Université de Strasbourg
M1 - Magistère
Statistique - études de cas

Ségolen Geffray
geffray@math.unistra.fr
Année 2017/2018

TP1: prise en main du logiciel R et illustration des propriétés d'un estimateur

---

Le but de ce TP est de proposer une illustration par simulations de certaines propriétés d'un estimateur $\widehat{\theta}_n$ d'un paramètre réel noté θ lorsque cet estimateur est déterminé à partir d'un échantillon $(X_1, ..., X_n)$ i.i.d. issu d'un modèle $(P_\theta)_{\theta\in\Theta}$.

Pour définir grossièrement le principe des simulations, il s'agit de générer des échantillons aléatoires selon des lois **entièrement** connues, y compris la valeur du ou des paramètre(s). Notons $(X_1^{\{m\}}, ..., X_n^{\{m\}})$ le $m^{\text{ème}}$ échantillon simulé pour $m \in \{1, ..., M\}$. On exploite ces M échantillons en appliquant la méthodologie étudiée comme on le ferait sur des données réelles c'est-à-dire en oubliant momentanément que l'on connaît les vraies valeurs. On calcule donc $\widehat{\theta}_n^{\{m\}}$ à partir de $(X_1^{\{m\}}, ..., X_n^{\{m\}})$, et ce, pour $m \in \{1, ..., M\}$. On analyse ensuite selon des critères appropriés les résultats de l'exploitation des échantillons que l'on compare aux vraies valeurs que l'on est censé obtenir. Nous nous concentrerons sur les qualités de précision et de consistance de l'estimateur ainsi que sur l'étude de la loi asymptotique de l'estimateur (correctement centré et normalisé).

Pour illustrer la précision d'un estimateur de θ à n fixé, on calculera son biais, sa variance et son risque quadratique ou plus exactement leur version empirique à savoir

$$\widehat{b_\theta}\left(\widehat{\theta}_n\right) = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\theta}_n^{\{m\}} - \theta\right) \overset{M\to\infty}{\to} \mathbb{E}_\theta\left[\widehat{\theta}_n - \theta\right] \text{p.s.}$$

$$\widehat{\text{Var}_\theta}\left(\widehat{\theta}_n\right) = \frac{1}{M}\left(\sum_{m=1}^{M}\left(\widehat{\theta}_n^{\{m\}}\right)^2 - \left(\sum_{m=1}^{M}\widehat{\theta}_n^{\{m\}}\right)^2\right) \overset{M\to\infty}{\to} \mathbb{E}_\theta\left[\left(\widehat{\theta}_n\right)^2\right] - \mathbb{E}_\theta\left[\widehat{\theta}_n\right]^2 = \text{Var}_\theta\left(\widehat{\theta}_n\right) \text{p.s.}$$

$$\widehat{R}_\theta\left(\widehat{\theta}_n\right) = \frac{1}{M}\left(\sum_{m=1}^{M}\left(\widehat{\theta}_n^{\{m\}} - \theta\right)^2\right) \overset{M\to\infty}{\to} \mathbb{E}_\theta\left[\left(\widehat{\theta}_n - \theta\right)^2\right] = R_\theta\left(\widehat{\theta}_n\right) \text{p.s.}$$

La loi forte des grands nombres valide les convergences presque-sûres annoncées pourvu que les moments d'ordre 1 et 2 impliqués existent et que les échantillons $(X_1^{\{m\}}, ..., X_n^{\{m\}})$ et donc les estimations $\widehat{\theta}_n^{\{m\}}$ aient été obtenus indépendamment les uns des autres pour $m \in \{1, ..., M\}$.

Pour illustrer la consistance de l'estimateur, on balaye différentes valeurs de n. On peut représenter sur un même graphique les valeurs médianes de l'estimateur obtenues pour les différentes valeurs de n ainsi que certains quantiles.

Pour illustrer la convergence en loi d'un estimateur, on peut tracer l'histogramme (ou bien une version lissée) des valeurs de l'estimateur (après centrage et normalisation adéquats) pour les différentes valeurs de n et superposer cet histogramme à la courbe de la loi obtenue lors de

l'étude théorique de l'estimateur.

Il est nécessaire de détailler le plan des simulations en explicitant le nombre d'échantillons générés noté M, la taille requise pour l'échantillon que l'on a noté n, les estimateurs étudiés et les scénarios choisis. Un scénario explicite le choix de la loi P_θ dont est issu l'échantillon $(X_1^{\{m\}}, ..., X_n^{\{m\}})$ pour $m \in \{1, ..., M\}$ ainsi que la **vraie** valeur du paramètre θ.
Le tableau ci-dessous résume ces choix.
Rappelons que l'on note $\overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

loi	paramètres	θ	estimateur	biais	comportement asymptotique
$\mathcal{N}(m, \sigma^2)$	$m = 0$, $\sigma^2 = 1$	m	$\widehat{m}_n = \overline{X}_n$	0	$\sqrt{n}\widehat{m}_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$
$\mathcal{N}(m, \sigma^2)$	$m = 0$, $\sigma^2 = 10$	m	$\widehat{m}_n = \overline{X}_n$	0	$\sqrt{n}\widehat{m}_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, 10)$
$\mathcal{N}(m, \sigma^2)$	$m = 0$, $\sigma^2 = 1$	σ^2	$\widehat{\sigma}_n^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$	0	$\sqrt{n}(\widehat{\sigma}_n^2 - 1) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 2)$
$\mathcal{N}(m, \sigma^2)$	$m = 0$, $\sigma^2 = 1$	σ^2	$\widehat{\sigma}_n^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$	$-\dfrac{1}{n}$	$\sqrt{n}(\widehat{\sigma}_n^2 - 1) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 2)$
$\mathcal{U}(0, a)$	$a = 1$	a	$\widehat{a}_n = 2\overline{X}_n$	0	$\sqrt{n}(\widehat{a}_n - 1) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \dfrac{1}{3})$
$\mathcal{U}(0, a)$	$a = 1$	a	$\widehat{a}_n = \max(X_1, ..., X_n)$	$-\dfrac{1}{n+1}$	$n(1 - \widehat{a}_n) \xrightarrow{\mathcal{D}} \mathcal{E}(1)$
$\mathcal{P}(\lambda)$	$\lambda = 0.8$	λ	$\widehat{\lambda}_n = \overline{X}_n$	0	$\sqrt{n}(\widehat{\lambda}_n - 0.8) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 0.8)$
$\mathcal{P}(\lambda)$	$\lambda = 10$	λ	$\widehat{\lambda}_n = \overline{X}_n$	0	$\sqrt{n}(\widehat{\lambda}_n - 10) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 10)$

On utilisera les valeurs $M = 1000$ et $n = 20, 50, 100, 200, 500$.

Le travail à effectuer peut se mettre sous la forme de l'algorithme suivant:
pour $n = 20, 50, 100, 200, 500$,

- pour $m = 1, ..., M$,
  - simuler un échantillon i.i.d. $(X_1^{\{m\}}, ..., X_n^{\{m\}})$ selon la loi P_θ
  - calculer $\widehat{\theta}_n^{\{m\}}$
- calculer $\widehat{b}_\theta\left(\widehat{\theta}_n\right)$, $\widehat{\mathrm{Var}}_\theta\left(\widehat{\theta}_n\right)$ et $\widehat{R}_\theta\left(\widehat{\theta}_n\right)$
- déterminer la médiane empirique de $\widehat{\theta}_n^{\{1\}}, ..., \widehat{\theta}_n^{\{M\}}$, les valeurs minimum et maximum ainsi que les quantiles empiriques d'ordre 0.025, 0.25, 0.75 et 0.975. Placer ces valeurs en ordonnée sur un graphique comportant n en abscisse.
- calculer les valeurs correctement centrées et normalisées de $\widehat{\theta}_n^{\{m\}}$ au vu de la convergence en loi à illustrer. Tracer l'histogramme de ces nouvelles valeurs. Superposer la courbe de la loi limite théorique.

En réalité, le logiciel `R` peut être assez lent lorsqu'il s'agit d'effectuer des boucles. Dans la mesure du possible, il faut donc éviter de lui demander de faire des boucles en préférant une programmation vectorielle.
Il est également souhaitable de ne pas saturer la mémoire de l'ordinateur en stockant le minimum possible de données.