Resampling techniques for statistical modeling

Gianluca Bontempi

Département d'Informatique

Boulevard de Triomphe - CP 212

http://www.ulb.ac.be/di

Rationale

- All the methods presented so far are strongly based on two assumptions:
  1. the form of distribution underlying the data is parametric and known. The only uncertainty concerns the values of some parameters.
  2. we focus on parameters θ (e.g. the mean) and corresponding estimators $\hat{\boldsymbol{\theta}}$ (e.g. the average) for which the analysis is simple.

Rationale(II)

- For a generic parameter in a parametric setting, the distribution of the statistic $\hat{\boldsymbol{\theta}}$ cannot be easily defined.
- For a generic problem, when only a set of observations are available, the parametric assumption cannot be made without strongly biasing the procedure.
- Two alternatives:
  1. make simpler the complex problem, with the risk of oversimplifying.
  2. use computer intensive techniques.
- Note that, also when the configuration is parametric, computer intensive methods could still be useful to assess the robustness of conclusions drawn from a parametric analysis.

Estimation of arbitrary statistics

- Consider a set D_N of N data points sampled from a one-dimensional distribution of a r.v. $\mathbf{x}$.
- Suppose that our quantity of interest is the mean of $\mathbf{x}$. It is straightforward to compute the estimate $\hat{\mu}$ of the mean, the bias and the variance of the estimator:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \text{Bias}[\hat{\boldsymbol{\mu}}] = 0, \qquad \text{Var}\,[\hat{\boldsymbol{\mu}}] = \frac{1}{N(N-1)}\sum_{i=1}^{N}(x_i - \hat{\mu})$$

- Consider now another quantity of interest, for example the median or a mode of the distribution. While it is easy to have an estimate of these quantities, their accuracy is difficult to be computed.
- In other terms, **given an arbitrary estimator $\hat{\theta}$, the analytical form of the variance Var $\left[\hat{\theta}\right]$ and the bias Bias$[\hat{\theta}]$ is typically not available.**

Example: the patch data

- Eight subjects wore medical patches designed to infuse a certain hormone in the blood.
- Each subject had his hormone levels measured after wearing three different patches: a placebo, an “old” patch and a “new” patch.
- The goal is to show bioequivalence. In other terms the Food and Drug Administration (FDA) will approve the new patch for sale only if the new patch is bioequivalent to the old one.
- The FDA criterion is

$$\theta = \frac{|E(\text{new}) - E(\text{old})|}{E(\text{old}) - E(\text{placebo})} \leq 0.2$$

Example: the patch data (II)

subj	plac	old	new	z=old-plac	y=new-old
1	9243	17649	16449	8406	-1200
2	9671	12013	14614	2342	2601
3	11792	19979	17274	8187	-2705
...	...	...	...	...	...
8	18806	29044	26325	10238	-2719
mean:				6342	-452.3

Estimate

$$\hat{\theta} = t(\hat{F}) = \frac{|\bar{y}|}{\bar{z}} = \frac{452.3}{6342} = 0.07$$

Does it satisfy the FDA criterion?

What about its accuracy, bias, variance?

Jacknife

- The **jacknife** (or **leave-one-out**) resampling technique aims at providing a computational procedure to estimate the variance and the bias of a generic estimator $\hat{\boldsymbol{\theta}}$.
- The technique was first proposed by Quenouille in 1949.
- The technique is based on removing samples from the available dataset and recalculating the estimator.
- It is a general-purpose tool which is easy to implement and solves a number of problems.

Jacknife for the mean

In order to show the theoretical foundation of the jacknife, we apply first this technique to the estimator $\hat{\mu}$ of the mean .
Suppose we have available a dataset D_N. Let us remove the ith sample from D_N and let us calculate the **leave-one-out (l-o-o) mean** estimate from the $N-1$ remaining samples

$$\hat{\mu}_{(i)} = \frac{1}{N-1}\sum_{j\neq i}^{N} x_j = \frac{N\hat{\mu} - x_i}{N-1}$$

Observe that the following relation holds

$$x_i = N\hat{\mu} - (N-1)\hat{\mu}_{(i)}$$

that is, we can calculate x_i if we know both $\hat{\mu}$ and $\hat{\mu}_{(i)}$.

Jacknife for an arbitrary statistic

Suppose that we wish to estimate some parameter θ as a very complex statistic of the N data points

$$\hat{\theta} = f(D_N) = f(x_1, x_2, \ldots, x_N)$$

We first compute

$$\hat{\theta}_{(i)} = f(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N)$$

which is called the ith **jacknife replication** of $\hat{\theta}$. Then by analogy with the relation holding for the mean we define the i-th **pseudovalue** by

$$\eta_{(i)} = N\hat{\theta} - (N-1)\hat{\theta}_{(i)}$$

These psudovalues assume the same role as the x_i in calculating the sampled mean.

The jacknife estimate of bias

Hence the **jacknife estimate** of θ is given by

$$\hat{\theta}_{\text{jk}} = \frac{1}{N}\sum_{i=1}^{N}\eta_{(i)} = \frac{1}{N}\sum_{i=1}^{N}\left(N\hat{\theta} - (N-1)\hat{\theta}_{(i)}\right) = N\hat{\theta} - (N-1)\hat{\theta}_{(\cdot)}$$

where

$$\hat{\theta}_{(\cdot)} = \frac{\sum_{i=1}^{N}\hat{\theta}_{(i)}}{N}$$

Since Bias$[\hat{\boldsymbol{\theta}}] = E[\hat{\boldsymbol{\theta}}] - \theta$, if we put Bias$_{\text{jk}}[\hat{\boldsymbol{\theta}}] = \hat{\theta} - \hat{\theta}_{\text{jk}}$, we obtain the **jacknife estimate of the bias of $\hat{\boldsymbol{\theta}}$**

$$\text{Bias}_{\text{jk}}[\hat{\boldsymbol{\theta}}] = (N-1)(\hat{\theta}_{(\cdot)} - \hat{\theta})$$

Note that in the particular case of a mean estimator (i.e. $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\mu}}$), we see that we obtain, as expected, Bias$_{\text{jk}}[\hat{\boldsymbol{\mu}}] = 0$.

The jacknife estimate of variance

An estimate of the variance of $\hat{\boldsymbol{\theta}}$ can be obtained from the sample variance of the pseudo-values

$$\mathrm{Var}_{\mathrm{jk}}[\hat{\boldsymbol{\theta}}] = \mathrm{Var}\left[\hat{\boldsymbol{\theta}}_{\mathrm{jk}}\right] =$$

$$= \frac{\mathrm{Var}\left[\eta_{(i)}\right]}{N} = \frac{\sum_{i=1}^{N}\left(\eta_{(i)} - \hat{\theta}_{\mathrm{jk}}\right)^2}{N(N-1)} = \left(\frac{N-1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^2\right)$$

Note that in the particular case of a mean estimator (i.e. $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\mu}}$), it can be shown that

$$\mathrm{Var}_{\mathrm{jk}}[\hat{\boldsymbol{\theta}}] = \mathrm{Var}\left[\hat{\boldsymbol{\mu}}\right] = \frac{\sum_{i=1}^{N}(x_i - \hat{\mu})^2}{N(N-1)}$$

Consideration on jacknife

- The major motivation for jacknife estimates is that they reduce bias.
- However, the jacknife can fail miserably is the statistic $\hat{\boldsymbol{\theta}}$ is not smooth (i.e. small changes in data cause small changes in the statistic).
- An example of non-smooth statistic forwhich the jacknife works badly is the median.

Bootstrap: why this name?

- Boostrap is a data-based simulation method for statistical inference.
- The use of the term *bootstrap* derives from the phrase *to pull oneself up by one's bootstrap*, widely thought to be based on one of the eighteenth century Adventures of Baron Munchausen, by R.E. 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.
- It is not the same as the term "bootstrap" used in computer science to "boot" a computer from a set of core instructions (though the derivation is similar).

The bootstrap

- The bootstrap was introduced in 1979 as a computer-based method for estimating the standard error of an estimator $\hat{\boldsymbol{\theta}}$.
- It enjoys the advantage of being completely automatic.
- The bootstrap estimate of standard errors requires no theoretial calculations and is available no matter how mathematicaly complicated the estimator $\hat{\boldsymbol{\theta}}$ may be.
- The bootstrap is extensively based on the plug-in principle.

The setting of the problem

- Consider a dataset D_N, whose sample values are the outcomes of iid r.v.s $\mathbf{z}_i$ whose probability density is $p_{\mathbf{z}}$.
- We want to make inference about $\theta = t(F)$ which is a characteristic of the population.
- We have chosen an estimator $\hat{\boldsymbol{\theta}}$.
- Our attention is focused on questions concerning the probability distribution of $\hat{\boldsymbol{\theta}}$. For example we would like to know about
  - its bias,
  - its standard error
  - its quantiles
  - its confidence values and so on.

The parametric case

- Suppose that $z \sim F_{\mathbf{z}}(z, \psi)$ and that we are interested in a generic parameter $\theta = t(F)$.
- Note that we have distinguished θ from ψ in order to have a more general setting. For example if $z \sim \mathcal{N}(\mu, \sigma^2)$, it could be that $\psi = \mu$ and $\theta = \log \mu$.
- Suppose we have an estimate $\hat{\psi}$ and a corresponging fitted distribution $F(z, \hat{\psi})$. We will define $\mathbf{z}^* \sim F(z, \hat{\psi})$.
- Suppose we want to calculate the distribution (e.g. bias or variance) of $\hat{\boldsymbol{\theta}}$. For a generic $\hat{\boldsymbol{\theta}}$ this operation is arduous.
- The idea is to estimate $\hat{\boldsymbol{\theta}}$ by using a simulated dataset.

The parametric case (II)

- Let us consider the dataset

$$D_{(b)} = \{z_1^*, z_2^*, \ldots, z_N^*\}$$

obtained by iid sampling the **known distribution** $\mathbf{z}^* \sim \hat{F}(z, \hat{\psi})$. $D_{(b)}$ is called the **bootstrap sample.**

- Consider B repetitions of the $D_{(b)}$ sampling.
- Let us define by

$$\hat{\theta}_{(b)} = t(D_{(b)}) \qquad b = 1, \ldots, B$$

the statistic computed by $D_{(b)}$, $b = 1, \ldots, B$. $\hat{\theta}_{(b)}$ is called **bootstrap replication** of $\hat{\theta}$.

- The statistical properties of $\hat{\boldsymbol{\theta}}$ can now be calculated on the basis of the distribution of $\hat{\boldsymbol{\theta}}_{(b)}$.

Bootstrap estimate of the variance

The bootstrap estimate of the variance of the estimator $\hat{\boldsymbol{\theta}}$, is the variance of the set $\hat{\theta}_{(b)}$, $b = 1, \ldots, B$.

$$\text{Var}_{\text{bs}}[\hat{\boldsymbol{\theta}}] = \text{Var}\left[\hat{\boldsymbol{\theta}}_{(b)}\right] = \frac{\sum_{b=1}^{B}(\hat{\theta}_{(b)} - \hat{\theta}_{(\cdot)})^2}{(B-1)}$$

where

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

If $\hat{\theta} = \hat{\mu}$, for $B \to \infty$ the bootstrap estimate $\text{Var}_{\text{bs}}[\hat{\boldsymbol{\theta}}]$ converges to the variance $\text{Var}\,[\hat{\boldsymbol{\mu}}]$.

Bootstrap estimate of bias

Let $\hat{\boldsymbol{\theta}}$ bt the estimator based on the original sample D_N and

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

Since Bias$[\hat{\boldsymbol{\theta}}] = E[\hat{\boldsymbol{\theta}}] - \theta$, the bootstrap estimate of bias is

$$\text{Bias}_{\text{bs}}[\hat{\boldsymbol{\theta}}] = \hat{\theta}_{(\cdot)} - \hat{\theta}$$

Then, since

$$\theta = E[\hat{\boldsymbol{\theta}}] - \text{Bias}[\hat{\boldsymbol{\theta}}]$$

the **bootstrap bias corrected** estimate is

$$\hat{\theta}_{\text{bs}} = \hat{\theta} - (\hat{\theta}_{(\cdot)} - \hat{\theta}) = 2\hat{\theta} - \hat{\theta}_{(\cdot)}$$

Bootstrap confidence interval

- **Standard bootstrap confidence limits** are based on the assumption that the estimator $\hat{\boldsymbol{\theta}}$ is normally distributed with mean θ and variance σ^2.
- Taking the bootstrap estimate of variance, an approximate $100(1-\alpha)\%$ confidence interval is given by

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\mathrm{Var}_{\mathrm{bs}}[\hat{\boldsymbol{\theta}}]} = \hat{\theta} \pm z_{\alpha/2}\sqrt{\frac{\sum_{b=1}^{B}(\hat{\theta}_{(b)} - \hat{\theta}_{(\cdot)})^2}{(B-1)}}$$

- An improved interval is given by using the bootstrap bias corrected estimate of θ. The interval becomes

$$\hat{\theta}_{\mathrm{bs}} \pm z_{\alpha/2}\sqrt{\mathrm{Var}_{\mathrm{bs}}[\hat{\boldsymbol{\theta}}]} = 2\hat{\theta} - \hat{\theta}_{(\cdot)} \pm z_{\alpha/2}\sqrt{\frac{\sum_{b=1}^{B}(\hat{\theta}_{(b)} - \hat{\theta}_{(\cdot)})^2}{(B-1)}}$$

Bootstrap percentile confidence interval

- A more direct approach for constructing a $100(1-\alpha)\%$ confidence interval is to use the upper and lower $\alpha/2$ values of the bootstrap distribution.
- The approaches using the full bootstrap fistribution are often referred to as **percentile confidence limits.**
- If $\hat{\theta}_{L,\alpha/2}$ denotes the value such that only a fraction $\alpha/2$ of all bootstrap estimates are inferior to it, and likewise $\hat{\theta}_{H,\alpha/2}$ is the value exceeded by only $\alpha/2$ of all bootstrap estimates, then an approximate confidence interval is given by

$$[\hat{\theta}_{L,\alpha/2}, \hat{\theta}_{H,\alpha/2}]$$

also called the **Efron's percentile confidence limit.**

Studentized bootstrap statistic

Recall that for $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$ with μ and σ^2 unknown, the $(1-\alpha)$ interval of confidence of μ is

$$[\hat{\mu} - t_{(N-1,\alpha/2)}\hat{\sigma}/\sqrt{N}, \hat{\mu} + t_{(N-1,\alpha/2)}\hat{\sigma}/\sqrt{N}]$$

For a generic distribution of $\mathbf{z}$ the Student $\mathcal{T}$ distribution is no more appropriate. At its place, the bootstrap method can be used to approximate the distribution of the random variable

$$\frac{\hat{\boldsymbol{\theta}} - \theta}{\hat{\boldsymbol{\sigma}}/\sqrt{N}}$$

The idea is to estimate the above variable through the **Studentized bootstrap statistic**

$$\mathbf{s}_{(b)} = \frac{\hat{\boldsymbol{\theta}}_{(b)} - \hat{\theta}}{\sqrt{\text{Var}\left[\hat{\boldsymbol{\theta}}_{(b)}\right]}}$$

Studentized bootstrap confidence interval

Given B bootstrap replications we obtain the empirical distribution of $\mathbf{s}_{(b)}$.

Hence the studentized bootstrap confidence interval for θ is

$$[\hat{\theta} - s_{(\alpha/2)}\hat{\sigma}/\sqrt{N}, \hat{\mu} + s_{(\alpha/2)}\hat{\sigma}/\sqrt{N}]$$

where $s_{(\alpha/2)}$ is the $\alpha/2$ upper point of the empirical distribution of $\mathbf{s}_{(b)}$.

The nonparametric case

- Suppose that we have no parametric model but that wa can sample in an iid manner a set D_N from a **completely unknown** distribution function.
- We will use the empirical distribution function $\hat{F}$ just as we would with a parametric model.
- Since the empirical $\hat{F}$ puts equal probabilities on each original data value, each simulated dataset $D_{(b)}$ is a random sample taken **with replacement** from D_N.
- This resampling procedure is called the **nonparametric bootstrap**.
- It differs from the parametric case only for the generation of the B datasets $D_{(b)}$.

The nonparametric bootstrap algorithm

Suppose we want to compute the standard error of an estimator $\hat{\boldsymbol{\theta}}$. The algorithm has the following steps:

1. Select B independent bootstrap samples $D_{(b)} = \{z_1^*, z_2^*, \ldots, z_N^*\}$, $b = 1, \ldots, B$, each consisting of N data values drawn with replacement from D_N.
2. Evaluate the bootstrap replication corresponding to each bootstrap sample

$$\hat{\theta}_{(b)} = t(D_{(b)}) \qquad b = 1, \ldots, B$$

3. Estimate the standard error $\sqrt{\text{Var}\left[\hat{\boldsymbol{\theta}}\right]}$ by the sample standard deviation $\sqrt{\text{Var}\left[\hat{\boldsymbol{\theta}}_{(b)}\right]}$ of the B replications.

Practical example

- Consider a dataset containing the service hours between failures of the air-conditioning equipment in a Boeing aircraft

$$D_N = \{3, 5, 7, 18, 43, 85, 91, 98, 100, 130, 230, 487\}$$

- We wish to estimate the mean or its reciprocal, the failure rate. Then in this practical example $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\mu}}$.
- We could adopt a parametric approach and suppose that the probabilistic distribution underlying the data is exponential
- Note that in this case it is known that

$$\text{Bias}[\hat{\boldsymbol{\mu}}] = 0, \qquad \text{Var}\,[\hat{\boldsymbol{\mu}}] = \hat{\mu}^2/N$$

- Let us see the relation between empirical and exponential distribution function ($\hat{\lambda} = 1/\hat{\mu}$).
- R script `airco.R`.

Exponential distribution

A continuous random variable $\mathbf{z}$ is said to be **exponentially distributed** with rate $\lambda > 0$ (also $\mathbf{z} \sim \mathcal{E}(\lambda)$) if its probability density function is given by

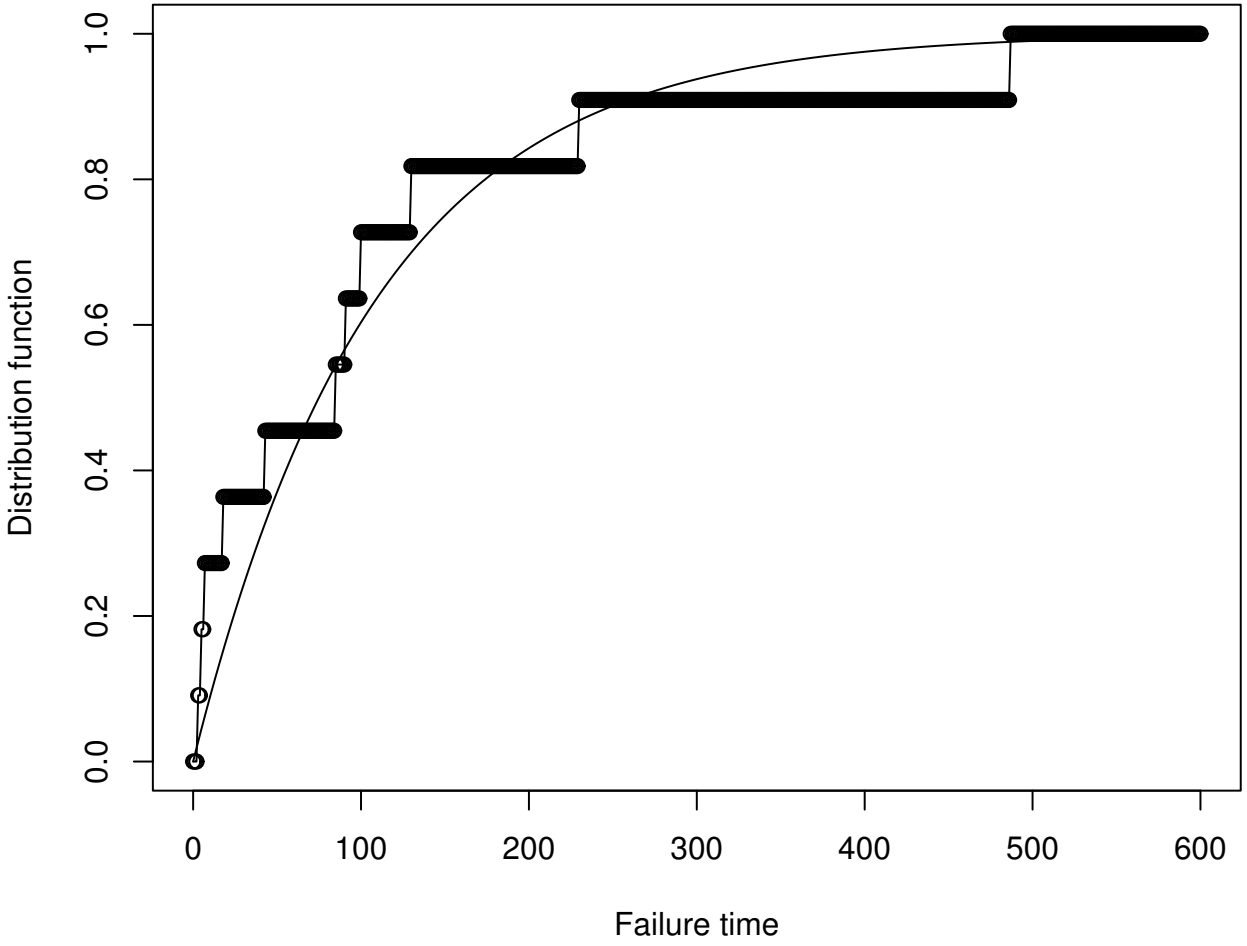
$$p_{\mathbf{z}}(z) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

- The mean of $\mathbf{z}$ is $1/\lambda$.
- The variance of $\mathbf{z}$ is $1/\lambda^2$.
- It can be considered as continuous approximation to the geometric distribution.
- Like the geometric distribution it satisfies the **memoryless property**

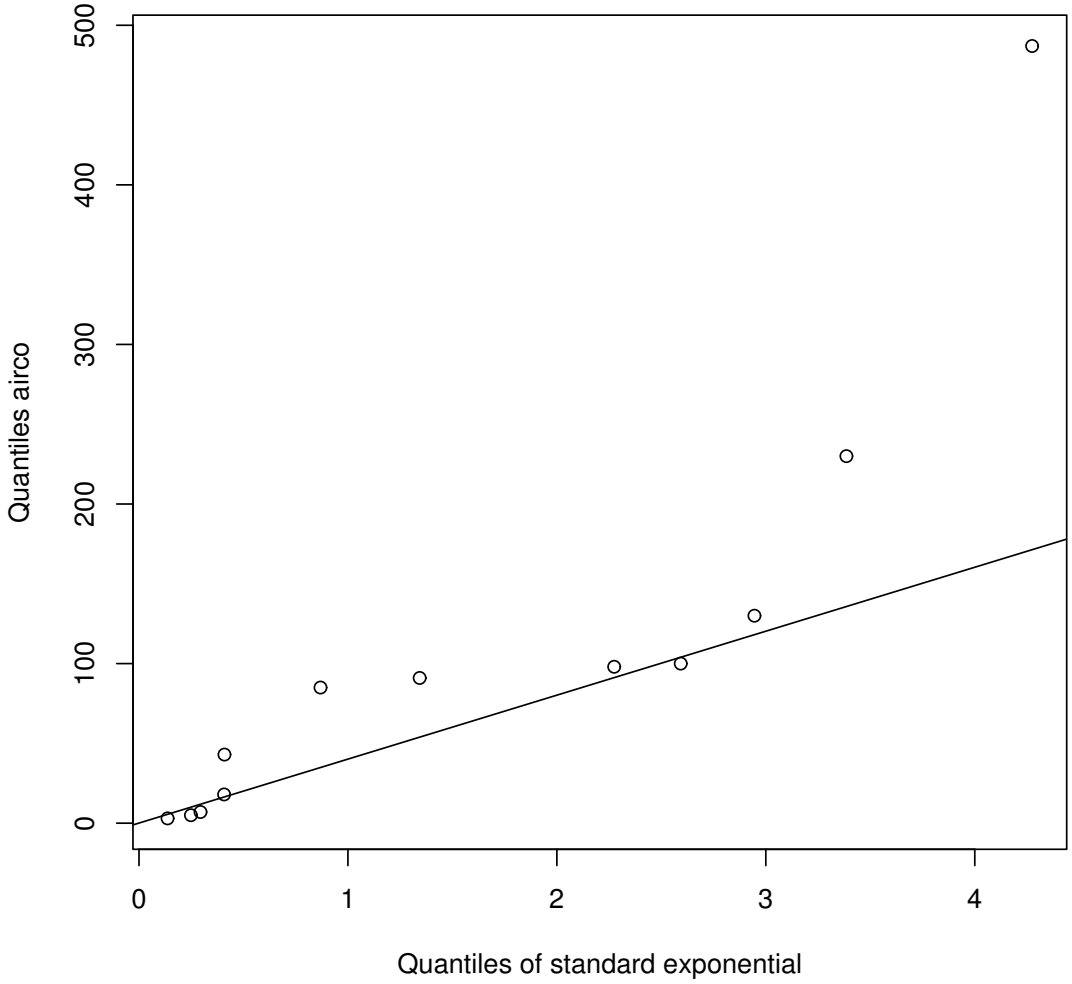
$$\text{Prob}\,\{\mathbf{z} \geq z_1 + z_2 | \mathbf{z} = z_1\} = \text{Prob}\,\{\mathbf{z} \geq z_2\}$$

- It is used to describe physical phenomena (e.g. radioactive decay time or failure time).

Ex: empirical and exponential distribution



Ex: empirical and exponential quantiles



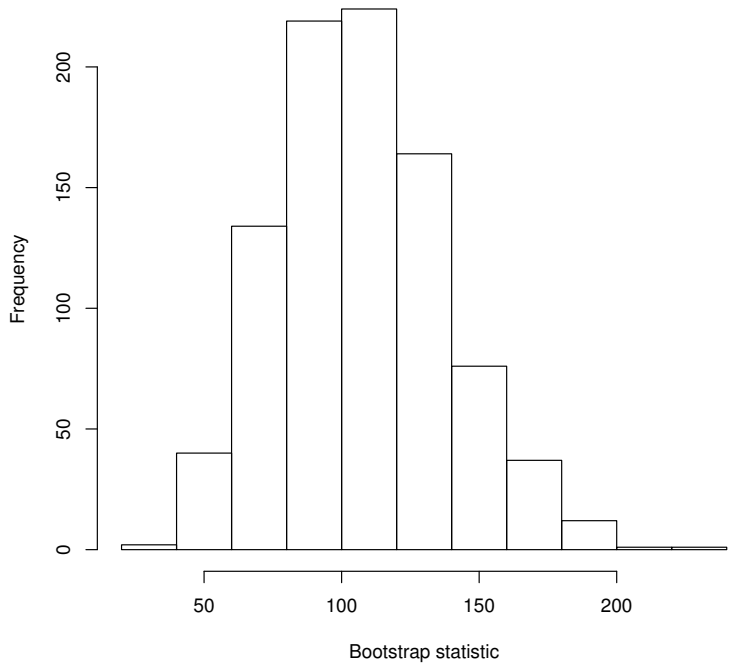
Ex: parametric bootstrap

- Consider as statistic $\hat{\boldsymbol{\theta}}$ the sample average $\hat{\boldsymbol{\mu}}$.
- Assume we know that the distribution is exponential but that we have not access to the analytical expressions of bias and variance of $\hat{\boldsymbol{\mu}}$.
- We perform parametric bootstrap to estimate these quantities.
- We sample B times the exponential distribution $\mathcal{E}(1/\hat{\mu})$.
- In the following we report the outcome of a set of bootstrap experiments.

Ex: distribution of $\hat{\boldsymbol{\theta}}_{(b)}$

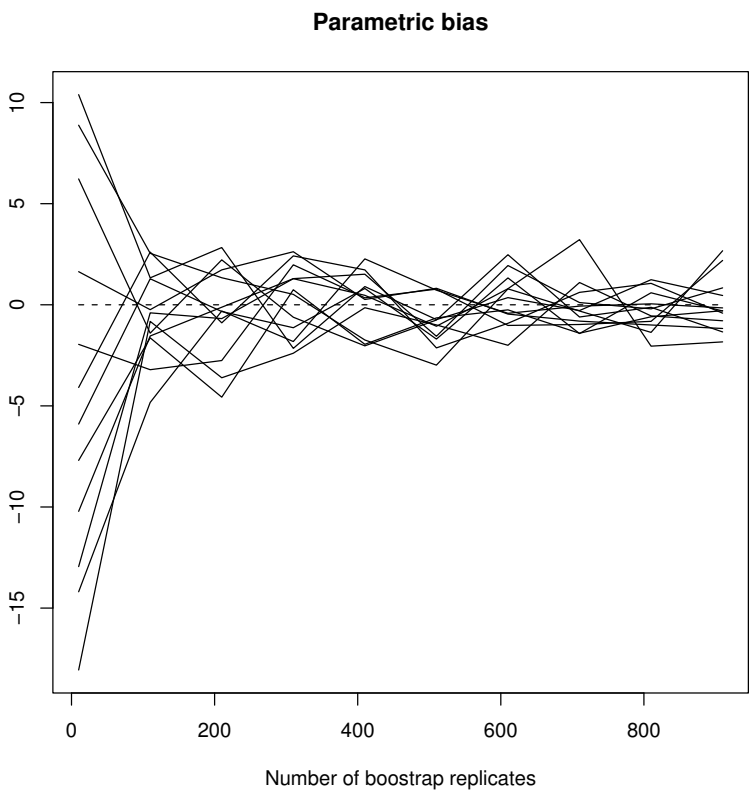
Consider the statistic $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\mu}}$.

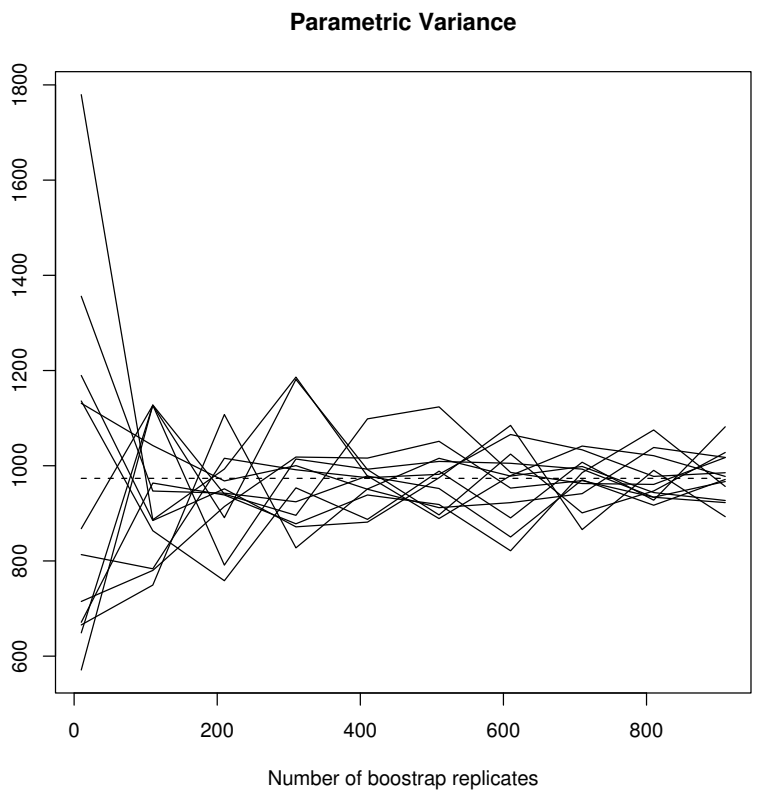
The distribution of $\hat{\boldsymbol{\theta}}_{(b)}$ on the basis of $B = 100$ of bootstrap parametric repetitions is represented by the following histogram



Ex: parametric bootstrap

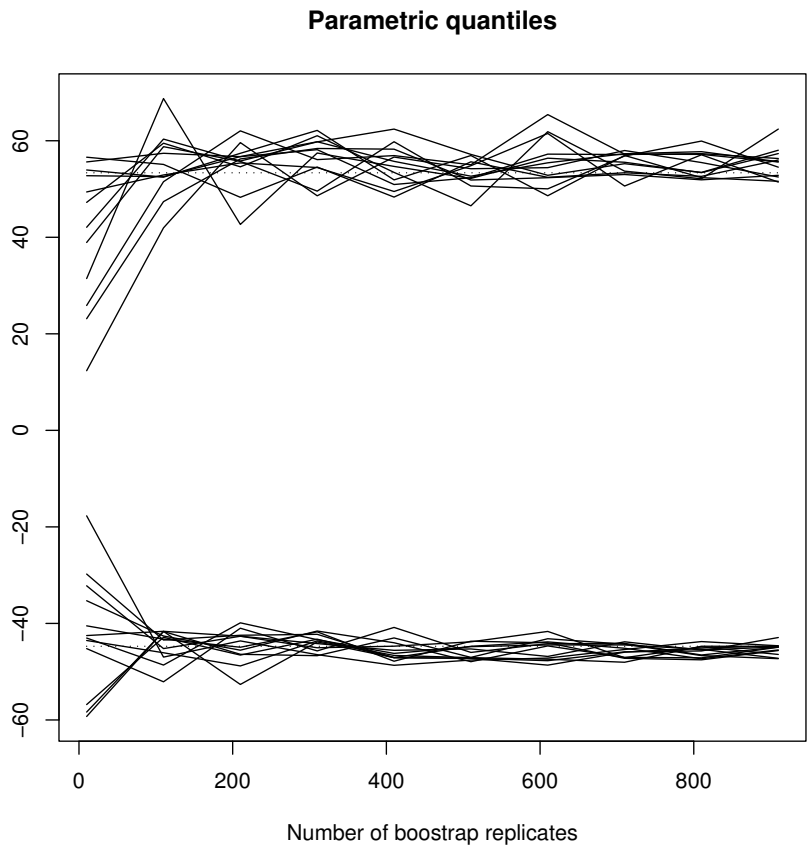
We report the bias and the variance of $\hat{\boldsymbol{\theta}}$ obtained in different random experiments. Each experiment is based on a set of bootstrap repetitions obtained by varying the number B.





Ex: parametric bootstrap

We report the quantiles of $\hat{\boldsymbol{\theta}}$ obtained in different random experiments. Each experiment is based on a set of bootstrap repetitions obtained by varying the number B.

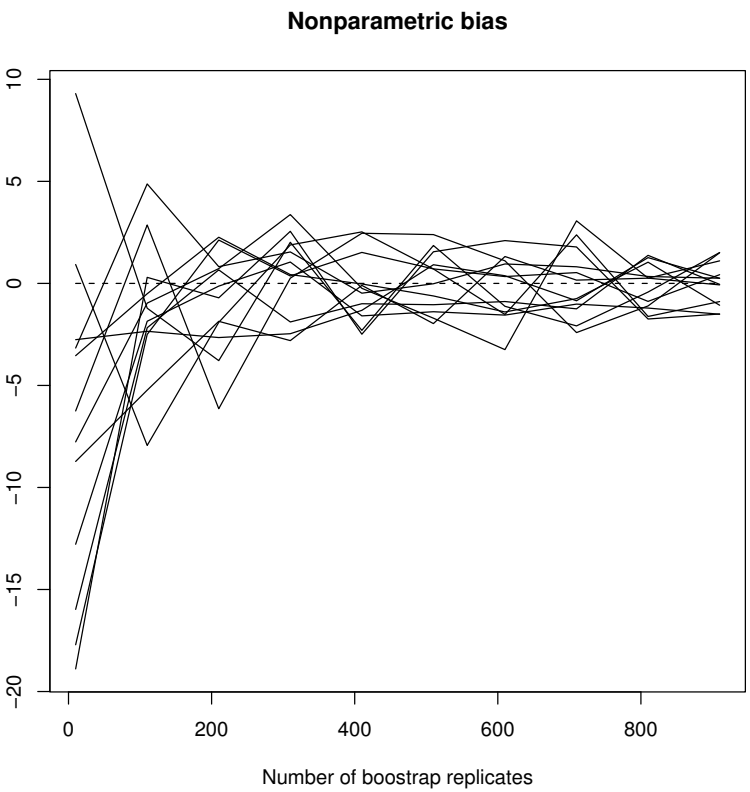


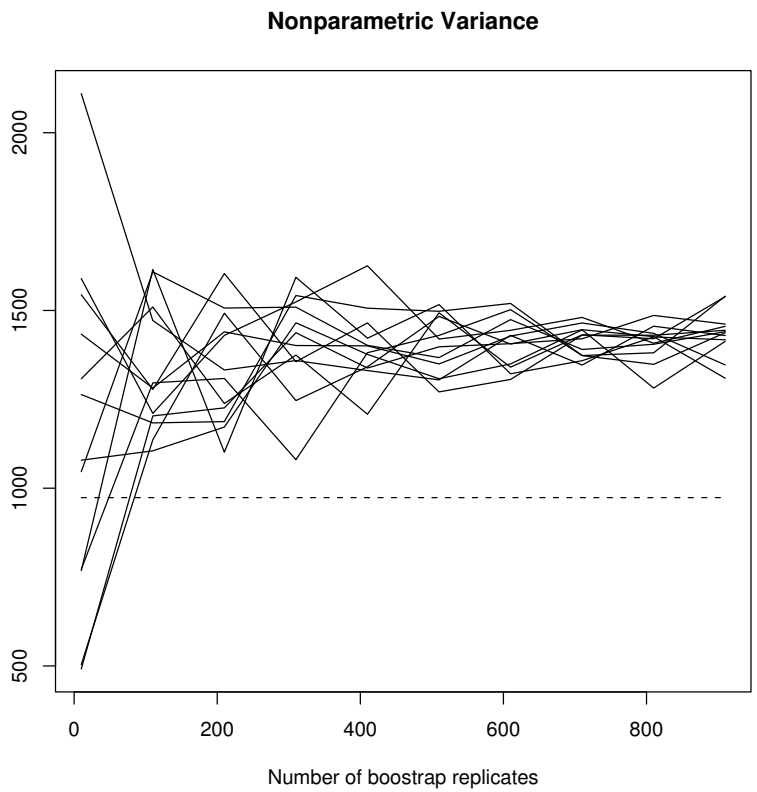
Dotted lines: theoretical quantiles under the hypothesis of exponential model.

Ex: nonparametric bootstrap

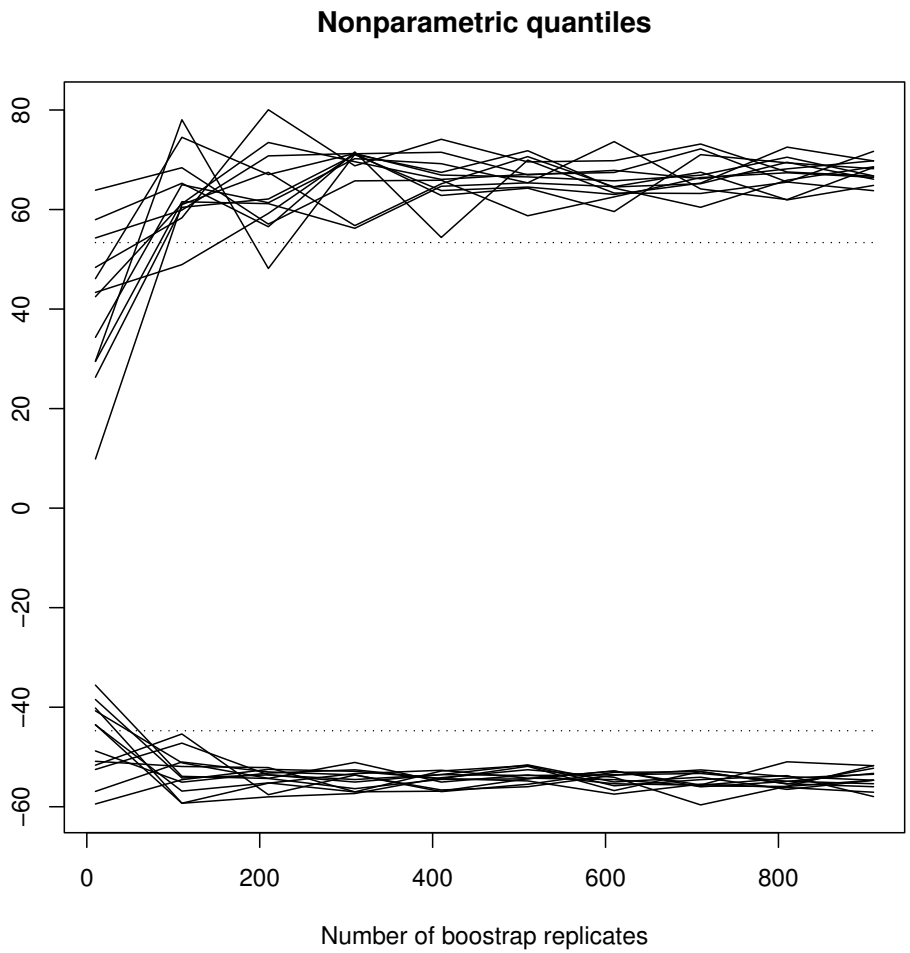
- Assume we do not know that the distribution is quasi exponential.
- We perform nonparametric bootstrap to estimate bias and variance of the estimator $\hat{\boldsymbol{\mu}}$.
- We sample B times with replacement the dataset D_N.
- We report the bias, the variance and the quantiles for different repetitions each having a set of B nonparametric trials.

Ex: nonparametric bootstrap





Ex: nonparametric bootstrap



Note that the nonparametric quantiles do not converge to the theoretical quantiles (dotted lines) based on the hypothesis of exponential model.

Parametric and nonparametric bootstrap

- For continuous data, a major difference between parametric and nonparametric resampling lies in the discreteness of the latter. Under nonparametric resampling the $\hat{\boldsymbol{\theta}}_{(b)}$ will have always a discrete distribution.
- However, given N distinct samples there are up to

$$B_{\text{max}} = \binom{2N-1}{N-1} = \frac{(2N-1)!}{N!(N-1)!}$$

  different bootstrap samples and consequently up to B_{max} possible values of $\hat{\theta}_{(b)}$ depending on the smoothness of t. For example for $N = 11$, $B_{\text{max}} = 352716$.
- For many practical applications, the effects of the discreteness are likely to be fairly minimal.

The bootstrap principle

- What we would like to know in an estimation problem is the distribution of $\hat{\boldsymbol{\theta}} - \theta$.
- What we have in bootstrap is a Monte Carlo approximation to the distribution $\hat{\boldsymbol{\theta}}_{(b)} - \hat{\theta}$.
- The key idea of the bootstrap is that for N sufficiently large we expect the two distributions to be nearly the same.
- In other terms the variability of $\hat{\boldsymbol{\theta}}_{(b)}$ (based on the empirical distribution) around $\hat{\theta}$ is expected to be similar (or mimic) the variability of $\hat{\boldsymbol{\theta}}$ (based on the true distribution) around θ.
- There is good reason to believe this will be true for large N, since as N gets larger and larger, the empirical $\hat{F}(\cdot)$ converge to $F(\cdot)$ (see the Glivenko-Cantelli theorem for iid samples).
- This idea is sometimes referred to as the *bootstrap principle*.

Error in resampling methods

- The error in resampling methods is generally a combination of **statistical error** and **simulation error**.
- Statistical error is due to the difference between the underlying distribution F and the empirical distribution $\hat{F}$. The magnitude of this error depends on the choice of $t(F)$.
- The use of rough statistics $t(F)$ (e.g. unsmooth or unstable) can make the resampling approach behave wildly. Example of nonsmooth statistics are sample quantiles and the median.
- The simulation error is due to the use of empirical (Monte Carlo) properties of $t(F)$ rather than exact properties.
- Simulation error decreases by increasing the number B of bootstrap replications.

Convergence of bootstrap estimate

In general terms for iid observations, we require

1. the convergence of $\hat{F}$ to F (satisfied by the Glivenko-Cantelli theorem) for $N \to \infty$;
2. an estimator such that the estimate $\hat{\theta}$ is the corresponding functional of the empirical distribution.

$$\theta = t(F) \to \hat{\theta} = t(\hat{F})$$

   This is satisfied for sample means, standard deviations, variances, medians and other sample quantiles.

3. a smoothness condition on the functional. This is not true for extreme order statistics such as the minimum and the maximum values.

When might the bootstrap fail?

So far we have assumed that the dataset D_N is iid sampled from a distribution F.

In some non conventional configurations, bootstrap might fail. For example

- Incomplete data (survival data, missing data).
- Dependent data (e.g. variance of a correlated time series).
- Dirty data (outliers)

For a critical view on bootstrap, see the publication *Exploring the limits of bootstrap* edited by Le Page and Billard which is a compilation of the papers presented at a special conference of the Institute of Mathematical Statistics hels in Ann Arbor, Michigan, 1990.