Outils statistiques M2
Fiche 1 : bootstrap

Responsable d'UE : Davide Giraudo

24 septembre 2023

1 Exercice 1 : Implémentation et illustration des deux approximations

Dans cet exercice, on considère un échantillon x tiré suivant une loi exponentielle $\mathcal{E}(0.3)$ et on s'intéresse à l'estimation de sa médiane m.

1. Écrire des fonctions qui, à partir de x, déterminent les médianes de B échantillons bootstrap (à partir desquels on pourra estimer m et x un intervalle de confiance de m) dans les cas suivants :
   (a) bootstrap non-paramétrique à la main (sans utiliser de fonction R pré-écrite)
   (b) bootstrap non-paramétrique en utilisant la fonction `boot`.
2. (a) Comparer, à l'aide d'un boxplot, la distribution empirique de l'échantillon $(\widehat{m_1}, \dots, \widehat{m_B})$ correspondant à $B = 500$ estimations obtenues sur des échantillons indépendants de taille $n = 20$, avec la distribution empirique de $(\widehat{m_1}, \dots, \widehat{m_B})$ obtenu par bootstrap sur un échantillon.
   (b) Commentez. Quelle est l'approximation illustrée par ce graphique ? L'erreur diminue-t-elle lorqu'on augmente B ?
3. (a) Comparer les distributions empiriques des échantillons bootstrap pour $n = 100, 500, 1000$.
   (b) Commentez. Quelle est l'approximation illustrée par ce graphique ?

2 Exercice 2 : Bootstrap paramétrique vs bootstrap non paramétrique

On considère une variable X d'espérance μ et un échantillon x i.i.d. de loi X.

1. Écrire une procédure (soit à la main, soit en utilisant la fonction `boot`) qui donne un intervalle de confiance bootstrap de μ à un niveau $(1 - \alpha)$ dans les cas suivants :

(a) bootstrap non-paramétrique

(b) bootstrap paramétrique en supposant que X suit une loi exponentielle

(c) bootstrap paramétrique en supposant que X suit une loi de Poisson

(d) bootstrap paramétrique en supposant que X suit une loi normale

2. Écrire un programme qui génère 100 échantillons $(x_1, \ldots, x_{15})$ de loi de Poisson de paramètre $\lambda = 1.2$ et qui calcule sur chaque échantillon les bornes inférieures et supérieures des différents intervalles de confiance. Déterminer la largeur empirique moyenne de chaque type d'intervalle de confiance ainsi que la probabilité de couverture empirique de chaque type d'intervalle de confiance. Que constatez-vous ?

Recommencer avec une loi Gamma de paramètre de forme (shape) égal à 1.2 puis 0.8 pour un paramètre d'échelle (scale) égal à chaque fois à 1.

3 Exercice 3 : comparaison des espérances de vie entre fumeurs et non-fumeurs

On considère les échantillons suivants, qui donnent les âges de décès de vingt fumeurs et vingt non-fumeurs.

Non-fumeurs : 84, 73, 73, 83, 80, 67, 91, 76, 90, 70, 70, 81, 78, 68, 64, 82, 91, 72, 84, 66
Fumeurs : 94, 65, 71, 83, 47, 55, 96, 57, 57, 64, 77, 97, 51, 50, 48, 41, 71, 59, 86, 71

1. Tester l'égalité des espérances de vies entre les deux conditions à l'aide d'un test basé sur le bootstrap. On précisera les hypothèses, la latéralité et la statistique de test choisie.

On s'intéresse dorénavant au ration des espérances de vie, à savoir, si X_{F} et X_{NF} désignent les âges de décès chez les fumeurs et les non-fumeurs respectivement,

$$\theta = \frac{\mathbb{E}\left[X_{\mathrm{F}}\right]}{\mathbb{E}\left[X_{\mathrm{NF}}\right]}. \tag{3.1}$$

2. Déterminer une estimation bootstrap du biais et de la variance de θ.
3. Déterminer un intervalle de confiance à 95% de θ.

4 Exercice 4 : bootstrap et régression linéaire

On considère un problème de régression linéaire gaussienne

$$Y_i = X_i\beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}\left(0, \sigma_i^2\right). \tag{4.1}$$

On dispose d'un échantillon $((X_i, Y_i))_{i=1}^n$. On considère deux manières d'obtenir des intervalles de confiance pour l'estimation de β :

1. Considérer les X_i comme aléatoires et faire des échantillons bootstrap en tirant avec remise des couples (X_i, Y_i).
2. Considérer les X_i comme constants et seuls les résidus comme aléatoires. On peut alors estimer les résidus par $\widehat{\varepsilon_i} = Y_i - X_i\widehat{\beta}$, créer des échantillons bootstrap des résidus et définir $Y_i^{*b} = X_i^{*b}\widehat{\beta} + \varepsilon^{*b}$.

Dans les deux cas, chaque échantillon bootstrap ainsi construit contient n couples (X, Y) qui permettent d'obtenir une estimation bootstrap de β.

1. Charger les données airquality. Après avoir supprimé les lignes ayant une valeur manquante, estimer les paramètres d'une régression linéaire du taux d'ozone en fonction du rayonnement solaire, du vent et de la température. Donner un intervalle de confiance à 95% des coefficients de β avec chacune des méthodes précédentes.
2. Reprendre la question précédente sur des données homoscédastiques ($\sigma_i = \sigma$ pour tout i), puis sur des données fortement hétéroscédastiques (avec des σ_i d'un ordre de grandeur plus grand que $X\beta$). Commentez les résultats données par chacune des méthodes.

5 Exercice 5 : majoration de la première erreur du bootstrap

On considère une variable aléatoire X telle que $\mathbb{E}\left[|X|^3\right] < \infty$. On considère que le paramètre θ à estimer est la moyenne et on note

$$F_n(t) = \mathbb{P}\left(\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \leqslant t\right), \quad \widehat{F}_n(t) = \mathbb{P}\left(\sqrt{n}\left(\widehat{\theta}_n^* - \widehat{\theta}\right) \leqslant t\right). \tag{5.1}$$

les probabilités d'avoir des erreurs d'estimation supérieures à $t/\sqrt{n}$ dans le monde réel et dans le monde bootstrap. On note Φ_σ la fonction de répartition de la loi normale centrée en 0 et d'écart-type σ. Le but de cet exercice est de montrer que

$$\sup_{t\in\mathbb{R}} \left|F_n(t) - \widehat{F}_n(t)\right| = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right). \tag{5.2}$$

Pour cela, on admet le théorème de Berry-Esseen qui borne la convergence du TCL :

Théorème 1 (Berry-Esseen)**.** *Soit X une variable aléatoire d'espérance μ, de variance σ^2 et telle que $\mu^3 := \mathbb{E}\left[|X-\mu|^3\right] < \infty$. On considère un échantillon de n tirages indépendants, de moyenne $\overline{X_n}$. Alors*

$$\sup_{z\in\mathbb{R}} \left|\mathbb{P}\left(\sqrt{n}\frac{\overline{X_n} - \mu}{\sigma} \leqslant z\right) - \Phi_1(z)\right| \leqslant \frac{33}{4}\frac{\mu_3}{\sigma^2\sqrt{n}}. \tag{5.3}$$

1. Démontrer que

$$\sup_{t\in\mathbb{R}} |F_n(t) - \Phi_\sigma(t)| = O\left(\frac{1}{\sqrt{n}}\right). \tag{5.4}$$

2. Démontrer que

$$\sup_{t\in\mathbb{R}}\left|\widehat{F_n}\left(t\right)-\Phi_{\widehat{\sigma}}\left(t\right)\right|=O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right). \tag{5.5}$$

3. Démontrer que

$$\sup_{t\in\mathbb{R}}\left|\Phi_{\sigma}\left(t\right)-\Phi_{\widehat{\sigma}}\left(t\right)\right|=O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right). \tag{5.6}$$

On pourra commencer par montrer que $|\widehat{\sigma}-\sigma|=O_{\mathbb{P}}\left(n^{-1/2}\right)$.