Practical n°5

stoehr@ceremade.dauphine.fr

Notations and Objectives. Consider $x_1, \ldots, x_n$ observations of some phenomenon or experiment. We assume that our observations come from a statistical model $\mathscr{P} = \{f_\theta, \theta \in \Theta\}$, *i.e.* we assume that $x_1, \ldots, x_n$ are drawn from the distribution f_θ for an unknown parameter θ. Under the assumed statistical model, a goal is to find the value of θ such that f_θ has most likely generated the observed data. Put in other words we want to build an estimator $\widehat{\theta}_n = T(X_1, \ldots, X_n)$ that approximate the unknown parameter θ (or more generally a function of θ) and that has some properties (*e.g.*, unbiased, consistent). In the first part we study an estimator for a statistical model used in biology.

The practical also deals with the parametric bootstrap method. This method is different than the bootstrap method seen in Practical n°4 (sometimes referred as empirical or non-parametric bootstrap) but shares similarities as it is also a resampling method. The difference between those two methods is the source of the bootstrap sample $x_1^\star, \ldots, x_n^\star$. While the bootstrap method makes no assumption about the underlying distribution ($x_1^\star, \ldots, x_n^\star$ are drawn from $\widehat{F}_n$), the parametric bootstrap generates samples $x_1^\star, \ldots, x_n^\star$ from the parametrized distribution $f_{\widehat{\theta}_n}$ where $\widehat{\theta}_n$ is a statistic that estimates the unknown parameter θ.

Exercise *(Hardy–Weinberg model).* The haptoglobine has 3 different possible configurations AA, aa, aA. Plato, *et al.* (1964) observed the haptoglobine type on a sample of $n = 190$ people:

Genotype	AA	aa	aA
Count	10	112	68

When the genes frequencies are at equilibrium, frequency of each configuration only depends on an unknown parameter $\theta \in]0,1[$ and the statistical model, referred to as Hardy–Weinberg model, is given by

$$\mathbb{P}[X = \mathrm{AA}] = (1-\theta)^2, \quad \mathbb{P}[X = \mathrm{aa}] = \theta^2, \quad \text{and} \quad \mathbb{P}[X = \mathrm{aA}] = 2\theta(1-\theta).$$

1. *(Bonus)* Show that the statistical model used satisfies the equilibrium assumptions, that is the gene configurations frequencies remain constant from generation to generation.

Part 1 – Estimating θ. For $X_1, \ldots, X_n$, *i.i.d.* random variables distributed according the statistical model, the unknown parameter θ can be estimated using the following estimator

$$\widehat{\theta}_n = 1 - \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\{X_k = \mathrm{AA}\}} - \frac{1}{2n}\sum_{k=1}^{n} \mathbb{1}_{\{X_k = \mathrm{aA}\}}.$$

2. *(Bonus)* Is $\widehat{\theta}_n$ the maximum likelihood estimator?

3. Compute the value of $\widehat{\theta}_n$ for the observed data.

4. Show that $\widehat{\theta}_n$ is an unbiased estimator of θ that converges in probability to θ.

5. **(a)** Show that for all $t \in \mathbb{R}_+$

$$\mathbb{P}\left[\left|\sqrt{\frac{2n}{\widehat{\theta}_n(1-\widehat{\theta}_n)}}\left(\widehat{\theta}_n - \theta\right)\right| \leq t\right] \underset{n\to+\infty}{\longrightarrow} \mathbb{P}[|Z| \leq t], \quad \text{where} \quad Z \sim \mathcal{N}(0,1),$$

and deduce an asymptotic confidence interval for θ with level $1-\alpha$.

(b) Compute the 95% asymptotic confidence interval for θ with the observed data.

6. **(a)** Determine an asymptotic confidence interval for θ with level $1-\alpha$ using the delta method.

(b) Compute the 95% asymptotic confidence interval for θ with the observed data.

Part 2 – Parametric bootstrap. We assumed that we have a model parametrized by θ. Since we have a consistent estimator $\widehat{\theta}_n$ of θ, we can apply the parametric bootstrap. Algorithm are the same than in Practical n°4, except for the resampling step: $x_1^\star, \ldots, x_n^\star$ are drawn from

$$\mathbb{P}[X = \text{AA}] = (1-\widehat{\theta}_n)^2, \quad \mathbb{P}[X = \text{aa}] = \widehat{\theta}_n^2, \quad \text{and} \quad \mathbb{P}[X = \text{aA}] = 2\widehat{\theta}_n(1-\widehat{\theta}_n).$$

7. Use parametric bootstrap with $N = 1000$ bootstrap samples to

(a) estimate the bias of $\widehat{\theta}_n$ relative to θ,

(b) give a 95% empirical bootstrap confidence interval for θ,

(c) give a 95% percentile bootstrap confidence interval for θ.

8. Compare the results with the results from Part – 1.

Part 3 – To go further. When we use parametric bootstrap, we trust the statistical model (rightly or wrongly?). In this part, we do not make assumptions anymore about a specific model and a parametrized underlying distribution. The parametric bootstrap hence does not apply but the empirical bootstrap does as it only requires the knowledge of the empirical distribution based on our observed data.

9. Use empirical bootstrap with $N = 1000$ bootstrap samples to

(a) estimate the bias of $\widehat{\theta}_n$ relative to θ,

(b) give a 95% empirical bootstrap confidence interval for θ,

(c) give a 95% percentile bootstrap confidence interval for θ.

10. Compare the results with the results from Part – 1 and Part – 2.