Université Paris Dauphine
Département MIDO
Licence 3 – Statistical modelling

2022–2023

Practical n°4

stoehr@ceremade.dauphine.fr

Notations and Objectives. Let $X_1, \dots, X_n$ be a sequence of *i.i.d.* random variables with an **unknown** cumulative distribution function F and denote $x_1, \dots, x_n$ realisations of these random variables. The practical focuses on properties (bias, variance or mean squared error) of some statistics $U = T(X_1, \dots, X_n)$. The problem is that U is a random variable for which we solely know the realisation $u = T(x_1, \dots, x_n)$ corresponding to the dataset. We cannot derive the properties of interest of U as F is unknown.

Bootstrap is a resampling method which allows to study the properties of the random variable U by relying on the empirical distribution $\widehat{F}_n$ related to $x_1, \dots, x_n$. The bootstrap principle states that the empirical distribution is a good approximation of the true distribution F. U thus behaves approximately like $U^\star = T(X_1^\star, \dots, X_n^\star)$, where $X_1^\star, \dots, X_n^\star$ are *i.i.d.* random variables distributed according to $\widehat{F}_n$. The bootstrap principle can then be describe as follows

- **(a)** compute $u = T(x_1, \dots, x_n)$ from the sample $x_1, \dots, x_n$ (this serves as a reference);
- **(b)** generate $x_1^\star, \dots, x_n^\star$ a resample of the data, *i.e.* a sample from $\widehat{F}_n$;
- **(c)** compute $u^\star = T(x_1^\star, \dots, x_n^\star)$ from the resample $x_1^\star, \dots, x_n^\star$;
- **(d)** estimate the property of interest using u and $u^\star$.

Exercise. The `faithful` dataset is a publicly available data set about 272 consecutive eruptions of the Old Faithful geyser in Yellowstone National Park in Wyoming (USA). We are interested in the statistic corresponding to the median length of an eruption, *i.e.*, $U = \text{median}(X_1, \dots, X_n)$.

1. Load the data into your workspace and represent their distribution.

2. Compute u the median length of an eruption.

Part 1 – Bias and mean squared error.

Definition. (*General definition*)

The bias of an estimator $U = T(X_1, \dots, X_n)$ relative to m is defined as

$$B(T(X_1, \dots, X_n), m) = \mathbb{E}[T(X_1, \dots, X_n) - m].$$

$U = T(X_1, \dots, X_n)$ is said to be an unbiased estimator of m if $B(T(X_1, \dots, X_n), m) = 0$.

Here m denotes the median of the distribution F, *i.e.*, $\mathbb{P}[X_1 \leq m] = 0.5$. We face two issues since F is unknown: m is unknown and the expected value cannot be computed. The idea to estimate the bias is to use $u = T(x_1, \dots, x_n)$ computed from the sample as a reference value for m and to approximate the

expected value by a Monte Carlo estimator with respect to the empirical distribution $\widehat{F}_n$, *i.e.*,

$$\frac{1}{N}\sum_{k=1}^{N} T(X_{1k}^{\star},\dots,X_{nk}^{\star}) := \frac{1}{N}\sum_{k=1}^{N} U_k^{\star},$$

where $X_{ik}^{\star}$, $(i,k) \in [\![1,n]\!] \times [\![1,N]\!]$, are *i.i.d.* random variables distributed according to $\widehat{F}_n$.

3. Estimate the bias of $U = \text{median}(X_1,\dots,X_n)$ relative to m using $N = 1000$ bootstrap samples.

Definition. (*General definition*)

The mean squared error of an estimator $U = T(X_1,\dots,X_n)$ relative to m is defined as

$$\text{MSE}(T(X_1,\dots,X_n),m) = \mathbb{E}\left[(T(X_1,\dots,X_n)-m)^2\right].$$

4. With a reasoning similar to the bias, estimate the mean squared error of $U = \text{median}(X_1,\dots,X_n)$ relative to m using $N = 1000$ bootstrap samples.

Part 2 – Confidence intervals. The point estimate $u = T(x_1,\dots,x_n)$ change each time the data change. A key question is then to quantify the variation of $U = T(X_1,\dots,X_n)$ to provide uncertainty measure, that is to estimate confidence intervals for m.

Definition. (*General definition*)

A confidence interval for m with confidence level $1-\alpha$, $\alpha \in]0,1[$, is an interval I such that $\mathbb{P}[m \in I] \geq 1-\alpha$.

In general, we are interested in the variation of $U = T(X_1,\dots,X_n)$ around m, that is we would like to know the distribution of $U-m$. If we knew this distribution, we could compute quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ (respectively of order $\alpha/2$ et $1-\alpha/2$) of $U-m$ and we would have

$$\mathbb{P}\left[q_{\alpha/2} \leq U-m \leq q_{1-\alpha/2}\right] = \mathbb{P}\left[m \in \left[U-q_{1-\alpha/2}, U-q_{\alpha/2}\right]\right] = 1-\alpha.$$

The bootstrap principle states that we can estimate the distribution of $U-m$ by the distribution of $U^{\star}-u = T(X_1^{\star},\dots,X_n^{\star}) - T(x_1,\dots,x_n)$ to get an estimate of the confidence interval, **called empirical bootstrap confidence interval**.

5. Give a 90% empirical bootstrap confidence interval for m, using $N = 1000$ bootstrap samples.

If we believe that the distribution of $U^{\star} = T(X_1^{\star},\dots,X_n^{\star})$ based on a particular sample $x_1,\dots,x_n$ is a good approximation of the distribution of $U = T(X_1,\dots,X_n)$, we could build a confidence interval based on the distribution of $U^{\star}$, *i.e.*, $\left[q_{\alpha/2}^{\star}, q_{1-\alpha/2}^{\star}\right]$, where $q_{\beta}^{\star}$ denotes the quantile of order $\beta \in]0,1[$ of $U^{\star}$, **called percentile bootstrap confidence interval**.

6. Give a 90% percentile bootstrap confidence interval for m, using $N = 1000$ bootstrap samples.

7. What limitations could we encounter with this bootstrap percentile method?

Part 3 – Improve your understanding. We consider in this part that the statistic of interest is the empirical mean $\overline{X}_n$. Denote μ the expected value of the distribution F.

8. Using the bootstrap principle, estimate $\mathbb{P}\left[|\overline{X}_n - \mu| > 3\right]$.