Statistical Methods for Data Science

Lesson 19 - Empirical bootstrap.

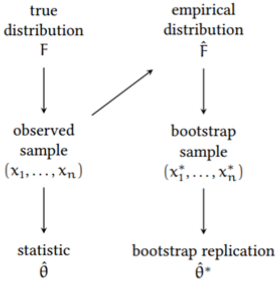
Salvatore Ruggieri

Department of Computer Science
University of Pisa
salvatore.ruggieri@unipi.it

Bootstrap principle

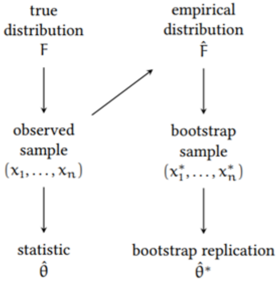
- Let $X_1, \ldots, X_n \sim F$ be a random sample
  - with *unknown distribution* F
- Estimator $T = h(X_1, \ldots, X_n)$, e.g., $\bar{X}_n = (X_1 + \ldots + X_n)/n$
- From a dataset $x_1, \ldots, x_n$, we can
  - derive a point estimate $\hat{\theta} = h(x_1, \ldots, x_n)$
  - or, derive an estimate $\hat{F}$ of F
- From $\hat{F}$ we can generate (a lot of) *bootstrap samples* $x_1^*, \ldots, x_n^*$
  - as realizations of $X_1^*, \ldots, X_n^* \sim \hat{F}$

  and then (a lot of) bootstrap point estimates $\hat{\theta}^* = h(x_1^*, \ldots, x_n^*)$
- By the LLN, the empirical distribution of $\hat{\theta}^*$ will approximate the distribution of $T^* = h(X_1^*, \ldots, X_n^*)$ and then of T



BOOTSTRAP PRINCIPLE. Use the dataset $x_1, x_2, \ldots, x_n$ to compute an estimate $\hat{F}$ for the "true" distribution function F. Replace the random sample $X_1, X_2, \ldots, X_n$ from F by a random sample $X_1^*, X_2^*, \ldots, X_n^*$ from $\hat{F}$, and approximate the probability distribution of $h(X_1, X_2, \ldots, X_n)$ by that of $h(X_1^*, X_2^*, \ldots, X_n^*)$.

Empirical bootstrap



- How to derive $\hat{F}$ from $x_1, \ldots, x_n$?
- If we know nothing about F, use the empirical distribution:

  $$\hat{F}(a) = F_n(a) = \frac{|\{i \in 1,\ldots,n \mid x_i \leq a\}|}{n}$$ [**Glivenko-Cantelli Thm**]
- How to generate a bootstrap sample $x_1^*, \ldots, x_n^*$?
  - x_i^* is chosen randomly from $\hat{F}$
  - i.e., x_i^* s chosen randomly from $x_1, \ldots, x_n$ (our dataset)
- Hence, a bootstrap dataset $x_1^*, \ldots, x_n^*$ is obtained by *random sampling with replacement*!
- Often the bootstrap approximation of the distribution of T will improve if we somehow normalize T by relating it to a corresponding feature of the "true" distribution.
  - rather than approximating the distribution of $\bar{X}_n$ by the one of $\bar{X}_n^*$
  - better to approximate $\bar{X}_n - \mu$ by $\bar{X}_n^* - \mu^*$, where $\mu^* = \bar{x}_n = (x_1^* + \ldots + x_n^*)/n$

[See remarks 18.1 and 18.2 of textbook]

Empirical bootstrap

EMPIRICAL BOOTSTRAP SIMULATION (FOR $\bar{X}_n - \mu$). Given a dataset $x_1, x_2, \ldots, x_n$, determine its empirical distribution function F_n as an estimate of F, and compute the expectation

$$\mu^* = \bar{x}_n = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

corresponding to F_n.

1. Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from F_n.
2. Compute the centered sample mean for the bootstrap dataset:

$$\bar{x}_n^* - \bar{x}_n,$$

   where

$$\bar{x}_n^* = \frac{x_1^* + x_2^* + \cdots + x_n^*}{n}.$$

Repeat steps 1 and 2 many times.

- Use the empirical distribution of $\delta^* = \bar{x}_n^* - \bar{x}_n$ for estimating
  - $\delta = \bar{x}_n - \mu$ as mean(δ^*)
  - and then $\mu = \bar{x}_n -$ mean(δ^*) with bias $\bar{x}_n - (\bar{x}_n -$ mean(δ^*)) $=$ mean(δ^*)

See R script

Empirical bootstrap

EMPIRICAL BOOTSTRAP SIMULATION (FOR $\bar{X}_n - \mu$). Given a dataset $x_1, x_2, \ldots, x_n$, determine its empirical distribution function F_n as an estimate of F, and compute the expectation

$$\mu^* = \bar{x}_n = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

corresponding to F_n.

1. Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from F_n.
2. Compute the centered sample mean for the bootstrap dataset:

$$\bar{x}_n^* - \bar{x}_n,$$

where

$$\bar{x}_n^* = \frac{x_1^* + x_2^* + \cdots + x_n^*}{n}.$$

Repeat steps 1 and 2 many times.

- Use the empirical distribution of $\delta^* = \bar{x}_n^* - \bar{x}_n$ for estimating
  - confidence interval (c_l, c_u) for $\delta = \bar{x}_n - \mu$ as $(q_{\alpha/2}, q_{1-\alpha/2})$ of δ^* distribution
  - $c_l \leq \delta = \bar{x}_n - \mu \leq c_u$ implies $\bar{x}_n - c_u \leq \mu \leq \bar{x}_n - c_l$, i.e. c.i. for μ is $(\bar{x}_n - c_u, \bar{x}_n - c_l)$

See R script

Empirical bootstrap

`boot.ci` method in R confidence intervals:

- `type='basic'`: $(\bar{x}_n - q_{1-\alpha/2}, \bar{x}_n - q_{\alpha/2})$ with quantiles over the distribution of δ^*
- `type='perc'`: $(q_{\alpha/2}, q_{1-\alpha/2})$ with quantiles over the distribution of $\bar{x}_n^*$
- `type='norm'`: $(\bar{x}_n - q_{1-\alpha/2}, \bar{x}_n - q_{\alpha/2})$ with quantiles over $N(mean(\delta^*), var(\delta^*))$
- `type='bca'`: bias correction and acceleration

Empirical bootstrap

`boot.ci` method in R confidence intervals:

- `type='stud'`: $(\bar{x}_n - q_{1-\alpha/2}\frac{s_n}{\sqrt{n}}, \bar{x}_n - q_{\alpha/2}\frac{s_n}{\sqrt{n}})$ with quantiles over the distribution of t^*

EMPIRICAL BOOTSTRAP SIMULATION FOR THE STUDENTIZED MEAN. Given a dataset $x_1, x_2, \ldots, x_n$, determine its empirical distribution function F_n as an estimate of F. The expectation corresponding to F_n is $\mu^* = \bar{x}_n$.

1. Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from F_n.
2. Compute the studentized mean for the bootstrap dataset:

$$t^* = \frac{\bar{x}_n^* - \bar{x}_n}{s_n^*/\sqrt{n}},$$

   where $\bar{x}_n^*$ and s_n^* are the sample mean and sample standard deviation of $x_1^*, x_2^*, \ldots, x_n^*$.

Repeat steps 1 and 2 many times.

See R script

Empirical bootstrap

- Bootstrap approach applies to **any** estimator, not only the mean
- Example 1: the German Tank problem
- Example 2: linear regression coefficients

See R script

An application of empirical bootstrap

- Bootstrap principle: the empirical distribution of $\delta^* = \bar{x}_n^* - \bar{x}_n$ approximates the distribution of $\delta = \bar{x}_n - \mu$
- Application: estimate $P(|\bar{X}_n - \mu| > 1)$ as the fraction of δ^* such that $|\delta^*| > 1$
- How good is the approximation?

See R script