Feuille de TP n°5 – Quantiles empiriques – Test de Kolmogorov-Smirnov

1 Convergence des quantiles empiriques

Soit μ une mesure de probabilité sur $\mathbb{R}$ de fonction de répartition F continue. Le quantile d'ordre p, noté k_p est défini par $F^{-1}(p)$. Soit $X_1, \dots, X_n$ un échantillon de loi μ. Le quantile empirique d'ordre p est défini par $X_{([np])}$.

1. Illustrer le fait que, pour $p \in]0,1[$,
$$X_{([np])} \underset{n\to\infty}{\longrightarrow} k_p \quad \text{p.s.}$$
2. Illustrer le fait que, pour $p \in]0,1[$,
$$\sqrt{n}\big(X_{([np])} - k_p\big) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{p(1-p)}{f(k_p)^2}\right).$$
3. Choisissons μ égale à la loi exponentielle de paramètre 1. Illustrer la convergence p.s. de $X_{(1)}$ vers 0. À quelle vitesse a lieu cette convergence, c'est-à-dire quelle est la bonne renormalisation v telle que $v(n)X_{(1)}$ converge en loi vers une limite non triviale ? Illustrer ce résultat. Comment le généraliser ?

Exercice 1. On considère le modèle de translation suivant : soit f une densité de probabilité strictement positive sur $\mathbb{R}$ telle que la médiane et la moyenne soient égales à 0. Les observations proviennent d'une loi μ_θ de densité $f_\theta(\cdot) = f(\cdot - \theta)$ pour un certain $\theta \in \mathbb{R}$.

1. Proposer deux estimateurs de θ en reliant leurs propriétés à celles de f.
2. Discuter leurs avantages respectifs dans les cas suivants :
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad f(x) = \frac{1}{2} e^{-|x|} \quad \text{et} \quad f(x) = \frac{1}{\pi} \frac{1}{1+x^2}.$$
3. Dans les deux premiers cas, quel est l'estimateur du maximum de vraisemblance ?

2 Fonction de répartition empirique

Soit $X_1, \dots, X_n$ un échantillon de loi μ sur $\mathbb{R}$ (et de fonction de répartition F). On définit la fonction de répartition empirique F_n de l'échantillon par :
$$\forall x \in \mathbb{R}, \quad F_n(x) \overset{\text{déf.}}{=} \frac{1}{n} \#\{X_i \leq x,\ i = 1, \dots, n\}.$$

1. Illustrer le fait que $F_n(x) \to F(x)$ p.s. pour tout $x \in \mathbb{R}$ en superposant la fonction de répartition F et plusieurs fonctions de répartition empiriques.
2. Montrer que
$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = \max_{1\leq i\leq n} \left[\max\left(\left| \frac{i}{n} - F(X_{(i)}) \right|, \left| \frac{i-1}{n} - F(X_{(i)}) \right| \right) \right],$$
où $X_{(i)}$ désigne la statistique d'ordre i. En déduire que cette statistique est libre (*i.e.* sa loi ne dépend pas de μ).
3. Illustrer le théorème de Glivenko-Cantelli qui assure que
$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \underset{n\to\infty}{\longrightarrow} 0 \quad \text{p.s.}$$

4. Illustrer le théorème de Kolmogorov-Smirnov qui assure que la statistique

$$K_n = \sqrt{n} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$$

converge en loi. On pourra pour cela tracer un histogramme d'un grand nombre de réalisations $(K_n^{(i)})_i$ de K_n pour n assez grand. On pourra aussi montrer que les quantiles empiriques de $(K_n^{(i)})_i$ fournissent une bonne approximation des quantiles de la loi de Kolmogorov-Smirnov. Si $k_{1-\alpha}$ et α sont liés par la relation

$$\mathbb{P}(K \geq k_{1-\alpha}) = \alpha,$$

alors $k_{0,9} = 1,22$, $k_{0,95} = 1,36$ et $k_{0,99} = 1,63$.

Exercice 2 (Pour aller plus loin)**.** Soit $(X_n)_n$ une suite de v.a. i.i.d. de loi uniforme sur $[0,1]$. On note

$$H_n(x) = \sqrt{n}(F_n(x) - F(x).$$

1. Calculer
$$\mathbb{E}(H_n(x)), \quad \mathbb{V}(H_n(x)) \quad \text{et, pour } x < y, \quad \mathbb{E}(H_n(x)H_n(y)).$$
2. En déduire que, pour tous $0 < x_1 < x_2 < \cdots < x_k < 1$,
$$(H_n(x_1), \ldots, H_n(x_k)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, K) \quad \text{avec} \quad Kij = (x_i \wedge x_j)(1 - x_i \vee x_j).$$
3. Écrire un programme qui génére un échantillon de grande taille de v.a. uniformes et trace H_n pour plusieurs valeurs de n pour illustrer que le fait que
$$(H_n(x))_{x\in[0,1]} \xrightarrow[n\to\infty]{\mathcal{L}} (B(x))_{x\in[0,1]}, \quad \text{où } B \text{ est un pont brownien } (???).$$

3 Tests non paramétriques

Theorème 3. *Soit $(X_n)_n$ une suite de v.a. i.i.d. de fonction de répartition F continue. On note F_n la fonction de répartition empirique de l'échantillon de taille n.*

$$nI_n = n\int (F_n(x) - F(x))^2 \, dF(x) \xrightarrow[n\to\infty]{\mathcal{L}} \sum_{k=1}^{\infty} \frac{Y_k}{(\pi k)^2},$$

où les v.a. $(Y_k)_k$ sont i.i.d. de loi $\chi^2(1)$.

Formule pratique. Tout comme pour la statistique de Kolmogorov-Smirnov, cette intégrale n'est en fait qu'une somme :

$$nI_n = \frac{1}{12n} + \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - F(X_{(i)}) \right)^2.$$

Pour $n \geq 100$, on peut utiliser les approximations suivantes :

$$\begin{cases} \mathbb{P}(\sqrt{n}D_n < 1.223) = 0.9 \quad \mathbb{P}(\sqrt{n}D_n < 1.358) = 0.95 \quad \mathbb{P}(\sqrt{n}D_n < 1.629) = 0.99 \\ \mathbb{P}(nI_n < .74346) = 0.99 \quad \mathbb{P}(nI_n < .46) = 0.95 \quad \mathbb{P}(nI_n < .34730) = 0.9. \end{cases}$$

►► On teste si un échantillon suit la loi exponentielle de paramètre 1. Tracer les courbes de puissance en fonction d'un paramètre de perturbation pour comparer les deux tests.
►► Essayer avec d'autres lois pour voir si les résultats s'inversent.

Remarque 4. Aucun résultat théorique n'existe sur la comparaison des puissances : la simulation est donc importante pour dégager le test le plus intéressant.