \mathrm{se}_{\hat{F}}(\hat{\theta}^*).$$

$$\widehat{Var}_{\hat{F}}(\hat{\theta}^*) = \left[\frac{1}{n^n}\sum_{j=1}^{n^n}(\hat{\theta}^*_j - \hat{\theta}^*_{(\cdot)})^2\right]$$

$$= \sum_{j=1}^{m} w_j(\hat{\theta}^*_j - \hat{\theta}^*_{(\cdot)})^2$$

BOOTSTRAP

Bootstrap estimate of the variance.

Finally, we obtain the Bootstrap estimate of the variance (and standard error) of $\hat{\theta}$:

$$\widehat{\mathrm{Var}}_B = \mathrm{Var}_{\hat{F}}(\hat{\theta}^*)$$

$$\widehat{\mathrm{se}}_B = \mathrm{se}_{\hat{F}}(\hat{\theta}^*).$$

$$\widehat{Var}_{\hat{F}}(\hat{\theta}^*) = \left[\frac{1}{n^n}\sum_{j=1}^{n^n}(\hat{\theta}^*_j - \hat{\theta}^*_{(\cdot)})^2\right]$$

$$= \sum_{j=1}^{m} w_j(\hat{\theta}^*_j - \hat{\theta}^*_{(\cdot)})^2$$

The m distinct datasets do not have the same probability of being extracted, but such probability cam be computed and is denoted here as w_j.

BOOTSTRAP

Bootstrap estimate of the variance.

Finally, we obtain the Bootstrap estimate of the variance (and standard error) of $\hat{\theta}$:

$$\widehat{\mathrm{Var}}_B = \mathrm{Var}_{\hat{F}}(\hat{\theta}^*)$$

$$\widehat{\mathrm{se}}_B = \mathrm{se}_{\hat{F}}(\hat{\theta}^*).$$

$$\widehat{Var}_{\hat{F}}(\hat{\theta}^*) = \left[\frac{1}{n^n}\sum_{j=1}^{n^n}(\hat{\theta}^*_j - \hat{\theta}^*_{(\cdot)})^2\right]$$

$$= \sum_{j=1}^{m} w_j(\hat{\theta}^*_j - \hat{\theta}^*_{(\cdot)})^2$$

The number m of distinct datasets is lower than n^n, but can be however really big!

Bootstrap estimate of the variance.

Algorithm.

- Repeat B times:
  - Draw a Bootstrap sample $\mathbf{x}_b^*$
  - Evaluate the Bootstrap replication $\hat{\theta}_b^* = s(\mathbf{x}_b^*)$
- Estimate the variance with the variance of the B replications:

$$\widehat{\mathrm{Var}}_B = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta}_{(\cdot)}^*)^2$$

with $\hat{\theta}_{(\cdot)}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$.

BOOTSTRAP

Bootstrap estimate of the variance.

Algorithm.

- Repeat B times:
  - Draw a Bootstrap sample $\mathbf{x}_b^*$
  - Evaluate the Bootstrap replication $\hat{\theta}_b^* = s(\mathbf{x}_b^*)$
- Estimate the variance with the variance of the B replications:

$$\widehat{\mathrm{Var}}_B = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}_b^* - \hat{\theta}_{(\cdot)}^*)^2$$

with $\hat{\theta}_{(\cdot)}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^*$.

$$\boxed{\lim_{B \to \infty} \widehat{\mathrm{Var}}_B = \mathrm{Var}_{\hat{F}}(\hat{\theta}^*)}$$

BOOTSTRAP

$$\lim_{B\to\infty} \widehat{\mathrm{Var}}_B = \mathrm{Var}_{\hat{F}}(\hat{\theta}^*)$$

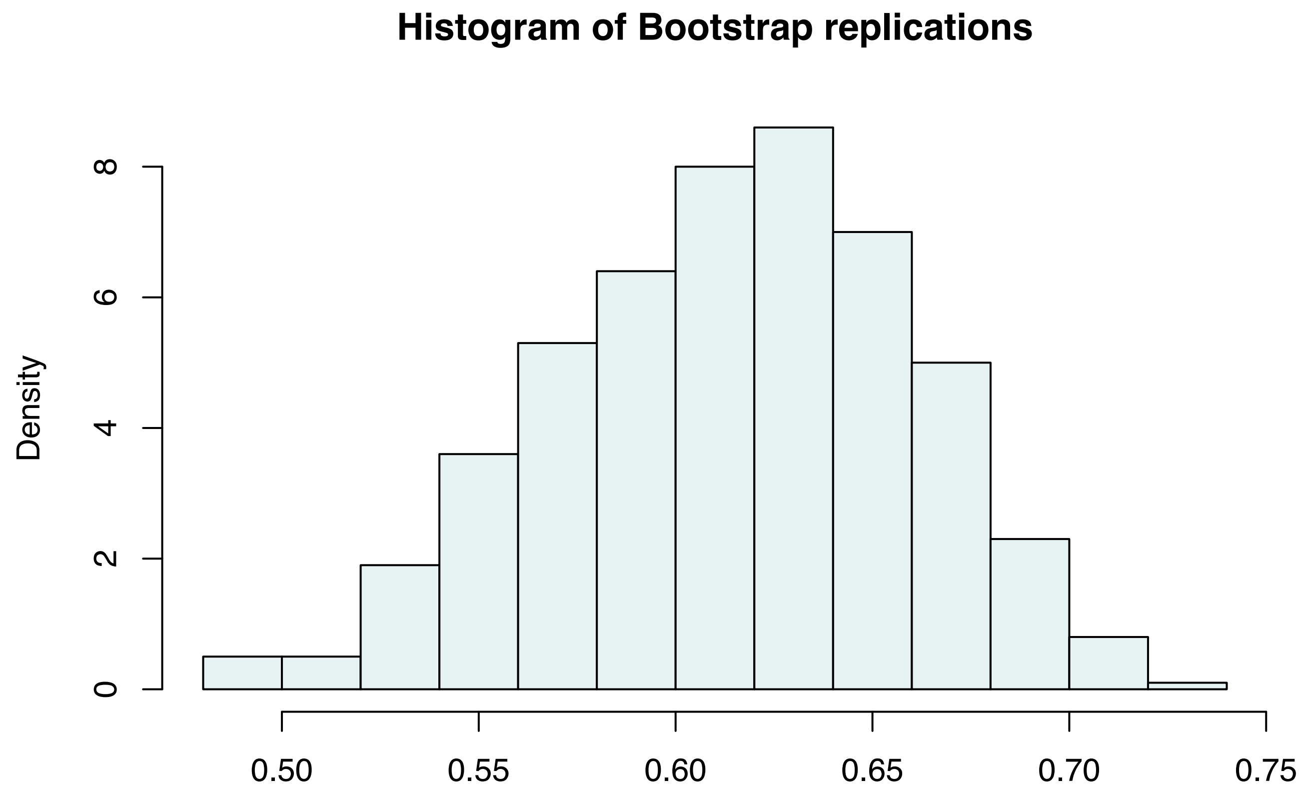
How many bootstraps?

- Nowadays often feasible to use a very big number (B>1000).
- Need R ≥ 100 for point estimate of bias, variance, etc.
- Need R ≫ 100, prefer R ≥ 1000 to estimate tail quantiles (they will be needed for 95% confidence intervals).

BOOTSTRAP

Example: test score data.

$$\hat{\theta} = 0.6191, \quad \widehat{\text{se}}_B = 0.0451, \quad \widehat{\text{Bias}}_B = -0.0051 \quad (B = 500)$$



BOOTSTRAP

Bootstrap estimate of the bias.

The bias of $\hat{\theta}$ can also be estimated with the Bootstrap. A simple estimate is the following:

$$\widehat{\text{Bias}}_B = \hat{\theta}^*_{(\cdot)} - \hat{\theta}$$

The Bias can be estimated through the same algorithm presented before, by using the average of B independent Boostrap replications $\hat{\theta}^*_{(\cdot)} = \frac{1}{B}\sum_{i=1}^{B} \hat{\theta}^*_b$.

A better Bootstrap estimate of the bias.

It applies only when $\hat{\theta} = t(\hat{F})$ (plug-in estimator). For the Bootrtap sample $\mathbf{x}^*$, and for all $j = 1, \dots, n$, define P_j^* as the proportion of units in the bootstrap sample that equals the jth original data point:

$$P_j^* = \#\{x_i = x_j\}/n$$

The quantities P_j^* can be collected in the resampling vector $\mathbf{P}^* = (P_1^*, \dots, P_n^*)$. Clearly, for each Bootstrap sample, $\sum_{j=1}^{n} P_j^* = 1$.

Now, $\hat{\theta}^*$ can be thought as a function of $\mathbf{P}^*$:

$$\hat{\theta}^* = T(\mathbf{P}^*).$$

BOOTSTRAP

A better Bootstrap estimate of the bias.

Similarly, we can define the resampling vector of the original data as

$$\mathbf{P}^0 = \left(\frac{1}{n}, \dots, \frac{1}{n}\right).$$

And since $\hat{\theta}$ is the plug-in estimator:

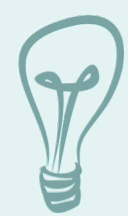
$$T(\mathbf{P}^0) = t(\hat{F}) = \hat{\theta}.$$

Idea: compare $\mathbf{P}^0$ with the distribution of $\mathbf{P}^*$:

$$\overline{\mathbf{P}}^* = \frac{1}{B}\sum_{b=1}^{B}\mathbf{P}_b^*$$

$$\widehat{\text{Bias}}_B = \hat{\theta}^* - T\left(\overline{\mathbf{P}}^0\right)$$

$$\overline{\text{Bias}}_B = \hat{\theta}^* - T\left(\overline{\mathbf{P}}^*\right)$$



BOOTSTRAP

A better Bootstrap estimate of the bias.

$$\overline{\mathbf{P}}^* = \frac{1}{B}\sum_{b=1}^{B} \mathbf{P}_b^*$$

$$\widehat{\text{Bias}}_B = \hat{\theta}^* - T\left(\overline{\mathbf{P}}^0\right)$$

$$\overline{\text{Bias}}_B = \hat{\theta}^* - T\left(\overline{\mathbf{P}}^*\right)$$

- $\widehat{\text{Bias}}_B$ is the estimate that we first defined. It is easier to compute and works for every estimator (not necessarily the plug-in estimator).
- Both estimates of the bias converge to the quantity $\text{Bias}_\infty = \text{Bias}_{\hat{F}}$ as $B \to \infty$.
- The convergence is faster for $\overline{\text{Bias}}_B$.

PARAMETRIC BOOTSTRAP

In some cases we can assume that data follow a parametric distribution, but we don't know the parameters of such distribution.

In this case we can use parametric Bootstrap:

Monte Carlo method	Parametric Bootstrap	Nonparametric Bootstrap

Amount of information about F →

PARAMETRIC BOOTSTRAP

Parametric Bootstrap is based on the direct computation of:

$$\mathrm{Var}_{\hat{F}_{Par}}(\hat{\theta}^*)$$

Where $\hat{F}_{Par}$ is an estimate of F derived from a parametric model.
For instance, if we assume $X_i \sim N(\mu, \sigma^2)$, we can estimate the parameters and then obtain $\hat{F}_{Par} \sim N(\hat{\mu}, \hat{\sigma}^2)$.
Bootstrap samples are now generated from $\hat{F}_{Par}$, and finally we can evaluate $\hat{\theta}$ on the Bootstrap samples and compute the variance:

$$\mathbf{x} \quad \rightarrow \quad \hat{F}_{Par} \quad \rightarrow \quad \mathbf{x}^* \quad \rightarrow \quad \hat{\theta}^* = s(\mathbf{x}^*) \quad \rightarrow \quad \mathrm{Var}_{\hat{F}_{Par}}(\hat{\theta}^*)$$

PARAMETRIC BOOTSTRAP

Algorithm.

- Choose a parametric distribution F_{Par} for the data.
- Estimate the parameters of the distribution using the sample $\mathbf{x}$, obtaining $\hat{F}_{Par}$.
- Repeat B times:
  - Draw a Bootstrap sample $\mathbf{x}_b^*$ from $\hat{F}_{Par}$.
  - Evaluate the Bootstrap replication $\hat{\theta}_b^* = s(\mathbf{x}_b^*)$.
- Estimate the variance with the variance of the B replications:

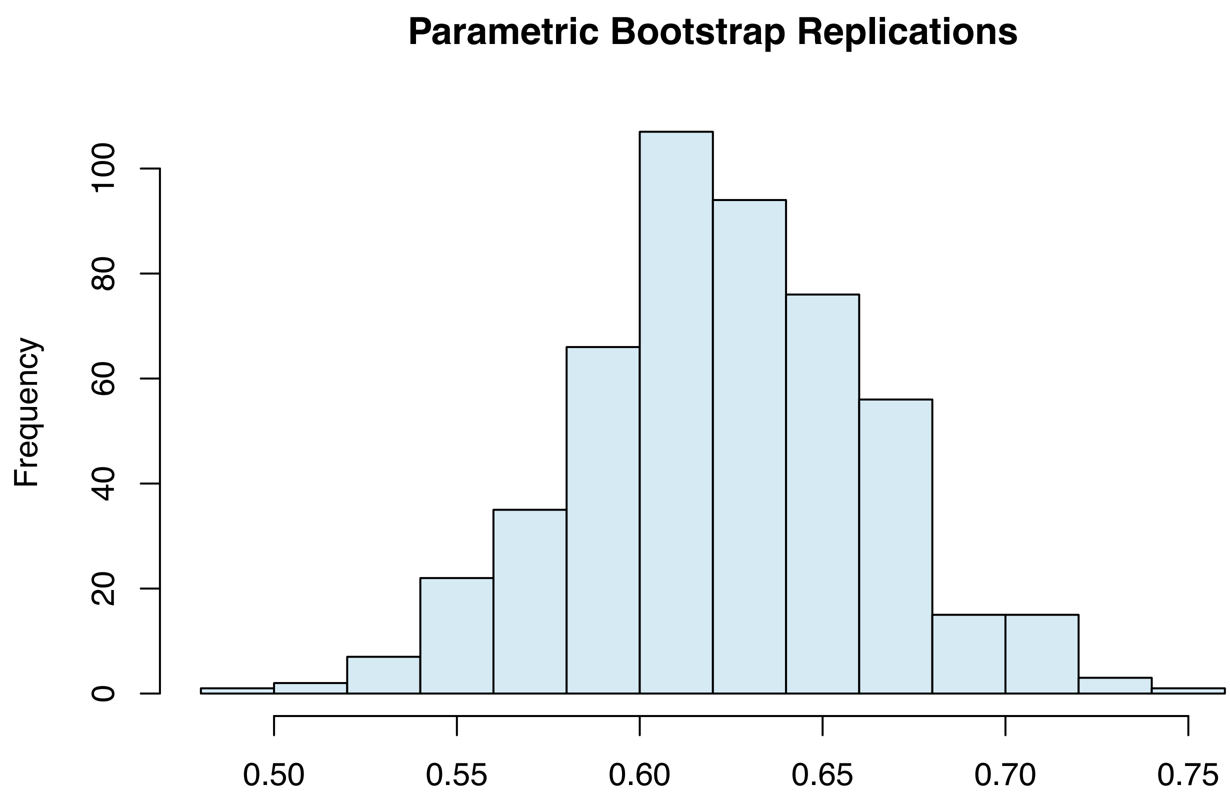
$$\widehat{\mathrm{Var}}_B = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta}_{(\cdot)}^*)^2$$

with $\hat{\theta}_{(\cdot)}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$.

PARAMETRIC BOOTSTRAP

Example: test score data.

$$\hat{\theta} = 0.6191, \quad \widehat{\mathrm{se}}_B = 0.0402, \quad \widehat{\mathrm{Bias}}_B = 0.0042 \quad (B = 500)$$

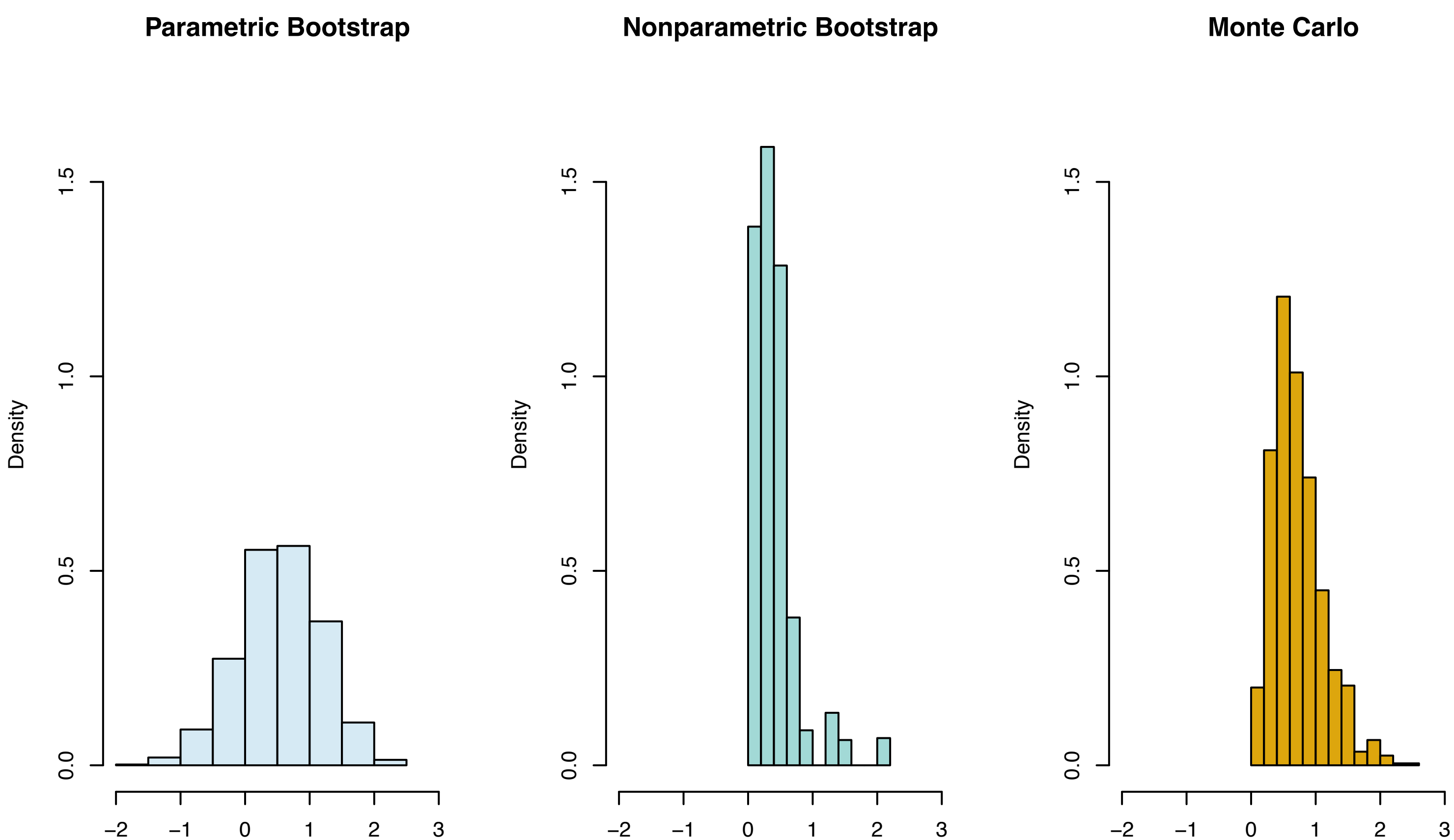


PARAMETRIC VS NONPARAMETRIC BOOTSTRAP

- Parametric bootstrap is useful when we have some knowledge about data distribution.
- The knowledge about the distribution reduces the variance of the estimate of the distribution function, giving better results.
- However, if the assumption about the data distribution is not met, parametric bootstrap can be biased.
- Nonparametric Bootstrap is not biased and more flexible. It does not require any assumption on the data distribution.
- However, nonparametric bootstrap can give poor estimates in some cases (e.g., when the support of the data distribution depends on the parameter that we need to estimate).

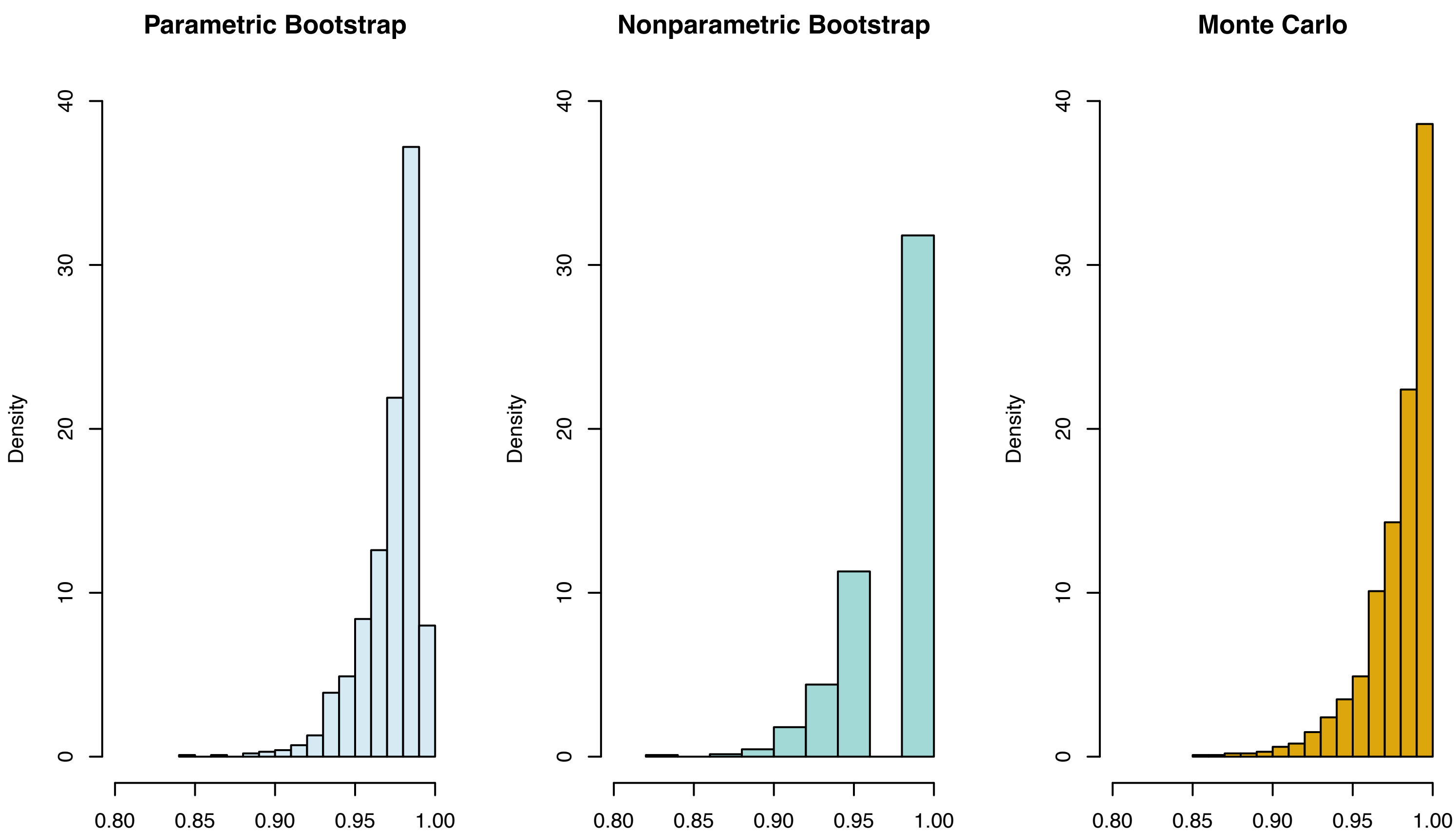
EXAMPLE OF PARAMETRIC BOOTSTRAP FAILURE

Misspecification of the parametric distribution: data generated from Exponential distribution and parametric Bootstrap based on Normal distribution.



EXAMPLE OF NONPARAMETRIC BOOTSTRAP FAILURE

The domain of the data distribution depends on theta. E.g., estimating the upper bound of the domain of a Uniform distribution on (0,1):

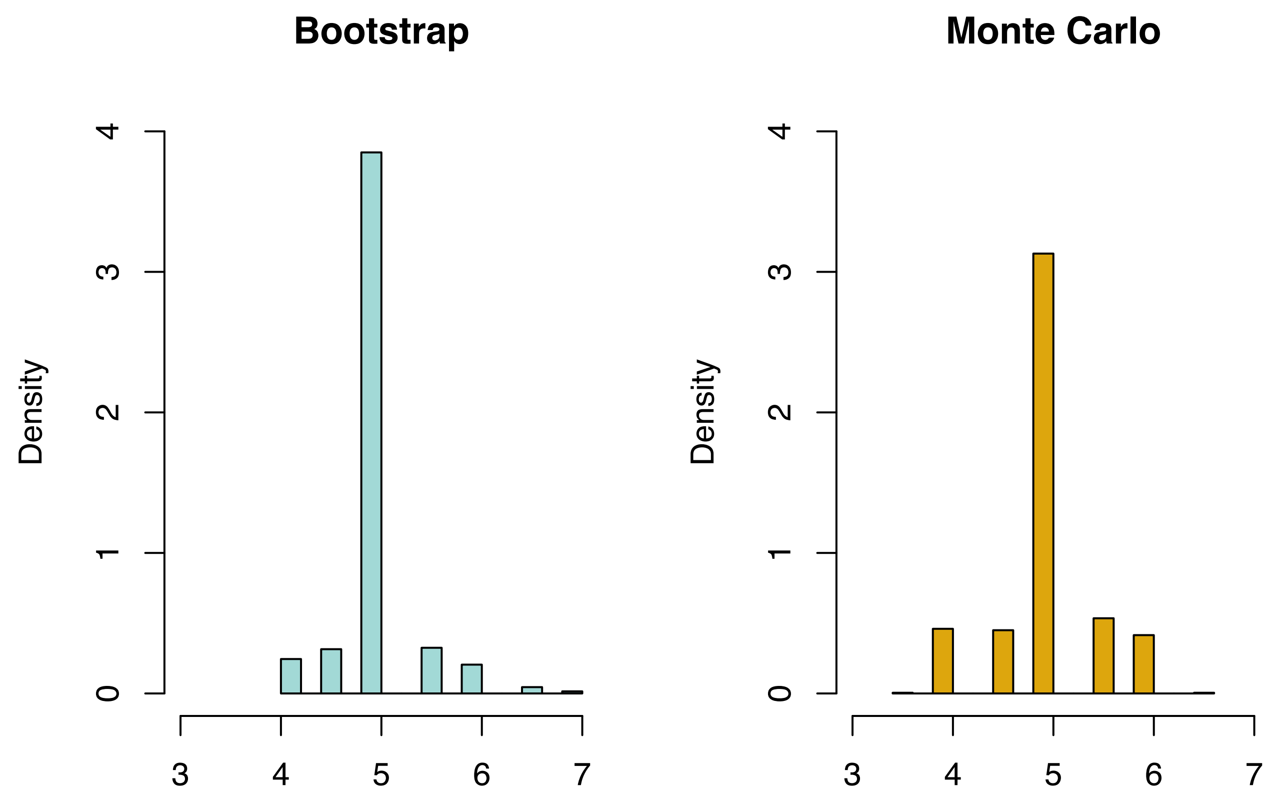


JACKNIFE VS BOOTSTRAP

- Jacknife is computationally more efficient than Bootstrap (it is based on only simulating n data sets).
- Bootstrap provides in general more reliable estimates. They tend to agree if the statistic is linear, or if it has a smooth expression.
- Jacknife fails for non-smooth statistics (e.g.,median).

JACKNIFE VS BOOSTRAP

We apply the Bootstrap to estimate the standard error of the sample median.



BOOTSTRAP FOR MORE COMPLEX DATA STRUCTURES

The bootstrap was described for a one-sample model:

- Individual data points can be numbers or more complex objects (vectors, matrices, functions, images, ...).
- Data are produced from a single distribution F.
- The Bootstrap can be applied to more general data structures.

BOOTSTRAP FOR MORE COMPLEX DATA STRUCTURES

The bootstrap was described for a one-sample model:

- Individual data points can be numbers or more complex objects (vectors, matrices, functions, images, …).
- Data are produced from a single distribution F.
- The Bootstrap can be applied to more general data structures.

General Bootstrap algorithm:

- We start from an unknown probability model P generating data $\mathbf{x}$.
- We find the (nonparametric or parametric) estimate $\hat{P}$ of the unknown model P.
- We generate Bootstrap samples $\mathbf{x}^*$ from the estimated model $\hat{P}$, and use them to evaluate the standard error, bias, and distribution of a quantity of interest θ.

UNIVERSITÀ CATTOLICA del Sacro Cuore

TWO-SAMPLE PROBLEM

Assume that we observe two samples of data $\mathbf{z} = (z_1, \ldots, z_m)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ (e.g., treatment and control). Denote as F and G the distributions of z_j and y_i, respectively, and assume that F and G are independent.
Assume that we are interested in evaluating the mean difference between the two groups:

$$\theta = \mathbb{E}[F] - \mathbb{E}[G] = \mu_z - \mu_y.$$

- The unknown probability model is $P = (F, G)$, with F and G independent.
- The plug-in estimator of θ is $\hat{\theta} = \overline{Z} - \overline{Y}$.
- $\hat{P} = (\hat{F}, \hat{G})$, being $\hat{F}$ $(\hat{G})$ the empirical distribution of data z_i (y_i), and $\mathbf{x} = (\mathbf{z}, \mathbf{y})$.
- The Bootstrap samples can be computed as $\mathbf{x}^* = (\mathbf{z}^*, \mathbf{y}^*)$, where $\mathbf{z}^*$ $(\mathbf{y}^*)$ is a sample of size m (n) drawn from the distribution $\hat{F}$ $(\hat{G})$.
- The bootstrap replication is then $\hat{\theta}^* = \frac{1}{m}\sum_{i=1}^{m} z_i^* - \frac{1}{n}\sum_{i=1}^{m} y_i^*$.
- The variance, standard deviation, bias, and distribution of $\hat{\theta}$ can be evaluated resampling B times from $\hat{P}$.

REGRESSION MODELS

Consider a linear regression model

$$x_i = (\underbrace{\mathbf{c}_i}_{\text{(1 x p) vector of covariates}}, \underbrace{y_i}_{\text{response}}) \quad i = 1, \ldots, n.$$

REGRESSION MODELS

Consider a linear regression model

$$x_i = (\mathbf{c}_i, y_i) \quad i = 1, \ldots, n.$$

In linear regression we assume:

$$\mu_i = \mathbb{E}[y_i|\mathbf{c}_i] = \mathbf{c}_i\boldsymbol{\beta} = \sum_{j=1}^{p} c_{ij}\beta_j.$$

This is true for the probabilistic model:

$$y_i = \mathbf{c}_i\boldsymbol{\beta} + \varepsilon_i \quad i = 1, \ldots, n$$

Where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ is a random sample from a distribution F such that

$$\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0} \quad \mathrm{Var}[\boldsymbol{\varepsilon}] = \sigma^2\mathbb{I}_{p\times p}$$

REGRESSION MODELS

The OLS estimator of the vector $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{c}_i \boldsymbol{\beta})^2 = (C'C)^{-1} C' \mathbf{y}$$

that is an unbiased estimator with standard error

$$\widehat{\mathrm{se}}(\hat{\beta}_j) = \hat{\sigma} \sqrt{[(C'C)^{-1}]_{j,j}}$$

REGRESSION MODELS

The OLS estimator of the vector $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{c}_i \boldsymbol{\beta})^2 = (C'C)^{-1} C' \mathbf{y}$$

that is an unbiased estimator with standard error

$$\widehat{\mathrm{se}}(\hat{\beta}_j) = \hat{\sigma} \sqrt{[(C'C)^{-1}]_{j,j}}$$

To apply the Bootstrap in this case we need to find an estimate of the model $P = (\boldsymbol{\beta}, F)$, and generate Bootstrap samples from the estimated model.

Idea: Estimate $\boldsymbol{\beta}$ with the OLS estimator and F with the empirical distribution of regression residuals.

REGRESSION MODELS

Bootstrap Algorithm for regression.

- Estimate $\boldsymbol{\beta}$ with the OLS estimate $\hat{\boldsymbol{\beta}}$.
- Compute the residuals $\hat{\varepsilon}_i = y_i - \mathbf{c}_i\hat{\boldsymbol{\beta}}$.
- Estimate the empirical distribution $\hat{F}$ of the residuals $\hat{\varepsilon}_i$: $\hat{F}$ gives probability $1/n$ to each residual $\hat{\varepsilon}_i$.
- The estimated model is now $\hat{P} = (\hat{\boldsymbol{\beta}}, \hat{F})$.
- Repeat B times:
  - Generate a Bootstrap sample $\mathbf{x}_b^* = (\mathbf{c}_i, y_i)$ from $\hat{P}$, with $\mathbf{y}_b^* = \mathbf{c}_i\hat{\boldsymbol{\beta}} + \varepsilon_{i_b}^*$.
  - Compute the Bootstrap replication $\hat{\beta}_b^* = (C'C)^{-1}C'\mathbf{y}^*$.

REGRESSION MODELS

Consider a linear regression model

$$x_i = (\mathbf{c}_i, y_i) \quad i = 1, \ldots, n.$$

Given that the pairs $(\mathbf{c}_i, y_i)$ are sampled from model P, another option is Bootstrapping directly the pairs!

$$\mathbf{x}^* = \{(c_{i1}, y_{i1}), (c_{i2}, y_{i2}), \ldots, (c_{in}, y_{in})\}$$

Random sample drawn with replacement from (1,2,...,n)

REGRESSION MODELS

Consider a linear regression model

$$x_i = (\mathbf{c}_i, y_i) \quad i = 1, \dots, n.$$

Given that the pairs $(\mathbf{c}_i, y_i)$ are sampled from model P, another option is Bootstrapping directly the pairs!

$$\mathbf{x}^* = \{(c_{i1}, y_{i1}), (c_{i2}, y_{i2}), \dots, (c_{in}, y_{in})\}$$

- Bootstrapping residuals works well if the linear model assumed for the regression is correct, and if the terms ε_i have the same distribution.
- Bootstrapping pairs is based on less assumptions: it is more robust to misspecifications of the model.

REGRESSION MODELS

The OLS estimator of the vector $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{c}_i \boldsymbol{\beta})^2 = (C'C)^{-1} C' \mathbf{y}$$

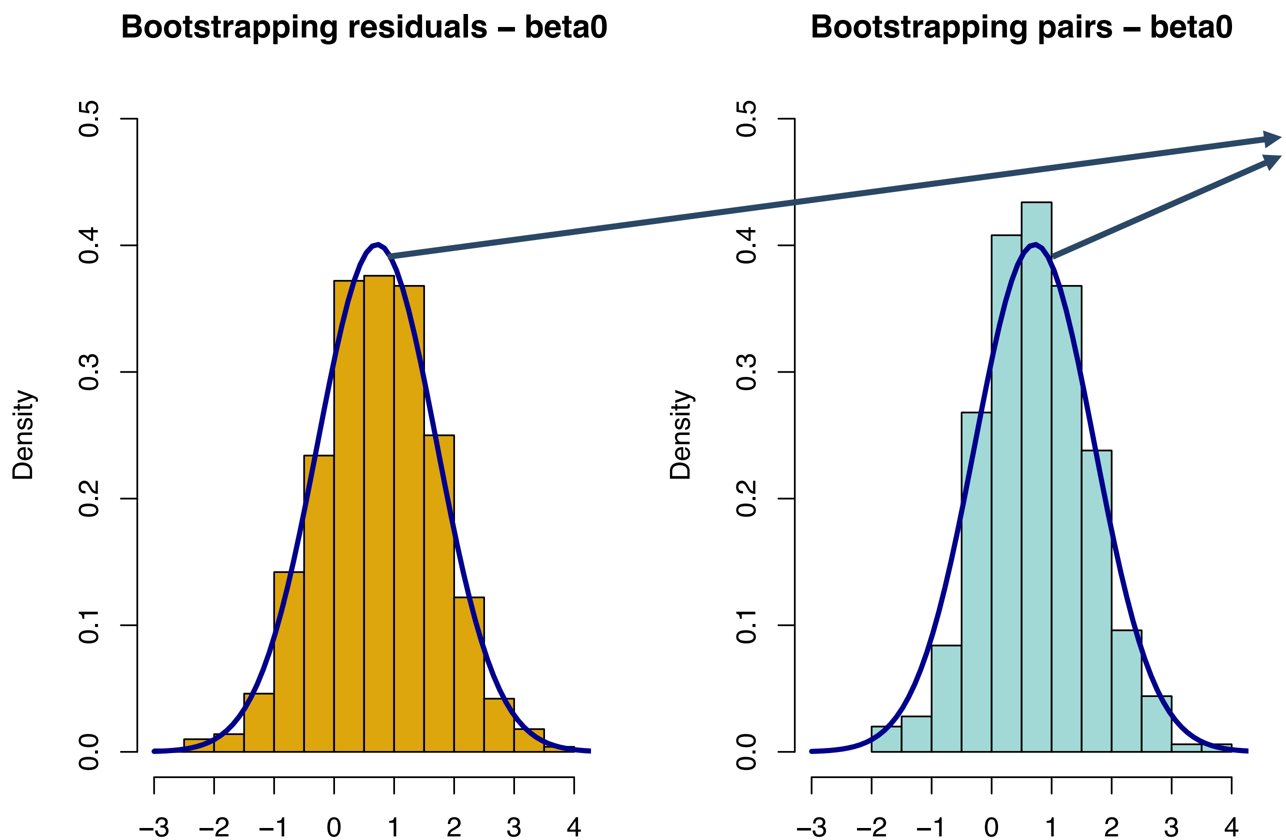
that is an unbiased estimator with standard error

$$\widehat{\text{se}}(\hat{\beta}_j) = \hat{\sigma} \sqrt{[(C'C)^{-1}]_{j,j}}$$

Why do we need Bootstrap to evaluate the standard error?

- Point estimation of the standard error can be computed in the classical way, without making distributional assumptions on the distribution of the residuals
- Inference (confidence intervals and tests) on beta is done in the classical way assuming normality.
- In the following lecture we will see how to make inference based on Bootstrap replications.

Estimating the intercept of a correctly specified regression model with Normal errors.



Theoretical distribution of the OLS estimator under Normality

REGRESSION MODELS

Estimating beta1 of a correctly specified regression model with Normal errors.