TP3 : Tests par rééchantillonnage

ENSAE 3ème année, 2017

Applications du bootstrap et autres techniques de rééchantillonnage (cours de M. Roquain)

Cadre général : test non paramétrique de la moyenne

Soit $\mathcal{X}_n = (X_1, \dots, X_n)$ i.i.d. de fonction de répartition $F_\theta(x) = G(x-\theta)$ pour G une fonction de répartition d'une loi centrée et $\theta \in \mathbb{R}$. On cherche à tester

$$H_0\text{: "}\theta = 0\text{" contre } H_1\text{: "}\theta \neq 0\text{"}$$

à l'aide de la statistique de test

$$T_n(\mathcal{X}_n) = n^{1/2}\bar{X}_n.$$

On construit un test de niveau α en rejetant H_0 si $|T_n| > c_\alpha$, pour une certaine constante c_α, éventuellement dépendante de $\mathcal{X}_n$.

Exercice 0 (Création de votre fichier)

Créer votre fichier `nom_prenom_TP3.R`. N'oubliez pas d'effectuer des commentaires dans votre fichier (que vous utilisiez un notebook ou pas).

Exercice 1 (Test standard)

On suppose que la loi de G est connue et se trouve être une loi $\mathcal{N}(0,1)$.

1) Pour $\alpha = 5\%$, donner la valeur de c_α et tracer la fonction puissance du test (c'est-à-dire la probabilité de rejet) pour $n = 10$, puis, sur le même graphe, pour $n = 1000$. Ajouter une légende et commenter.

2) Charger le jeu de données "data1forTP3.dat" disponible sur ma page http://etienne.roquain.free.fr/teaching.html. Calculer la p-valeur du test en utilisant le modèle gaussien. Sur ce jeu de données, rejette t-on H_0 au niveau 5% ?

Exercice 2 (Test par bootstrap)

A présent, on ne suppose plus la loi de G est gaussienne.

1) Pour le jeu de données "data1forTP3.dat", calibrer maintenant $c_\alpha(\mathcal{X}_n)$ en utilisant une approximation bootstrap de la loi de $T_n(\mathcal{X}_n)$ *sous H_0*. Comparer au c_α obtenu avec le modèle gaussien et commenter.

2) Calculer la p-valeur du test par bootstrap et comparer à l'exercice 1. Que se passe t-il si B est trop petit ?

3) (Plus dur !) Approcher à présent la fonction puissance du test au niveau α par la technique du "double bootstrap". Plus précisément, la probabilité de rejet de H_0 sous la moyenne θ est approchée par la probabilité

$$\mathbb{P}\left(T_n(\mathcal{X}_n^\star) > c_\alpha(\mathcal{X}_n^\star) - n^{1/2}\theta \middle| \mathcal{X}_n\right) + \mathbb{P}\left(T_n(\mathcal{X}_n^\star) < -c_\alpha(\mathcal{X}_n^\star) - n^{1/2}\theta \middle| \mathcal{X}_n\right)$$

dans lequel le bootstrap se fait toujours sous H_0. Comparer à la fonction puissance issue du cadre Gaussien.

Exercice 3 (test par subsampling)

Répondre aux questions 1), 2) de l'exercice 2 avec une approche par subsampling. Commenter. Notez bien que la calibration du test par subsampling ne nécessite pas de rééchantillonner sous H_0 (!).

Exercice 4 (test par symétrisation)

On suppose en plus que la loi définie par G *est symétrique* (par rapport à 0).

1) Générer $(\varepsilon_1, \ldots, \varepsilon_n)$ des variables aléatoires i.i.d. qui sont des *signes aléatoires*, c'est-à-dire qui valent -1 ou $+1$ avec la même probabilité.

2) Si $\mathcal{Y}_n = (Y_1, \ldots, Y_n)$ sont i.i.d. de loi de student à 5 degrés de liberté, quelle est la loi de l'échantillon $\mathcal{Y}_n^\star = (\varepsilon_1 Y_1, \ldots, \varepsilon_n Y_n)$? Illustrer ceci avec un graphe.

3) Répondre aux questions 1), 2) de l'exercice 2 avec une approche par symétrisation.